数学の参考書

Mathematics Reference Book

Notes on Mathematics

by

Desync

$\boxed{\mathcal{WMX}}$

Warwick Mathematics Exchange

Copyright

History

First Edition: 2022-08-19
Current Edition: 2025-06-22

Dedication

To anyone going through
the pain of a maths degree

Contents

List of Figures

List of Tables

Preface

This document is a work-in-progress. Please see the content outline in chapter 1 for a list of material I intend to cover.

TO DO:

- Reformat all of Analysis I to use theorem environments.
- Add glossary/list of notation. (Done for MVC alone due to MVC notation gore, and for some of topology.)
- Write chapter on introductory calculus.
- Write new full chapter on set theory.
- Finish chapter on number theory.
- Finish chapter on combinatorics.
- Finish chapter on the lambda calculus.
- Finish chapters on homology/cohomology.

This document is intended to cover a wide variety of topics useful for beginning to study mathematics. As with many things in maths, there are many different perspectives with which various topics can be viewed. In fact, that is one of the things that makes maths so powerful – the ability to use axioms as interfaces between various theories. But, if something here is confusing or just doesn't quite click for you, I encourage you to research further on the topic.

Even if you don't fully understand something, just having seen it at all will make it easier the second time around. Rather than it being some new mysterious concept to learn from scratch, it'll be a familiar face (even if only vaguely so). The more broadly you research, the more often you'll run into these collisions, and the more tightly everything will connect together.

Everything here is probably more detail than you probably need, in particular the chapter(s) on foundations, which is probably only useful to a select few logicians and category/set theorists. However, it is helpful to at least see how the foundations of mathematics are built (as well as alternative foundations to traditional material set theory). It is not expected that you will understand everything on first reading.

Additionally, topics less favoured by the author (i.e., analysis, calculus) have much less prose, and are generally a curated list of theorems and examples, while others, like abstract algebra and category theory, are almost entirely prose and discussions.

This document has been written in such a way that the chapters are intended to be read in order, like a book. However, wherever important definitions have been buried deep within previous chapters, they will be repeated if highly relevant to the current content. Additionally, topics in mathematics are not linear, and chapters will often refer to each other. Because of this, the table of contents above, and any inline references are all hyperlinked for your convenience. Additionally, if a term is unfamiliar to you, take a quick look through the index at the back, which is also hyperlinked.

Disclaimer: I make *absolutely no guarantee* that this document is complete nor without error. Any opinions expressed in this document are my own and do not necessarily represent the views of any other individual or insitution.

Notes on Formatting

New terminology will be introduced in *italics* when used for the first time. Named theorems will also be introduced in *italics*. Important points will be **bold** or will be underlined.

Where relevant, differing notational conventions will be brought up in the sections they are used.

Acknowledgements

I would like to acknowledge the extraordinary amount of intuition and inspiration I have gained from the content created by Grant Sanderson at *3Blue1Brown*. Many sections, particularly the ones on linear algebra, are broadly adapted from his animated works.

I also offer my thanks and apologies to professor James Aspnes, from whose excellent computer science notes most of the first chapter on symbolic logic is taken from, and on which the structure of various other chapters is based.

My gratitude to Michael Stevens of *Vsauce* for inspiring my lifelong curiosity of anything and everything I encounter. In particular, the section on transfinite numbers in the introductory set theory chapter is also based heavily on the video *How To Count Past Infinity*.

Some of the parts of the chapters on abstract and linear algebra were adapted from notes and lectures by Dr Nicholas Jackson and Dr Christian Böhning; the later parts of analysis from notes by Jose Rodrigo; topology from notes by James Robinson and Richard Sharp, and lectures from Rohini Ramadas; and vector and multivariable calculus/analysis from notes and lectures by Siri Chongchitnan and Mario Micallef, respectively (though I am sure that Mario disapproves highly of my notational conventions, just as I disapprove of his).

My thanks also to my good friend and computer scientist Musab Guma'a for his support in various topics in theoretical computer science, and whose notes on algorithms, program verification, and formal languages I have often referred to.

Thanks also to my good friends and analysis enjoyers Ryan Tay and Aris Mercier, from whom I have stolen many examples of contour integration. (Conversely, I did the Tikz diagrams for them, so I consider this a fair trade.)

Much of the information presented in this document was broadly researched from Wikipedia, and many of the proofs inspired by or checked against ProofWiki. Many thanks to the hardworking editors and writers of those respective sites.

Authors

This document was written by Kit '*Desync*' L.

Please send me a PM on Discord `@.desync`, a message in the WMX server, or an email to liurjk@gmail.com for any corrections.

If you found these notes helpful and want to support me, you can buy me a coffee!

Chapter 1

Introduction

"If you want to build a ship, don't drum up people to collect wood, and don't assign them tasks and work, but rather, teach them to long for the endless immensity of the sea."

— Antoine de Saint-Exupéry

1.1 Motivation

Why should you study mathematics?

The hope is that, given that you're reading this already, and are probably already enrolled into or are applying for a mathematics course, you will have your own personal answer to this. However, if you'll spare me some time to ramble, I'll offer you a short discussion of my own perspective on this.

To begin with, I'd like to start with people's view of maths. A perspective I'd seen a lot in the tutorials that I'd run, and I'm sure that any teachers reading this can agree, is that a lot of students perceive maths to be; here's a bunch of problems; solve them; and getting the correct answer is the only thing that matters. Or alternatively, that maths is "completely useless" in the real world.

And that's as far as many students go, in terms of learning more about maths. It's not about forming connections between different aspects of maths. Or even as anything more than a tool to help calculate prices. I mean, why would anyone need to solve a quadratic outside of a maths exam?

Should we stop forcing students to learn about quadratics and trig and whatnot?

I'm going to say something that might surprise you; I think that sounds like a reasonable idea. However, before you send me an angry email, let me finish; I'll explain in due course.

Now, maths. The stuff we do in primary or most of secondary education isn't really what I'm talking about, and it's not like I'll be going into the specifics of advanced abstract algebra or category theory or whatever, because that's not the point either, as much as I like those topics.

To be honest, I find that modern education, at least up to secondary level or so (and at least, in my own personal experience), is really good at making maths boring. It's the cycle of just solving more things over and over again. But, it's only boring for the same reason that learning an alphabet would be boring to a fifteen year old.

The stuff we do in school for maths is basically learning the alphabet in order for you to write. The difference is, we learn the alphabet when we're four or five, and we don't find it boring then, because we're dumb four or five year olds and don't know any better.

Specifically, the problem with maths is that we're bad at it. Numbers and computation are things we can't see or touch, and our brains aren't optimised to deal with them. Rather than learning the proper foundations of modern maths at four or five, we have to delay it by a decade, and that's one of the reasons we find it boring; it's just the foundations, and we're taught it at such an older age. It's basic stuff. And I mean basic in the foundational meaning there, not "easy" – because it isn't. I mean, consider what it's like learning the Greek alphabet or Japanese syllabaries with an English background; that's certainly not easy, but it's basic in the same sense. It's the boring stuff like learning how to write your letters properly before you can truly manipulate the language. In the same way, you have to know how to manipulate formulae instinctively before you do anything else more advanced, much like how you don't actually have to think about every single letter every time you write or type something down.

When people see someone who's good at maths, they sometimes assume that they must just be incredibly intelligent or some kind of natural genius. Maybe it does come easier to some people than others, and maybe it doesn't, but that's not the point. Someone who decides to pursue maths, no matter the level, doesn't necessarily find it easy – they just enjoy how difficult it is; the puzzle solving aspect of it. They enjoy the moment of understanding when it all clicks in place; whether that's a six year old spotting patterns in multiplication tables, or a sixteen year old just finally getting how the chain rule works.

People who do programming will identify with this well – I mean, if you program on any kind of regular basis, you'll know exactly what I mean. Once in the past, I'd spent six hours debugging a program I was working on before realising the problem was that I missed a colon somewhere in the five thousand odd lines of code.* And, just like working on a maths problem and not getting the answer, I was frustrated and irritated. I'd like to say that I'd learnt from that mistake, and I haven't done it again since, but that'd just be a sad lie. But after the dust settled, and everything was finished; the feeling of satisfaction once the program was finally working is hard to describe. Rather than the empty satisfaction from spending five seconds on an easy problem, I chose to continue working away at this one thing. And in the end, it did what I wanted.

And I'd like to think that this is why people do this kind of thing. I mean, the easiest way to put a football† into the goal, is probably to pick it up and throw it. But we don't do that. We follow rules that make life hard for ourselves – because it's more fun that way. It's not about scoring the goal, or getting the question right the first time. It's the experience you go through. But just as playing sports well requires a certain level of physical proficiency, doing maths requires a certain level of perseverance, and of course, the background knowledge required in the first place. Also, that's all the sports you're getting from me for the rest of this book.

So, I said that people are not good at maths, and that's true. People aren't inherently good at manipulating abstract ideas like they are with languages. We're subject to all sorts of biases as fallacies; because they're what worked for us to survive in the past.

Let's do a few examples. Let's say that I bet you £10 on a fair coin flip; if it's heads, I get your £10, and if it's tails, you get my £10. Most people won't take that bet. Fair enough, given that the expected value is zero.

But what about my £11 against your £10? Most people still won't take that bet, despite it having a positive expected value. Even if I offer £20 against your £10, a lot of people will still not take that bet, despite the expected value of the bet now being £10 in their favour.

This particular fallacy is called loss aversion. People tend to weigh losses about twice as highly as gains. And this makes some sense from an evolutionary standpoint. If you're in a forest, and you have some amount of food, it does make sense that you'd value having that food much higher over, say, using it as bait in the hopes of getting more in return. A prehistoric human willing to take that bet, probably ends

* More realistically, the problem was that I needed a better linter that would point out simple missing colons in the first place.

† "Soccer" to any North American readers.

up hungrier on certain days than someone not taking that risk and having a more consistent supply of food. That consistency is more valuable, so, we're risk averse.

Now, what if I offered that £11 against your £10 again, but we play the game a hundred times instead (or any other choice of large number). People are seemingly even less willing to play this version, but mathematically, you are effectively guaranteed to gain money this time around.*

It's not just probability either; another example where our cognitive biases lead us astray would be how people tend to put a lot of effort into getting £5 off of a £10 purchase, but not a lot into getting £5 off of a £1,000 purchase. It's £5 saved either way, but people tend to think proportionally; £5 from £10 is 50% off, while £5 from £1,000 is basically nothing. But in both situations, it's the same amount saved. And again, this makes sense – the question "is there one lion over there, or two?" is very different than, "are there 101 lions or 102?". Proportions matter for much of our prehistoric lives, and it's hard to fight against it, despite it not mattering in modern matters like money.

Moreover, maths is in some senses even more abstract than language, which was essential for every person to learn for survival. Being unable to keep track of exactly how much food you had, etc. – probably made life harder. Being unable to communicate – killed. Mathematical reasoning wasn't particularly at the top of the importance tree for much of our history. Note, I'm not an evolutionary anthropologist, so take all of this with a grain of salt, but you get my point.

However, in our modern technological world, it's become increasingly important that the average person is able to do maths – having a good foothold in logical thinking is increasingly becoming a necessity.

So, when we teach children about quadratics and trigonometry, it's not really about the x squareds or the triangles and circles. Mathematics isn't about the numbers – not really. It's about the ways we eliminate bias and logical fallacies; and the methodical thinking and the approach you take to solving problems – and not just random problems on an exam either. One of the greatest tools one learns from mathematics is the ability to abstract away unimportant details and process information in a logical manner. Once you learn to think mathematically, you'll have a different perspective on everything, and not just numbers or shapes.

That doesn't mean you won't get it wrong sometimes. You'll run into dead ends – getting the occasional answer wrong, or still screwing up with missing colons in code – but you shouldn't stop because of that. As a mathematician, you'll probably get it wrong than most people, but hopefully only because you're asking more questions and trying more problems than most. But you shouldn't fear those dead ends, because they're what you learn the most from. I see people afraid to do maths, because they're scared that they'll do it incorrectly. And, yes, that's a fair point when you're staring down the barrels of a scary exam paper. I get it. But that doesn't mean you shouldn't try. Everybody starts somewhere.

When you reduce the discussion to "nobody actually uses quadratics", removing that level of maths from the curriculum might sound like a good idea.

When I said that sounds like a reasonable idea, I mean it. It *sounds* reasonable – just like many other ideas that sound reasonable when you only explore the surface details. But if you look deeper at what mathematics actually teaches, you can see what a terrible loss that would be.

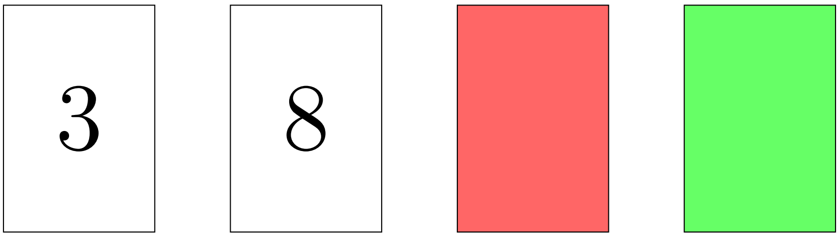
1.2 Mathematics is Hard

As mentioned above, the human brain is not at all optimised to perform formal mathematical reasoning, which is why most of mathematics has only been developed relatively recently, and why even basic arithmetic, or even counting, takes years of studying.

However, mathematical reasoning is very similar to legal and social reasoning, which we are often much happier to do.

* See the weak law of large numbers, or Chebyshev's inequality.

Say you are presented with these four cards:



Each card has a number on one side, and a colour on the other. Which card(s) must be turned over to verify the proposition that if a card shows an even number on one face, then its opposite face is green?

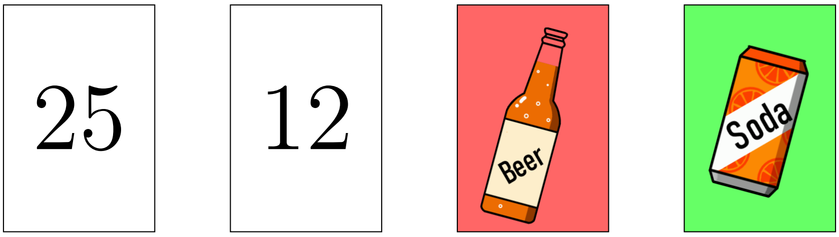
Take a moment to consider this problem* before continuing on.

The statement says that *if* a card shows an even number on one side, *then* the opposite face is green. The statement does *not* make any claims about what happens when a card has an odd number on it, nor does it make any restrictions on what other cards may bear the green colour. This means that the 3 card and the green card do not need to be checked, because the rule is not relevant in these cases.

The only way for the proposition to be false is if there is a card with both an even number *and* a red face; so, we need to check the 8 card to see if it has a red face, and also the red card to check if its other face is not even.

If that seemed somewhat confusing, consider this next problem:

Each of the following cards has an age on one side, and a beverage on the other.



Which card(s) must be turned over to verify the proposition that if if you are drinking alcohol, then you must be over 18?

This time, the problem is easy; make sure the person under 18 isn't drinking anything alcoholic, and check that person drinking beer isn't under 18. The 25 year old can drink anything, and the soda can be drank by anyone.

We intuitively understand that while under-18s can only drink non-alcoholic beverages, they aren't the only ones who *can* drink them, and similarly, while over-18s can drink alcoholic beverages, they don't *have* to.

Formally, these two problems are identical: there are two possible states for the front and back of each card, and the proposition to be verified relates the two in equivalent ways. The difference is that the latter problem is based on a real life situation, while the first is much more abstract, making the logical structure more difficult to understand for most people. Various other similar alternative formulations

* This problem is known as the *Wason selection task* or the *four-card problem*, and is a famous puzzle in deductive reasoning.

of this problem into other situations about policing a social rule are answered overwhelmingly correctly when compared to the abstract problem, which was answered correctly less than 10% of the time in study it was originally posed in.

Similarly, there is very little structural difference between the two sentences,

- $x \in S \to x + 1 \in S$ – if x is in S, then $x + 1$ is in S.
- If x is of royal blood, then x's child is of royal blood.

Because the first is about boring numbers and sets, while the second is about an interesting and historically important social construct, most people have a much easier time when deducing that to show that someone is a royal, we need to start from some known royal, then follow the line of descendents, than they have when deducing that to show a number is part of a set S, we need to start with some known element of S, then repeatedly add 1, despite the underlying logic being identical.

These kinds of questions tend to produce correct responses when posed in the context of social relations, particularly when it's about a social rule about exchange, involving benefits or restrictions that only certain people can claim, such as alcoholic beverages and being of sufficient age.

Part of this difference may be due to familiarity: we are taught all of these social constructs from a young age. In mathematics, this familiarity comes from practice, but importantly, also from constant exposure to different areas of mathematics. The wider your breadth of knowledge, the more likely it is that a new piece of information will correlate with something you already know, making it easier to retain.

But another reason, is that we're hardwired to understand social rules, given our history of highly structured social hierarchies, especially when compared to other animals.[*] As shown above, we are generally very good at policing these rules, given the right context.

There are two main things you need in order to become good at mathematics. That is, the creative kind of mathematics that involves logical connections and writing proofs, rather than kind that involves the mechanical application of rules and methods, as is generally the form that is taught in schools. One of them is *mathematical maturity* – which, among other things, is the ability to read and write the brief but concise language that mathematicians use to communicate information. Importantly, this includes the usage of clear notation when useful, and its omission when not. That is, use words, and don't "symbol spaghetti" unnecessarily. Again, this is learned through practice and exposure.

The other, is to learn how to activate the parts of your brain responsible for policing social and legal rules – the parts that make the second formulations of the problems above much easier to answer. To do this, it may be helpful to get a little angry, and imagine that finishing that annoying proof, or finally learning that damned definition is the only thing that that will prevent your worst enemy from unfairly taking – *stealing* – some prize or reward that *you* rightfully *deserve*. Every time you're stuck on a problem, curse this enemy, then show how much better and cleverer than them you are by finishing it off. If you do not have an arch-nemesis, I am happy to serve as such[†] until you find a suitable person yourself. However you choose to do this, this part of your brain can certainly become a powerful tool for computation and logical thinking.

1.3 Exercising

In some sense, university mathematics follows on from college/school, simply continuing the topics covered there. On the other hand, university mathematics diverges heavily in style, with a much larger focus on abstraction and rigourous proofs, rather than just mechanically applying rules in sequence.

Whatever your reason, you need to be fully engaged with whatever you're studying. Your brain, optimised to save as much cognitive effort as possible, is smart enough to know when you don't really care about

[*] Again, I'm not an evolutionary anthropologist.

[†] As your arch-nemesis, I will act the part by ignoring any emails sent my way.

something, and will begin to think about other things if you are not engaged with your work. If you find this happening, go and do something else for a while – forcing yourself to work isn't helpful – you won't get anything done now, you likely won't be productive later on, and you'll probably trick yourself into thinking that you spent your time well, making you more likely to skip working later.

For this reason, there are exercises throughout this text. As much as we all hate "the proof is left as an exercise to the reader", this is to prompt you to actually practise the material being taught. The best time for practice is just after being taught – not the day or week after. If you at least start working on something right now, it makes it far easier to pick back up when you inevitably forget it later than if you never practised at all. It is recommended to do, or at least start to attempt, the exercises promptly upon reaching them, as they are generally placed immediately following the material required to answer them, and will act as a self-check that you actually understood the section, rather than passively consuming the text, then assuming you've properly internalised everything. Mathematics, as they say, is not a spectator sport.

An additional section of exercises is also included at the end of each chapter, with these questions generally covering the entire chapter, or otherwise requiring application of knowledge from various sections at once. Some questions will require additional research, outside of this document. This is again to encourage you to develop a wide breadth of knowledge.

This document is an introduction to various mathematical topics, but is not a study guide. You need to find the best way for you to study, yourself.

1.4 Prerequisites

Ideally, I'd like anyone at any level of mathematical education to be able to pick up something from this document, but, as you and I both have finite time of existence, a line has to be drawn somewhere. Unfortunately, for most of the later topics covered in this document, that line happens to lie roughly somewhere past the level of basic fluency with complex numbers.

Of course, most of the symbolic logic in the first chapter is highly foundational in nature, and therefore requires little prior knowledge. For these sections, only some mathematical maturity and concentration is required.

However, to comfortably read through everything in this document, the following is a list of loose prerequisites. But again, it should be emphasised that it is not expected that you should understand everything on first reading.

Foundations.

- Basic propositional logic: if... then..., simple converse, negation, contrapositive, etc.
- Competence with reading symbolic predicate logic.
- Competence with routine algebraic manipulations; completing the square, partial fractions, etc.
- Definition of basic structures: sets, relations, functions, images, etc.
- Familiarity with basic proof techniques; direct proof, contradiction, contrapositive, construction, etc.
- Basic set operations and proofs.
- Circular and hyperbolic trigonometric functions.
- Common theorems and results in analysis, such as the factor/remainder theorems, intermediate value theorem, Cauchy-Schwarz inequality, etc.

- Taylor's theorem (without error bounding) for functions of single real variables.
- Simple and strong induction.
- Polar, spherical, and cylindrical coordinates.
- Complex numbers.
  - Complex arithmetic.
  - Geometry of complex numbers. (Interpretation of complex multiplication as rotation.)
  - Euler's identity.
- Parametrisation of simple 2D curves, vector lines, planes and half-spaces.

Calculus.

- Standard differentiation techniques.
- Standard integration techniques.
- Finding critical and stationary points.
- Solving first and second order linear ODEs.
- Solving coupled first-order ODEs.
- Familiarity with recurrence relations.
- Some familiarity with local phase portraits.

Linear Algebra.

- Matrix-vector and matrix-matrix multiplication. (We will rederive definitions these from geometric principles, so not a strict prerequisite per se, but having prior computational experience is certainly helpful.)
- Dot and cross products.
- Determinants and inverses.
- Interpretation of matrix columns as basis vectors.

Statistics.

- Venn and Euler diagrams.
- Naïve set theory.
- Common probability distributions, such as binomial, normal, and poisson.

Mechanics.

- Vectors and vector calculus.

Discrete.

- Reasonable competence with common programming patterns, including recursion and iteration, but not of more advanced techniques like dynamic programming/memoization or anything to do with functional programming (if you know what a monad is, please get in touch).
- Knowledge of imperative paradigms, such as procedural and basic object-oriented programming, but *not* of declarative paradigms, such as functional or logic programming.
- Familiarity with programming concepts, such as control flow, local/global scoping, variable binding, etc. (these last two notions are particularly applicable in mathematics.)

- Simple data structures, such as stacks, queues, linked lists, etc.
- Basic combinatorics.
- Simple algorithms on graphs and networks.
- Basic familiarity with concepts in asymptotic analysis.

1.5 Content Outlines

The author heavily prefers abstract algebra, topology, symbolic logic, foundations, and theoretical computer science over physics, mechanics, statistics, analysis, or calculus. Because of this, not much of those topics outside of the very basics is covered, and as a side effect, almost no applied mathematics is included outside of examples.

The following are not necessarily in order, pedagogical or otherwise.

1.5.1 Mathematical Logic

The basic building blocks of mathematics. If axioms are syntax and grammar, then these are words, which which we build up the rich language of mathematics.

- Symbolic logic.
- Propositional logic.
- Predicate logic.
- Axioms, theories, and models.
- Proofs.
- Provability.

1.5.2 Set Theory

The most common choice for the foundations of mathematics. As we will learn, everything in mathematics is really a set. Even numbers, whether natural, real or complex. Functions are sets too. Really.

- Naïve set theory.
- Predicates and sets.
- Set operations.
- Algebra of sets.
- Set comprehension.
- Axiomatic set theory.
- Countable and uncountable sets.
- Constructing all of mathematics from sets.
- The von Neumann hierarchy of sets

1.5.3 Relations

The first step towards building a function out of sets. Also massively useful in every part of mathematics in the form of equivalence relations. Also helpful for defining the real numbers, as well as ordinal numbers for those of us wanting to work with transfinite values.

- Equivalence relations.
- Equivalence classes.
- Orders: total orders, partial orders, lattices, well-orders, order-types, and ordinals.

1.5.4 Functions

Ubiquitous throughout maths. Sets are much more dull without them. In some axiomatisations (see chapter on category theory), they're more fundamental than sets themselves.

- Functions as sets.
- Injectivity, surjectivity and bijectivity.
- Cardinality: countable and uncountable sets.

1.5.5 Iterated Notation

Generalising binary operations to arbitrary arity.

- Formal Definitions.
- Indexing and scoping.
- Double summations.
- Closed forms and standard results.
- Products.
- Set operations.
- Logical operations.

1.5.6 Induction

Taking advantage of the recursive definition of the natural numbers to prove statements indexed over the naturals.

- Simple induction.
  - Base cases.
  - Multiple counters.
- Strong induction.
- Backward-forward induction.
- Transfinite induction.

1.5.7 Number Theory

The study of integers and integer-valued functions.

- Modular arithmetic.

- Primes and divsibility.
- (Extended) Euclid's algorithm and inverses.
- Chinese Remainder Theorem.
- Fermat's Little Theorem.
- Euler's Theorem.
- Galois Fields.
- Induction and recursion.

1.5.8 Abstract Algebra

Analysing the structure of sets and operations themselves, with less focus on the specifics. The importance of the concept of an isomorphism cannot be understated.

- Groups.
- Rings.
- Fields.
- Field extensions.
- Galois Theory.
- Categories.

1.5.9 Linear Algebra

Applicable everywhere we have a notion of adding two things together to get a third, or scaling them. A sub-branch of abstract algebra.

- Vectors.
- Cross and dot products.
- Linear Maps.
- Scalars and fields.
- Vector spaces.
- Change of basis.
- Abstract vector spaces.

1.5.10 Real Analysis

Formalising the definition of limits, convergence, differentiability and integrability.

- General and real inequality axioms.
- Sequences and limits.
- Completeness.
- Boundedness.
- Axioms of the real numbers.
- Bolzano-Weierstrass theorem.

- Cauchy Completeness.
- Alternating and general series.
- Riemann's Rearrangement theorem.
- Intermediate Value Theorem.
- Extreme Value Theorem.
- Mean Value Theorem.
- Power series.
- Limits of functions.
- Derivatives.
- L'Hôpital's rule.
- Cauchy's Mean Value Theorem.
- Taylor's Theorem.

1.5.11 Advanced Real Analysis

More and stronger forms of continuity. Formalising the Riemann integral.

- Riemann integration.
- Improper integration.
- Pointwise convergence.
- Uniform convergence.
- Series of functions.
- Space filling curves.

We will *not* cover any material on functional analysis, harmonic analysis, Fourier analysis, variational analysis, integral operators, Banach spaces, or Hilbert spaces.

1.5.12 Topology

The study of continuity and continuous maps. The foundations of analysis.

- Normed spaces.
- Metric spaces.
- Lebesgue spaces.
- Open and closed sets.
- Topological spaces.
- Closures, interiors and boundaries.
- The Hausdorff property.
- Continuity and homeomorphisms.
- Uniform and Lipschitz continuity.
- Open covers and compactness.

- Tychonov's theorem.
- Heine–Borel theorem.
- The fundamental group.
- Path-connectedness.
- Connectedness.
- Completeness.
- Quotient and product topologies.
- Retractions.
- Liftings.
- Homotopy.
- Brouwer's fixed point theorem.
- Borsuk-Ulam theorem.

1.5.13 Algebraic Topology

Studying topological spaces with algebraic invariants. Now we're getting somewhere.

- Seifert-van Kampen theorem.
- Projective spaces.
- CW complexes.
- Homology theory.
- Δ-complexes and simplicial homology.
- Chain complexes.
- Singular homology.
- Jordan curve theorem.
- Manifolds.
- Homology and orientation.
- Cellular homology.
- Cohomology.
- Poincaré duality.

1.5.14 Calculus

Seeing how sensitive functions are to nudges in their inputs. Very helpful for physics, particularly in the form of differential equations.

- Geometric and physical intuitions.
- Derivative notations.
- Differentiation from first principles.
- Implicit differentiation.

- Power, Chain and Product rules.
- Fundamental theorem of calculus.

1.5.15 Differential Equations

Seeing how sensitive functions are to nudges in their inputs. Very helpful for physics, particularly in the form of differential equations.

- Differential Equations.
- Existence and Uniqueness of solutions.
- Recurrence Relations.
- Systems of Linear ODEs.
- Variation of Parameters.
- Laplace transformations.

1.5.16 Vector Calculus

Calculus, but now with more dimensions and with strange new quantities and matrices.

- Curves and Parametrisations.
- Parametric surfaces.
- Vector Calculus.
- Frenet-Serret frame.
- Multivariable scalar-valued functions.
- Double and triple integration.
- Change of coordinate systems.
- Vector fields.
  - Divergence and curl.
  - Surface integrals.
  - Divergence theorem.
  - Line integrals.
  - Circulation and Stokes' theorem
  - Green's theorem.

1.5.17 Complex analysis

Despite the name, most of these functions are simpler and far more well-behaved than their real counterparts.

- Complex differentiability.
- Analytic functions.
- Holomorphic functions.
- Meromorphic functions.

- Cauchy-Riemann conditions.
- Picard's theorem.
- Cauchy's integral theorem.
- Contour integration.
- Laurent series.
- Liouville's theorem.

1.5.18 Combinatorics

Counting things, but formalised. Very useful for... everything really; "One starts out in life trying to do mathematics, and winds up doing combinatorics." — Ian Macdonald.

- Pigeonhole principle.
- Enumerative combinatorics.
- The twelvefold way.
- The inclusion-exclusion principle.
- Boole's inequality.
- Bell numbers.
- Stirling numbers.
- Catalan numbers.
- Young diagrams.
- Recurrence relations.
- Generating functions.
- Graph theory.
- Ramsey theory.

1.5.19 Complexity Analysis

Very helpful for computer scientists and logicians. Allows us to quantify how bad and inefficient all of our algorithms are. Not to be confused with *complex analysis.*

- Landau symbols (big O notation, and related).
- Master Theorem.
- Complexity classes.
- Polynomial time solvability.
- Boolean satisfiability.

1.5.20 Combinatorial Optimisation

Attempting to find an optimal solution when the search space for a problem is so large that exhaustive searches and optimal algorithms are intractable.

- Graph theory.

- Optimisation.
  - Linear programming.
  - Simplex algorithm.
  - Critical path analysis.
  - Transportation and allocation.
- Independence systems and matroids.
- Approximation algorithms.
- Lovász local lemma.

1.5.21 Graph Theory

The study of abstract structures known as graphs. The basis of many data structures in computer science. Helpful in the same way that abstract algebra is.

- Definitions: graphs, paths, circuits.
- Searching algorithms.
- Route inspection.
- Edge matchings.
- Vertex covers.
- Hall's condition.
- Ramsey theory.
- Travelling salesman problem.
- Network flow.

1.5.22 Probability & Statistics

Makes you a better gambler.

- Sample Spaces.
- Conditional Probability.
- Independence.
- Law of Total Probability.
- Bayes' Theorem.
- Expected Value.
- Variance.
- Probability Distributions.
- Probability Mass Functions and Probability Density Functions.
- Binomial, Poisson, Normal, Negative Binomial and Geometric distributions.
- Law of Large Numbers.
- Central Limit Theorem.

- Binomial approximations.
- Probabilistic inequalities.
  - Markov's Inequality.
  - Chebychev's Inequality.
  - Chernoff bounds.
- Jointly distributed random variables.
- Game theory.
- Decision Analysis.

1.5.23 Lambda Calculus

If you think functions are cooler than sets. Closely related to type theory and category theory.

- Prefix notation.
- Lambda terms.
- Free and bound variables.
- Data types.
- Church numerals.
- Recursion.
- Combinators.
- Simply-typed lambda calculus.
- Type theory.

1.5.24 Category Theory

If your abstract and universal algebras aren't abstract or universal enough.

- Categories.
- Duality.
- Commutative diagrams.
- Categorical isomorphisms.
- Functors.
  - Covariance and contravariace.
  - Full and faithful functors.
  - Adjoint functors.
  - Hom functors.
  - Representable functors.
- Natural transformations.
  - Vertical and horizontal compositions.
  - Interchange law.

  - Equivalence of categories.
- The Yoneda lemma.
- Universal properties.
- Universal elements.
- Cones and cocones over a diagram.
  - Limits and colimits as universal cones.
  - Products and coproducts.
  - Pullbacks and pushouts.
  - Limit representations.
  - Functoriality of limits.
- Completeness.
- Adjunctions.
  - Units and counits.
  - Limits and colimits as adjoints.
  - Calculus of adjunctions.
- Ends and coends.
- Monads.
  - Monadic adjunctions.
  - Monadic functors.
  - Free algebras.
- Subobjects and power objects.
- Monoidal categories.
- Cartesian and cocartesian monoidal categories.
- Internalisation.
  - Internal homs.
  - Monoidal closed and cartesian closed categories.
  - Internal categories.
- Topoi.
  - The Elementary Theory of the Category of Sets.
  - Internal logics.
  - Intuitionistic higher-order theories.
  - Categorical semantics and type theories.
  - Sheaves and Grothendieck topoi.
- Higher category theory.
  - 2-categories.

  - Infinity categories.
  - Groupoids.
- All Concepts are Kan Extensions

Chapter 2

Mathematical Logic

"Contrariwise, if it was so, it might be; and if it were so, it would be; but as it isn't, it ain't. That's logic."

— Lewis Carroll, *Through the Looking-Glass*

Reality is messy and complicated, but we need a way to be able to talk about it. In particular, numbers are strange abstract things we can't see or touch, and can even be infinite in size, depending on the computation in question.

So, we need a language that allows us to reason and manipulate these things that we can't really comprehend, but we also want this language to follow strict well-defined rules that everyone can agree on. Two different people using the same set of rules should always come to the same conclusions, given the same starting information.

We could create a new framework for every new problem we encounter, but modern mathematics has mostly settled into using one set of standard rules, and, taking inspiration from the first great virtue[*] of a programmer, we are going to *steal their code.*

2.1 Symbolic Logic

Concept		Model		Theory
piles of rocks				
herds of sheep	→	$\mathbb{N} = \{0,1,2,\ldots\}$	→	$\forall x, \exists y : y = x + 1$
tally marks				

We want to model a concept or something we see with something simpler that encapsulates the important part, and only the important parts, of whatever is being modelled. In the example above, the central concept to be modelled is counting – the things actually being counted doesn't matter, so that information is not included in our model. The problem is, if the model is very big or complex, such as, say, every single model used in modern maths, how do we know that what we say is valid?

2.1.1 Axioms, Models and Inference Rules

The most commonly used approach is to create a list of statements called *axioms* which we define to be true, and a list of *inference rules* that let us derive new statements from existing statements. Together, axioms and inference rules generate a *theory* consisting of all the statements that can be constructed

[*] Laziness, Impatience, Hubris

from the axioms by applying the inference rules. All the statements within a theory that are not axioms are called *theorems*.

Example.

- All men are mortal (axiom);
- Socrates is a man (axiom);
- If "all A are B" and "C is A", then "C is B" (inference rule);
- Therefore, Socrates is mortal (theorem).

△

We can't do anything further with these axioms with our inference rule, so these three statements form our entire theory about Socrates, men and mortality.

Theories describe *models*, collections of objects and relations between those objects. For any given theory, there may be many models that are consistent with it. For example, a model that includes both mortal Socrates and an immortal Hyperion is consistent with the theory above, because the theory doesn't say anything about Hyperion or things that are not men.

If we define too many axioms, we can start to get inconsistencies: "All men are mortal; all gods are not mortal; all gods are men; Socrates is a man." immediately leads to a contradiction between the axioms. Obviously, this example is rather simple, but for a different set of axioms, it may take a while for the inconsistencies to start showing up. Thus, we try to use as few axioms as possible.

On the other hand, insufficiently many axioms underconstrains the model. To try construct the natural numbers, we might say that 0 is a number, and that any number x has a "successor", denoted $S(x)$. While the natural numbers satisfy these axioms, it isn't the only model that works. Our model could just consist of 0, with $S(0) = 0$. This obviously is not what we want, so our next axiom could be that $\forall x\,(S(x) \neq 0)$, but we could get stuck with $S(0) = 1 = S(1)$. Adding in $S(x) = S(y)$ if and only if $x = y$, then we get the natural numbers as desired, but also some extras, say $\alpha, S(\alpha) = \beta, S(\beta) = \alpha$.

One thing to note is that an axiom we come up with isn't more likely to be true if it better explains or predicts what we observe in some physical universe. Axioms are true because we say they are, and their consequences and the theories we build on top of them just become what we observe. For example, while the axioms above include extra numbers as well as what we usually mean by naturals, there's nothing inherently wrong with having the extra numbers. We could build a new branch of maths around this new theory.

We don't have to fit our theories to a physical universe, whose behaviours and underlying laws would be the same whether we were here or not. When we define a set of axioms, we create a new universe, ourselves. If the axioms we declare to be true lead us to contradictions or paradoxes, we can tweak the axioms, write new ones, or just drop them entirely – or, we could just refuse to allow ourselves to do the things that cause the paradoxes; in a lot of contexts, we don't allow ourselves to divide by zero.

2.1.2 Standard Axiom Systems and Models

That being said, there are certain properties that we like our axiomatic systems to have. Since most of modern mathematics is built around one axiomatic system, it's useful to prove your results in that system, or at least, in an equivalent axiomatic system, rather than some arbitrary system of your own, so that other people can apply your results to their own theories.

One such property is *consistency*. A theory is *consistent* if it cannot simultaneously prove P and not-P for all P. An inconsistent theory can prove anything to be true,* given that P and not-P are true, so

* This is called the *principle of explosion*: $P, \neg P \vdash Q$: if P and not-P are true, then the statement $P \vee \neg P \to Q$, or "If P or not-P, then Q", says that Q is true.

consistency is basically a required property for a theory to be of any use.*

As we saw above, it is rather tricky to exactly nail down the right axioms to do what we want. We didn't even manage to define the naturals properly, and we haven't even discussed what inference rules we use.

Fortunately, mathematicians and logicians have been working on this for a very long time, and they've managed to define most thing we care about in ways that are both consistent and useful to work with. So, rather than defining our own axiom systems and models from the ground up, we'll copy and paste their axiom systems.

Almost all of modern mathematics fits within one of the following models:

- The natural numbers $\mathbb{N}$, defined using the *Peano axioms*. If all you want to do is count, add or multiply, this is generally sufficient.
- The integers, $\mathbb{Z}$. The naturals, but now we can subtract. Division is still problematic.
- The rational numbers $\mathbb{Q}$. Now we can divide. But what about $\sqrt{2}$?
- The real numbers $\mathbb{R}$. Now we have $\sqrt{2}$. But what about $\sqrt{-1}$?
- The complex numbers $\mathbb{C}$. We are pretty much done here.
- There are further extensions to these number systems, such as the hyperreals, hyperbolic numbers, hypercomplex numbers (of which Hamilton's quaternions are an example), p-adic numbers, transfinite numbers, and more, but we'll draw the line here, as these number systems are too specialised for an introductory document such as this one. At most, we may have a peek at some of these later.
- The von Neumann universe of sets. Defined using the axioms of set theory, the universe of sets contains a rich variety of sets, which include, among a *lot* of other things, structures equivalent to all of the above systems.

  This is generally what is used in modern mathematics, the idea being that we start with sets, and define everything else in terms of sets. If we have a good set of axioms that we trust are consistent with sets, then everything we construct from sets should also be consistent. The only problem is in doing the construction; we've already seen how nuanced and difficult this can be – if not careful, the structures we create may not be what we think they are, like our "natural numbers" from before including a bunch of extra numbers.
- We also have completely alternative systems to set theory, such as second-order or higher-order logics, lambda calculus or (topoi) category theory, but these won't be covered in detail here, as, while they provide interesting alternative ways to look at structures, to the end user who just wants to use those structures, they generally don't allow you to do anything that you couldn't already do with sets.

  Lambda calculus is, however, very important for the study and implementation of computation theory and functional programming, and category theory is an extension and unification of all abstract algebras, and more. Category theory and the lambda calculus have their own dedicated chapters, §51 and §50, respectively.

The two main systems we will discuss are *propositional logic*, and *predicate logic*. Propositional logic is concerned with models that contain statements which are either true or false called *propositions*. We can

* Some systems actually do allow for statements to be both true and false. This is usually not useful due to the principle of explosion mentioned previously, but some devious logicians have managed to create axioms for systems that allow for statements to be both true and false, while also preventing logical explosions. Such systems are called *paraconsistent logics*. One reason as to why such a thing is helpful is in the resolution of certain paradoxes. For example, [This proposition is false] (the Liar's paradox), or ["yields falsehood when preceded by its quotation" yields falsehood when preceded by its quotation] (Quine's paradox) aren't problematic propositions in paraconsistent logics.

connect propositions together and write a few equations, but for this most part, there isn't much else to do here.

Predicate logic, on the other hand, allows us to use constants which represent object in our model, and *predicates*, which is a kind of function that takes an object, and returns true or false, but can also be thought of as properties that objects can have or not have.

Alternative names for propositional and predicate logics are *zeroth-order* and *first-order* logics, and of course, there are higher-order logics as well, which allow us to be even more expressive with our statements. However, for most of modern mathematics, first-order logic is sufficient.

2.2 Propositional Logic

In *propositional logic* or *zeroth-order logic*, we deal with statements called *propositions* and *logical connectives* between them. Propositions cannot contain variables, and are therefore either always true, or always false. We also use the symbols, $\top$ and $\bot$, or 1 and 0, for true and false, respectively.

Propositions:

- $2 + 2 = 4$ (always true).
- $2 + 2 = 5$ (always false).
- "Socrates is a man" (always true).
- "Socrates is a dog" (always false).

Non-propositions:

- $x + 2 = 4$ (either true or false, depending on the value of x).
- $0x = 0$ (always true, but not a proposition because it contains a variable).
- $0x = 1$ (always false, but still not a proposition).
- "Socrates" (this is an object and doesn't have a truth value by itself).

Notably, in propositional logic, the proposition "Socrates is a man" is an indivisible atom of truth or falsity that says nothing about "Socrates" or "[being] a man" individually. Because it is an indivisible statement, we can represent the whole proposition with a single letter, for example, p. We cannot, however, represent either individual part alone.

Such an indivisible proposition is called an *atom*, an *atomic formula* or a *literal*. Literals can also be divided into positive and negative *polarities*, where a negative literal is the negation of a positive literal; i.e., "p" is a positive literal, and "$\neg p$" is a negative literal. Positive and negative literals are also called each other's *complementary* literals.

2.2.1 Logical Connectives

Propositions in isolation are not very interesting. So much so that we often don't even consider specific propositions, and just refer to general ones with letters, often p and q. We can make these propositions slightly more interesting by combining them with logical connectives into *compound propositions.*

- *Negation* or *NOT* – the negation of p is written as $\neg p$ or sometimes $\bar{p}$. It is false when p is true, and true when p is false. This is pretty much the same as in normal conversation.
- *(Inclusive) Disjunction*, *Join* or *OR* – the disjunction of p and q is written $p \vee q$, and is true if at least one of p and q is true.

Note that this is different than how we often use "or" in normal conversation: if I were to, completely truthfully, say "You will give me your wallet, or I will stab you with this rusty kitchen knife", you would be understandably quite upset if you handed me your wallet and still get stabbed. However, to a logician mugger, this would be entirely justified, as the first part of a true inclusive disjunction being true doesn't preclude the second from also being true.

- *Exclusive Disjunction* or *XOR* – the exclusive disjunction of p and q is written as $p \veebar q$ or $p \oplus q$, and is true if exactly one of p or q is true. Exclusive disjunction is not often used in classical logic, but has many important applications, particularly in computing and finite field algebra.

  To indicate exclusive disjunction, we sometimes use the wording, "either p or q", to distinguish it from inclusive disjunction. Now, if you are ever being mugged by a logician, you know what to ask to clarify your chances of being stabbed.

- *Conjunction*, *Meet* or *AND* – the and of p and q is written as $p \wedge q$, and is true when both p is true and q is true. This one is generally the same as in common speech.

- *Material Implication* or *Material Conditional* – This is perhaps the most important connective for proofs, corresponding to the "If... then..." pattern of speech. The implication of p and q is written $p \to q$ or $p \Rightarrow q$. p is called the *antecedent* or *premise*, and q the *consequent* of the implication. The implication is true when (p is true and q is true), or when p is false. In fact, the only way for $p \to q$ to be false is if p is true, but q is false, so another way to write this is $\neg p \vee q$.

  $p \to q$ being true when p is false but q is true often causes some surprise; after all, if p is false, then how can it claim any credit for q being true? Both statements being false also leading to the compound being true also seems somewhat suspect.

  This surprise might be because in ordinary language, we usually aren't interested in implications where the first proposition is known to be false, so we don't usually think to assign them any truth values. However, one reason why it's nice to define the truth values in this way is that we often use the implication symbol in this way. For example, we should all agree that the proposition,

  $$\forall x \in \mathbb{Z} : (x > 1) \to (x^2 > 2)$$

  is true. (We haven't met the mysterious symbol, $\forall$, yet, but here, it means "For all integers x, the proposition ... holds.". It isn't a valid symbol within propositional logic, but we'll soon move onto predicate logic, where we will encounter it frequently.)

  The statement contains infinitely many implications – one for each integer – so included within it is the statement $0 > 1 \to 0^2 > 2$, where the antecedent and consequent are both clearly false, but we still say that the proposition is true overall. Because of this, we define $p \to q$ to be true whenever p is false, regardless of the value of q.

  Implication can also be thought of as a promise "if you do p, then I will do q". If you fail to do p, then regardless of whether I do q or not, I will not have broken my promise – so an implication is always true if the antecedent is false.

  In ordinary language, we often interpret "if... then..." to be the much stronger *biconditional* where it otherwise carries connotations of causality. This is another reason why we define our terms so stringently in mathematics and logic, due to natural language being rife with hidden rules and assumptions.[*]

  I could once again, entirely truthfully, say, "If the moon is made of green cheese, then the world will end at midnight". It may sound like I know of some mechanism by which a green-cheese moon will cause the end of the world, but I am simply making a trivially true statement by starting with a false premise and violating the implicit assumption that a statement in a conversation should mean something and not just be an exercise in logic.

[*] Search up "Grice's maxims" or "the cooperative principle" for an interesting discussion on this topic.

Conditional propositions like this where the antecedent is false, are called *vacuous truths*, because the proposition is true while not really saying anything meaningful – in particular, we can't infer anything about the truth value of the consequent from a vacuous truth. These can sometimes cause seemingly incoherent statements to be true. For example, the proposition "All the lights in the room are turned on *and* turned off" is true, if there are no lights in the room to begin with. In the equation above, the proposition as a whole is considered to be true non-vacuously, since some integers are indeed greater than 1 and the proposition still holds for them, but we would say that the *cases* where $x < 1$ are vacuously true.

Alternative wordings to "if p then q" include; "p is *sufficient* for q", because knowing p is true is sufficient information to tell us that q is true; or "q is *necessary* for p", because q being true is guaranteed by p being true (or equivalently, it is impossible to have p be true without q also being true). We will generally use the "if... then..." pattern in this document, but sufficient and necessary are commonly used in other fields.

- *Material Equivalence*, *material biconditional* or *XNOR* – If both $p \to q$ and $q \to p$, such that p and q always share the same truth values, then we write $p \leftrightarrow q$ or $p \Leftrightarrow q$, and say p holds *if and only if* q holds.

  Again, however, this is purely a logical proposition, and no causality between p and q has to be enforced. For example, the compound proposition, "The moon is made of green cheese if and only if $2 + 2 = 5$" is true, despite the lack of connection between green-cheesiness and faulty arithmetic, purely because both sub-propositions are false.

  Alternatives to "p if and only if q" include; "q is necessary and sufficient for p", which is a combination of the two alternative wordings for material implication; "p precisely/exactly when q"; or the abbreviation, "p iff q". This last alternative, "iff", is sometimes regarded as unsuitable for formal writing, so a style guide should be consulted before it is used in such a setting. We will continue to use "if and only if" in this document.

Given that there are sixteen possible ways to associate two Boolean inputs to four binary outputs, you might think we've skipped over a few in the list above. But, it turn out that a lot of logical connectives are equivalent, just with swapped arguments, or other similar redundancies.

In fact, every logical connective can be expressed purely in terms of NAND (a logical connective that's basically a NOT gate glued to the output of an AND gate), and this is actually how most computer hardware is built, since it's a lot cheaper to make a lot of one gate, than fewer distinct gates. If a set of logical connectives can express all possible logical connectives, the set is *functionally complete.*

While it would be possible for us to continue using only NAND operations, or with also including the other unlisted binary connectives, it's a lot more human readable if we just stick with a few common connectives, particularly if those connectives have close analogues in ordinary language.

The list above is just of what is commonly used, but other logical connectives certainly are available.

2.2.1.1 Precedence

Now, a short sidenote should be made about *precedence*. Just as with arithmetic operations, we like to reduce the number of brackets necessary to disambiguate an expression, and we do so by introducing precedence rules, or perhaps more familarly, *order of operations.*

In many programming languages and compilers, the order is $\neg$, $\wedge$, ($\veebar$), $\vee$, $\to$, $\leftrightarrow$. There is good reason for this; some notations for $\wedge$ (conjunction) are multiplicative, while $\vee$ (disjunction) is additive, corresponding to how probabilities of events are calculated,[*] and assigning a higher precedence to conjunction

[*] The probability of the events A and B both happening is equal to the product of the probabilities, while the sum represents the probability that A or B happens, assuming A and B are disjoint.

is analogous to multiplication having higher precedence than addition in ordinary arithmetic, making logic statements easier to parse.

However, this is not a universal convention. Particularly in mathematics, we often assign $\wedge$ and $\vee$ the same precedence, due to something called *duality*,[*] which effectively says that $\wedge$ and $\vee$ are symmetric under certain transformations. Assigning the two the same precedence emphasises this symmetry, and is particularly nice to use when studying logic symbols as an algebraic structure.

There isn't a commonly agreed upon standard convention for the precedence of XOR in both programming and symbolic logic, since it is not as often used, so it is safest to just use some extra brackets to clarify formulae using it.

Both disjunction and conjunction are associative, so $p \vee q \vee r$ is the same as $p \vee (q \vee r)$ and as $(p \vee q) \vee r$, and similarly, $p \wedge q \wedge r$ is the same as $p \wedge (q \wedge r)$ and as $(p \wedge q) \wedge r$.

Implication, however, is not associative, the convention being that the operation *binds to the right* or is *right-associative*[†] such that $p \to q \to r$ is read as $p \to (q \to r)$.

As with everything in mathematics, any convention can be used, as long as it is clearly stated and doesn't annoy the target audience too much. If you are writing a programming manual, you might want to stick with the order given above. In this document, we will use the mathematics convention of assigning $\wedge$ and $\vee$ the same precedence, and clarifying with brackets.

2.2.2 Truth Tables

To fully specify the function of logical connectives, we give *truth tables*, in which every combination of truth values for inputs is assigned a truth value as an output.

Here is the truth table for negation:

p	$\neg p$
0	1
1	0

and a table for the rest of the binary connectives:

p	q	$p \vee q$	$p \veebar q$	$p \wedge q$	$p \to q$	$p \leftrightarrow q$
0	0	0	0	0	1	1
0	1	1	1	0	1	0
1	0	1	1	0	0	0
1	1	1	0	1	1	1

[*] Informally, there is nothing inherently more correct about our choice of symbols in Boolean algebra than any other. We could just as easily have named what we call 0 and 1 to say, α and β, and, as long as we do so consistently, our working would still be valid Boolean algebra, albeit with some cosmetic differences.

However, suppose we renamed 0 to 1, and 1 to 0. We would be operating on the same values, since we still only have 0's and 1's, but this relabelled Boolean algebra would not be identical to the original, as we now find that $\wedge$ in the new system behaves like how $\vee$ did in the original; there is a discernable difference between the new and old systems, despite both operating on 0's and 1's. If we also interchange $\wedge$ and $\vee$, then the new system is indistinguishable from the original, the only difference being that rows in truth tables may swap place.

When values or operations can be paired up in such a way that they can be exchanged and leave the structure of the algebra unchanged, we call them *dual*, so 0 and 1 are dual, as are $\wedge$ and $\vee$. The De Morgan duality principle asserts that Boolean algebra is unchanged when all dual pairs are interchanged.

More generally, the principle of duality is a statement about partially ordered sets and has connections to many other branches of maths, in particular, order and category theory. For instance, the De Morgan duality can be explained as a group of automorphisms, swapping certain Boolean functions around.

As another note, if we replace $\wedge$ with $\cap$, $\vee$ with $\cup$, $\neg$ with ^{C}, 0 with $\varnothing$, and 1 with U, we get set algebra, and it behaves in completely the same way as logic statements – it turns out that the algebra of sets and boolean logic are isomorphic algebraic structures.

[†] Note that while associativity is a *property* of an operation that can be proved, being left-associative or right-associative is a notational *definition*. Our convention could just as equally define implication to be left-associative.

We note that there are $2^2 = 4$ different combinations of the truth values of p and q, so this table has 4 rows. In general, there will be 2^n rows for propositions with n literals. For example,

$$(p \to r) \lor (q \to r) \leftrightarrow (p \land q) \to r$$

$(p$	$\to$	$r)$	$\lor$	$(q$	$\to$	$r)$	$\leftrightarrow$	$(p$	$\land$	$q)$	$\to$	r
0	1	0	**1**	0	1	0	**1**	0	0	0	**1**	0
0	1	1	**1**	0	1	1	**1**	0	0	0	**1**	1
0	1	0	**1**	1	0	0	**1**	0	0	1	**1**	0
0	1	1	**1**	1	1	1	**1**	0	0	1	**1**	1
1	0	0	**1**	0	1	0	**1**	1	0	0	**1**	0
1	1	1	**1**	0	1	1	**1**	1	0	0	**1**	1
1	0	0	**0**	1	0	0	**1**	1	1	1	**0**	0
1	1	1	**1**	1	1	1	**1**	1	1	1	**1**	1
0	1	0	2	0	1	0	3	0	1	0	2	0

The different typefaces are purely for visual contrast, and the additional row at the bottom shows the order in which the columns would be filled in if the table were to be drawn by hand. The columns labelled 0 contain the truth assignments, given to p, q and r, then the columns labelled 1 are calculated from the 0 columns, then 2 from 1, and so on.

You may also notice that we appear to assign truth values to connectives, which is slightly cursed, but this is just a shorthand for assigning truth values to entire expressions, including arguments; i.e., the 1 in the second column doesn't mean that the connective "$\to$" is true, it means that the proposition "$0 \to 0$" is true. This just saves us from having to draw a new table every layer up.

In propositional logic, we can think of each row of a truth table as a model for the proposition, since the only things we can describe in propositional logic is whether a proposition is true or not. Constructing a truth table then corresponds to generating all possible models.

Checking if a particular proposition is true in this way is a simple version of *model checking*: running through all possible models for a proposition, and seeing if the statement we want to prove holds in all models. This works well in propositional logic because the list of models is just the list of combinations of truth values for p, q, r, etc., and we can easily fill in further columns using the simple rules for each logical connective.

For predicate logic, this becomes much more difficult as typical predicate systems contain infinitely many models, many of which are also infinitely large. Here, we rely on proofs constructed using inference rules.

Because the central column consists entirely of 1's, we conclude that the proposition always holds, so the left and right side are equivalent.

A compound proposition that is true like this regardless of the truth values of the propositions it contains is called a *tautology*. Similarly, if the compound proposition is always false, it's a *contradiction*. These two concepts are negations of each other.

One useful class of tautologies are *logical equivalences*, which are tautologies of the form $P \leftrightarrow Q$, where P and Q are compound propositions. In this case, we write $P \equiv Q$.

Note that logical equivalences are distinct from material equivalences. The material equivalence of propositions p and q, $p \leftrightarrow q$ is itself another proposition within the same *object language* – within the same axiom system – as p and q individually. In particular, the truth value of $p \leftrightarrow q$ may differ between models. The logical equivalence of p and q, however, is a statement in *metalanguage*. p and q are logically equivalent if and only if they share the same truth value in *every* model.

Whenever two propositions are logically equivalent, they may be exchanged within a logic statement. This is useful because logical equivalence in Boolean formulae is functionally the same as equality in algebraic formulae. For example, if we know $p \wedge \neg p \equiv 0$ and $0 \wedge q \equiv 0$, then we can immediately transform $(p \wedge \neg p) \wedge q \equiv 0 \wedge q \equiv 0$ without having to do any real work.

To prove that two statements P and Q are logically equivalent, we either construct a truth table, or we transform P to Q using previously obtained logical equivalences.

For example,

p	$\neg p$	$p \wedge \neg p$	0
0	1	0	0
1	0	0	0

The last two columns are the same for all choices of p, so we know that $p \wedge \neg p \equiv 0$.

Similarly,

q	0	$0 \wedge q$
0	0	0
1	0	0

the second and third columns shows that $0 \wedge q \equiv 0$.

2.2.3 Logical Equivalences

Many commonly used logical equivalences have been given names, some of which have been listed below:[*]

$\neg\neg p \equiv p$	Double negation or involution law
$p \wedge q \equiv q \wedge p$ $p \vee q \equiv q \vee p$	Commutativity laws
$p \wedge (q \wedge r) \equiv (p \wedge q) \wedge r$ $p \vee (q \vee r) \equiv (p \vee q) \vee r$	Associativity laws
$p \wedge (q \vee r) \equiv (p \wedge q) \vee (q \wedge r)$ $p \vee (q \wedge r) \equiv (p \vee q) \wedge (q \vee r)$	Distributive laws
$p \wedge 1 \equiv p$ $p \vee 0 \equiv p$	Identity laws
$p \vee 1 \equiv 1$ $p \wedge 0 \equiv 0$	Domination laws
$p \vee p \equiv p$ $p \wedge p \equiv p$	Idempotency laws

[*] You'll notice that almost all of these come in pairs. This is due to the principle of duality discussed in a previous footnote. The main thing to remember is that, given a true statement, interchanging 0 with 1 and $\wedge$ with $\vee$ returns another true statement.

$$\begin{aligned} p \vee \neg p &\equiv 1 \\ p \wedge \neg p &\equiv 0 \end{aligned}$$

Negation laws

$$\begin{aligned} \neg(p \wedge q) &\equiv \neg p \vee \neg q \\ \neg(p \vee q) &\equiv \neg p \wedge \neg q \end{aligned}$$

De Morgan's laws

$$\begin{aligned} p \vee (p \wedge q) &\equiv p \\ p \wedge (p \vee q) &\equiv p \\ p \leftrightarrow 0 &\equiv \neg p \\ p \leftrightarrow 1 &\equiv p \\ p \veebar 0 &\equiv p \\ p \veebar 1 &\equiv \neg p \\ p \rightarrow 0 &\equiv \neg p \\ p \rightarrow 1 &\equiv 1 \\ 0 \rightarrow p &\equiv 1 \\ 1 \rightarrow p &\equiv p \end{aligned}$$

Absorption laws

$$p \rightarrow q \equiv \neg p \vee q \qquad \text{Equivalence of implication and disjunction}$$

$$p \rightarrow q \equiv \neg q \rightarrow \neg p \qquad \text{Contraposition}$$

$$p \leftrightarrow q \equiv (p \rightarrow q) \wedge (q \rightarrow p) \qquad \text{Expansion of material equivalence}$$

$$p \leftrightarrow q \equiv \neg p \leftrightarrow \neg q \qquad \text{Inverse of material equivalence}$$

$$p \leftrightarrow q \equiv q \leftrightarrow p \qquad \text{Commutativity of material equivalence}$$

Some of these deserve some special mention – in particular, the contrapositive.

For any conditional statement $p \rightarrow q$, say, "If *it is raining*, then *I wear a coat.*" we have four related propositions:

- *Negation* (the *logical complement*): $\neg(p \rightarrow q)$

  "*It is not the case that* if *it is raining* then *I wear a coat*", or "If *it is raining*, then *sometimes I do not wear a coat.*"

  The truth value of the negation is always the opposite of the original statement. If the negation is true, then the original statement is false, and vice versa.

- *Contraposition* (the *contrapositive*): $\neg q \rightarrow \neg p$

  "If *I don't wear a coat*, then *it is not raining.*"

  The truth value of the contrapositive is the same as the original proposition; a statement is equivalent to its contrapositive.

- *Inversion* (the *inverse*): $\neg p \rightarrow \neg q$:

  "If *it is not raining*, then *I do not wear a coat.*"

  The truth value of the inverse is unrelated to the original proposition, as demonstrated by the example – the original statement merely states that I wear a coat if it is raining, and makes no claims as to what I wear when it is not raining.

- *Conversion* (the *converse*): $q \rightarrow p$

> "If *I wear a coat*, then *it is raining*."
>
> The converse of a statement is actually the contrapositive of the inverse, and so shares the same truth value as the inverse (which is again, unrelated to the truth value of the original statement) – the original statement only asserts that I wear a coat when it is raining, not that I *only* wear a coat when it is raining.

The negation laws $p \lor \neg p \equiv 1$ and $p \land \neg p \equiv 0$ are also known as *the law of the excluded middle* and *the law of non-contradiction*, respectively. The law of the excluded middle states that at least one of a statement and its negation is true, while the law of non-contradiction states that a statement and its negation cannot both be true. Together, they imply that for every proposition, either the proposition or its negation is true, but not both (that is to say, that $p \veebar \neg p$ is a tautology). This is not to be confused with the stronger *law of bivalence*, which states that every proposition is either true or false.[*]

These two laws are what allow us to do case analysis, where we prove that Q holds by proving both $P \to Q$ and $\neg P \to Q$.

For example, we can prove that there exists irrational numbers a and b such that a^b is rational. We know that $\sqrt{2}$ is irrational, so consider the number,

$$\sqrt{2}^{\sqrt{2}}$$

By the law of the excluded middle, we know that this number is either rational or irrational. If it is rational, then the proof is complete, and

$$a = \sqrt{2} \text{ and } b = \sqrt{2}$$

Otherwise, let $a = \sqrt{2}^{\sqrt{2}}$, and $b = \sqrt{2}$ such that,

$$\begin{aligned} a^b &= \left(\sqrt{2}^{\sqrt{2}}\right)^{\sqrt{2}} \\ &= \sqrt{2}^{(\sqrt{2}^2)} \\ &= \sqrt{2}^2 \\ &= 2 \\ &\in \mathbb{Q} \end{aligned}$$

and we are done.

With all of these logical equivalences in mind, instead of the massive unwieldy table from before, we could prove $(p \to r) \lor (q \to r) \equiv (p \land q) \to r$ using,

$$\begin{aligned} (p \to r) \lor (q \to r) &\equiv (\neg p \lor r) \lor (\neg q \lor r) && \text{Equivalence of material implication and disjunction} \\ &\equiv \neg p \lor \neg q \lor r \lor r && \text{Associativity and commutativity of } \lor \\ &\equiv \neg p \lor \neg q \lor r && \text{Absorption law} \\ &\equiv \neg(p \land q) \lor r && \text{De Morgan's law} \\ &\equiv (p \land q) \to r && \text{Equivalence of material implication and disjunction} \end{aligned}$$

[*] We also have to be careful about vague predicates. For example, consider a red and green striped shirt. Is the proposition, "the shirt is red", true? The truth value depends on what we mean by "is". If by "is", we mean, "contains parts that are coloured", then the proposition is true. If "is" means "is completely coloured..." then the proposition is false.

2.2.4 Exercises

1. Let p be the proposition "x is a human", q be the proposition "x is mortal", and r be the proposition "x is Socrates". Formalise the following statements into propositonal logic:
   (a) If x is Socrates, then x is a human.
   (b) x being human is a sufficient condition for x being mortal.
   (c) x being mortal is a necessary condition for x being Socrates.
   (d) Either x is human and x is mortal or neither x is human nor x is mortal.
2. Identify whether each of the following statements are propositions or not.
   (a) $25 = 25$.
   (b) $x^2 \geq 0$.
   (c) π.
   (d) $2 < 1$.
   (e) $p \to q$.
   (f) $\wedge$.
   (g) $3 + x^2 = 4$.
   (h) "The proposition p is true."
   (i) "Either it will rain tomorrow, or it will rain today."
   (j) $a^2 + b^2 = c^2$.
3. Negate the statement, "Every proposition is either true or false."
4. Socrates says, "If I am guilty, I must be punished. I am not guilty, therefore I must not be punished." Is this argument logically sound?
5. Compute the truth table of $\big((p \vee q) \to r\big) \to \neg(q \wedge p \wedge \neg r)$.
6. Compute the truth table of $(p \vee q) \wedge \neg(p \wedge q)$ and identify which common logical connective this formula represents.
7. Draw a truth table to verify that the distributive law $(p \vee (q \wedge r)) \leftrightarrow ((p \vee q) \wedge (p \vee r))$ is a tautology.
8. Verify that the rule of inference $(p \wedge (p \to q)) \to q$ ("*modus ponens*") is a tautology using logical equivalences.
9. Construct the contrapositive of the statement, "It is not the case that if it will rain today, then I will wear a coat."

2.3 Predicate Logic

This isn't of too much interest to us, so we move on to *predicate logic* or *first-order logic*. While propositional logic deals with simple declarative proposition, predicate logic additionally covers *predicates* and *quantification*. A predicate takes an object as an argument, and evaluates to true or false.

For example, consider the statements "Socrates is a philosopher" and "Plato is a philosopher". These are both valid propositions in both propositional and predicate logic, but in propositional logic they are completely unrelated statements, and could be reduced down to the atoms p and q. However, the predicate "is a philosopher" appears in both statements, sharing the common structure of "x is a philosopher", and we could denote this as $P(x)$ (sometimes, for single letter predicates, we drop the brackets and simply write Px). The *variable* x is instantiated as "Socrates" in the first statement, and "Plato" in the second. Note that x is a *free* variable – x by itself doesn't represent anything, and is simply a placeholder variable for the predicate to take.

While the predicate above only takes a single variable, in general, they can take several. For example, $Q(x,y)$ = "x is the teacher of y" and $R(x,y,z)$ = "$x + y + z = 0$" are valid predicates. A predicate without a specific variable is *not* a proposition, as it does not have a fixed truth value, so $P(x)$, $Q(x,y)$, and $R(x,y,z)$ are not propositions. If specific values are given for all the variables, then it is a proposition again, and we can talk about that proposition being true (P(Socrates) and Q(Socrates,Plato) are true) or false (R(1,2,3) is false).

Statements involving predicates can also be connected using logical connectives in predicate logic. For example, the predicate, $S(x)$="x is a scholar", could be connected to $P(x)$ in the first-order formula, $P(x) \to S(x)$, or, "If x is a philosopher, then x is a scholar.". The truth of this formula depends both on x, and on the interpretation of the predicates "is a scholar" and "is a philosopher".

2.3.1 Quantification

Rather than talking about specific values of variables, we also want to be able to say when a proposition is true for several different values of their arguments. For this, we *bind* the variables with *quantifiers*. The two main quantifiers we use are the *universal quantifiers* and *existential quantifiers*.

In quantifying a proposition, the free variable becomes *bound*. For example, in the statement $\forall x \exists y$: $Q(x,y,z)$, x and y are bound, while z is free: the truth value of the proposition depends entirely on z.

2.3.1.1 Universal Quantification

The universal quantifier, $\forall$ ("For all" or "For every"), states that a statement is true for all values of the bound argument, within some *universe of discourse.*[*] For example, we could apply the universal quantifier to x in the previous formula to get, $\forall a : P(x) \to S(x)$, or, "For every x, if x is a philosopher, then x is a scholar". Here, the universe not specified, but due to the predicates implicitly being statements that apply to people, the universe could be, for example, the set of all people.

The proposition $\forall x : (x \neq 1) \to (x^2 \neq 1)$ is ambiguous, however, if no universe is identified. If the universe is the set of real numbers, then the proposition is false, with $x = -1$ being a counterexample. On the other hand, if the universe is the set of natural numbers, then the proposition is true. We can explicitly define the universe of discourse with set membership notation, for example $\forall x \in \mathbb{N} : (x \neq 1) \to (x^2 \neq 1)$. This is really shorthand for $\forall x : x \in \mathbb{N} \to (x \neq 1 \to x^2 \neq 1)$ or $\forall x : (x \in \mathbb{N} \wedge x \neq 1) \to x^2 \neq 1$, but the short form makes it clearer that the intent of $x \in \mathbb{N}$ is to restrict the domain of x. This is a form of *syntactic sugar* – notation to make things easier to express and read ("sweeter" for human use).

Universal quantification is equivalent to an infinite conjunction over the entire universe. For example, $\forall x \in \mathbb{N} : P(x)$ is the same thing as $P(0) \wedge P(1) \wedge P(2) \wedge \cdots$. While infinite expressions like this are not allowed in predicate logic, it's a helpful reminder that universal quantifiers require *every* value of the bound variable to be true, not just some or most of them.

2.3.1.2 Existential Quantification

We can also use the existential quantifier $\exists$, ("There exists... such that..."), which asserts that a statement is true for at least one value of the variable. Returning to the example from before, $\exists x P(x)$ means "There exists (at least one) x such that x is a scholar". This quantified proposition is true, as demonstrated by the existence of Socrates. As with universal quantification, we can explicitly identify the universe of discourse with set membership.

[*] If you are wondering why we need a universe of discourse, it is to prevent certain paradoxes from occuring, among other reasons. This is discussed in more depth in §4.2.1.

Existential quantification is equivalent to an infinite disjunction over the entire universe. For example, $\exists x \in \mathbb{N} P(x)$ is the same thing as $P(0) \lor P(1) \lor P(2) \lor \cdots$. Again, no infinite expressions allowed, but it's another helpful reminder of what the symbol means.

2.3.1.3 Unique Existential Quantification

The *unique existential* or *uniqueness* quantifier, $\exists!$ ("There exists exactly one... such that..."), functions almost exactly identically to the existential quantifier, but only allows for a single value of the bound variable to be true. So, $\exists! x : P(x)$ means "there exists exactly one x such that x is a philosopher", which is false, as Socrates and Plato are both philosophers.

To prove unique existence, we not only have to show that a proposition holds for some x, but that it also doesn't hold for every other x.

Unique existential quantification can also be expressed in terms of universal and existential quantifications:

$$\begin{aligned} \exists! x : P(x) &\equiv \exists x : P(x) \land (\neg \exists y : P(y) \land (y \neq x)) \\ &\equiv \exists x : P(x) \land (\forall y : P(y) \rightarrow (y = x)) \end{aligned}$$

Another definition that neatly separates the notions of existence and uniqueness into two clauses at the expense of brevity is,

$$\equiv \exists x : P(x) \land \forall y \forall z [(P(y) \land P(z)) \rightarrow (y = z)]$$

Unique existential quantification is equivalent to an infinite XOR over the entire universe, so $\exists! x \in \mathbb{N} : P(x)$ is the same thing as $P(0) \veebar P(1) \veebar P(2) \veebar \cdots$.

Unique existential quantification is rarely used, compared to universal and existential quantification.

2.3.1.4 Scope of Quantifiers

As well as being restricted the a universe of discourse, we can further restrict the scope of a quantifier with more general predicates using implications or conjunctions. For example, given that the universe of discourse has already been specified to be the natural numbers, the statement $\forall x > 1 : x^2 > 2$ is short for $\forall x : (x > 1) \rightarrow (x^2 > 2)$. Similarly, the statement $\exists x > 0 : x^2 = 4$ is short for $\exists x : x > 0 \land x^2 = 4$. Note that restrictions on universal quantifiers are expressed with implications, while restrictions on existential quantifiers are expressed with conjunctions.

2.3.1.5 Negation of Quantifiers

$$\begin{aligned} \neg \forall x P(x) &\equiv \exists x \neg P(x) \\ \neg \exists x P(x) &\equiv \forall x \neg P(x) \end{aligned}$$

These are effectively the quantification versions of De Morgan's laws.

For example, if you want to prove that not all men are mortal ($\neg \forall h : \text{Mortal}(h)$), you only need to find one man that is not mortal ($\exists h : \neg \text{Mortal}(h)$), but if you want to prove that no man is mortal ($\neg \exists h : \text{Mortal}(h)$), you need to show that all men are not mortal ($\forall h : \neg \text{Mortal}(h)$).

2.3.1.6 Nested Quantifiers

The statements bound by a quantifier itself can contain quantifiers. For example, the statement "there is no largest prime number" could be written as,

$$\neg \exists x : \text{Prime}(x) \land (\forall y : y > x \rightarrow \neg \text{Prime}(y))$$

or, "There does not exist an x such that x is prime and that all y's greater than x are not prime." Or, shorter still (but not equivalently),

$$\forall x \exists y : \text{Prime}(y) \wedge (y > x)$$

or, "For any x, there is a greater y that is prime."

Note that order matters for nesting quantifiers. Let the universe of discourse be the set of real numbers. Then,

$$\forall x \exists y : y > x$$

or, "For all numbers x, there exists a number y greater than x" is true: for any x, we can certainly find a y that is greater than it. In contrast,

$$\exists y \forall x : y > x$$

or, "There exists a number y such that every x is less than y." clearly isn't true (at least, not over the set of real numbers).

The first statement says that $\exists y : y > x$ is true, regardless of what x is, while the second says that there is some y such that $\forall x : y > x$ is true.

One way to think about nested quantification is like a game between two players, alternating turns picking values for the quantified variables, with the adversary starting. Additionally, we assume that the adversary plays perfectly, picking the worse possible values for us, if relevant. It's no use trying to prove $\exists y \forall x : y > x$ if we pick $y = 1$ and our adversary nicely picks $x = 0$. We want our adversary to be a smug smartass, picking whatever value *doesn't* work for us, in this example, say, $x = 2$.

If we can always win the game, then the statement is true. For example, to prove $\forall x \exists y : y > x$, we let our adversary pick some real x, and we have to try pick a y such that y is greater than x. Picking $y = x + 1$ (our choice of value is allowed to depend on what has already been picked, as they are now fixed from the perspective of the inner statements) works, regardless of what x is, so the statement is true.

Of course, we can nest quantifiers deeper than 2 layers. One definition we will see later on in the real analysis section is the definition of convergence for a sequence, (a_n):

$$\Big[(a_n) \to a\Big] \coloneqq \Big[\forall \varepsilon > 0 : \exists N > 0 : \forall n > N : |a_n - a| < \varepsilon\Big]$$

Now we can interpret nested quantifiers, we can easily unravel what the right hand side means:

- The adversary picks a value for ε greater than 0;
- We pick a value of N;
- They pick a value of n greater than N;
- If $|a_n - a| < \varepsilon$, we win.

So, if we wanted to prove that the sequence $a_n = \frac{1}{n}$ converges to 0, we,

- Choose $\varepsilon > 0$;
- Let $N > \frac{1}{\varepsilon}$;
- Let $n > N$;
- $a_n = \frac{1}{n} < \frac{1}{N} < \varepsilon$, so $|a_n - 0| < \varepsilon$, as required.

Exercises.

- Prove that the sequence $a_n = \frac{1}{\sqrt{n}}$ converges to 0.
- Prove that the sequence $a_n = n$ does not converge.

2.3.1.7 Higher-Order Logics

In first-order logic, variables always refer to things and never to predicates. One effect of this is that predicates cannot be quantified over in first-order logic. For example, the variable in $\text{Cube}(a) =$ "a is a cube" could be quantified in first-order logic as $\exists x, \text{Cube}(x)$.

However, we cannot quantify predicates; "$\exists P : P(b)$" is not a valid sentence of first-order logic, but it is a valid sentence of second-order logic. Here, P is a *predicate variable* and represents the set of objects with property P.

In first-order logic, there is no way to identify the set of all objects that satisfy some given predicates, but this is possible in second-order logic. For example, the set of all objects that are cubes or tetrahedrons could be represented in second-order logic as,

$$\exists P \forall x : Px \leftrightarrow (\text{Cube}(x) \vee \text{Tetra}(x))$$

We can also assert properties about this set in second-order logic. For instance, the following says that this set of cubes and tetrahedrons does not contain any spheres.

$$\forall P \forall x : (Px \leftrightarrow (\text{Cube}(x) \vee \text{Tetra}(x))) \rightarrow \neg \exists x : Px \wedge \text{Sphere}(x)$$

We can't however, have variables for predicates of predicates. For example, we can't say that there is a property $\text{Shape}(P)$ that is true for the predicates P: Cube, Tetra and Sphere. Doing so requires third-order logic.

As well as higher-order logics, other completely different alternative models such as lambda calculus or topoi category theory exist, but first-order logic is sufficient machinery for most of modern mathematics.

2.3.2 Function Symbols

A *function symbol* looks a lot like a predicate, but instead of returning a truth value, it returns another object. Function symbols may take any number of arguments. The number of arguments a function symbol takes is called its *arity*. A zero-arity function symbol is also called a *constant*. For example, in the formula, "2+2=5", there are three constants, "2", "2" and "5", a two-arity function, "+", and a very special predicate, "=".

Function symbols allow us to populate our universe without having to include various axioms about things existing. The convention is that anything we can name exists. An example is the construction of the natural numbers with the Peano axioms.

We start with a single axiom, stating that the constant function symbol 0 exists. Then, we apply the successor function, S to obtain $0, S(0), S(S(0)), S(S(S(0))), \ldots$ (which we often relabel with the usual, 0,1,2,3,...), and now our universe has an entire infinity of natural numbers without us ever having to define a new axiom for each next number.

Note that two objects constructed in different ways aren't guaranteed to be distinct. To check whether two objects are the same, we use the equality predicate.

2.3.3 Equality

The equality predicate, $=$, is typically included as a standard part of predicate logic. If $x = y$, then x and y represent the same element in the universe of discourse.

The equality predicate satisfies,

- the *reflexivity axiom*: $\forall x : x = x$;

- the *substitution axiom schema*[*]: $\forall x \forall y : (x = y) \to (Px \leftrightarrow Py)$, where P is any predicate;
- the *symmetry property*: $\forall x \forall y : x = y \to y = x$;
- the *transitivity property*: $\forall x \forall y \forall z : x = y \wedge y = z \to x = y$.

The last two properties are not axioms, as they can be proved from the first two, but are listed as they are useful in proofs, and, together, they classify equality as a type of *equivalence relation*, which is discussed in §4.4.2.

The substitution axiom schema also immediately gives a *substitution rule* which states that $x = y, Px \vdash Py$[†], or "Given $x = y$ and Px, we can deduce Py." Almost every proof you will have previously seen in elementary algebra will have consisted purely of repeated applications of this rule.

Exercise. Prove the symmetry and transitivity properties of equality from the reflexivity axiom and substitution axiom schema.

2.3.4 Formal Languages & Structures

It is helpful to distinguish between the symbols we write from the ideas we bind to those symbols.

A *formal language* is a list of symbols called an *alphabet* that concatenate into strings called *sentences* or *formulae* according to some *grammar* or *syntax* rules. If a formula is part of a formal language, it is *well-formed*.

A theory of any particular logic system (the set of all possible formulae) forms a formal language. A formal language by itself only specifies the syntax of formulae, and not their semantics.

As an example, the following list of rules describes a formal language, L, over the alphabet, $\Sigma = \{0,1,2,3,4,5,6,7,8,9,+,=\}$:

- Every non-empty string that does not contain "+" or "=" and does not start with "0" is in L.
- The string "0" is in L.
- A string containing "=" is in L if and only if there is exactly one "=", and it separates two valid strings of L.
- A string containing "+" but not "=" is in L if and only if every "+" in the string separates two valid strings of L.
- No string is in L unless implied by the previous rules.

The string "$12 + 34 = 5678$" is in L, but the string "$+ = 12 =$" is not. This formal language expresses natural numbers, well-formed additions and well-formed addition equalities, but it only expresses what they look like (syntax), and not what they mean (semantics). For example, the rules do not indicate that the symbol "0" represents the number zero, the symbol "+" means addition, "$12 + 34 = 5678$" is false, etc.

To assign semantics to a formal language, we use an *interpretation function*.

In propositional logic, an interpretation function assigns a truth value to each propositional symbol. But, as it turns out, every language in propositional logic is equivalent, since the only differences between different propositional logics is in what propositions we assign to each letter, i.e., "*p*" might represent "Socrates is a philosopher" in one logic, but "The moon is made of green cheese" in another, but they

[*] An axiom *schema* is the generalisation of an axiom – an axiom only contains variables representing objects, but an axiom schema can contain variables representing formulae. An axiom schema represents an infinite number of regular axioms – in this case, the axiom schema is equivalent to having an individual axiom for each possible predicate that P could represent.

[†] We will discuss the mysterious symbol $\vdash$ in more detail soon.

function identically as a symbol – the only difference would be what truth value it is assigned by the interpretation function, which is not part of the formal language. Another way to say this is that there is only one zeroth-order language.

More generally, the alphabet of a formal language is further partitioned into *logical* and *non-logical* symbols. A logical symbol is a symbol that is agreed to always have the same meaning. For example, $\forall$ always means "For all..." and $\wedge$ always means conjunction. In contrast, non-logical symbols are things like predicate symbols and function symbols that only have meaning when assigned one.

A *signature* lists the non-logical symbols of an alphabet, alongside some syntactic information about those symbols. For each non-logical symbol, the signature identifies the symbol as a constant symbol, a function symbol, or a predicate symbol. In the latter two cases, the signature also identifies the arity of the symbol (we can also define constants as zero-arity functions and combine the two sets together).

In predicte logic, an interpretation function further provides the *extension* of formulae in a language – the list of (ordered) arguments that make the formula true. For instance, an interpretation function could take the predicate symbol Pa (for "a is a philosopher") and assign it the extension $\{s\}$ (for "Socrates"), indicating that Ps is true. Note, however, that this is all the interpretation function does; it assigns the extension $\{s\}$ to the non-logical predicate symbol Pa, and says nothing about what the predicate symbol Pa or constant symbol s actually stand for, because the underlying logic doesn't care. The interpretation also does not have anything to say about logical connectives, such as $\wedge$, or any other logical symbol.

Unlike in propositional logic, there are many distinct first-order languages, each defined by its signature. Given a signature, σ, we call the corresponding formal language the *set of σ-formulae.*

To assign a value to all the sentences of a first-order language, the following information is needed:

- A non-empty universe of discourse, D.
- For every constant symbol, an element of D as its interpretation.
- For every n-ary function symbol, a function $D^n \to D$ as its interpretation.
- For every n-ary predicate symbol, a subset of D^n as its extension.

or in other words, a universe of discourse, an interpretation function and a signature. These things together form a *structure* (of signature-σ), or a σ-structure.

A structure with an interpretation in which all the sentences of a particular theory is true is called a *model.*

In general, we can't hope to find all possible models of a given theory as there are usually infinitely many, but they are still useful to us: if we can find a model of a particular theory, then the existence of the model demonstrates that the theory is consistent; and if we can find a model of the theory in which some additional statement outside the theory, S, doesn't hold, then we can demonstrate that S is not provable from the theory (if T is the list of axioms that define the theory, then $\neg(T \vdash S)$.)

2.3.4.1 Examples

The theory defined by the sole axiom $\neg\exists x$ has exactly one model – it's empty. Now, consider the theory defined by the sole axiom $\exists!x$, or, $\exists x \forall y : y = x$. It also has exactly one model, but now with a single element. We can also force a model to have exactly k elements using a single axiom. For example, $k = 3$ is given by $\exists x_1 \exists x_2 \exists x_3 \forall y : y = x_1 \vee y = x_2 \vee y = x_3 \wedge x_1 \neq x_2 \wedge x_2 \neq x_3 \wedge x_3 \neq x_1$. These elements are indistinguishable, so there is again exactly one model, with 3 indistinguishable elements.

Suppose we have a predicate, P, and include the axiom $\exists x : P(x)$. Now, we have infinitely many models: take any non-empty model, and include at least one of its elements in the extension of P. If we have a model with two elements, x and y with Px and $\neg Py$, we see that $\forall x : Px$ does not hold, so $\exists x : Px$ is not sufficient to prove $\forall x : Px$, since we have an example of a model where $\exists x : Px$ holds but $\forall x : Px$ doesn't. On the other hand, an empty model satisfies $\forall x : Px \equiv \neg\exists x : Px$, but not $\exists x : Px$.

Now, suppose we have a function symbol S and constant symbol 0, and the single Peano axiom, $\forall x \forall y : Sx = Sy \to x = y$. The natural numbers, $\mathbb{N}$, and the integers, $\mathbb{Z}$, are both models for this system, but so is the set of integers $\bmod n$, $\mathbb{Z}_n$, for all n. In each of the models, every element has a unique predecessor as demanded by the axiom. Adding in the next Peano axiom, $\forall x : Sx \neq 0$ eliminates $\mathbb{Z}$ and $\mathbb{Z}_n$, as those sets contain elements that have 0 as a successor. But we don't eliminate a model that consists of two or three copies of $\mathbb{N}$ sitting next to each other, but only one contains the "official" 0 as a symbol, or even a model that consists of a single copy of $\mathbb{N}$, along with any number of copies of $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{Z}_n$.

2.4 Proofs

A *proof* is a way to derive statements from other statements; we begin with axioms, theorems or lemmata*, and *premises* or *antecedents* P, and use *inference rules* to derive the *conclusion* or *consequent*, Q. Anything not proven within a proof is called a *hypothesis*.

When we can prove a conclusion Q from premises P_1,P_2,..., we say that Q is *deducible* or *provable* from P_1,P_2,..., and we write $P_1,P_2,\ldots \vdash Q$, using the *turnstile* symbol, $\vdash$. If Q is provable using only inference rules without premises, we can write $\vdash Q$. Note not all the premises need be required to prove the conclusion; the list of premises merely needs to be sufficient to do so, as extra premises can always remain unused. Alternative wording to "Q is provable from P" include; "P (logically) entails Q"; "Q is (logically) entailed by P"; and "P yields Q".

Note that provability, $\vdash$, is quite distinct from implication, $\to$: if our inference rules are not sufficiently strong, it may be true that $P \to Q$ is true, but Q is not provable from P. On the other hand, if our inference rule are too strong (for example, strong enough to prove even false things), then we could have $P \vdash Q$, but $P \to Q$ be false.

We use $\vdash$ when we want to talk about whether a proof can exist in some logical system, while $\to$ is a logical connective. Because $\vdash$ discusses the provability, which is outside the scope of a theory, it is a metalogical symbol, while $\to$ is inside, making it a symbol in the object language of the logical system.

There are a couple properties of formal systems of interest:

- *Decidability* – a theory, T, is decidable if there exists a finite terminating procedure to determine whether $T \vdash \varphi$, where φ is any formula in the language.
- *Syntactic Completeness* – T is syntactically complete if every formula φ in the language of T is either provable or disprovable: at least one of $T \vdash \varphi$ and $T \vdash \neg\varphi$ holds.
- *Semantic Completeness* – T is semantically complete if every true statement is provable.
- Consistency – T is consistent if there is no formula φ such that both φ and $\neg\varphi$ are provable. Using our new terminology, we can also say that T is consistent if it has a model.

Gödel's completeness theorem shows a correspondence between semantic truths and syntactic provability in predicate logic; that is, every consistent first-order theory has a model and is therefore semantically complete.

However, we also have *Gödel's incompleteness theorems*, a pair of theorems which put other limitations on what logic systems can do. The first incompleteness theorem states that any axiom system for predicate logic that is consistent and is also powerful enough to represent arithmetic is syntactically incomplete: there are true statements, S_T, that cannot be proven within T. The second incompleteness theorem

* A *lemma* (plural lemmata) is a theorem that is intended not as an end result, but as a tool to prove another theorem. For this reason, they are also sometimes called *helping theorems* or *auxiliary theorems.* While on the topic, a *corollary* is a theorem, often of less importance, that can be immediately deduced from a prior, more notable theorem.

extends this result by showing that S_T can be the sentence that expresses the consistency of T – so any axiom system we choose for a first-order system will always be incomplete and cannot prove its own consistency. Turing later showed that such an axiom system is also undecidable.

Fortunately, logicians have come up with a standard list of inference rules that are just strong enough to prove everything we want and are not any stronger.

A proof is *valid* if and only if its conclusion is true whenever the premises are true.

For example, the proof,

- All men are mortal.
- Socrates is a man.
- Therefore, Socrates is mortal.

is valid, not because it has a true conclusion and premises, but because the conclusion is true whenever the premises are true.

The proof,

- All fish can fly.
- Socrates is a fish.
- Therefore, Socrates can fly.

is equally valid, despite having false premises and a false conclusions.

In contrast, the proofs,

- All men are immortal.
- Socrates is a man,
- Therefore, Socrates is mortal.

and

- All apples are plants.
- All fruits are plants.
- Therefore, all apples are fruits.

despite having true conclusions, are *invalid*, because their conclusions are not logically deduced from the premises. In the second invalid proof, both the premises and conclusion happen to be true in our universe, but still is not a valid argument as the conclusion does not follow from the premises.

Validity is not to do with truth values, per se, but is to do with logical deduction. Another way to phrase validity, is that the formula as a whole holds in all possible structures.

If a valid proof also has true premises, then it is *sound*: $P \vdash Q$ implies that $P \rightarrow Q$ is a tautology. Conversely, if a logic system has every tautology as a theorem, it is *(semantically) complete*: $P \rightarrow Q$ implies that $P \vdash Q$.

If every sentence in a formal system is sound/complete, then the system itself is said to be sound/complete.

2.4.1 Inference Rules

Our main source of inference rules is based off of tautologies of the form $S_1 \wedge S_2 \wedge \cdots \rightarrow Q$. Given such a tautology, there is a corresponding inference rule that allows us to assert Q given that S_1,$S_2 \ldots$ all hold. We can do so by showing that S_n is an axiom/theorem/premise, or by proving them from other axioms/theorems/premises.

The most important inference rule, *modus ponens*, is of this form, based on the tautology, $(p \land (p \to q)) \to q$. This is what we use to prove,

1. If it is raining, then I wear a coat. (We assert $p \to q$ is true)
2. It is raining. (And also that p is true)
3. Therefore, I wear a coat. (So modus ponens applied to 1 and 2 gives us q)

As well as horizontally with the turnstile symbol, inference rules can be given in the form,

$$\begin{array}{c} \text{Premise 1} \\ \text{Premise 2} \\ \text{Premise 3} \\ \cdots \\ \hline \text{Conclusion} \end{array}$$

or

$$\frac{\text{Premise 1} \qquad \text{Premise 2} \qquad \text{Premise 3} \qquad \cdots}{\text{Conclusion}}$$

with the horizontal line acting like a higher order version of $\vdash$: it lets us combine proofs into bigger proofs.

Many important inference rules in classical propositional logic have been given names. Some of these have been listed below.

$p \vdash p \lor q$	Addition or disjunction introduction
$p \to q, r \to q, p \lor r \vdash q$	Disjunction elimination
$p, q \vdash p \land q$	Conjunction introduction
$p \land q \vdash p$	Simplification or conjunction elimination
$p, p \to q \vdash q$	Modus ponens
$\neg q, p \to q \vdash \neg p$	Modus tollens
$p \to q, r \to s, p \lor r \vdash q \lor s$	Constructive dilemma
$p \to q, r \to s, \neg q \lor \neg s \vdash \neg p \lor \neg r$	Destructive dilemma
$p \to q \vdash p \to (p \land q)$	Absorption
$p \to q, p \to \neg q \vdash \neg p$	Negation introduction
$p \to q, q \to r \vdash p \to r$	Hypothetical syllogism
$p \lor q, \neg p \vdash q$	Modus tollendo ponens or disjunctive syllogism
$p, \neg(p \land q) \vdash \neg q$	Modus ponendo tollens
$p \lor q, \neg p \lor r \vdash q \lor r$	Resolution

It isn't really necessary to remember all the names, except perhaps modus ponens. Most of the other rules can be derived from modus ponens combined with some tautologies or logical equivalences anyway. For instance, the addition inference rule is just modus ponens applied to p and the tautology $p \to (p \lor q)$.

The first four just let us pack and unpack variables from various connectives. Modus ponens and modus tollens let us apply implications. They are contrapositives of each other, so only one has to be memorised. Constructive and destructive dilemma allows us to replace the variables in conjunctive statements given certain implications. Absorption allows us to introduce conjunctions to proofs. Hypothetical syllogism just states that implication is transitive, while disjunctive syllogism allows us to replace disjunctions if we know one of the premises is false. Disjunctive syllogism is another rule that can be written in terms of modus ponens, this time through the logical equivalence, $p \lor q \equiv \neg p \to q$. For resolution to be useful, we first have to discuss *normal forms*.

2.4.1.1 Conjunctive Normal Forms

A compound proposition is in *conjunctive normal form* or *CNF* if it is a conjunction of one or more *clauses*, where a clause is a disjunction of atoms; it is an AND of OR statements. A compound proposition is similarly in *disjunctive normal form* or *DNF* if it is the disjunction of one or more clauses, where a clause is a conjunction of atoms; it is an OR of AND statements.

Propositions in CNF:

- p
- $(p \lor \neg q) \land r$
- $(p \lor q) \land (\neg p \lor r) \land q \land (\neg q \lor \neg r)$
- $p \land \neg q \land r \land t \land \neg u \land v$

Propositions not in CNF:

- $(p \land q) \land (q \lor r)$
- $(p \lor q) \land (q \to \neg r) \land (\neg p \lor r)$
- $(p \lor (q \land r)) \land (p \lor \neg r)$

Interchanging $\land$ and $\lor$ above gives examples of clauses in and not in DNF.

Using the equivalence of material implication and disjunction, along with De Morgan's laws and the distributive laws, it is possible to rewrite any compound proposition in a normal form. However, applying these laws blindly does not necessarily produce the simplest normal form for a compound proposition.

For example,

$$\begin{aligned}
(P \to Q) \land (\neg P \to Q) &\equiv (\neg P \lor Q) \land (P \lor Q) \\
&\equiv (\neg P \land P) \lor (\neg P \land Q) \lor (Q \land P) \lor (Q \land Q) \\
&\equiv 0 \lor (\neg P \land Q) \lor (Q \land P) \lor Q \\
&\equiv (\neg P \land Q) \lor (Q \land P) \lor Q
\end{aligned}$$

Inspecting the clauses closer, we see that Q controls the value of the entire expression, so a simpler CNF for the proposition is just Q.

$$\equiv Q$$

We should really draw out a truth table to prove this formally, but it should be clear enough that this is true.

The CNF in particular is useful because it supports resolution well. We can construct proofs from CNF formulae by looking at occurences of some proposition and its negation and resolving them, generating a new clause called a *resolvent*. For example,

$$\begin{aligned}
&(P \lor Q) \land (P \lor \neg R) \land (\neg P \lor Q) \land (\neg Q \lor R) \\
\vdash &(P \lor Q) \land (P \lor \neg R) \land (\neg P \lor Q) \land (\neg Q \lor R) \land Q \\
\vdash &(P \lor Q) \land (P \lor \neg R) \land (\neg P \lor Q) \land (\neg Q \lor R) \land Q \land R \\
\vdash &(P \lor Q) \land (P \lor \neg R) \land (\neg P \lor Q) \land (\neg Q \lor R) \land Q \land R \land P \\
\vdash &P,Q,R
\end{aligned}$$

Resolution is not as useful for humans, but, due to its mechanical simplicity, is very popular with computer theorem provers. This topic is covered in more detail in a later chapter on automated theorem proving and program verification.

2.4.2 Implication and Natural Deduction

As discussed earlier, provability and implication are distinct, the former being metalogical in nature, and the latter being a logical statement within the logical system.

However, due to our choice of premade inference rules stolen from other clever logicians, these two notions often coincide.

For example, suppose that $P \to Q$ is provable without any assumptions, so we can write,

$$\vdash P \to Q$$

Because extraneous premises are allowed, we may also write,

$$\begin{gathered} P \vdash P \to Q \\ P \vdash P,P \to Q \end{gathered}$$

And, apply modus ponens to the right hand side, we get,

$$P \vdash Q$$

so we can go from $\vdash P \to Q$ to $P \vdash Q$. In this sense, provability is weaker than implication: assuming modus ponens, provability holds whenever implication does. This fact isn't used much, since we are generally more interested in implication, but can we go the other way?

2.4.2.1 The Deduction Theorem

A proof normally shows that, given a set of axioms, Γ, if a set of premises, $P_1,P_2,\ldots,P_n$ holds, then the conclusion, Q, holds. To use this result later, it's useful to be able to package the proof up as the implication, $P_1 \land P_2 \land \cdots \land P_n \to Q$.

In other words, we want to go from $\Gamma,P_1,P_2,\ldots,P_n \vdash Q$ to $\Gamma \vdash (P_1 \land P_2 \land \cdots \land P_n) \to Q$.

The *deduction theorem* is a metatheorem – a theorem about the logical system itself – that says exactly that: if Q is deducible from a set of premises, $\Gamma,P_1,P_2,\ldots,P_n$, then the implication $(P_1 \land P_2 \land \cdots \land P_n) \to Q$ is deducible from Γ alone.

In the special case that Γ is the empty set, then the deduction theorem says that $P \vdash Q$ implies $\vdash P \to Q$.

The proof of the deduction theorem depends on the logic system and the set of inference rules we start with, and is already rather complex just for propositional logic, so it is omitted. The main idea is that there is an efficient algorithm that extracts a proof of the desired implication given the proof of Q given the premises.

The deduction theorem does not apply if any of the premises contain free variables. Fortunately, for most of the things we like to work with, this is usually not the case.

2.4.2.2 Natural Deduction

In practice, we don't refer to the deduction theorem directly, instead adding a new inference rule,

$$\frac{\Gamma,P \vdash Q}{\Gamma \vdash P \to Q}$$

which says that, if we can prove Q with premises Γ and P, then we can prove $P \to Q$ with premise Γ.

This type of inference rule where we track the assumptions behind every particular result is called *natural deduction*, and was invented to make inference rules function more like how actual mathematical proofs do, as opposed to the modus-ponens-only method that modern logicians had previously been using.

The rule above, in particular, is called *implication introduction*. The corresponding implication elimination is just modus ponens,

$$\frac{\Gamma \vdash P \to Q \qquad \Gamma \vdash P}{\Gamma \vdash Q}$$

We can also rewrite the list from above like this, (also further explaining the introduction and elimination names), as well as some extras:

Introduction Rules	Elimination Rules
$\dfrac{\Gamma, P \vdash Q \quad \Gamma, P \vdash \neg Q}{\Gamma \vdash \neg P}\,(\neg\mathrm{I})$	$\dfrac{\Gamma, \neg P \vdash Q \quad \Gamma, \neg P \vdash \neg Q}{\Gamma \vdash P}\,(\neg\mathrm{E})$
$\dfrac{\Gamma \vdash P}{\Gamma \vdash \neg\neg P}\,(\neg\neg\mathrm{I})$	$\dfrac{\Gamma \vdash \neg\neg P}{\Gamma \vdash P}\,(\neg\neg\mathrm{E})$
$\dfrac{\Gamma \vdash P \quad \Gamma \vdash Q}{\Gamma \vdash P \land Q}\,(\land\mathrm{I})$	$\dfrac{\Gamma \vdash P \land Q}{\Gamma \vdash P}\,(\land\mathrm{E_L}) \quad \dfrac{\Gamma \vdash P \land Q}{\Gamma \vdash Q}\,(\land\mathrm{E_R})$
$\dfrac{\Gamma \vdash P}{\Gamma \vdash P \lor Q}\,(\lor\mathrm{I_L}) \quad \dfrac{\Gamma \vdash Q}{\Gamma \vdash P \lor Q}\,(\lor\mathrm{I_R})$	$\dfrac{\Gamma \vdash P \lor Q \quad \Gamma \vdash \neg Q}{\Gamma \vdash P}\,(\lor\mathrm{E_L}) \quad \dfrac{\Gamma \vdash P \lor Q \quad \Gamma \vdash \neg P}{\Gamma \vdash Q}\,(\lor\mathrm{E_R})$
	$\dfrac{\Gamma \vdash P \to Q \quad \Gamma \vdash R \to Q \quad \Gamma \vdash P \lor R}{\Gamma \vdash Q}\,(\lor\mathrm{E})$
$\dfrac{}{\Gamma \vdash \top}\,(\top\mathrm{I})$	No $\top$ elimination
No $\bot$ introduction	$\dfrac{\Gamma \vdash \bot}{\Gamma \vdash Q}\,(\bot\mathrm{E})^*$
$\dfrac{\Gamma, P \vdash Q}{\Gamma \vdash P \to Q}\,(\to\mathrm{I})$	$\dfrac{\Gamma \vdash P \to Q \quad \Gamma \vdash P}{\Gamma \vdash Q}\,(\to\mathrm{E_L}) \quad \dfrac{\Gamma \vdash P \to Q \quad \Gamma \vdash \neg Q}{\Gamma \vdash \neg P}\,(\to\mathrm{E_R})$
$\dfrac{\Gamma \vdash P \to Q \quad \Gamma \vdash Q \to P}{\Gamma \vdash P \leftrightarrow Q}\,(\leftrightarrow\mathrm{I})$	$\dfrac{\Gamma \vdash P \leftrightarrow Q}{\Gamma \vdash P \to Q}\,(\leftrightarrow\mathrm{E_L}) \quad \dfrac{\Gamma \vdash P \leftrightarrow Q}{\Gamma \vdash Q \to P}\,(\leftrightarrow\mathrm{E_R})$
$\dfrac{\Gamma \vdash Py}{\Gamma \vdash \forall x : Px}\,(\forall\mathrm{I})$	$\dfrac{\Gamma \vdash \forall x : Px}{\Gamma \vdash Py}\,(\forall\mathrm{E})$
$\dfrac{\Gamma \vdash Py}{\Gamma \vdash \exists x : Px}\,(\exists\mathrm{I})$	$\dfrac{\Gamma \vdash \exists x : Px}{\Gamma \vdash Py}\,(\exists\mathrm{E})$

2.4.3 Inference Rules for Equality

The equality predicate is special in that it allows for the *substitution* inference rule,

$$x = y, Px \vdash Py$$

and we can also assert $x = x$ directly:

$$\vdash x = x$$

* The Q here is not a mistake: it's saying that if you can prove $\bot$ from Γ, then you can prove any statement you want – this is the principle of explosion written out in the language of natural deduction.

If we didn't want to include the substitution rule as an inference rule, we could instead represent it as an axiom schema,

$$\forall x \forall y : (x = y \wedge Px) \rightarrow Py$$

2.4.4 Inference Rules for Quantified Propositions

In terms of natural deduction, these rules are the introduction and elimination rules for $\forall$ ad $\exists$.

2.4.4.1 Universal Generalisation

If y is an arbitrarily selected variable in the universe of discourse, then

$$\frac{\Gamma \vdash Py}{\Gamma \vdash \forall x : Px}$$

which is to say, if we know that Py is true, and we have assumed nothing about y, then Px is true for all x.

2.4.4.2 Unversal Instantiation

Conversely, we have,

$$\forall x : Qx \vdash Qc$$

which says that a specific statement about c is provable from a general statement about all possible values x, that is, if P holds for all elements in the universe of discourse, and c is an element of the universe of discourse, then Pc holds.

For example, "Given that all humans are mortal, it follows that the human called Socrates is mortal."

2.4.4.3 Existential Generalisation

Essentially the opposite of universal instantiation,

$$Pc \vdash \exists x : Px$$

this says that, if we want to show that Px holds for at least one x, and we know that Pc holds, then we can use c as our example of x. This style of proof is *proof by construction* or *proof by example.*

For instance, if we are asked to prove that there exists an even prime number, we can produce the example 2, and this is sufficient.

Despite the name, these proofs are not always constructive. For example, earlier (§2.2.3) we proved that there exists two irrational numbers a and b such that a^b is rational, but we only did so by producing two candidates for a^b and proving that at least one of them is rational – we never actually identified one single object which makes the statement true.*

Constructive proofs are generally more useful than non-constructive proofs, because the constructed example often has additional useful properties, or is helpful in other contexts. In some schools of thought and logic systems, non-constructive proofs aren't even considered proofs.

One particular example of this is intuitionistic logic. In this system, the law of the excluded middle also does not hold, nor does double negation: a proposition is only true when directly proved to be so. Rather than preserving truth-values, operations in intuitionistic logic instead preserve *justification* with respect to *evidence.*

* It turns out that not only is $\sqrt{2}^{\sqrt{2}}$ irrational, it is also trancendental. The proof of this is, however, non-trivial – see the Gelfond-Schneider theorem. Fortunately, picking $a = \sqrt{2}$ and $b = \log_2 9$ gives us an easy constructive proof with $a^b = 3$.

Intuitionistic logic is often objected to due to this lack of law of excluded middle and double negation elimination – two central rules of classical logic. Hibert himself wrote, “Taking the principle of excluded middle from the mathematician would be the same, say, as proscribing the telescope to the astronomer or to the boxer the use of his fists. To prohibit existence statements and the principle of excluded middle is tantamount to relinquishing the science of mathematics altogether.”

However, intuitionistic logic does have applications. It has been proven that, given a constructive proof that an object exists, an algorithm for generating examples can be constructed from the proof – constructive proof systems and computation models are really the same type of mathematical objects. See the *Curry-Howard correspondence* for more.

2.4.4.4 Existential Instantiation

$$\exists x : Px \vdash Pc$$

with the restriction that the symbol c has not been used previously.

This says that because we know that P holds for at least one element of the universe, we can give it a name, say, c.

We do this whenever we say “Let x be some number such that... holds...”, assuming we know that whatever is supposed to hold does indeed have values that work.

2.4.5 Proof Techniques

A proof technique is a framework to guide you along proving certain types of statements. This doesn’t mean you don’t have to think about what you’re doing, but it’ll give you a idea of what you should be trying.

The following table gives techniques to try for proving $A \to B$, mostly classified by the structure of B.

To prove $A \leftrightarrow B$, prove $A \to B$ and $B \to A$ separately.

Name	When	Assume	Conclude	Description
Direct proof	First strategy to try.	A	B	Apply inference rules and work from A to B. May be helpful to work backwards from B at the same time and meet in the middle.
Contraposition	$B = \neg Q$	$\neg B$	$\neg A$	Apply any other strategy to show $\neg B \to \neg A$, then use the contraposition rule.
Contradiction	$B = \neg Q$	$A \wedge \neg B$	$\perp$	Apply any other strategy to prove both P and $\neg P$ for any P and invoke the law of non-contradiction to conclude that the premise must therefore be false.
Construction	$B = \exists x : Px$	A	Pc	Pick a c and use any other strategy to prove Pc holds.
Counterexample	$B = \neg\forall x : Px$	A	$\neg Pc$	Pick a c and use any other strategy to prove $\neg Pc$ holds. This is identical to proof by construction, except we've applied De Morgan's laws to the quantifiers.
Universal Generalisation	$B = \forall x : Px \to Qx$	A, Pc	Qc	Assume A and Pc hold for some arbitrary c. Apply any other strategy to prove that Qc holds, but do not assume any extra information about c outside of what Pc gives you. For example, if Pc is "c is even", you can use the fact that 2 divides c evenly, but *not* the assumption that, for example, c is positive or that c is composite, because Pc doesn't give you enough information to deduce that.
Universal Instantiation	$A = \forall x : Px$	A	B	Pick a specific c and use any other strategy to prove $Pc \to B$. Because A holds for all x, this time, we're allowed to pick a specific c and use all its properties.
Case Analysis	$A = C \vee D$	C,D	B	Assume C and prove B. Then, assume D and prove B. Effectively, A tells you that at least one of C and D is true. If both individually imply B, then A certainly does too.
Induction	$B = \forall x \in \mathbb{N} : Px$	A	$P(n)$ and $\forall n \in N : P(n) \to P(n+1)$	Discussed in §5.

2.4.6 An Example

Real proofs that humans write are usually written as a combination of natural language and formal logical notation. However, it should be possible to translate any natural language proof into a purely logical one. There is little reason to do so, but as mathematicians, that has never stopped us.*

2.4.6.1 Axioms for Even Numbers

Using the Peano axiom convention of writing numbers in terms of the successor function, we can define even numbers using the following axioms:

A_1: $\forall x : Ex \leftrightarrow (x = 0 \vee (\exists y : Ey \wedge x = SSy))$;

A_2: $\forall x : 0 \neq Sx$;

A_3: $\forall x \forall y : Sx = Sy \rightarrow x = y$.

where we interpret Ex to mean that x is even. The first axiom concerns even numbers, while the last two are extra properties about S which will be required later.

2.4.6.2 A Theorem

Theorem 2.4.1. *The following propositions are true;*

1. $E0$;
2. $\neg E(S0)$;
3. $E(SS0)$;
4. $\neg E(SSS0)$;
5. $E(SSSS0)$;

Proof.

1. Directly true from axiom A_1.
2. We prove this statement through contradiction. Suppose $E(S0)$ holds, so either $S0 = 0$, or $S0 = SSy$ for some y such that Ey holds. The first case contradicts A_2, while applying A_3 to the second case gives $0 = Sy$ which again contradicts A_2. It follows that the original assumption that $E(S0)$ is true does not hold.
3. We prove this through existential instantiation. $SS0 = 0$ contradicts A_2, so, from A_1, $E(SS0)$ holds only if there exists y such that Ey holds and $SS0 = SSy$. Let $y = 0$. Ey then holds from A_1, and $SS0 = SSy$ as desired, so the statement holds.
4. From A_2, we know $SSS0 = 0$ does not hold, so $E(SSS0)$ holds only if there exists y with Ey and $SSS0 = SSy$. Applying A_3 twice gives $SSS0 = SSy \leftrightarrow S0 = y$, but we already know $\neg E(S0)$, so $\neg E(SSS0)$.
5. Since $E(SS0)$ and $SSSS0 = SS(SS0)$, $E(SSSS0)$.

■

We can do these proofs again but in formal logic notation, explicitly writing down our inference rules. Because it takes so much space, we do this only for the proof of $\neg E(S0)$.

[*] Please do not actually do this. “Symbol spaghetti” is considered bad form exactly because it’s so unreadable.

The goal of the proof is to show $A_1,A_2,A_3 \vdash \neg E(S0)$. Because they'll come up together every line, we'll group up the axioms into $\Gamma = \{A_1,A_2,A_3\}$. Our strategy is to show $\Gamma \vdash E(S0) \to Q$ for some Q such that $\Gamma \vdash \neg Q$, then apply modus tollens to get the desired $\Gamma \vdash \neg E(S0)$.

Proof.

$\Gamma \vdash E(S0) \leftrightarrow (S0 = 0 \vee \exists y : Ey \wedge S0 = SSy)$	$\forall$E applied to A_1
$\Gamma \vdash E(S0) \to (S0 = 0 \vee \exists y : Ey \wedge S0 = SSy)$	$\leftrightarrow$ E and $\wedge$E.
$\Gamma,E(S0) \vdash S0 = 0 \vee \exists y : Ey \wedge S0 = SSy$	$\to$ E
$\Gamma,E(S0) \vdash \neg(S0 = 0)$	$\forall$E applied to A_2
$\Gamma,E(S0) \vdash \exists y : Ey \wedge S0 = SSy$	Combine previous two steps with $\vee\text{E}_\text{L}$
$\Gamma,E(S0) \vdash Ez \wedge S0 = SSz$	$\exists$E
$\Gamma,E(S0) \vdash S0 = SSz$	$\wedge\text{E}_\text{L}$
$\Gamma,E(S0) \vdash S0 = SSz \leftrightarrow 0 = Sz$	$\forall$E applied to A_3
$\Gamma,E(S0) \vdash S0 = SSz \to 0 = Sz$	$\leftrightarrow$ E and $\wedge$E.
$\Gamma,E(S0) \vdash 0 = Sz$	$\to \text{E}_\text{L}$ applied to $S0 = SSz$ and $S0 = SSz \to 0 = Sz$
$\Gamma \vdash E(S0) \to 0 = Sz$	$\to$ I
$\Gamma \vdash \neg(0 = Sz)$	$\forall$E and A_2
$\Gamma \vdash \neg E(S0)$	$\to \text{E}_\text{R}$

■

One thing of note is how $E(S0)$ moves in front of the turnstile in the middle of the proof. This is a common technique, and is what is happening behind the scenes whenever a natural language proof says "Suppose P holds...".

Using P, or specifically in this case, $E(S0)$, as an assumption saves us from repeatedly writing "if P, then...", and is what allows us to separate out the variables from $P \to Q$ and apply inference rules to Q.

2.4.6.3 A More General Theorem

So far, we have only proved results about a few specific numbers, but we can prove some results about infinitely many numbers.

Theorem 2.4.2. $\forall x : Ex \to E(SSSSx)$*: for all* x*, if* x *is even, then* $SSSSx$ *is also even.*

We write the proof with inference rules again, combining or omitting some less interesting steps.

Proof.

$\Gamma,Ex \vdash (\exists y : Ey \wedge SSx = SSy) \to E(SSx)$	A_1, $\forall$E, $\vee\text{E}_\text{L}$
$\Gamma,Ex \vdash Ex$	Any premise is provable
$\Gamma,Ex \vdash SSx = SSx$	Reflexivity of $=$
$\Gamma,Ex \vdash Ex \wedge SSx = SSx$	$\wedge$I applied to previous two steps
$\Gamma,Ex \vdash \exists y : Ey \wedge SSy = SSx$	Let $y = x$
$\Gamma,Ex \vdash E(SSx)$	Modus ponens
We have shown $Ex \to E(SSx)$.	
Now do all of this again to show $E(SSx) \to E(SSSSx)$ (omitted).	
$\Gamma \vdash E(SSSSx)$	$\to$ I
$\Gamma \vdash Ex \to E(SSSSx)$	$\to$ I

$$\Gamma \vdash \forall x : Ex \to E(SSSSx) \qquad \forall\mathrm{I}$$

■

If we were to write out the skipped parts, it might make sense to first prove a lemma $\forall x : Ex \to E(SSx)$, then just apply it to x twice.

The assertion that an even x exists (Ex is to the left of $\vdash$ for most of the proof) does a lot for us: it both introduces a new symbol x which we use for universal generalisation ($\forall$E), and the assumption that it is even allows us to use the deduction theorem ($\to$ E). Note that we can't apply universal generalisation until Ex is no longer an assumption (it is moved to the right of $\vdash$), because universal generalisation only works if x is not used in the assumptions.

2.4.6.4 Another Claim

Something we know about the natural numbers is that, if x is even, then $x + 1$ is odd, and vice versa. Formally, this is written as,

Claim. $\forall x : Ex \leftrightarrow \neg E(Sx)$: *for all x, x is even if and only if Sx is not even.*

Unfortunately, our axiom system is not sufficiently strong to prove this claim.

Here is a model that satisfies the axioms, but for which the claim fails:

- Include the natural numbers in our model, 0, $S0$, $SS0$, etc.
- Include an extra unnatural number, u, such that $u = Su$ and Eu.

Including this extra number doesn't violate any axioms. A_1 is satisfied since $Eu \leftrightarrow E(SSu)$ holds, since both u and SSu are even; A_2 is satisifed, since $0 \neq Su = u$; and A_3 is satisfied since $Sx = Sy \leftrightarrow x = y$ holds whenever x and y are both natural or both unnatural, and also if one is natural and the other is not, since $x \neq y$ and $Sx \neq Sy$ would both then hold.

But, Eu holds and $E(Su) = Eu$, which also holds, contradicting our claim. So, if we want the successor of any even number to be odd, we need a stronger set of axioms.

What we are really missing here is the *axiom schema of induction*, which says that $(P(0) \wedge \forall x : P(x) \to P(Sx)) \to \forall x : P(x)$: if a proposition holds for 0, and $P(x)$ implies that $P(Sx)$ also holds, then P holds for all x.

Chapter 3

Iterated Notation

"A problem well stated is a problem half-solved."

— Charles Kettering

Almost all of the operations we have been working with are binary, and only take two arguments. In maths, it is not uncommon for us to chain, or *iterate*, these operations together.

If f is a general binary operation, and $x_1,x_2,\dots x_n$ is a sequence of valid arguments, then we write $f/(x_1,x_2,\dots x_n)$ to indicate the iteration of f over that sequence.[†]

For some special binary operations, however, we have some more common standard notation.

3.1 Summation

Perhaps the most commonly iterated binary operation is addition. Given a sequence, $x_1,x_2,\dots x_n$, its sum, $x_1 + x_2 + \cdots + x_n$ is written as the *summation*,

$$\sum_{i=a}^{b} x_i$$

or, written inline as, $\sum_{i=a}^{b} x_i$. The large symbol is an elongated capital Greek letter sigma, for *sum.* The variable i is the *index (of summation)*, a is the *lower bound* or *lower limit*, and b is the *upper bound* or *upper limit.* In this expression, i is a bound variable, while a and b are free (§2.3.1).

This is essentially the mathematical notation for a `for` loop; we loop through all values of i between a and b, including both endpoints, summing up the body of the summation for each value of i. That is, the expression,

$$\sum_{i=a}^{b} f(i)$$

means, set i equal to a, then substitute it into the expression f to obtain $f(a)$, increment i by 1, then repeat until i is equal to b, so,

$$\sum_{i=a}^{b} f(i) = f(a) + f(a+1) + f(a+2) + \cdots + f(b-1) + f(b)$$

[†] You may recognise this notation as the second-order function, *fold* or *reduce* in computer science.

If $b < a$, then the summation evaluates to zero. For instance,

$$\sum_{i=2}^{1} \frac{\sin(\zeta(-i))\operatorname{erf}(\exp i)}{\Gamma\left(\sqrt[i]{\pi}\right)} = 0$$

This rule doesn't generally matter, and only really comes up as a boundary case for more general formulae. For example, the sum of the first n naturals can be given by,

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

This formula still holds if $n = 0$, returning zero due to this rule. The equality still holds for $n = -1$ as well, but not for $n \leq -2$. The proof of this formula is commonly done by weak induction (§5.1).

Summation notation is used both to save space – it is easier to write $\sum_{i=1}^{100} i^2$ than to write $1 + 4 + 9 + 16 + \cdots + 100\,000$ – and also for clarity – note that the reader had to assume what the $\cdots$ symbol indicated, while in summation notation, the expression is unambiguously interpreted as the sum of the first 100 square numbers.

3.1.1 Formal Definition

A summation is more formally defined as the recurrence,

$$\sum_{i=a}^{b} f(i) = \begin{cases} 0 & b < a \\ f(a) + \sum_{i=a+1}^{b} + f(i) & b \geq a \end{cases}$$

So, we recursively compute a summation by repeated extracting the bottom value out from the summation. For example, we would evaluate this summation as,

$$\begin{aligned} \sum_{i=0}^{3} i^2 &= 0^2 + \sum_{i=1}^{3} i^2 \\ &= 0^2 + 1^2 + \sum_{i=2}^{3} i^2 \\ &= 0^2 + 1^2 + 2^2 + \sum_{i=3}^{3} i^2 \\ &= 0^2 + 1^2 + 2^2 + 3^2 + \sum_{i=4}^{3} i^2 \\ &= 0^2 + 1^2 + 2^2 + 3^2 + 0 \\ &= 14 \end{aligned}$$

This definition readily applies to negative integer bounds,

$$\begin{aligned} \sum_{i=-2}^{1} i^2 &= (-2)^2 + \sum_{i=-1}^{1} i^2 \\ &= (-2)^2 + (-1)^2 + \sum_{i=0}^{1} i^2 \\ &= (-2)^2 + (-1)^2 + 0^2 + \sum_{i=1}^{1} i^2 \end{aligned}$$

$$\begin{aligned}
&= (-2)^2 + (-1)^2 + 0^2 + 1^2 + \sum_{i=2}^{1} i^2 \\
&= (-2)^2 + (-1)^2 + 0^2 + 1^2 + 0 \\
&= 6
\end{aligned}$$

This definition of a summation is also technically defined for non-integer bounds as well,

$$\begin{aligned}
\sum_{i=-\frac{1}{2}}^{\frac{9}{4}} i^2 &= \left(-\frac{1}{2}\right)^2 + \sum_{i=\frac{1}{2}}^{\frac{9}{4}} \\
&= \left(-\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \sum_{i=\frac{3}{2}}^{\frac{9}{4}} \\
&= \left(-\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{3}{2}\right)^2 + \sum_{i=\frac{5}{2}}^{\frac{9}{4}} \\
&= \left(-\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{3}{2}\right)^2 + 0 \\
&= \frac{11}{4}
\end{aligned}$$

but this is confusing, and very uncommon. In particular, the lower bound is almost always an integer value. However, the upper bound will occasionally be a division, or some other expression that doesn't return a round integer. For instance, you could imagine that we might want to sum up all the natural numbers less than half of some variable, which we would write as,

$$\sum_{i=0}^{\frac{n}{2}} i$$

and the $b < a$ returning 0 handles the possibly non-integer upper bound. However, we often include a floor or ceiling function (§34) to clarify where the summation should terminate.

If $b - a$ is an integer, then we can also evaluate a summation by extracting terms from the top if it is more convenient.

Lemma. *If $b - a$ is an integer, then,*

$$\sum_{i=a}^{b} f(i) = \begin{cases} 0 & b < a \\ f(a) + \sum_{i=a}^{b-1} + f(i) & b \geq a \end{cases}$$

Proof. If $b < a$, then the summation returns a zero, as before. Otherwise, $b \geq a$. We induct (§5.2) on $b - a$.

If $b - a = 0$, then,

$$\begin{aligned}
\sum_{i=a}^{b} f(i) &= f(a) \\
&= f(b) \\
&= f(b) + \sum_{i=a}^{b-1} f(i)
\end{aligned}$$

Else, if $b - a > 0$, then $b - a > b - a - 1 = b - (a + 1)$, and,

$$\begin{aligned}\sum_{i=a}^{b} f(i) &= f(a) + \sum_{i=a+1}^{b} f(i) \\ &= f(a) + f(b) + \sum_{i=a+1}^{b-1} f(i) \\ &= f(b) + \sum_{i=a}^{b-1} f(i)\end{aligned}$$

where the first and last equality hold by the recursive definition of summation, and the middle equality holds by the strong induction hypothesis. ■

Again, while this lemma holds for non-integer bounds as long as they have an integer difference, in practice, we will generally only encounter summations where both bounds are integers.

3.1.2 Scope

3.1.3 Linearity

3.1.4 Index Variables

3.1.5 Indexing Sets

3.1.6 Series

3.1.7 Double Sums

3.1.8 Closed Forms

3.1.9 Standard Sums

3.2 Products

3.3 Sets

3.4 Intersections

3.5 Unions

3.6 Logical Connectives

Chapter 4

Introduction to Set Theory

> *"The Axiom of Choice is necessary to select a set from an infinite number of pairs of socks, but not an infinite number of pairs of shoes."*
>
> — Bertrand Russell, *Introduction to mathematical philosophy*

This chapter is intended as a brief and informal introduction to set theory, with a focus on concepts that appear frequently in other fields of mathematics. For a more in-depth and formalised treatment of the subject, see the following chapter, §6.

4.1 Introduction

Set theory is the branch of mathematical logic that studies *sets*, which are, informally, "collections of other objects". Set theory is the most common choice for the foundational system of mathematics, the idea being, if you can write everything else in terms of sets – numbers, relations, functions – and you have a consistent theory about sets, then you can be sure that everything built from sets also behaves consistently. If symbolic logics are the syntax of mathematics, then set theory forms its alphabet.

One nice thing about set theory is that it only requires one additional predicate over the standard ones packaged with predicate logic – the *membership* or *element* predicate, $\in$, where $x \in S$ means that x is an element of the set S. S can also be completely identified by a list of all x that satisfies $x \in S$, and every other predicate in set theory can be defined in terms of $\in$ and logical connectives.

We begin with *naïve set theory*, a non-formalised theory discussed in natural language. In ordinary naïve set theory, any collection of objects can form a set. Unfortunately, not restricting the definition of a set in this way causes some paradoxes, which are not ideal for something we want to use as the foundation of mathematics, so we use axiomatic set theory instead, where we only use sets which we can prove to exist from the axioms. There are several consistent* and completely incompatible ways to axiomatise set theory, but the most popular choice is *Zermelo-Fraenkel set theory with Choice*, or *ZFC*.

4.2 Naïve Set Theory

Naïve set theory is the informal version of set theory that uses natural language to describe sets and operations on them.

* Again, Gödel's incompleteness theorem means that any sufficiently complicated first-order logic system (this includes almost all axiomatic set theories) cannot be proved consistent from within the theory itself, but it is generally believed that the most common axiomatic systems are consistent, as they do exclude some paradoxes.

A set is a *well-defined* collection of objects called *elements* or *members*.

Notice, however, that this definition of a set doesn't actually tell us how sets are formed, and what operations on a set will produce another set.

As previously mentioned, several paradoxes occur if we do not restrict the definition of a set, which we do with axioms in axiomatic set theory. However, naïve set theory is not necessarily inconsistent either, if it specifies the sets allowed to be considered with *definitions*. Unfortunately, "well-defined" by itself doesn't sufficiently and unambiguously guarantee what does and does not constitute a set.

For this section, "well-defined" is interpreted as an intention to rule out inconsistencies: the paradoxes of naïve set theory only occur in specific contexts, and is usually irrelevant to the usually simpler context in which we're working.

Throughout this section, we will point out some problems in naïve set theory and name the axioms in ZFC which resolve them, but will not discuss them yet.

4.2.1 Specifying Sets

The simplest way to specify a set of objects is to list them between curly brackets – this is known as defining a set *extensionally*. For instance, 2 is in the set $\{1,2,3\}$ – the set containing 1, 2 and 3 – and is also in the set of even integers. The set of even integers is clearly infinite, so there is no requirement that sets need be finite in size. Sets can also be empty – there is a unique set with no elements called the *empty set*, denoted $\varnothing$, but we could also write the empty set as $\{\}$.

This notation can also be informally abused by writing something like, {dogs} to indicate "the set of all dogs" with the word itself acting as a descriptor, but this set, as written, could be interpreted by mathematicians to be "the set containing the single element "dogs"."

We also have the *universe of discourse*, denoted U, which is the set of everything relevant for whatever we're working on. We use the symbol $\in$ to denote membership and $\notin$ for its negation, so we could write $2 \in \{1,2,3\}$, or $4 \notin \{1,2,3\}$. We can also reverse the symbol and write $\ni$ for "has an element", so $\{1,2,3\} \ni 2$.

In principle, elements can be anything we want. For instance, this is a (somewhat unusual and not very prototypical) set containing a wide variety of elements:

$$\{512,\{e\},\text{dogs},\{1,2\},\{36,\{\{5,3\},\{1,\text{cats}\},8\},\varnothing,\{7\}\}\}$$

But, we usually just care about only having mathematical objects as elements, such as numbers, or even other sets, which themselves will contain other elements. For instance, the set above contains $\{1,2\}$ and $\{36,\{\{5,3\},\{1,\text{cats}\},8\},\varnothing,\{7\}\}$ which are sets, the latter of which contains even more sets. An element of a set which is not itself a set is called a *urelement*, so 512 and "dogs" are urelements in the set above. In contrast, a *hereditary* or *pure* set, is a set whose elements are hereditary sets – that is, all elements of a hereditary set are themselves sets, and all elements of elements are sets, and so on.

Note that $\in$ only represents direct membership, so the set $\{\{1\}\}$ is "the set containing the set containing 1", but doesn't contain 1 itself, so $\{1\} \in \{\{1\}\}$ but $1 \notin \{\{1\}\}$. That is, membership is not transitive. There is also no standard notation for being an element of an element, or other nested memberships.

Two sets, A and B are *equal* when the contain precisely the same elements (in axiomatic set theory, this is the axiom of extensionality) – that is, every element of A is an element of B, and every element of B is an element of A. This means that a set is completely determined by its elements, not its description, so the set with elements 2, 3 and 5 is equal to the set of all prime numbers less than 6. If A is equal to B, then we write $A = B$ as you might expect.

Another side effect of this definition of equality is that sets are *unordered* and elements are *unique*, which means that we don't care about the order of the objects, and we don't care about duplicates, so, these sets are all equal:

$$\{1,2,3\} = \{3,2,1\} = \{1,1,1,3,1,3,2,3,3,2\}$$

As well as defining sets extensionally, we can define them *intensionally* by writing $\{x : P(x)\}$ or sometimes, $\{x \mid P(x)\}$[*] which denotes the set containing all objects for which the condition P holds. This notation is called *set-builder notation* or *set comprehension*, particularly in the context of functional programming.[†] Some variants of this notation include,

- $\{x \in A : P(x)\}$: the set of all x that are members of A such that P holds. For example, $\{x \in \mathbb{Z} : x \text{ is even}\}$ is the set of all even integers. In axiomatic set theory, this is justified by the *axiom schema of specification.*
- $\{F(x) : x \in A\}$: the set of all objects obtained by putting the members of the set A into the formula F. For example, $\{2x : x \in \mathbb{Z}\}$ is also the set of even integers. In axiomatic set theory, this is justified by the *axiom of replacement.*
- $\{F(x) : P(x)\}$: the most general form of set-builder notation, denoting the set of objects obtained by putting the members of the set of all objects such that P holds into the formula F.

So if we wanted to write down the set of all dogs, we should really write $\{x : x \text{ is a dog}\}$.

The idea that any predicate P generates a set is called *unrestricted comprehension*. Unfortunately, this seemingly reasonable idea happens to be too strong, and directly leads to contradictions. For example, *Russell's paradox*: if $R = \{x : x \notin x\}$, then $R \in R \leftrightarrow R \notin R$.[‡] In axiomatic set theory, this problem is solved by the *axiom schema of specification* (which is alternatively known as the axiom schema of restricted comprehension for this reason).

Sometimes, it is useful to be able to refer to multiple sets at once. We call a collection of sets a *family of sets*. A family isn't necessarily a set, because it may contain duplicates which we care about.

4.2.2 Set Operations

The *union* of two sets A and B is the set of all objects which are elements of A or of B, or or both (axiom of union), and is denoted $A \cup B$. Unions are the set-theoretic version of $\wedge$.

The *intersection* of A and B is the set of all objects which are both in A and in B, denoted $A \cap B$. Intersections correspond to $\vee$.

The *relative complement* of B relative to A or the *set-theoretic difference* of A and B (note the different order), is the set of all objects that are in A but not in B, written as $A \setminus B$ or sometimes as $A - B$.[§]

[*] This latter notation can cause issues when P includes absolute values, norms, or other similar notations. For example, $\{x : \|x\| > 0\}$ is slightly clearer than $\{x \mid \|x\| > 0\}$.

On the other hand, if our predicate or elements were functions and we needed to write down the domain and codomain within the set (this happens if, for example, we are indexing by the domain or codomain in some way), then $\{\iota : A \hookrightarrow X \mid A \subseteq X\}$ is clearer than $\{\iota : A \hookrightarrow X : A \subseteq X\}$.

One other problem with : is that sometimes quantifiers (and type annotators) are separated with :. However, quantifiers essentially never appear on the left side of a set written in set-builder notation, and are moreover always on the right side of a $\forall$ or $\exists$ symbol, so ambigous cases are effectively non-existent.

In any case, both notations are very popular, so either is acceptable, but we will generally use the former.

[†] Some high-level programming languages, such as Haskell (which is a purely functional language) and Python have similar built-in methods called *list comprehension*, which is almost the same thing as set comprehension, except that order does matter for lists, and duplicates can be included. In maths, we call ordered sets that can contain duplicate elements *tuples*, and sets which can contain just duplicate elements *multisets*. Note, however, that "tuple" in Python refers to an unmodifiable list, which is a very different data structure. Tuples in mathematics must also have finite cardinality. In computer science, this also happens to be the case, but is less of a design decision than a limitation of our physical reality.

[‡] Other related problems include the *barber paradox* and the *Grelling-Nelson paradox*, which can both be rephrased in terms of set-theory and Russell's paradox.

[§] We highly discourage this latter notation because $A - B$ could refer to $\{a - b : a \in A, b \in B\}$ (the *Minkowski difference* or *geometric difference* of A and B), and similarly, we don't write $A + B$ for $A \cup B$, because this could refer to $\{a + b : a \in A, b \in B\}$ (the Minkowski sum or *dilation* of A and B). $A \setminus B$, on the other hand, has far fewer notational

In contrast to the relative complement, we also have the *absolute complement* of A, which is the set of objects that are not in A, corresponding to the notion of negation. There are various notations for this, including A^C, A' and $\bar{A}$. Note that the set of objects being considered must implicitly belong to the universe set, so another way to write this is $U \setminus A$.

The *symmetric difference* of A and B is the set of all objects which are in A or in B, but not both, written as $A \triangle B$. Symmetric differences correspond to $\veebar$.

Given two sets, A and B, A is a *(non-strict) subset* of B if every element of A is also an element of B, and is written $A \subseteq B$. Note that if $A = B$, this still holds, so every set is a subset of itself. If $A \neq B$ but $A \subseteq B$, we say that A is a *proper* or *strict* subset, denoted as $A \subset B$. As with $\in$, we can reverse these symbols to represent *supersets*.* Subsets are the set-theoretic version of $\rightarrow$.

Note that in natural language, there are two ways for a set to be "in" another set: $A \in B$ and $A \subseteq B$. These two notions of being inside another set do not always coincide – it is possible for one of these to be true, and the other false.

For example,

- $A = \{1\}$, $B = \{\{1\}\}$. $A \in B$, but $A \not\subseteq B$, because the set $\{1\}$ is in B, but 1 is not, so A contains elements that are not in B.
- $A = \{1\}$, $B = \{1,2,3\}$. $A \subseteq B$ (and $A \subset B$ as well), but $A \notin B$, because B doesn't contain an element $\{1\}$.

We will try use the wording "is in" for $\in$ and "is contained in" for $\subseteq$ in natural language definitions and discussions, but it is safest to refer any the symbolic definitions included.

The *power set* of a set A is the set of all subsets of A, and is written $\mathcal{P}(A)$.

The *cardinality* of a set A is the number of elements in A, and is written $|A|$. The cardinality of the power set of A is $|\mathcal{P}(A)| = 2^{|A|}$. This is because every subset either contains or doesn't contain any particular element, so each element is a binary choice, and there are $2^{|A|}$ possible ways to include or not include the $|A|$ elements, with each selection of inclusions giving a different unique subset.

If the intersection of two sets is the empty set, the two sets are *disjoint*. If a family of sets are all pairwise disjoint, they are *mutually exclusive*. If the union of two sets A and B is equal to a set C, then we say A and B *cover* or *exhaust* C. If the sets A,B,C,... are mutually exclusive and cover a set, S, then they form a *partition* of S and we say that S is *partitioned* by A,B,C,....

We can write all of these set operations and relations in terms of logical connectives in set-builder notation:

- $A \cup B = \{x : x \in A \vee x \in B\}$;
- $A \cap B = \{x : x \in A \wedge x \in B\}$;
- $A \setminus B = \{x : x \in A \wedge x \notin B\}$ or $\{x \in A : x \notin B\}$;
- $A \triangle B = \{x : x \in A \veebar x \in B\}$;

collisions.

* There are other different and incompatible notations, the most popular of which are listed below

- $A \subseteq B$ for (non-strict) subset, and $A \subset B$ for strict subset (this what we are using).
- $A \subset B$ for (non-strict) subset, and $A \subsetneq B$ for strict subset.

However, there is good reason to use the convention we have selected for this document: $\subseteq$ and $\subset$ as we have defined them correspond nicely with the inequality symbols $\leq$ and $<$, respectively.

For example, $x \leq y$ means that x may or may not equal y, but $x < y$ means that x certainly does not equal y. Similarly, $X \subseteq Y$ means that X may or may not equal Y, but $X \subset Y$ means that X certainly does not equal Y.

Additionally, as we will see later in §4.5, our choice of convention matches up very nicely with certain set-theoretic constructions of the natural numbers.

These notations are not universal, however, so any convention can be used as long as it is clearly stated.

- $A \subseteq B \leftrightarrow \forall x : x \in A \rightarrow x \in B$;
- $A' = A^C = \bar{A} = U \setminus A = \{x \in U : x \notin A\}$ or $\{x : x \notin A\}$.

and we can see some similarities between the symbols $\cap$ and $\cup$ and $\wedge$ and $\vee$. This is not a coincidence.

4.2.3 Proofs with Sets

So far, we only really have three things we can prove about sets – given S and T, show:

- $S \in T$;
- $S \subseteq T$;
- $S = T$

and their negatives. We can also show $S \subset T$, but this is just the same as showing $S \subseteq T$ and $S \neq T$.

Showing $S \in T$ requires unwrapping the structure of T and seeing what conditions are required for something to be an element of T. Showing $S \subseteq T$ is the same as showing an element is in T, but now we prove it for every element of S. Showing $S = T$ is generally done by showing $S \subseteq T$ and $T \subseteq S$. The former says $\forall x : x \in S \rightarrow x \in T$, while the latter says $\forall x : x \in T \rightarrow x \in S$, which together imply $\forall x : x \in S \leftrightarrow x \in T$, which satisfies the definition of equality.

Because $S \nsubseteq T$ and $S \neq T$ are existential statements, rather than universal ones, they generally have simpler proofs. For example, disproving $S \subseteq T$ only requires a counterexample: pick an element of S and show it isn't in T.

There are a couple of useful standard results, which are helpful in other proofs, so we'll package them together as a lemma:

Lemma 4.2.1. *The following statements hold for all sets S and T and predicates P:*

1. $S \supseteq S \cap T$;
2. $S \subseteq S \cup T$;
3. $S \supseteq \{x \in S : P(x)\}$;
4. $S = (S \cap T) \cup (S \setminus T)$.

Proof.

1. Let x be in $S \cap T$. Then, $x \in S$ and $x \in T$ by the definition of $S \cap T$, so we have $x \in S$. As the choice of x was arbitrary, by $\forall$I, $\forall x \in S \cap T$, we have $x \in S$, so $S \supseteq S \cap T$, as required.
2. Let $x \in S$, so $x \in S \vee x \in T$ holds, satisfying the definition of $S \cup T$.
3. Let x be in $\{x \in S : P(x)\}$. Then, by the definition of set comprehension, $x \in S \wedge Px$ holds. By $\wedge\mathrm{E_L}$, we have $x \in S$.
4. To show equality, we need to show that each side is a subset of each.

   First, let $x \in S$. If $x \in T$, then $x \in (S \cap T)$; if $x \notin T$, then $x \in (S \setminus T)$. By the law of the excluded middle, we know at least one of these is the case, so $x \in (S \cap T) \cup (S \setminus T)$. Because x was arbitrary, $S \subseteq (S \cap T) \cup (S \setminus T)$.

   Now, let $x \in (S \cap T) \cup (S \setminus T)$. If $x \in (S \setminus T)$, then $x \in S$ and $x \notin T$; if $x \in (S \cap T)$, then $x \in S$ and $x \in T$. In either case, $x \in S$. As x was arbitrary, $(S \cap T) \cup (S \setminus T) \subseteq S$.

   Because both $S \subseteq (S \cap T) \cup (S \setminus T)$ and $(S \cap T) \cup (S \setminus T) \subseteq S$, $S = (S \cap T) \cup (S \setminus T)$, as required.

■

Using similar arguments, we can translate across the other properties of $\wedge$ and $\vee$.

$$(A^C)^C = A \qquad \text{Double negation or involution law}$$

$$\begin{aligned} A \cup B &= B \cup A \\ A \cap B &= B \cap A \end{aligned} \qquad \text{Commutativity laws}$$

$$\begin{aligned} (A \cup B) \cup C &= A \cup (B \cup C) \\ (A \cap B) \cap C &= A \cap (B \cap C) \end{aligned} \qquad \text{Associativity laws}$$

$$\begin{aligned} A \cup (B \cap C) &= (A \cup B) \cap (A \cup C) \\ A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) \end{aligned} \qquad \text{Distributive laws}$$

$$\begin{aligned} A \cup \varnothing &= A \\ A \cap U &= A \end{aligned} \qquad \text{Identity laws}$$

$$\begin{aligned} A \cup U &= U \\ A \cap \varnothing &= \varnothing \end{aligned} \qquad \text{Domination laws}$$

$$\begin{aligned} A \cup A &= A \\ A \cap A &= A \end{aligned} \qquad \text{Idempotency laws}$$

$$\begin{aligned} A \cap A^C &= \varnothing \\ A \cup A^C &= U \end{aligned} \qquad \text{Negation laws}$$

$$\begin{aligned} (A \cup B)^C &= A^C \cap B^C \\ (A \cap B)^C &= A^C \cup B^C \end{aligned} \qquad \text{De Morgan's laws}$$

$$\begin{aligned} A \cup (A \cap B) &= A \\ A \cap (A \cup B) &= A \end{aligned} \qquad \text{Absorption laws}$$

$$\begin{aligned} \varnothing^C &= U \\ U^C &= \varnothing \end{aligned} \qquad \text{Complement laws}$$

Compare and contrast with §2.2.3.

4.3 Axiomatic Set Theory

As we saw earlier, there are problems with allowing anything to be an element of sets. To resolve this, we define axioms, and only use sets we can prove to exist from those axioms.

Unless you are a logician or a set theorist working at this foundational level, knowing the axioms comprehensively is very much unnecessary. In practice, you just need to know what sets you can construct, and the things you can do to existing sets to get other sets.

For most use cases, the following axioms are sufficient:

- *Axiom of extensionality*: every set is determined by its elements.
- *Axiom of pairing*: if a and b are objects, then there exists a set containing only a and b.
- *Axiom schema of separation*: the elements of a set which satisfy a predicate is also a set.
- *Axiom of the power set*: the set of all subsets of a set is itself a set.
- *Axiom of the union*: for every set, T, there is another set, $\cup T$ that contains as elements precisely all the elements of the elements of T.
- *Axiom of infinity*: there exists an infinite set.

These axioms form *Zermelo set theory*,[*] denoted Z^-. The language of Z^- includes the membership relation $\in$, the equality relation $=$, and an extra unary predicate that identifies whether an object is a set or a urelement.

Most of these axioms translate fairly closely into *Zermelo-Fraenkel* axiomatisation, or ZF. Fraenkel's contribution to ZF was the *axiom of replacement*, and von Neumann later added the *axiom of regularity*.

As we will see later, without replacement, certain infinite sets cannot be constructed. Virtually all results in all branches in mathematics hold in the absence of regularity, but including it makes working with ordinal numbers easier, and it allows you to do induction in infinite sets larger than the naturals. Regularity also prevents sets from being elements of themselves, as well as infinitely descending chains of sets.

Finally, the *axiom of choice* is added to ZF to form *Zermelo-Fraenkel set theory with Choice*, or *ZFC*. The exact phrasing of the axioms included with ZFC depend on the author, but these different axiomatisations will all be equivalent. For example, the axiom of regularity is sometimes replaced with the *axiom of induction*, the axiom schema of replacement with the *axiom schema of collection*, and more. Some of these also depend on the choice of underlying logic as well – for instance, the regularity and induction axioms are equivalent under ordinary first-order logic, but are not under intuitionistic logic. There is a vast variety of axiomatisations that may or may not be equivalent depending on logic systems and various other factors, leading to a rich collection of internally consistent but incompatible theories. These different systems will not be discussed here, and we will continue with ZFC with classical first-order logic.

All formulations of ZFC imply that at least one set exists, but some authors also include an extra axiom that more directly asserts the existence of a set, for instance, the *axiom of the empty set*, $\exists x \forall y : y \notin x$ which implies that the empty set exists. However, these kinds of axioms are now often excluded because, in the semantics of first-order logic, the universe of discourse must be non-empty, so it is already a theorem of first-order logic that something exists (this is usually expressed as $\exists x : x = x$), and because everything in ZFC is a set, this therefore implies that a set exists. Furthermore, the axiom of infinity asserts that an infinite set exists, again implying that a set exists, so an additional axiom of existence is unnecessary.

4.3.1 Axiom of Extensionality

Two sets are equal (are the same set) if they have the same elements:

$$\forall x \forall y [\forall z (z \in x \leftrightarrow z \in y) \rightarrow x = y]$$

That is, a set is determined uniquely by its members.

4.3.2 Axiom of Regularity (or Foundation)

Every non-empty set, x, contains a member y such that x and y are disjoint:

$$\forall x [\exists a (a \in x) \rightarrow \exists y (y \in x \land \neg \exists (z \in y \land z \in x))]$$

[*] Mostly. Some of these axioms have been shortened and rephrased to be easier to explain.

Once we have proven the existence of the empty set (either by defining it to exist, depending on the system, or through the use of the axiom schema of specification), we can rewrite this as,

$$\forall x(x \neq \varnothing \to \exists y(y \in x \land y \cap x = \varnothing))$$

This axiom, along with the axiom of pairing, implies that no set is an element of itself.

Another notable consequence of this axiom is that there are no urelements in ZF: every object in the universe of discourse is a set – and specifically, a hereditary set. You might think that is rather strange for something we use as the foundations of maths given that things like 1 or π exist, and they don't look like sets, but it turns out that we can define numbers, and everything else, in terms of sets. We do so under the axiom of infinity, and later in §4.5.

4.3.3 Axiom Schema of Specification (or Separation/Restricted Comprehension)

The subset of a set z obeying a formula $\varphi(x)$ with a free variable x, $\{x \in z : \varphi(x)\}$, always exists. Formally, if φ is a formula in the language of ZF, x,z,w_1,w_2,…,w_n are free variables, and y is not free in φ, then:

$$\forall z \forall w_1 \forall w_2 \ldots \forall w_n \exists y \forall x[x \in y \leftrightarrow (x \in z \land \varphi(x,w_1,w_2,\ldots,w_n,z))]$$

This axiom is what resolves Russell's paradox: we can only construct *subsets* that obey formulae, so objects like $\{x : \varphi(x)\}$ are not sets.

This axiom is redundant in some other axiomatisations of ZF that include an existence axiom, in that, it follows from the axiom schema of replacement combined with the axiom of the empty set. On the other hand, the axiom schema of specification can be used to prove the existence of the empty set, once at least one set is known to exist, which we do know from the axiom of infinity.

We do this by using a property φ which no set has. For example, if w is any existing set, then the empty set can be constructed as,

$$\varnothing = \{u \in w : u \in u \land \neg(u \in u)\}$$

So the axiom of the empty set is implied by this set of axioms. Furthermore, the axiom of extensionality implies that the empty set is unique.

The main alternative logic to first-order logic is *type theory*, which corresponds closely to data types in computer science. Rather than restricting what sets can exist with axioms, type theory modifies the underlying logic so that self-referential sets are not problematic – in particular, ZFC is defined both by the rules of first-order logic, and its own axioms. Type theories, however, do not have axioms, and are defined entirely by their rules of inference. For instance, in ZFC, Russell's paradox is axiomatically resolved as above, while in type theory, the predicate $x \notin x$ itself is disallowed by inference rules down at the logic level. The axiom of specification and axiom of choice, in particular, are theorems (they can be proven) in type theory.

4.3.4 Axiom of Pairing

If x and y are sets, then there exists a set that contains exactly x and y as elements.

$$\forall x \forall y \exists z \forall w[w \in z \leftrightarrow (w = x \lor w = y)]$$

The axiom of pairing is actually redundant in ZF because it follows from the axiom schema of replacement, given that we already have a set with at least two elements, which is assured by either the axiom of infinity, or the axiom schema of specification combined with the axiom of the power set applied twice to any set.

4.3.5 Axiom of Union

The union of the elements of the elements of a set exists. That is, for any set of sets $\mathcal{F}$, there exists a set A containing every element that is a member of some member of $\mathcal{F}$:

$$\forall \mathcal{F} \exists A \forall Y \forall x[(x \in Y \land Y \in \mathcal{F}) \to x \in A]$$

Although this doesn't directly assert the existence of $\cap\mathcal{F}$, we can construct it from A using the axiom schema of specification.

Additionally, combined with the axiom of pairing, this implies that, for any two sets, there is a set (called the union) that contains exactly the elements of the two sets.

4.3.6 Axiom Schema of Replacement

The image of any set under any definable mapping is also a set – replacing every member of a set with something else gives another set.

Formally, let φ be any formula in the language of ZF, x,y,A,w_1,w_2,$\ldots$,w_n are free variables, and B is not free in φ, then:

$$\forall w_1 \ldots \forall w_n \forall A([\forall x \in A \exists! y \varphi(x,y,w_1,\ldots,w_n,A)] \to \exists B \forall y[y \in B \leftrightarrow \exists x \in A \varphi(x,y,w_1,\ldots,w_n,A)])$$

or, if φ represents a definable function f, A represents its domain, and $f(x)$ is a set for every $x \in A$, then the image of f is a subset of some set B.

4.3.7 Axiom of Infinity

Let $S(w)$ abbreviate $w \cup \{w\}$, where w is a set ($\{w\}$ is a valid set, obtainable by applying the axiom of pairing with $x = y = w$ to obtain $z = \{w\}$). Then, there exists a set X such that the empty set is a member of X, and whenever a set y is a member of X, then $S(y)$ is also a member of X. That is, there exists a set X with infinitely many members.

$$\exists X[\exists e(\forall z \neg(z \in e)) \land e \in X \land \forall y(y \in X \to S(y) \in X)]$$

The above definition includes a clause to define the empty set. If the empty set has been constructed in another way, such as with the axiom schema of replacement, the axiom of infinity can instead be stated as,

$$\exists X[\varnothing \in X \land \forall x \in X(S(x) \in X)]$$

The axiom of regularity is also required to show that all the members of X are distinct. Otherwise, X could be a finite cycle of sets, much like our attempted construction of the naturals from before.

In fact, this axiom is very closely related to the *von Neumann* construction of the natural numbers, and also looks rather similar to the Peano construction of the naturals. We give a brief overview of this now:

We first define 0 to be the empty set, then define each natural number to be S applied to the number before it.

Abbreviating $\{\}$ as $\varnothing$, we have,

- $0 = \varnothing$;
- $1 = 0 \cup \{0\} = \{0\} = \{\varnothing\}$;
- $2 = 1 \cup \{1\} = \{0,1\} = \{\varnothing,\{\varnothing\}\}$;
- $3 = 2 \cup \{2\} = \{0,1,2\} = \{\varnothing,\{\varnothing\},\{\varnothing,\{\varnothing\}\}\}$;
- $4 = 3 \cup \{3\} = \{0,1,2,3\} = \{\varnothing,\{\varnothing\},\{\varnothing,\{\varnothing\}\},\{\varnothing,\{\varnothing\},\{\varnothing,\{\varnothing\}\}\}\}$;

- $n = (n-1) \cup \{n-1\} = \{0,1,\ldots\} =$ a mess of nested brackets

This representation is useful because the cardinality of each set coincides with the number it represents. We also have $n \leq m$ if and only if $n \subseteq m$, and $n < m$ if and only if $n \subset m$ (or $n \in m$).

This construction of natural numbers using sets can then be extended to integers, rationals, reals and beyond with various even more complicated nested set structures, with usual arithmetic operations defined in terms of set operations. We discuss this in §4.5.

There are actually many sets that satisfy the requirements for X. The minimal such set is the transfinite von Neumann ordinal, ω, which is a superset of the natural numbers as defined above. The axiom schema of specification can be applied to remove extra elements, and this set is also unique by the axiom of extensionality.

Some finitist theories reject the axiom of infinity, accepting only finite cardinals. However, the axiom of infinity (alongside other strong cardinal axioms) is generally accepted in modern mathematics.

4.3.8 Axiom of the Power Set

For any set x, there exists a set y that contains every subset of x.

$$\forall x \exists y \forall z [\forall w (w \in z \to w \in x) \to z \in y]$$

or,

$$\forall x \exists y \forall z [z \subseteq x \to z \in y]$$

The axiom of power set appears in most axiomatisations of set theory, and is generally considered uncontroversial.

4.3.9 Axiom of Choice

This axiom turns ZF into ZFC.

For any set X of non-empty sets, there exists a *choice function* f that is defined on X and maps each set of X to an element of the set. That is, given a collection of non-empty sets, it is possible to construct another set by arbitrarily choosing one object from each set, even if the collection is infinite, or equivalently, the Cartesian product (§4.4.1) of a collection of non-empty sets is non-empty.

$$\forall X [\varnothing \notin X \to \exists f : X \to \bigcup X \quad \forall A \in X (f(A) \in A)]$$

For example, given the collection of sets of natural numbers, {3,5,9}, {2,5,7}, {4,5,9}, {1,2,8}, we could construct a new set by say, picking the smallest number from each set, to get {3,2,4,1}. In this case, “select the smallest number” is acting as our choice function. Even if we had infinitely many sets, the choice function will still work, and we will still get a set.

However, for certain collections (notably, the power set of the reals), no choice function is known, so we must invoke the axiom of choice to construct sets from those collections. Because of this, the axiom of choice has historically been controversial, because it doesn’t actually tell you what the choice function is, only that it exists. The axiom of choice is now generally accepted in most axiomatisations of mathematics, as there are many important and core theorems that rely on its use.

4.4 Ordered pairs

Because of the axiom of extensionality, sets are unordered and elements are unique. However, there are often times when we would prefer this to not be the case. We call ordered sets that can contain duplicates *tuples*, and sets which can just contain duplicate elements *multisets*. Tuples must have finite cardinality, but there is no restriction on the size of multisets.

For now, we will only be considering tuples as they are used far more frequently, but multisets are useful in combinatorics, where we might need repeated elements, but not care about order.

Because everything in ZFC is a set, we need to define some kind of set structure to represent tuples. There are many different ways to do this, some more suited for some logics, but the most commonly accepted way is to represent the ordered pair (a,b), also written as $\langle a,b\rangle$, as the structure $\{\{a\},\{a,b\}\}$ (this is the *Kuratowski construction*). We can prove that this representation satisfies all the properties we would want an ordered pair to have. We also call the elements of ordered pairs the first and second *coordinates*, *components*, or *entries*, or the left and right *projections* of the pair, to distinguish them from elements of ordinary sets.

4.4.1 Cartesian Products

With ordered pairs, we can define a new operation: the *Cartesian product* of two sets, A and B, written as $A \times B$, is the set $\{(x,y) : x \in A \wedge y \in B\}$, or, the set of ordered pairs where the first coordinate is from A and the second is from B.

If $|A| = n$ and $|B| = m$, then $|A \times B| = nm$. If A and B are von Neumann ordinals, this is how we actually define multiplication.

Given our selected encoding of ordered pairs, we can prove the existence of the Cartesian product using the axiom of the power set. $\mathcal{P}(A \cup B)$ contains all the sets of $\{x\}$ and $\{x,y\}$ we need, so $\mathcal{P}(\mathcal{P}(A \cup B))$ contains all the pairs $\{\{x\},\{x,y\}\}$ required. It also includes many other sets we don't care about, but they can be removed with the axiom of specification.

Because the result of the Cartesian product is a set of ordered pairs, the Cartesian product is not commutative. Strictly speaking, the Cartesian product as we have defined it is also never associative unless one of the involved sets is empty. For instance, if $A = \{a\}$, then $(A \times A) \times A = \big\{((a,a),a)\big\} \neq \big\{(a,(a,a))\big\} = A \times (A \times A)$. However, it is extremely convenient if we can treat this element as the triple (a,a,a), in which case the Cartesian product would be associative.

The main reason we can do this in practice is because the sets resulting from a string of Cartesian products being evaluated in any order are all equivalent in a "natural" manner – that is, that there are functions between these different orderings that compose "coherently". (For a precise meaning of the word *natural*, see §52.2.5; and for *coherence*, see §52.7.)

Informally, this coherence allows us to treat the elements of an n-fold Cartesian product as n-tuples rather than as sequences of nested pairs. Because of this, outside of set theory, the Cartesian product effectively *is* associative.

This also makes the following notation sensible: if we take the Cartesian product of a set with itself, say $A \times A$, we also write A^2, and continue similarly for repeated Cartesian products, $A^3,A^4,\ldots$ – without associativity, these expressions are not well-defined unless we stipulate an additional ordering of multiplication.

4.4.2 Relations

A (*binary* or *dyadic*) *relation* associates elements of one set, the *domain*, with elements of another set, the *codomain*. We represent a binary relation from the set X to Y as a set of ordered pairs (x,y) with $x \in X$ and $y \in Y$, and we say that an element $a \in X$ is *related to* an element $b \in Y$ if and only if the pair (a,b) is in the set of ordered pairs that define the relation. In this way, a relation from X to Y can be seen as a subset of $X \times Y$, or equivalently, as an element in $\mathcal{P}(X \times Y)$. If two elements, a and b are related by a relation, R, we use infix notation and write aRb.

Two elements, x and y of a set X are *comparable* with respect to a binary relation R if at least one of xRy or yRx is true. For example, if R is "divides", then $2R4$ holds, because 2 divides 4 is true, so 2 and 4 are comparable. However, neither $3R4$ and $4R3$ hold, so 3 and 4 are *incomparable* under R. If two elements x and y are comparable under a relation, we express this using the rather hilarious looking $x \overset{<}{\underset{>}{=}} y$. Similarly, if they are incomparable, we write $x \not\overset{<}{\underset{>}{=}} y$.

If the domain and codomain of a relation are the same set, the relation is called a *homogeneous relation* or an *endorelation*. One example of a (homogeneous) relation is the *identity relation* over a set X, which is the set $\mathrm{Id}_X = \{(n,n) \in X^2 : n \in X\}$, where every element is related only to itself. Equality $=$ is a type of identity relation. Another example is the relation $\leq$, which is the set $\{(n,m) \in \mathbb{N} : n \subseteq m\}$.

There are many properties of interest that relations can have. For a general binary relation R over sets X and Y, the relation can be:

- *Injective* or *left-unique*: $\forall x,z \in X \, \forall y \in Y : xRy \land zRy \to x = y$ – an element of the codomain is related to at most one element in the domain. Note that this does not require any element of the codomain to actually be related to something, only that, if it *is* related to something, it is to at most one thing.
- *Functional* or *right-unique*: $\forall x \in X \, \forall y,z \in Y : xRy \land xRz \to y = z$ – an element of the domain is related to at most one element in the codomain. This has a similar caveat to the above.
- *One-to-one*: injective and functional.
- *One-to-many*: injective and non-functional.
- *Many-to-one*: functional and non-injective.
- *Many-to-many*: neither functional nor injective.
- *Total* or *left-total*: $\forall x \in X \, \exists y \in Y : xRy$ – for all $x \in X$ – every element of the domain is related to at least one element in the codomain.
- *Surjective* or *right-total*: $\forall y \in Y \, \exists x \in Y : xRy$ – every element of the codomain is related to at least one element in the domain.

Injectivity and totality can be seen as inverse properties, as can functionality and surjectivity, as they swap values if the domain and codomain are interchanged.

There are more properties of interest for endorelations, but these are discussed later in §4.4.9.

4.4.3 Operations on Binary Relations

If R and S are binary relations over sets X and Y, then $R \cup S = \{(x,y) : xRy \lor xSy\}$ is the *union relation* of R and S over sets X and Y. That is, the union relation of R and S is the relation that contains pairs that are satisfied by either R or S.

Similarly, $R \cap S = \{(x,y) : xRy \land xSy\}$ is the *intersection relation*, which contains pairs that are satisfied by both R and S.

If R is a binary relation over sets X and Y, and S is a binary relation over Y and Z, we can *compose* R and S into $S \circ R = \{(x,y) : \exists y \in Y : xRy \land ySz\}$, the *composition relation* of R and S over sets X and Z.

If R is a binary relation over sets X and Y, then $R^\top = \{(y,x) : xRy\}$ is the *converse relation* of R over Y and X. A binary relation is equal to its converse if and only if it is symmetric.

If R is a binary relation over sets X and Y, then $\bar{R} = \neg R = \{(x,y) : \neg xRy\}$ is the *complementary relation* of R over X and Y.

4.4.4 Functions

A *function* is a binary relation which is both functional and total (§4.4.2) – so every element of the domain is related to some element in the codomain, and is related to at most one such element. To distinguish a function from an ordinary relation, we write $f : X \to Y$, where X and Y are the domain and codomain of the function, respectively, and for each $x \in X$, we write $f(x)$ to represent the unique $y \in Y$ that satisfies xfy, or equivalently, the unique $y \in Y$ such that $(x,y) \in f$.

The set of all functions from X to Y is written Y^X. This is a subset of $\mathcal{P}(X \times Y)$, so this exists by specification. When the domains are finite, we can uniquely determine a function just by listing its values, but when the domains are infinite, almost all functions cannot be written down. We cannot list its values, as such a list would be infinitely long, and we can't write down some formula to generate its values, because such a formula would almost always be infinitely long for an arbitrary function. However, this generally isn't a problem, because, for all practical purposes, we are usually only interested in functions with enough structure that they can be defined by a formula. For most of the functions we use, we define them not by listing an infinite set of ordered pairs, but by giving a formula to compute $f(x)$ from x. We can also define an *anonymous function* by writing the formula in terms of the elements of X, and not giving a name to the function at all. For example, the named function $f(x) = x^2$ can be written as the anonymous function $x \mapsto x^2$. We often think of functions as mapping values in X to values in Y, so we could also say that in the function above, x is *mapped* to x^2.

4.4.4.1 Surjections, Injections and Bijections

For a function $f : X \to Y$, we have another set of interest: its *image*. This is the set $\{f(x) : x \in X\}$, or, the subset of Y that has some element in X that is related to it.

A function that is additionally surjective or right-total, is a *surjection*. That is, every element of the codomain of a surjection has a source element in the domain that maps to it under the surjection. Another way to phrase surjectivity is that a function is surjective if its image is equal to its codomain.

For example, the function $f : \mathbb{N} \to \mathbb{N}$, $x \mapsto x + 1$ is not surjective, because the element 0 in the codomain does not have a source element that maps to it: that is, 0 is in the domain of f, but not in the image. However, $f : \mathbb{Z} \to \mathbb{Z}$, $x \mapsto x + 1$ *is* surjective, because for every element in the codomain has a source element in the domain.

A function that is additionally injective, or left-unique, is a *injection*. That is, unique elements in the domain are mapped to unique elements in the codomain – the function is one-to-one.

For example, the function $f : \mathbb{N} \to \mathbb{N}$, $x \mapsto x^2$ is injective, because unique elements in the domain are mapped to unique elements in the codomain. However, $f : \mathbb{Z} \to \mathbb{Z}$, $x \mapsto x^2$ is *not* injective, because some distinct elements in the domain are mapped to the same elements in the codomain, for instance, 1 and -1 are both mapped to 1.

A function that is both surjective and injective is *bijective*, or is a *bijection*. We also call bijections *one-to-one correspondences*, which is not to be confused with *one-to-one*, which just means injective and functional.

4.4.4.2 n-ary Functions

If $f : X \times Y \to Z$, then we can just write $f(x,y)$ for $f((x,y))$. In general, a function can have any arity, including zero. A function of arity n is just a function with domain $X_1 \times X_2 \times \cdots \times X_n$ to some codomain Y, and we similarly drop additional brackets and just write $f(x_1,x_2,\ldots,x_n)$.

4.4.4.3 Function Composition

The operations on relations as listed in §4.4.3 all translate over to functions, so we can similarly find compositions of functions.

Two functions, $f : X \to Y$ and $g : Y \to Z$ can be *composed* to give a new function from X to Z called a *composition*, written as $g \circ f$. We define $(g \circ f)(x) = g(f(x))$.

There are some special cases of interest. If the composition function of two functions f and g is the identity function, Id_X or $x \mapsto x$, then we say that g and f are *inverses.* In particular, if $g \circ f = f(g(x)) = \mathrm{Id}_x$, then g is a *left inverse* of f. If $f \circ g = f(g(x)) = \mathrm{Id}_X$, then g is a *right inverse* of f. If g is both a left and right inverse of f, then we just say it is *the* inverse of f.

A function has a left inverse if and only if it is injective, while a function has a right inverse if and only if it is surjective.

Because injectivity and surjectivity are statements about elements, we can move focus towards the function itself by defining injectivity and surjectivity as having a left and right inverse, respectively. This is useful in some fields, particularly in category theory, where we prefer not to look at elements of domains and codomains and are instead more interested in analysing structures of functions instead.

4.4.5 Endorelation Properties

A general endorelation R over a set X can also be:

- *Reflexive*: $\forall x \in X : xRx$ – every element is related to itself. For example, $x \leq x$ holds for all x.
- *Irreflexive*: $\forall x \in X : \neg xRx$ – every element is not related to itself. For instance, $<$ is irreflexive, because $x < x$ does not hold for any x.
- *Symmetric*: $\forall x,y \in X : xRy \to yRx$ – if x is related to y, then y is related to x, similar to commutativity. For example, if $x = y$, then $y = x$.
- *Antisymmetric*: $\forall x,y \in X : xRy \wedge yRx \to x = y$ – if x is related to y and y is related to x, then $x = y$. For example, if $x \leq y$ and $y \leq x$, then $x = y$, so $\leq$ is antisymmetric.
- *Asymmetric*: $\forall x,y \in X : xRy \to \neg yRx$ – if x is related to y, then y is not related to x. For example, $<$ is asymmetric because $x < y$ implies $\neg y < x$, but $\leq$ is not asymmetric. A relation is asymmetric if and only if it is both antisymmetric and irreflexive.
- *Transitive*: $\forall x,y,z \in X : xRy \wedge yRz \to xRz$ – if x is related to y and y is related to z, then x is related to z. For example, the game rock-paper-scissors, with relation R being “wins against” over the set $\{r,p,s\}$ is not transitive, because rRp and pRs but rRs does not hold.
- *Connected*: $\forall x,y \in X : x \neq y \to xRy \vee yRx$.
- *Strongly Connected*: $\forall x,y \in X : xRy \vee yRx$.

4.4.6 Preorders

4.4.6.1 Non-Strict Preorders

A (non-strict) *preorder* or a *quasiorder* is a relation that is reflexive and transitive. We often use $\prec$, $\lesssim$, or similar, to distinguish this type of relation from the more familiar partial orders, but $\leq$ is also common. If $a \prec b$, we say that *a precedes b*, or that *b covers a.*

A preorder is a superset of partial orders and equivalence relations, as defined below: an antisymmetric preorder is a partial order, while a symmetric preorder is an equivalence relation.

4.4.6.2 Strict Preorders

A *strict preorder* is a relation that is irreflexive and transitive. We often use $\prec$ or $<$ for this type of relation.

4.4.7 Partial Orders

4.4.7.1 Non-Strict Partial Orders

A *weak* or *non-strict partial order* is an antisymmetric non-strict preorder. That is, a relation that is reflexive, antisymmetric, and transitive. For example, $\leq$ is a non-strict partial order. The term *partial order* by itself can also refer to this type of non-strict partial order relation. A non-strict partial order is also called an *asymmetric preorder.*

4.4.7.2 Strict Partial Orders

A *strong* or *strict partial order* is an asymmetric strict preorder. That is, a relation that is irreflexive, antisymmetric and transitive. For example, $<$ is a strict partial order.

Irreflexivity and antisymmetry imply asymmetry, so an alternative definition replaces the requirement for antisymmetry with asymmetry, but asymmetry also implies irreflexivity, so either irreflexivity or asymmetry (but not both) can technically be completely omitted from this definition.

Irreflexivity and transitivity also imply asymmetry, so every strict preorder is a strict partial order. Because a strict partial order is defined to be a type of strict preorder, this means that the two relations are actually equivalent: a relation is a strict partial order if and only if it is a strict preorder.

4.4.8 Total Orders

Note that both strict and non-strict partial orders do not require every element to be comparable. In contrast, a *total order* is a partial ordering where every element is comparable: that is, they are additionally connected. Total orders can be strict or non-strict as above.

4.4.9 Equivalence Relations

An *equivalence relation* is a binary relation that is reflexive, symmetric, and transitive. If two elements are related by an equivalence relation, we say they are *equivalent*.

Let R be an equivalence relation over a set X. The *equivalence class* of an element $a \in X$ with respect to R is the set $[a]_R = \{x \in X : xRa\}$, or, the set of all elements of X that are equivalent to a. The equivalence class of a is sometimes just written $[a]$, if the equivalence relation is clear. The set of all equivalence classes partitions X.

If aRb, then we say a and b are equivalent *up to R.* For example, if our relation R is aRb if a is a permutation of b, then the tuples, (1,2,3) and (3,2,1) are said to be equivalent *up to ordering.* Equivalence relations are a mathematician's way of saying they don't care about some property. If you're just looking at the contents of some tuples, and you don't care about the order, you can just define an equivalence relation to put tuples with the same contents in the same equivalence class, as above, and just treat them as the same object. Later, in §10.2, we work with integers, but we only care about the remainder under division by some number, m, so we define an equivalence relation* to do exactly that.

4.4.10 Well-Founded Relations

A relation, R, is *well-founded* on a collection of sets, X, if every non-empty subset $S \subseteq X$ has a minimum element with respect to R, that is, some element $m \in S$ such that sRm does not hold for all $s \in S$.

In conjunction with the axiom of choice (the full axiom is not required, some weakened variants are sufficient), a relation is well-founded if it contains no infinitely descending chains. That is, there is no infinite sequence $x_0,x_1,x_2,\ldots \in X$ such that $x_{n+1}Rx_n$ for all natural n.

* We actually define a *congruence relation*, which is an equivalence relation that is also compatible with the operations on some structure. That is, performing an operation on two equivalent objects yields equivalent objects.

The axiom of regularity in ZFC asserts that all sets are well-founded.

There is a related notion of *converse well-founded* relations, where infinitely ascending chains are disallowed, but we will not be using these here.

4.4.11 Well-Ordering

A *well-ordering* is a total order that is well-founded. That is, a well-ordering on a set S is a total order on S such that every non-empty subset of S has a least element in this ordering. S, together with this well-order relation is then called a *well-ordered set*. Equivalently, a relation is a well-ordering if it is a well-founded total ordering.

Similarly to well-founded relations, no infinitely descending chains, $x_1 > x_2 > x_3 > \ldots$, can exist.

To show that this is implied by every set having a least element, suppose that a given total order has the least-element property. Then, given a supposedly infinitely descending chain, $x_1 > x_2 > x_3 > \ldots$, a least element, x_i exists. But then, $x_i \not> x_{i+1}$. Conversely, suppose some non-empty set S does not have a least element. Then, we can construct an infinitely descending chain by choosing any x_1 arbitrarily, then recursively generate a descending chain by selecting some element less than the smallest of $x_1, \ldots, x_i$ to be x_{i+1} for all i. Doing so requires the axiom of choice (§4.3.9), or Zorn's lemma (§5.4).

The useful property of well-orders, is that induction (§5) works on well-ordered sets. That is, if, P holds for the smallest element, m, in a set, S, and $P(x')$ for all $x' < x$ implies $P(x)$, then P holds for all $x \in S$. The proof is that if P doesn't hold for at least some x, then there is a least element $y \in \{y \in S : \neg P(y)\} \subseteq S$ for which P doesn't hold, which exists because S, and hence any subset of S, is well-ordered. But, if $y = m$, this contradicts that $P(m)$ holds. Otherwise, $y > m$, in which case P holds for all $x < y$, so $P(y)$ holds by the second assumption, which is a contradiction.

For sets that are not well-ordered, this argument generally does not work. For example, induction doesn't work on the integers, because there isn't a number negative enough to work as a base case for all integers. Even if we include a new minimum element, $-\infty$, we can't finish the inductive step because we can't find the minimum y in the set of integers excluding $-\infty$.

It is possible to have a well-ordering in an infinite set where some elements do not have predecessors. For instance, consider the ordering of $\{0,1\} \times \mathbb{N}$ given by $(a,b) \prec (x,y)$ if $a < x$, or $b < y$ if $a = x$. This is a well-ordering because no infinitely descending chains exist. However, the element, (1,0) does not have a predecessor. We call (1,0) a *limit ordinal* – a number with predecessors, but no direct predecessor. Induction can still be done on these sets (specifically, a generalisation of weak induction called *transfinite* or *well-founded* induction. See §5.4 for further discussion.), but more work is required to deal with these extra cases that don't occur in regular induction.

4.4.12 Lattices

A *lattice* is a partially ordered set, with ordering relation $\prec$, such that,

- Every pair of elements, x and y has a unique *supremum* (§34.3.2) or *meet*, written as $x \wedge y$, such that,
  - $(x \wedge y) \prec x$;
  - $(x \wedge y) \prec y$;
  - $z \prec (x \wedge y)$ for any z such that $z \prec x$ and $z \prec y$.
- Every pair of elements, x and y has a unique *infimum* or *join*, written as $x \vee y$, such that,
  - $x \prec (x \vee y)$;
  - $y \prec (x \vee y)$;

  - $(x \vee y) \prec z$ for any z such that $x \prec z$ and $y \prec z$.

The meet and join are dual: any true statement can be transformed into another true statement by interchanging meets and joins and inverting the direction of any orderings.

You may also recognise the symbols for meet and join from earlier symbolic logic. This is because those previous logic statements in Boolean algebra actually have this lattice structure. Boolean algebra also has an additional operator, the complement operator $\neg$, and has the property that meet and join also distribute over each other, making Boolean algebra a *complemented* and *distributive* lattice. Set algebra is also isomorphic to Boolean algebra.

Examples of lattices are,

- Any total ordering:
  - $x \wedge y$ is min(x,y);
  - $x \vee y$ is max(x,y).
- Subsets of a fixed set ordered by inclusion:
  - $x \wedge y$ is $x \cap y$;
  - $x \vee y$ is $x \cup y$.
- The divisibility relation on the positive integers (§10.1):
  - $x \wedge y$ is gcd(x,y);
  - $x \vee y$ is lcm(x,y).

4.4.13 Minimal & Maximal Elements

If for some x, $y \leq x$ only if $y = x$, then x is *minimal.* Or equivalently, x is minimal if there does not exist any y such that $y < x$. A partial order may have any number of minimal elements, including none. For example, the integers have no minimal element, the naturals have one minimal element, 0, and a set with k mutually incomparable elements has k minimal elements.

If an element x satisfies $x \leq y$ for all y, then x is a *minimum.* A partial order may have at most one minimum, such as 0 in the naturals, but can also have none at all, either because it contains an infinite descending chain like with the integers, or because it has more than one minimal element. Any minimum element is also minimal.

We define maximal and maximum elements similarly, as elements that are not less than any other element and elements that are greater than all other elements, respectively. Again, maximum elements are also maximal.

While these definitions seem similar, they are distinct, elements can be maximal, but not maximum. For example, consider the family of all subsets of $\mathbb{N}$ with at most three elements, ordered by $\subseteq$. Then, the set $\{0,1,2\}$ is a maximal element of this family, because it is not a subset of any larger set, but it is not a maximum, because it is not a superset of $\{3\}$ (and similarly for any other three-element set).

4.4.14 Exercises

- Prove that the definition of ordered pairs we chose ($(a,b) := \{\{a\},\{a,b\}\}$) satisfies the characteristic property, $(a,b) = (c,d) \leftrightarrow (a = c \wedge b = d)$.
- Prove that the Cartesian product of A and B is commutative only if $A = B$, or at least one of A and B is the empty set.
- Prove that the Cartesian product is not associative unless one of the involved sets is the empty set.

- Give an example of a symmetric and transitive relation which is not reflexive.
- Prove that if a function is right invertible, then the function is surjective (you will need to use the axiom of choice).
    - Prove that if $f : X \to Y$ has a right inverse, then this right inverse is injective.
    - Prove that if $f : X \to Y$ has a left inverse, then this left inverse is surjective.
- Let $S \subset \mathbb{R}^2$ be a unit circle (i.e. the set $\{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$).
    - Give an explicit example of a surjection $f : S \times \mathbb{R} \to \mathbb{C}$.
    - Give an explicit example of an injection $g : S \times \mathbb{R} \to \mathbb{C}$.
- Suppose n is a positive natural number. Define the relation R where xRy if n divides $b - a$. Prove that R is an equivalence relation. (This is the equivalence relation of numbers *modulo n*.)
- Suppose X is a set, and let R be an equivalence relation on X. Prove that the set of equivalence classes partition X. That is,
    - For all $a \in X$, $a \in [a]$;
    - For all $a,b \in X$, either $[a] = [b]$ or $[a] \cap [b] = \varnothing$.

4.5 Constructing the von Neumann Universe

4.5.1 The Naturals

If there exists a bijection between two sets, X and Y, then $|X| = |Y|$. We can use this fact to measure the size of arbitrary sets, as long as we have a standard list of sets we already know the cardinality. The most common choice is the *von Neumann ordinals*.

As in the definition of the axiom of infinity (§4.3.7), let $S(w)$ abbreviate $w \cup \{w\}$, where w is a set. We can use S to construct the naturals, $\mathbb{N}$: we define 0 to be the empty set, then define each natural number to be S applied to the number before it.

Abbreviating $\{\}$ as $\varnothing$, we have,

- $0 = \varnothing$;
- $1 = 0 \cup \{0\} = \{0\} = \{\varnothing\}$;
- $2 = 1 \cup \{1\} = \{0,1\} = \{\varnothing,\{\varnothing\}\}$;
- $3 = 2 \cup \{2\} = \{0,1,2\} = \{\varnothing,\{\varnothing\},\{\varnothing,\{\varnothing\}\}\}$;
- $4 = 4 \cup \{3\} = \{0,1,2,3\} = \{\varnothing,\{\varnothing\},\{\varnothing,\{\varnothing\}\},\{\varnothing,\{\varnothing\},\{\varnothing,\{\varnothing\}\}\}\}$;
- $n = (n-1) \cup \{n-1\} = \{0,1,\ldots\} =$ a mess of nested brackets

Natural numbers represented as hereditary sets like this are called von Neumann ordinals. There are ways to extend von Neumann ordinals to represent infinite quantities of different sizes.

This representation is useful because the cardinality of each set coincides with the number it represents. We also have $n \leq m$ if and only if $n \subseteq m$, and $n < m$ if and only if $n \subset m$ (or $n \in m$).

Addition and multiplication can then also all be defined in terms of the successor function, S, though it does get rather messy.

4.5.2 The Integers

The integers, $\mathbb{Z}$ can be represented as ordered pairs. We encode the pair (x,y) as the integer $x - y$. Because this representation isn't unique, i.e. $3-2$ and $4-3$ both represent the integer 1, we also include an equivalence relation $(a,b) = (c,d)$ if and only if $a+d = c+b$. So, a positive integer z is then represented as the equivalence class $[(z,0)]$, while a negative integer z is represented as the equivalence class $[(0,z)]$.

We can then define addition as $(a,b) + (c,d) = (a + c,b + d)$, where the internal addition is just natural number addition as defined with successor functions. Subtraction is then, $(a,b) - (c,d) = (a,b) + (d,c) = (a + d,b + c)$, and multiplication is, $(a,b) \cdot (c,d) = ((a \cdot c) + (b \cdot d),(a \cdot d) + (b \cdot c))$.

4.5.3 The Rationals

The rationals are then simple enough to extend from the integers: the rational number $\frac{p}{q}$ is encoded as the ordered pair (p,q), where p and q are integers. Again, we have duplicates, such as $(-a,b)$ and $(a,-b)$ representing the same rational, so we include another equivalence relation.

Arithmetic operations can then also be defined rather laboriously in terms of set operations.

4.5.4 The Reals

It turns out that the reals can be defined in several different ways. Because these constructions can be rather complicated, especially given that we're already several levels of nested ordered pairs in, for ease of use, we often just axiomatically define the real numbers as a type of algebraic structure called a field (§11.2), or even in terms of smaller structures (§12.10), but we can also explicitly construct it.

In real analysis (§34), one standard technique is *construction from Cauchy sequences* or *completion*. But the more set-theoretic way is to define the partial orderings $<$ and $\geq$, then encode the real number x as the pair of sets, $\{y \in \mathbb{Q} : y < x\}$ and $\{y \in \mathbb{Q} : y \geq x\}$. This is known as a *Dedekind cut*. A Dedekind cut is any pair of subsets, (A,B) of $\mathbb{Q}$, such that,

- $A \cap B = \mathbb{Q} \land A \cup B = \varnothing$ – A and B partition $\mathbb{Q}$;
- $\forall a \in A\, \forall b \in B : a < b$ – every element of A is less than every element of B;
- $\forall x \in A\, \exists y \in A : x < y$ – A does not contain a largest element.

For convenience, we may take A to represent the Dedekind cut (A,B), because B is completely determined by A. By doing this, we can think of a real number more intuitively as being represented by the set of all rational numbers smaller than it.

All the usual arithmetic operations can then be defined rather tediously through various total orderings and set operations.

This construction of the real numbers also allows us to easily obtain the *extended real number system*, a system useful in some areas of calculus where ±∞ are treated as numbers, by associating $-\infty$ with $A = \varnothing$ and ∞ with $A = \mathbb{Q}$. If we just want the regular real number system, we simply restrict A to be non-empty and not equal to $\mathbb{Q}$.

4.5.5 Calculating Cardinalities

The cardinality of the union of two disjoint sets is the sum of the cardinalities of the sets. That is, if A and B are disjoint, then $|A \cup B| = |A| + |B|$. These unions act as addition for cardinalities.

Similarly, the cardinality of the Cartesian product of two sets is the product of the cardinalities of the sets, so $|A \times B| = |A| \cdot |B|$. Even though the Cartesian product is not commutative, swapping the order of A and B just swaps each pair (a,b) with (b,a), which is a bijection, so $|A \times B| = |B \times A|$.

The set of all functions from B to A has size $|A|^{|B|}$. The power set, $\mathcal{P}(A)$ is a special case of this, where we map from the set $\{0{,}1\}$ to A, with each subset being encoded by an *indicator function* which maps each element to 0 or 1, depending on whether the element is included in the subset or not.

A set is *countable* if it has the same cardinality as a subset of the naturals, or equivalently, a set S is countable if there exists an injective function $f : S \to \mathbb{N}$. If the subset is infinite in size, then the set is *countably infinite*. Countable sets are also called *enumerable* or *denumerable*.

4.5.6 Cardinals & Ordinals

It is now convenient to discuss the difference between cardinal numbers and ordinal numbers, given the name of the von Neumann *ordinals*.

The cardinality of a set, as mentioned before, is the number of elements of the set, which we can measure with bijections.

We use cardinalities to talk about *amounts* – how many things there are. We usually use the natural numbers as cardinals. For example, in “5 apples”, and “3 sheep”, the natural numbers 5 and 3 are cardinal numbers.

But how many naturals are there? What is the cardinality of $\mathbb{N}$?

It can’t be a number in the naturals, because there’s always 1 plus that number, so clearly, whatever natural number we pick will never be enough.

Instead, we give this number the label, $\aleph_0$ (aleph null). $\aleph_0$ is the first infinite cardinal. $\aleph_0$ is how many natural numbers there are. It’s also how many even naturals there are.

This might be surprising, because the natural numbers are a superset of the even naturals, so it might seem like there should be twice as many, but, doubling every natural number obtains the set of even naturals. That is, a bijection – the doubling function – exists between the two sets, so the two sets have the same cardinality: $\aleph_0$.

Similarly, we have things like $\aleph_0 + \aleph_0 = \aleph_0$, as demonstrated by adding the even naturals to the odd naturals to obtain the natural numbers. We also have $\aleph_0 = \aleph_0 - 1$, as demonstrated by the function $x \mapsto x + 1$ forming a bijection between $\mathbb{N}$ and $\mathbb{N} \setminus \{0\}$.

$\aleph_0$ is also the number of rational numbers. Again, this seems surprising, since the rationals seem so much more dense (see §34.3 or §37.4.3) on the number line, but as famously shown by Cantor, there’s a way to form a bijection between the rationals and the naturals:

$$\begin{array}{cccccccccc}
\frac{1}{1} & \rightarrow & \frac{1}{2} & & \frac{1}{3} & \rightarrow & \frac{1}{4} & & \frac{1}{5} & \rightarrow \\
 & \swarrow & & \nearrow & & \swarrow & & \nearrow & & \cdots \\
\frac{2}{1} & & \frac{2}{2} & & \frac{2}{3} & & \frac{2}{4} & & \frac{2}{5} & \cdots \\
\downarrow & \nearrow & & \swarrow & & \nearrow & & & & \cdots \\
\frac{3}{1} & & \frac{3}{2} & & \frac{3}{3} & & \frac{3}{4} & & \frac{3}{5} & \cdots \\
 & \swarrow & & \nearrow & & & & & & \cdots \\
\frac{4}{1} & & \frac{4}{2} & & \frac{4}{3} & & \frac{4}{4} & & \frac{4}{5} & \cdots \\
\downarrow & \nearrow & & & & & & & & \cdots \\
\frac{5}{1} & & \frac{5}{2} & & \frac{5}{3} & & \frac{5}{4} & & \frac{5}{5} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}$$

Cantor’s zig-zag argument

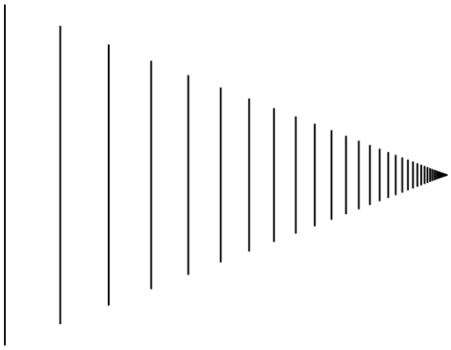
This bijects the positive rationals with the naturals, but the negative rationals can easily be interleaved with this sequence, giving the desired bijection. Cantor also showed that $\aleph_0 \cdot \aleph_0 = \aleph_0$ by constructing a bijection from $\mathbb{N}^2$ to $\mathbb{N}$ with his *pairing function*,

$$(x,y) \mapsto \frac{(x+y+1)(x+y)}{2} + y$$

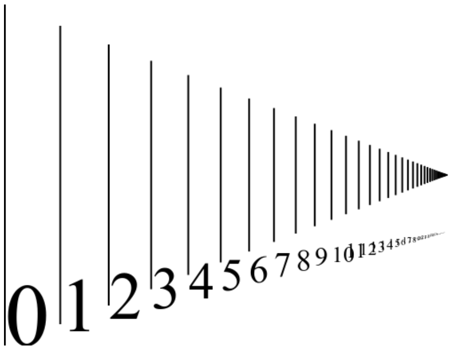
Through repeated applications of this pairing function, this also implies that $\mathbb{N}^{1000}$, for example, has the same cardinality as $\mathbb{N}$.

4.5.6.1 $\aleph_0$ & ω

If we draw a set of lines, each a fraction of the length of the previous, and a fraction of the distance from each last line, we can Zenoianly fit infinitely many lines in a finite space:

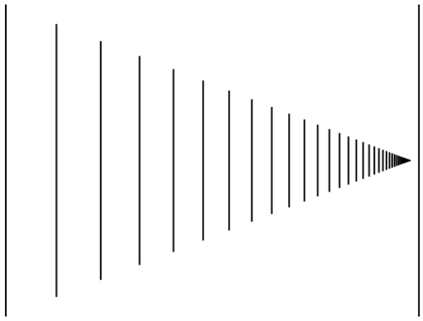


The number of lines here is equal to the number of natural numbers. We can demonstrate this by pairing each line up with a natural number.



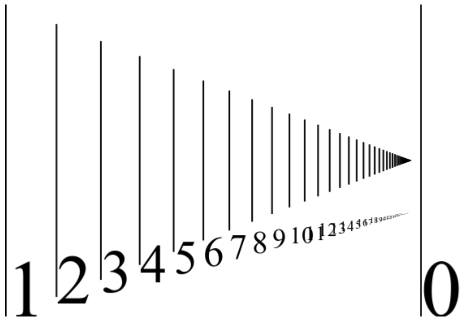
There is always a next line, but there is also always a next natural number, so this is a bijection. Both sets have cardinality $\aleph_0$.

But what happens if we do this?



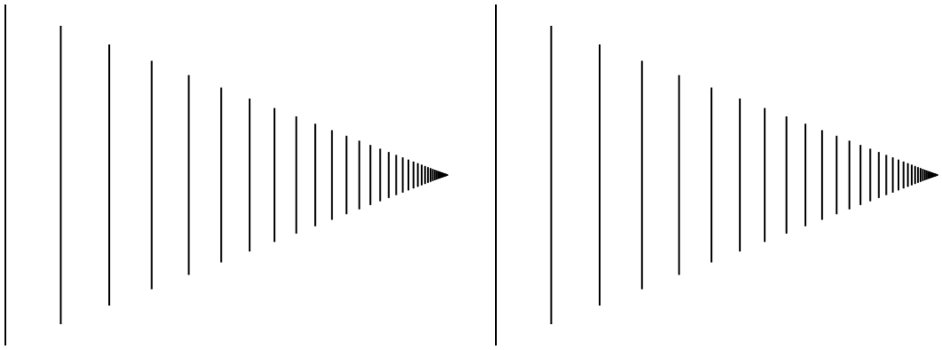
Do we have $\aleph_0 + 1$ lines?

No – there are still only $\aleph_0$ lines, because we can still show another bijection between this new set of lines, and the naturals. We do this by relabelling the lines, starting with the extra line, then continuing as before:

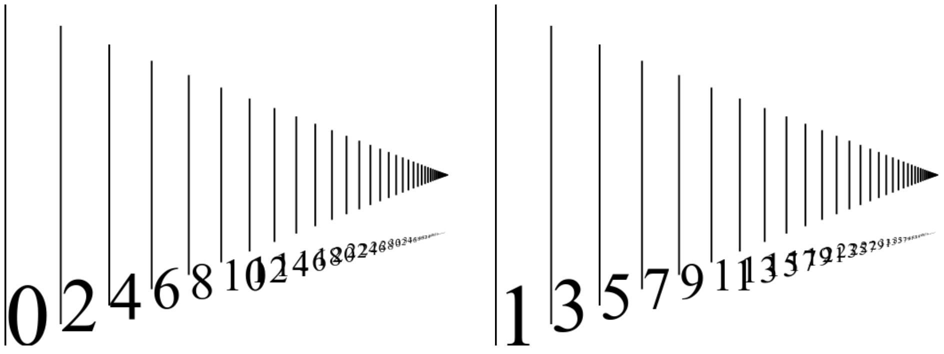


The extra line doesn't appear to affect the cardinality of this set of lines – the number of lines hasn't changed.

We could add two more lines, or three more, or four. The cardinality of this set will not change. We can even add another infinite $\aleph_0$ of more lines without changing the cardinality:



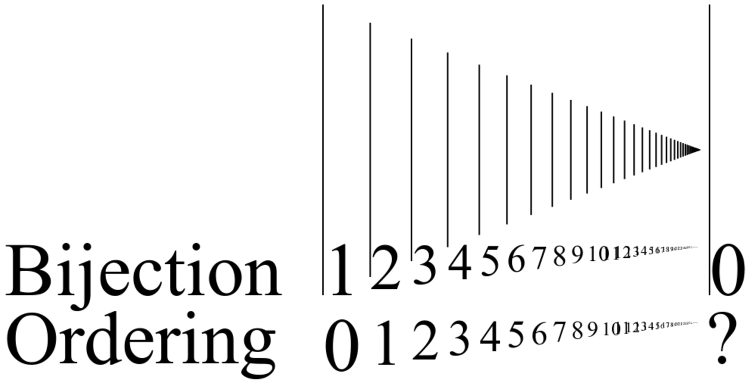
We can relabel the left half with the even naturals, and the right with the odd naturals, again showing a bijection.



However, while this does contain the same number of lines, something clearly does seem structurally different between this pair of infinities, and the single infinite group of lines we started with. If it's not the amount of stuff, what *is* the difference?

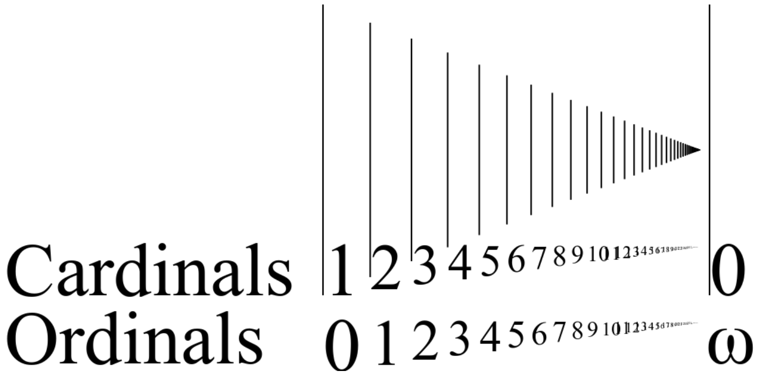
Let's revisit the situation with only one extra line after an $\aleph_0$ sized collection. Rather than forming a bijection in any way we want, what happens if we insist on labelling each line in the order they are drawn? With this requirement, we have to label the first line 0, and continue, left to right.

What number does the new line get?



For infinite quantities, labelling things in order is quite different from counting them. This new line doesn't add anything to the total, but in order to label it in the order it was drawn, we need a set of labels, a set of numbers, that extends past the naturals. We need the *ordinal numbers*.

The first transfinite ordinal is ω. If you came ωth place in a race, that would mean an infinite number of people finished the race, then you did.



Next, comes $\omega + 1$,[*] which perhaps doesn't look as much like a single number, but it is; like how $1 + i$ is a single complex number. However, while a complex number can be separated into real and imaginary parts, $\omega + 1$ is one indivisible atomic number.

Ordinal numbers label things in order – they aren't about how many things there are, they are about how those things are arranged – they tell us about the *order-type* of those things.

Two ordered sets (that is, sets equipped with a binary relation of partial order or stronger), X and Y, have the same order-type if there exists a bijection between them that preserves order.[†] So for any $a,b \in X$ with $a < b$, a bijection f must give $f(a) < f(b)$. Order-type also satisfies the definition of an equivalence relation, so sets that share order-type are in the same equivalence class.

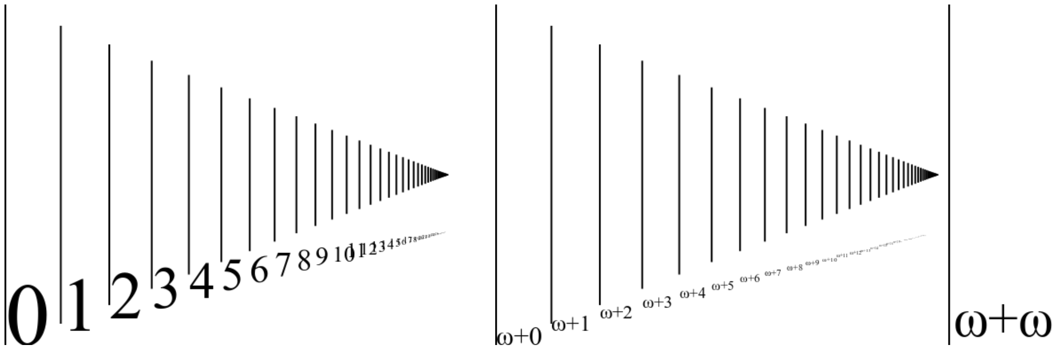
The order-type of a set is the first ordinal number not needed to label everything in the set. So the set $\{-12,\pi\}$ requires the labels 0 and 1 to label everything in the set, so the order-type of this set is 2.

For finite quantities, order-type and cardinality coincide. The order-type of all the naturals, $\aleph_0$, is ω.

Every time we add a new line, we don't increase the number of things, so the cardinality remains at $\aleph_0$, but the order-type of the sequence of lines changes to $\omega + 1$, or $\omega + 2$, and so on.

No matter how long an arrangement becomes, as long as it's well-ordered – every subset has a minimal element (§4.4.11) – the arrangement describes a valid ordinal number.

We can even go out to $\omega + \omega$, or $\omega \cdot 2$:



This corresponds to the well-ordering of the natural numbers given by defining all even numbers to be less than all odd numbers, and having usual ordering apply within the separated even and the odds:

$$0 \prec 2 \prec 4 \prec 6 \prec 8 \prec \ldots \prec 1 \prec 3 \prec 5 \prec 7 \prec 9 \prec \ldots$$

This well-ordered set has order-type $\omega + \omega$ – same as the set of line above. It is worth emphasising again that larger order-types do not correspond to more objects – These ordinals all refer to the same number of things, just arranged differently. The previous two examples still contain the same number of objects as the naturals: $\aleph_0$.

$\omega + 1$ isn't *bigger* than ω, it just comes *after*.

Also note that the axiom of infinity only guarantees the existence of one infinite set, and going out to $\omega + \omega$ would be creating a second. However, we can resolve this with the axiom schema of replacement (§4.3.6), by letting φ map $x \mapsto \omega + x$, so we know that $\omega + \omega$ is a valid set.

However, $\aleph_0$ is not the end. For instance, what is the cardinality of $\mathcal{P}(\mathbb{N})$? We know this set should have $2^{\aleph_0}$ subsets, but how many is that exactly?

[*] Note that addition is not commutative with transfinite ordinals. A line *after* ω gives $\omega + 1$, but a line *before* ω, is just ω, so $1 + \omega = \omega$.

[†] The function must be *monotonically increasing* (§34.2.1)

Imagine a list of all natural numbers.

	0	1	2	3	4	5	6	7	8	9	⋯

We now generate subsets of the naturals. We can represent sets with an indicator function, putting a 1 in the column of a natural number if the set contains that natural number, and a 0 otherwise. For example, the set of even naturals would be,

	0	1	2	3	4	5	6	7	8	9	⋯
	1	0	1	0	1	0	1	0	1	0	⋯

We continue, adding infinitely many subsets, labelling each one with a natural number as we go.

	0	1	2	3	4	5	6	7	8	9	⋯
0	1	0	1	0	1	0	1	0	1	0	⋯
1	0	0	1	1	1	1	0	1	0	1	⋯
2	0	0	0	1	0	1	0	1	0	0	⋯
3	1	1	1	1	0	1	1	0	1	1	⋯
4	0	1	1	0	0	0	0	1	0	1	⋯
5	1	0	1	0	1	1	0	0	1	0	⋯
						⋮					

If we can show that there is a set not on this list, that means we've exhausted all the natural numbers, yet still have elements left unlabelled, so the cardinality of this set must be larger than the set of naturals – a bigger infinity than $\aleph_0$.

We do this by checking the first element of the first subset, and doing the opposite to our new set. 0 is a member of this set, so it is not a member of our new set. Then, we move diagonally down to check if 1 is a member of the second subset. In this case, it isn't, so 1 is a member of our new set.

	0	1	2	3	4	5	6	7	8	9	⋯
0	1	0	1	0	1	0	1	0	1	0	⋯
1	0	0	1	1	1	1	0	1	0	1	⋯
2	0	0	0	1	0	1	0	1	0	0	⋯
3	1	1	1	1	0	1	1	0	1	1	⋯
4	0	1	1	0	0	0	0	1	0	1	⋯
5	1	0	1	0	1	1	0	0	1	0	⋯
						⋮					
?	0	1	1	0	1	0	⋯				

Continuing like this, our new set is guaranteed to be distinct from every other subset in at least one place by how we've defined it. Even if we somehow put this set back into the list, this process – *Cantor's diagonalisation argument* – can still be done.

More generally, we can prove that the power set of any set will always resist a one-to-one correspondence with the original set:

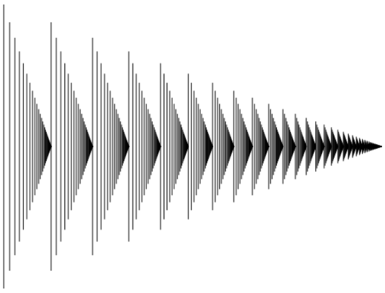
Let S be any set, and let $f : S \to \mathcal{P}(S)$ be a surjection. Let $A = \{x \in S : x \notin f(x)\}$, and suppose $A = f(y)$. Then, $y \in A \leftrightarrow y \notin A$, so f cannot exist. Since a surjective f cannot exist, it follows that no bijection can exist, so $|S| \neq |\mathcal{P}(S)|$.*

It turns out that the power set of the naturals, $2^{\aleph_0}$ has the same cardinality as the set of real numbers – we can construct a bijection between the two. This set of real numbers is sometimes called the *continuum*, and its cardinality is given the symbol $\mathfrak{C}$. Repeated applications of power sets will produce sets that

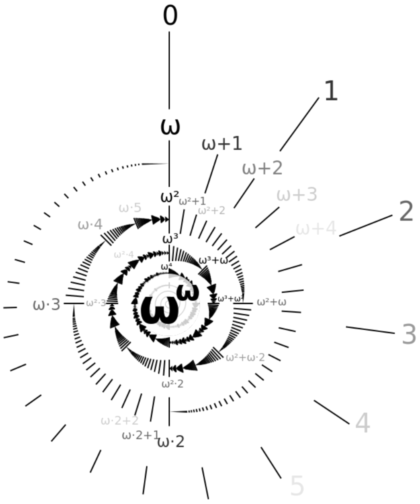
* Corollary: There are functions $f : \mathbb{N} \to \{0,1\}$ that cannot be computed by any computer program. Computer programs are finite sequences of a finite alphabet of possible instructions, while the set of functions f has size $2^{|\mathbb{N}|} = \mathcal{P}(\mathbb{N})$ which is uncountable, so no bijection can exist between the sets.

can't be put into one-to-one correspondence with the last, so we know there's an infinite chain of ever larger infinite cardinals.

Now, using the axiom schema of replacement, we can construct ever larger and larger sets. We can map $x \mapsto \omega \cdot x$, going up to $\omega \cdot \omega$, or ω^2, which looks something like this:



Now we map $x \mapsto \omega^x$ to reach ω^ω.



Then, use $x \mapsto \underbrace{\omega^{\omega^{\cdot^{\cdot^{\cdot^{\omega}}}}}}_{x}$ to reach $\underbrace{\omega^{\omega^{\omega^{\cdot^{\cdot^{\cdot}}}}}}_{\omega}$, and we run out of standard notation to use. But we just label this number ε_0, and continue from there. Replacement will still work, regardless of whether we have enough notation to write down these new infinities.

However, ε_0, and every other ordinal we can reach with replacement, while monstrously large, is still just an arrangement of $\aleph_0$ things. They're all still countable ordinals. Furthermore, these arrangements are all well-ordered, so they have an order-type – some ordinal that comes after them.

In this case, this ordinal is ω_1: the first uncountable ordinal. Now, because ω_1 comes after every order-type of $\aleph_0$ things, it must describe an arrangement of more things than the last aleph number. If it didn't, it would lie within the epsilon numbers, which directly contradicts its own definition as the order-type of those epsilon numbers. The cardinal number describing the number of things that ω_1 arranges is called $\aleph_1$.

It is not known how large the power set of the naturals is in relation to this cardinal. It can't lie between $\aleph_0$ and $\aleph_1$, because there aren't cardinals between them. It could be the case that $2^{\aleph_0} = \aleph_1$, which is the *continuum hypothesis*, but it could also be larger.

The continuum hypothesis has been proven to be consistent with ZFC. However, the negation of the continuum hypothesis has also been proven to be consistent with ZFC, so it is independent of, and cannot be proven from within ZFC.

4.5.6.2 ε_0 & Inaccessible Cardinals

Notice that, now we have reached $\aleph_1$, we have another set of subscripts to work with. With replacement, we can now reach $\aleph_\omega$, then $\aleph_{\omega^2}$, up to $\aleph_{\omega_{\omega_{\ddots}}}$, and we again run out of notation. The axiom schema of replacement allows us to make larger and larger leaps as we go along, each replacement granting access to even larger replacements. We can continue, reaching bigger and bigger infinities from below.

So, surely there can't be anything larger.

But, that's what we said about getting from the finite up to ω. We could axiomatically assert the existence of a cardinal so large, no amount of replacement or power setting on anything smaller could ever reach it from below. Such a number is called an *inaccessible cardinal*, because it can't be accessed from below. Within the numbers we have already reached, we can see traces of such a cardinal: $\aleph_0$.

All numbers less than $\aleph_0$ are finite, and no finite amount of replacement or power setting on finite sets can give anything but another finite amount. The only way we accessed $\aleph_0$ in the first place was by declaring that it existed with the axiom of infinity. And even further, set theorists have described many, many sets, far larger than inaccessible cardinals, each one requiring a new large cardinal axiom.

While there are infinitely many cardinals and infinitely many interesting things to cover, we unfortunately have very finite time, so this is where we'll draw the line.

Chapter 5

Induction

> *"Begin with the simplest examples."*
>
> — David Hilbert, *Methoden der mathematischen Physik*

Induction is a technique used to prove universal statements about some class of objects built from smaller objects. We do this by showing that if every object has a property given that smaller objects have that property, then every object in the class has that property. *Recursion* is a related idea, but for definitions – building up a class of objects by defining objects in terms of previous ones.

5.1 Simple Induction

The first form of induction you are likely familiar with is *simple* or *weak induction.* This is the form commonly used to prove that a statement holds for all natural numbers.

It is related to the recursive definition of the natural numbers. We have seen many definitions of the naturals, but all of them follow the same basic pattern:

- 0 is a natural number.
- If x is a natural number, then $x + 1$ is also a natural number.

For instance, the von Neumann ordinal construction operation isn't $x + 1$, but $x \cup \{x\}$, and our base object is $\varnothing$. For the Peano construction, the construction operation is $S(x)$, with base object 0.

So, all of these are *recursive definitions.* Later natural numbers $(x + 1)$ are defined in terms of ones we already have (x), using some given operation $(+1)$, and we start with a base object to begin building off of (0).

Because this recursive definition is built into the naturals, we can leverage this to our advantage. If we want to prove that a predicate, P, holds for a natural number, we can do this by showing that a natural number can't be constructed without having P be true. However, in doing so, we have also inadvertently shown that P holds for all naturals.

In simple induction, we do this by proving that $P(0)$ holds, and that $P(x)$ implies $P(x + 1)$. Then, P would hold for all natural numbers, built directly into each one by construction.

We express this idea formally as the *axiom of induction*,

$$\forall P : (P(0) \land \forall x : P(x) \to P(Sx)) \to \forall x : P(x)$$

In the above statement of the axiom, the predicate is being quantified over, so this axiom is stated in second-order logic, which is not ideal, given the first-order nature of ZFC. Because of this, we would normally use an *axiom schema of induction* instead, where we use declare a separate axiom for each possible predicate, but we can treat predicates as sets in ZFC, which *can* be quantified over:

$$\forall P(0 \in P \land \forall n \in \mathbb{N} : n \in P \to (n+1) \in P) \to \mathbb{N} \subseteq P$$

The above statement, and by extension, induction itself, is actually a theorem in ZFC, and not an axiom.

Regardless of our choice of axiomatisation, they all say the same thing: if a proposition holds for 0, and $P(x)$ implies that $P(x+1)$ also holds, then P holds for all natural x.

$P(0)$ is called the *base case*, where the rest of the proof is built up from. To prove that $P(x+1)$ follows from $P(x)$, we start by assuming that $P(x)$ holds, called the *induction hypothesis* (sometimes abbreviated IH), then prove that $P(x+1)$ follows, the *inductive step.*

As a classic example, let us prove Gauss' formula for the sum of consecutive naturals.

Theorem (Gauss Summation)**.** *For all n, $\sum_{i=0}^{n} i = 0 + 1 + 2 + \cdots + n = \frac{n(n+1)}{2}$.*

Proof. Let $P(n)$ be the statement that $\sum_{i=0}^{n} i = \frac{n(n+1)}{2}$. We induct on n.

$P(0)$ gives an empty sum on the left side, equalling zero, while the right side evaluates to $\frac{0(0+1)}{2} = 0$, so $P(0)$ holds, thus demonstrating the base case.

Then, for the induction hypothesis, we assume that $P(n)$ holds for some fixed $n \geq 0$.

Then,

$$\begin{aligned}
\sum_{i=0}^{n} i &= \frac{n(n+1)}{2} \\
\sum_{i=0}^{n} i + (n+1) &= \frac{n(n+1)}{2} + (n+1) \\
\sum_{i=0}^{n+1} i &= \frac{(n+1)n}{2} + \frac{(n+1)2}{2} \\
\sum_{i=0}^{n+1} i &= \frac{(n+1)(n+2)}{2} \\
\sum_{i=0}^{n+1} i &= \frac{(n+1)((n+1)+1)}{2}
\end{aligned}$$

which is exactly the statement $P(n+1)$, thus completing the inductive step.

As the base case has been shown to be true, and the inductive step has been established, the statement $P(n)$ holds for all natural numbers n by the principle of mathematical induction. ■

We were rather verbose in this proof, naming the base case, induction hypothesis and inductive step. This isn't strictly necessary, but explicitly using this terminology signals to the reader that you are attempting a proof by induction (as well as immediately saying "We induct on n.").

We haven't defined exponentiation over the naturals yet, but we could do it as,

$$x^0 = 1$$

$$x^{n+1} = x \cdot x^n$$

where $n \in \mathbb{N}$.

This is a recursive definition: for example, to compute 2^4, we expand the exponent out until the base case is reached, as $2^3 = 2 \cdot 2^2 = 2 \cdot 2 \cdot 2^1 = 2 \cdot 2 \cdot 2 \cdot 2^0$. This recursive definition makes natural exponentiation well suited to inductive proofs.

Theorem. *If $a > 1$, then $a^n > 1$ for all $n > 0$.*

Proof. Let $a > 1$, and let $P(n)$ be the statement that $a^n > 1$ when $n > 0$. We induct on n.

$P(0)$ holds vacuously as $n = 0$ gives $n \not> 0$ evaluating as true. Suppose $P(n)$ holds for some arbitrary fixed value of $n \geq 0$.

If $n = 0$, then $a^1 = a^0 \cdot a = 1 \cdot a = a > 1$. Otherwise, if $n > 0$, $a^{n+1} = a \cdot a^n > a \cdot 1 > 1$.

In either case, $P(n)$ implies $P(n+1)$. As the base case has been shown to be true, and the inductive step has been established, the statement $P(n)$ holds for all natural numbers n by the principle of mathematical induction. ■

5.1.1 Alternative Base Cases

As we saw in the last example, having a base case of $n = 0$, can sometimes require some annoying case analysis, or other similar workarounds if 0 isn't the first case where the predicate holds. In practice, we use a more general induction axiom that works for any integer base case,

$$\forall P(z_0 \in P \land \forall z \in \mathbb{Z} : z \in P \to (z + 1) \in P) \to \{z \in \mathbb{Z} : x \geq z_0\} \subseteq P$$

So proving $P(z_0)$ and $P(z) \to P(z + 1)$, gives $P(z)$ for all integers $z \geq z_0$. This is somewhat more powerful than our first axiomatisation of induction, because this form works for negative integers as well.

Intuitively, this axiom holds for the same reason the previous theorem holds – if $P(z_0)$ holds, and any larger integer can be reached by applying the +1 operation enough times, and each +1 operation preserves the truth value of P, then P holds for all integers greater than or equal to z_0.

To prove this formally, we do a simple change of variable and let $Q(n) = P(z - z_0)$, and use induction on Q – for the purposes of induction, a set of bounded below negative integers shifted along behaves identically to the naturals.

As an example of starting at a non-zero base case:

Theorem 5.1.1. *Let $n \in \mathbb{N}$. If $n \geq 4$, then $2^n \geq n^2$.*

Proof. Let $P(n)$ be the statement that $2^n \geq n^2$. $P(4)$ holds with equality, as $2^4 = 16 = 4^2$.[*] Suppose $P(n)$ holds for some arbitrary fixed $n \geq 4$.

Then,

$$2^{n+1} = 2 \cdot 2^n$$

[*] Extension tasks:

- What are all the integer solutions of $x^y = y^x$?
- What about real solutions?
- Complex solutions?
- What is $\lim_{t\to\infty} t^{\frac{1}{t-1}}$?
- What is the connection between the above limit and the equation in the first part?

$$\begin{aligned}
&\geq 2n^2 \\
&= n^2 + n^2 \\
&\geq n^2 + 4n \\
&= n^2 + 2n + 2n \\
&\geq n^2 + 2n + 1 \\
&= (n+1)^2
\end{aligned}$$

so $P(n+1)$ also holds, completing the inductive step.

As the base case has been shown to be true, and the inductive step has been established, the statement $P(n)$ holds for all natural numbers $n \geq 4$ by the principle of mathematical induction. ■

5.1.2 Validity of Recursive Definitions

Previously, we defined natural exponentiation recursively by defining a base case, $x^0 = 1$, and giving a rule to compute x^{n+1} given x^n. Using simple induction, we can show that these definitions work. That is, these definitions uniquely and exactly define the objects they are supposed to.

Lemma 5.1.2. *Let S be a set, let $g : S \to S$ be an function, and let $f : \mathbb{N} \to S$ satisfy,*

$$\begin{aligned}
f(0) &= x_0 \\
f(n+1) &= g(f(n))
\end{aligned}$$

Then, f is unique.

Proof. Suppose there exists some function $f' : \mathbb{N} \to S$ that also satisfies $f'(0) = x_0$ and $f'(n+1) = g(f'(n))$. We will show that $f'(n) = f(n)$ for all n by induction on n.

The base case holds as $f'(0) = x_0 = f(0)$. Now suppose $f'(n) = f(n)$ for some arbitrary fixed $n \geq 0$. Then, $f'(n+1) = g(f'(n)) = g(f(n)) = f(n+1)$, so f is unique. ■

5.1.3 Alternative Operations

We can also perform induction with a different operation than +1, say, +2. This will, however, only prove a statement for a subset of the naturals or integers.

Theorem 5.1.3. *$a \in \mathbb{Z}$ is odd if and only if a^2 is odd.*

Proof. Let $P(n)$ represent the statement that n is odd if and only if n^2 is odd. We induct on n.

$P(1)$ holds as 1 is odd and $1^2 = 1$ is also odd. Suppose $P(n)$ holds for some arbitrary fixed odd integer $n \geq 1$.

Then, $n+2$ is the sum of an odd and even integer, which is odd, and $(n+2)^2 = n^2+4n+4 = n^2+4(n+1)$, which, by the inductive hypothesis, is the sum of an odd and even number, which is odd. So, both $n+2$ and $(n+2)^2$ are odd, which is exactly the statement $P(n+2)$ ■

This proof is valid enough, but doesn't actually follow directly from the axiom of induction. After all, the construction of the naturals uses +1, so our induction using +2 will need a bit of work to be completely formally valid. However, these fixes are easy enough corollaries of regular induction: we just redefine a new proposition in terms of the old, encoding the variable, similar to translating the integers across. For this example, our encoding of n could be $m := \frac{n-1}{2}$, then we prove a similar statement on $Q(m)$ with regular induction.

This is, however, a fairly obvious and pedantic point, so this generally isn't included with an induction proof, since the related encoded induction should be obvious to the reader.

5.1.4 Multiple Counters & Base Cases

We can prove many properties of operations on natural numbers using induction. Here, we will prove various properties about addition in particular. The analogous properties for multiplication is left as an exercise for the reader.

For this section, we will use a Peano axiomatisation ofthe natural numbers. That is,

A1 $n + 0 = n$;

A2 $n + S(m) = S(n + m)$.

And we define the symbol 1 to represent the successor of 0, so $S(0) = 1$.

In particular, we have

Lemma 5.1.4. $S(a) = a + 1$

Proof.

$$\begin{aligned} S(a) &= S(a+0) && \text{A1} \\ &= a + S(0) && \text{A2} \\ &= a + 1 && \text{Definition of 1} \end{aligned}$$

■

so this definition of the symbol 1 matches with our intuition of succession over the naturals.

Theorem (Existence of Identity Element of Addition over the Naturals)**.** *0 is the identity element of addition over the naturals.*

Proof. From A1, we know 0 is a right identity for any natural n. We prove that 0 is also a left identity by induction on n.

Let $P(n)$ be the statement that $0 + n = n$. $P(0)$ holds from A1 as 0 is a right identity. Suppose $P(n)$ holds for some arbitrary fixed value of $n \geq 0$.

Then,

$$\begin{aligned} 0 + S(a) &= S(0 + a) && \text{A2} \\ &= S(a) && \text{IH} \end{aligned}$$

thus completing the inductive step, so P holds for all n by induction. ■

Theorem (Associativity of Addition over the Naturals)**.** *For all a,b,$c \in \mathbb{N}$, $(a + b) + c = a + (b + c)$*

Proof. Fix arbitrary a,$b \in \mathbb{N}$, and let $P(c)$ represent the statement that $(a + b) + c = a + (b + c)$. $P(0)$ holds from two applications of A1 as $(a + b) + 0 = a + b = a + (b + 0)$. Suppose $P(c)$ holds for some arbitrary fixed $c \geq 0$.

Then,

$$\begin{aligned} (a+b) + S(c) &= S((a+b)+c) && \text{A2} \\ &= S(a + (b+c)) && \text{IH} \\ &= a + S(b+c) && \text{A2} \\ &= a + (b + S(c)) && \text{A2} \end{aligned}$$

so $P(S(c))$ holds. It follows that P holds for all c by induction. ■

Now, this proof is somewhat different from the simple inductions we have seen before. We induct on two different variables to complete this proof, as well as using several base cases.

Theorem (Commutativity of Addition over the Naturals)**.** *For all* a,$b \in \mathbb{N}$, $a + b = b + a$.

Proof. Let $P(b)$ be the statement that $a + b = b + a$ for a fixed a. To prove $P(b)$, we will induct on b, but some preamble is required.

$P(0)$ holds immediately from the identity element property of 0 as proved above.

Now, we will show a second base case, $P(1)$, by induction.

Let $Q(a)$ be the statement that $a + 1 = 1 + a$. We induct on a. We have already proved $P(0)$, so 0 commutes with everything – in particular, with 1. So, $Q(0)$ gives $0 + 1 = 1 + 0$ which holds by $P(0)$, completing the base case for Q. Then,

$$\begin{aligned} S(a) + 1 &= S(a) + S(0) && \text{Definition of 1} \\ &= S(S(a) + 0) && \text{A2} \\ &= S((a + 1) + 0) && \text{Lemma 5.1.4} \\ &= S(a + 1) && \text{A1} \\ &= S(1 + a) && \text{IH} \\ &= 1 + S(a) && \text{A2} \end{aligned}$$

completing the induction on a, proving the base case $b = 1$. Now, suppose for all natural numbers a, P holds. We now induct on b.

$$\begin{aligned} a + S(b) &= a + (b + 1) && \text{Lemma 5.1.4} \\ &= (a + b) + 1 && \text{Associativity} \\ &= (b + a) + 1 && \text{IH} \\ &= b + (a + 1) && \text{Associativity} \\ &= b + (1 + a) && \text{Base case } b = 1 \\ &= (b + 1) + a && \text{Associativity} \\ &= S(b) + 1 && \text{Lemma 5.1.4} \end{aligned}$$

completing the induction on b. ■

For this proof, we had to prove a base case by induction, within the induction.

5.2 Strong Induction

Another form of induction is *complete* or *strong induction*, where we assume that the induction hypothesis holds for all naturals less than $n + 1$, and not just n itself. The name comes from the stronger *induction hypothesis*, because we are assuming more things. Strong induction is not, however, a stronger technique than simple or weak induction. They are in fact equivalent to both each other and to the axiom of induction.

Formally, instead of proving $\forall n : P(n) \to P(n + 1)$, we prove $\forall n : (\forall k \le n : Q(k)) \to Q(n + 1)$. But, if we let $P(n) \equiv \forall k \le n : Q(k)$, we see that this is exactly the same thing as simple induction, since, if $\forall k \le n : Q(k) \to Q(n + 1)$, it also implies $\forall k \le n + 1 : Q(k)$, giving us the original induction formula $\forall n : P(n) \to P(n + 1)$.

Any statement you can prove with weak induction, you can prove with strong induction, and vice versa. The reason we have a distinction between the two, however, is that often, one method will be significantly easier than the other for a human to read and write, though they will be logically equivalent.

As an example of strong induction, suppose we have an infinite supply of alien 4 and 5 pence pieces. We can prove by strong induction that we can reach any monetary value above 12 pence* using just these two denominations.

Reframing this question as a statement about linear combinations, we have,

Theorem 5.2.1. *For all $n \geq 12$, there exist $a,b \in \mathbb{N}$ such that $n = 4a + 5b$.*

Proof. Let $P(n)$ be the statement that there exist $a,b \in \mathbb{N}$ such that $n = 4a + 5b$. We provide 4 base cases:

- $4 \cdot 3 + 5 \cdot 0 = 12$
- $4 \cdot 2 + 5 \cdot 1 = 13$
- $4 \cdot 1 + 5 \cdot 2 = 14$
- $4 \cdot 0 + 5 \cdot 3 = 15$

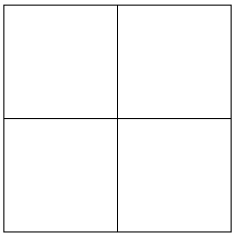
So $P(n)$ holds for $n = 12,13,14,15$. Now fix an arbitrary natural $n > 15$ and suppose that $P(k)$ holds for all $12 \leq k \leq n$.

Let $k = n - 4$. Because $n > 15$, it follows that $12 \leq n - 4 \leq n$, so $P(n - 4)$ holds by the inductive hypothesis. So, the sum $n - 4$ can be formed by some linear combination of the 4 and 5 pence pieces. Then, adding an additional 5 pence piece gives us $n + 1$, so, if $P(k)$ holds for all $12 \leq k \leq n$, it also holds for $P(n + 1)$, completing the inductive step. ■

Exercise. Prove the same statement with weak induction.

As shown above, strong induction is particularly helpful for statements which can increment by different amounts, or otherwise have multiple base cases. As another example, how many ways are there to cut a square up into smaller squares?

The most obvious non-trivial first thing to try is to cut a square into quarters, forming 4 smaller squares.

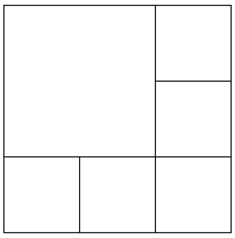


So, 4 is reachable. In fact, every time we have a valid division of the square, we can always divide one of the smaller constituent squares into 4, adding 3 to the total, so we've really reached all of $1 + 3n$.

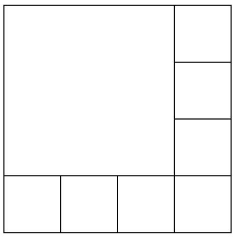
[*] The largest number we *can't* make, is then 11. 11 is then called the *Frobenius number* of 4 and 5. More generally, the Frobenius number of a set of numbers is the largest number that cannot be created as a linear combination of those numbers.

Rather famously, the Frobenius number for chicken nuggets at a certain fast food chain is 43, because they can only be bought in packs of 6, 9 and 20.

For two numbers, x and y, the Frobenius number is given by $xy - x - y$. For three or more numbers, no explicit formula is known. We do, however, have algorithms that can compute this number for any fixed numbers of denominations of coins, but if the number of denominations can be arbitrarily large, the problem is NP-hard.



So we've reached 6, and by extension, $6 + 3n$.



and now, we have 8, and $8 + 3n$

Clearly, the three cases now partition all of the naturals greater than 6. In fact, any number of squares greater than 6 can be created. The proof of this is extremely similar to the last strong induction proof and is left to the reader. As a non-induction extension task, prove that you cannot divide a square into 2, 3 or 5 smaller squares.

The contrapositive of strong induction is actually the method of *infinite descent*, an example of which is used in the proof of the irrationality of $\sqrt{2}$ in the section for real analysis (Lemma 34.3.1).

5.3 Backward-Forward Induction

In a previous section, §34.1.2, we proved the AM-GM inequality through a series of replacement of elements. Another way is through *backward-forward induction.* This technique is very rarely used, but allows us to use induction "backwards", by proving a statement holds for $n - 1$ from it holding for n. However, this only gives a finite number of cases where this works, so we need another "forward" induction to prove the statement for infinitely many natural numbers. We don't have to prove it for all naturals in the forward induction, because any gaps will be filled in by the backward induction.

Proposition (AM-GM inequality)**.** For any set, X, of cardinality $n \in \mathbb{N}$ containing the non-negative real numbers $x_1, x_2 ... x_{n-1}, x_n$, the inequality

$$\frac{1}{n}\sum_{i=1}^{n} x_i \geq \sqrt[n]{\prod_{i=1}^{n} x_i}$$

holds.

Proof. If the terms are all equal, $x_1 = x_2 = \cdots = x_n$, then; their sum is nx_1, so their arithmetic mean is x_1; and their product is x_1^n, so their geometric mean is x_1; so the proposition holds with equality in this case.

We need to prove that if the terms are not all equal, then the proposition still holds. Clearly, this can only be the case for $n > 1$. We divide this step into subcases.

If $n = 2$, then we have two terms, x_1 and x_2.

$$\frac{x_1 + x_2}{2} - \sqrt{x_1 x_2} = \frac{x_1 - 2\sqrt{x_1 x_2} + x_2}{2}$$

$$\begin{aligned} &= \frac{(\sqrt{x_1} + \sqrt{x_2})^2}{2} \\ &\geq 0 \end{aligned}$$

so $\frac{x_1+x_2}{2} \geq \sqrt{x_1 x_2}$ as desired.

Now, suppose $n = 2^k$, where $k \in \mathbb{Z}^+$. We induct on k. If $k = 1$, then $n = 2$, which holds from the case above. Now, suppose the statement holds for an arbitrary fixed value $k - 1 \geq 1$. Then, we prove the statement holds for k by subdividing the case with 2^k elements into two halves of size 2^{k-1} and applying the inductive hypothesis within each half, then applying the base case to the two halves themselves.

$$\begin{aligned} \frac{x_1 + x_2 + \cdots + x_{2^k}}{2^k} &= \frac{\frac{x_1+x_2+\cdots+x_{2^{k-1}}}{2^{k-1}} + \frac{x_{2^{k-1}+1}+x_{2^{k-1}+2}+\cdots+x_{2^k}}{2^{k-1}}}{2} \\ &\geq \frac{\sqrt[2^{k-1}]{x_1 x_2 \cdots x_{2^{k-1}}} + \sqrt[2^{k-1}]{x_{2^{k-1}+1} x_{2^{k-1}+2} \cdots x_{2^k}}}{2} \\ &\geq \sqrt{\sqrt[2^{k-1}]{x_1 x_2 \cdots x_{2^{k-1}}} \cdot \sqrt[2^{k-1}]{x_{2^{k-1}+1} x_{2^{k-1}+2} \cdots x_{2^k}}} \\ &= \sqrt[2^k]{x_1 \cdot x_2 \cdots x_{2^k}} \end{aligned}$$

Where, in the first inequality, the two sides are equal only if $x_1 = x_2 = \cdots = x_{2^{k-1}}$, and $x_{2^{k-1}+1} = x_{2^{k-1}+2} = \cdots = x_{2^k}$, in which case, both the arithmetic and geometric means of the first term would be equal to x_1, and similarly, both the arithmetic and geometric means of the first term would be equal to $x_{2^{k-1}+1}$. In the second inequality, the two sides are equal only if the two geometric means are equal. Since by assumption, the terms are not all equal, it is not possible for both inequalities to be equalities, so we know,

$$\frac{x_1 + x_2 + \cdots + x_{2^k}}{2^k} \geq \sqrt[2^k]{x_1 x_2 \cdots x_{2^k}}$$

as required.

Now, if n is not a natural power of 2, then it is clearly less than some natural power of 2, since the sequence of terms, $a_k = 2^k$ is not bounded above. So, without loss of generality, let k be some natural power of 2 such that $2^k > n$. Also let us label the arithmetic mean of our n terms as α, and let $x_{n+1} = x_{n+2} = \cdots = x_k = \alpha$, so our extra terms do not contribute towards changing the average.

$$\begin{aligned} \alpha &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{\frac{k}{n}(x_1 + x_2 + \cdots + x_n)}{k} \\ &= \frac{\left(1 - 1 + \frac{k}{n}\right)(x_1 + x_2 + \cdots + x_n)}{k} \\ &= \frac{x_1 + x_2 + \cdots + x_n + \left(\frac{k}{n} - 1\right)(x_1 + x_2 + \cdots + x_n)}{k} \\ &= \frac{x_1 + x_2 + \cdots + x_n + \left(\frac{k-n}{n}\right)(x_1 + x_2 + \cdots + x_n)}{k} \\ &= \frac{x_1 + x_2 + \cdots + x_n + \left(\frac{x_1+x_2+\cdots+x_n}{n}\right)(k - n)}{k} \\ &= \frac{x_1 + x_2 + \cdots + x_n + \alpha(k - n)}{k} \\ &= \frac{x_1 + x_2 + \cdots + x_n + x_{n+1} + x_{n+2} + \cdots + x_k}{k} \\ &\geq \sqrt[k]{x_1 x_2 \cdots x_n x_{n+1} x_{n+2} \cdots x_k} \\ &= \sqrt[k]{x_1 x_2 \cdots x_n \alpha^{k-n}} \end{aligned}$$

So,

$$\begin{aligned}\alpha^k &\geq x_1x_2\cdots x_n\alpha^{k-n}\\ \frac{\alpha^k}{\alpha^{k-n}} &\geq x_1x_2\cdots x_n\\ \alpha^n &\geq x_1x_2\cdots x_n\\ \alpha &\geq \sqrt[n]{x_1x_2\cdots x_n}\end{aligned}$$

giving the required inequality. ■

5.3.1 Well-Ordering

In ZFC, the axiom of induction and the well-ordering principle (§6.11.4) are equivalent. With the Peano axioms, however, they are not, as we will see.

Theorem (Induction Well-Ordering Equivalence)**.** *The axiom of induction is equivalent to the well-ordering principle.*

Proof. Suppose S is a set of natural numbers with no least element. 1 is not in S, as 1 is the least element of all naturals, and would therefore be the least element of a subset of the naturals. Let $Q = S'$, so 1 is in Q. Let n be an arbitrary natural number, and suppose that $\{1,2,3,\ldots n-1\}$ is in Q. If n is in S, then it would be the least element of S as every natural less than n is in Q, which is disallowed, so n must also be in Q. By strong induction, every natural number must be in Q, so S must be the empty set. Therefore, any non-empty subset of the naturals must have a least element. This completes the forward direction.

Let $P(n)$ be a statement on natural n , such that $P(1)$ is true and $P(n)$ implies $P(n+1)$. Let $S = \{x : \neg P(x)\}$. Suppose that S is not the empty set, such that m is the least element of S. As $P(1)$ is true, 1 is not in S, so $m \neq 1$, so $m > 1$. It follows that $m-1$ is a natural number. As m is the least element of S, $m-1$ is not in S. It follows that $P(m-1)$ is true. But $P(n)$ implies $P(n+1)$, so $P(m-1)$ implies that $P(m)$ is true, which is a contradiction. It follows that S is the empty set, so $P(n)$ holds for all natural numbers. This completes the backward direction, establishing the equivalence of induction and well-ordering. ■

While this proof is perfectly valid in ZFC, the problem is, $m-1$ is not guaranteed to be a unique and well-defined natural number in the Peano axiomatisation, so the backward direction of the proof does not hold in Peano. This actually makes induction strictly stronger than well-ordering, in the context of the other Peano axioms.

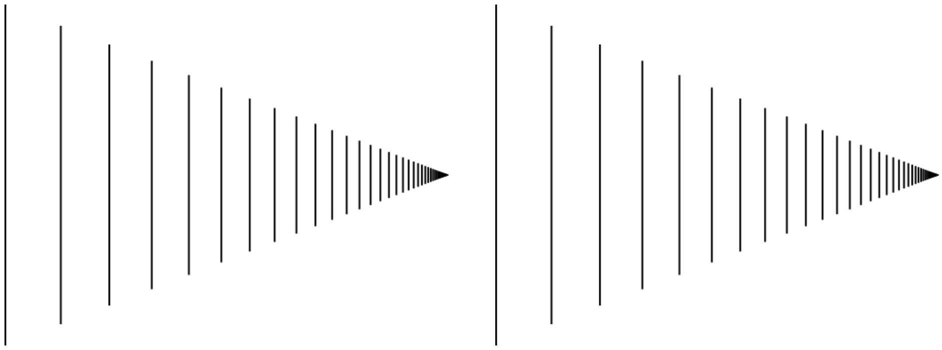
We can show this with a counter example of a well-ordered Peano set that doesn't obey the axiom of induction.

For our Peano axioms, we will suppose,

- Trichotomy (Lemma 11.3.2);
- For every natural number n, we have $n+1 > n$;
- For every natural number n, there does not exist a natural number between $n+1$ and n;
- No natural is less than 0.

The set, $\{(0,n) : n \in \mathbb{N}\} \cup \{(1,n) : n \in \mathbb{N}\}$ is well-ordered by the relation $(a,b) \prec (c,d)$ if $a < c$, or $b < d$ if $a = c$. Furthermore, this set also satisfies all the Peano axioms, where the Peano constant 0 is interpreted as $(0,0)$, and $S((x,n)) = S((x,n+1))$ for all $x \in \{0,1\}$ and $n \in \mathbb{N}$.

If this feels familiar, that's because it is! This setup is a more formal version of the odd-even ordering of the naturals from §4.5.6.1.



As an example of a predicate which does not follow the axiom of induction, let $P((x,n))$ represent the statement that $(x,n) = (0,0)$ or $(x,n) = S((y,m))$ for some $y \in \{0,1\}$ and $m \in \mathbb{N}$. The base case, $P((0,0))$ holds trivially, as does the induction step by the definition of the successor function, so $P((x,n)) \to P(S((x,n)))$, so P should hold for all pairs (x,n) with $x \in \{0,1\}$ and $n \in \mathbb{N}$. However, P fails to hold for (1,0). That is, in the well-ordered set of lines above, the first line in the right half is not the successor of any element in the set.

While the axiom of induction and the well-ordering principle are equivalent in many axiom systems, we need the additional axiom,

- Every natural number is either 0 or $n + 1$ for some natural number n.

for them to be equivalent with Peano axioms. Specifically, the axiom above combined with the first two axioms in the list above are equivalent to induction.

5.4 Transfinite Induction

Well-founded or *Noetherian induction* is the generalisation of regular mathematical induction to well-founded sets. If R is a well-founded relation on a set, X, and $P(x)$ is a predicate defined over elements x of X, then, to show that $P(x)$ holds for all $x \in X$, we only need to show that $P(y)$ holds for all y such that yRx.

That is,

$$\forall x \in X : [(\forall y \in X : yRx \to P(y)) \to P(x)] \to \forall x \in X : P(x)$$

Depending on set being used and the type of relation that R is, we have special cases of well-founded induction. For instance, if we have $(\mathbb{N},S)$, where S is the successor relation, we have simple induction. If we have $(\mathbb{N}, <)$, we have strong induction. In fact, we well-founded induction doesn't even have to operate on sets – it also works on proper classes, as long as the relation in question is set-like and well-founded.

There are other types of induction with various exotic sets and relations, such as *structural induction*, which allows us to prove statements about structures like trees and graphs, and is particularly important for computer science, or *ε-induction*, which allows us to prove statements hold for all sets.

But, one type of induction of particular note to us is *transfinite induction*, which is where R is a ordering on the class of ordinal numbers as we constructed in §4.5.6.1.

Let $P(\alpha)$ be a predicate defined over all ordinals α. Suppose that, if $P(\beta)$ holds for all $\beta < \alpha$ then $P(\alpha)$ also holds. Then, transfinite induction says that P holds for all ordinals.

Transfinite induction proofs often distinguish three cases:

- when α is a minimal (§4.4.13) element – that is, no element precedes α;

- when α has a direct predecessor – the set of elements which precedes α has a largest element.
- when α is a *limit ordinal* – there exists elements which precede α, but α itself has no direct predecessor.

ω is an example of a limit ordinal because any smaller ordinal will have a following ordinal generated by adding 1 that is still less than ω. ε_0 as we defined it is another limit ordinal, this time preceded by all ω numbers.

We can actually see all three cases in the example we constructed in the previous section. (1,0) is a limit ordinal because elements that precede it exist, but it does not have a direct predecessor; (0,0) is a minimal element as no elements that precede it exist; and every other element has direct predecessors.

One motivating example of transfinite induction is *Zorn's lemma.*

Theorem (Zorn's Lemma)**.** *Every partially ordered set that contains upper bounds for any totally ordered subsets must contain a maximal element.*

Zorn's lemma is required to prove a vast variety of results in various fields, from topology to metalogic. One notable theorem that Zorn's lemma proves is that every vector space, even infinite-dimensional ones, has a basis.

We provide a sketch of the proof of Zorn's lemma below.

Suppose Zorn's lemma is false, so there exists a partially ordered set, P with partial order $\prec$, such that every totally ordered subset has an upper bound, but there does not exist a maximal element of P. That is, for every $p \in P$, there exists another $q \in P$ such that $p \prec q$.

For each totally ordered subset, $S \subseteq P$, define $b(S)$ to be a function that returns an element of P that is bigger than every element in S. We know such a number always exists, because every totally ordered subset is assumed to have an upper bound. However, it aren't guaranteed that there is a smallest element of P that is bigger than every element in S, so there isn't an obvious way to pick such an element. So, we need to invoke the axiom of choice in order to define b.

Using the function, b, we generate an ordered list of elements, $a_0 \prec a_1 \prec a_2 \prec a_3 \prec \ldots \prec a_\omega \prec a_{\omega+1} \prec \ldots$. We pick the values of a_i with transfinite recursion. a_0 is selected arbitrarily (we know a_0 exists because the empty set has an upper bound in P, so P is non-empty), then for any ordinal, w, we choose $a_w = b(\{a_v : v < w\})$, recalling that an ordinal is the set of sets smaller than itself. Because the a_v are totally ordered, this definition is well-founded.

This sequence requires not only the natural numbers to index it, but all ordinals. In fact, this sequence is so long, it cannot possibly be a set – the class of all ordinals is larger than any set that we can construct in ZFC, so this sequence of elements of P is larger than P could possibly be, giving a contradiction.

5.5 Exercises

Unless otherwise specified, take n to be a positive integer.

1. Prove that if x and y are both positive, then $x < y$ implies $x^n < y^n$.
2. Prove that $\sum_{i=1}^{n} \frac{1}{i^2} \leq 2 - \frac{1}{n}$.
3. Prove that $6^n - 1$ is divisible by 5.
4. Prove that $n^2 - 1$ is divisible by 8 for all odd integers n.
5. Prove that $2^{n+1} > n^2$.
6. Prove that $n! > 3^n$ for $n \geq 7$.
7. Prove that $n^3 + n$ is even for all integers n with a pair of inductions.
8. Prove that

$$\sum_{i=1}^{n} (2i-1)^2 = \frac{4n^3 - n}{3}$$

9. Consider the sequence (a_n) recursively defined by $a_{n+1} = \sqrt{a_n + 2}$, with initial term $a_0 = 1$.
   (a) Show that $2 - a_{n+1} = \frac{2-a_n}{2+\sqrt{2+a_n}}$.
   (b) Using this identity, prove by induction that

$$2 - a_n \leq \frac{1}{(2+\sqrt{3})^n}$$

   (c) Prove that $1 \leq a_n \leq 2$.
   (d) Show that $(a_{n+1} - \sqrt{2}) = \frac{(a-\sqrt{2})^2}{2a_n}$.
   (e) Using this identity, prove by induction that

$$\left| a_n - \sqrt{2} \right| \leq \frac{1}{2^{2^n}}$$

   and conclude that (a_n) converges to $\sqrt{2}$.

10. Consider the sequence (a_n) recursively defined by $a_{n+1} = 2a_1a_n - a_{n-1}$ for $n \geq 1$ with initial terms $a_0 = 1$ and $a_1 = \cos\theta$ for some fixed constant θ. Prove that $a_n = \cos(n\theta)$ for all $n \geq 0$.
11. Show that if $n \geq 2$, and the following radical is nested n times:

$$\underbrace{\sqrt{1+\sqrt{1+\sqrt{1+\sqrt{1+\cdots}}}}}_{n \text{ times}}$$

then the expression is irrational.

12. Prove that for any non-negative integer n, we have

$$\int_0^\infty t^n e^{-t}\, dt = n!$$

13. Prove that

$$\sqrt{n} \leq \sum_{i=1}^{n} \frac{1}{\sqrt{k}} \leq 2\sqrt{n} - 1$$

14. Prove the following:

   (a) $\displaystyle\sum_{i=1}^{n} i = \frac{1}{2}n(n+1)$.

   (b) $\displaystyle\sum_{i=1}^{n} i^2 = \frac{1}{6}n(n+1)(2n+1)$.

   (c) $\displaystyle\sum_{i=1}^{n} i^3 = \frac{1}{4}n^2(n+1)^2$.

   (d) $\displaystyle\sum_{i=1}^{n} i^4 = \frac{1}{30}n(n+1)(2n+1)(3n^2+3n-1)$.

   (e) $\displaystyle\sum_{i=1}^{n} i^5 = \frac{1}{12}n^2(n+1)^2(2n^2+2n-1)$.

15. Generalising the previous question, prove that

$$\sum_{i=1}^{n} \prod_{j=0}^{m-1} (i+j) = \frac{1}{m+1} \prod_{j=0}^{m} (n+j)$$

   for all positive integers n and m.

16. Let $H_n = \sum_{i=1}^{n} \frac{1}{n}$ be the nth partial sum of the harmonic series. Show that $H_{2n} \geq H_n + \frac{1}{2}$ for all n, then show by induction that $H_{2^n} \geq 1 + \frac{n}{2}$ for all n and conclude that the harmonic series diverges.

17. Let F_n denote the nth Fibonacci number, defined by $F_n = F_{n-1} + F_{n-2}$, where $F_0 = 0$ and $F_1 = F_2 = 1$. Prove that:

   (a) $\displaystyle\sum_{i=0}^{n} F_i = F_{n+2} - 1$ for $n \geq 1$.

   (b) $\displaystyle\sum_{i=0}^{n} F_{2i} = F_{2n+1} - 1$ for $n \geq 1$.

   (c) $\displaystyle\sum_{i=0}^{n} F_{2i-1} = F_{2n}$ for $n \geq 1$.

   (d) $\displaystyle\sum_{i=0}^{n} F_i^2 = F_n F_{n+1}$ for $n \geq 1$.

   (e) $F_n = \frac{\phi^n - \psi^n}{\sqrt{5}}$, where $\phi = \frac{1+\sqrt{5}}{2}$ is the golden ratio, and $\psi = \frac{1-\sqrt{5}}{2} = -\frac{1}{\phi} = \phi - 1$.

   (f) $F_{m+n+1} = F_m F_n + F_{m+1} F_{n+1}$ for all m,$n \geq 0$.

18. Prove that multiplication is associative over the naturals from the Peano axioms.

19. Prove that multiplication is commutative over the naturals from the Peano axioms.

20. Find all analytic functions f such that there exists a complex number z_0 such that

$$f(z) = z_0 + zf(z^2)$$

by following these steps:

(a) Use analyticity of f to express $zf(x^2)$ as a power series with coefficients (a_n).

(b) Find an expression for $f(z) - zf(z^2)$ as the difference of two power series and compare coefficients to form recurrence relations for (a_n).

(c) Rearrange these recurrence relations and prove, using backward-forward induction, that $a_n = z_0$ if n is of the form $2^i - 1$ for some non-negative integer i, and that $a_n = 0$ otherwise.

(d) Hence conclude that f must be of the form $z_0 \sum_{j=0}^{\infty} z^{2^j - 1}$.

Chapter 6

Set Theory

> *"People think of axioms as laws you have to follow or true things you have to assume and I think neither of these perspectives is correct. It's more accurate to think of axioms as a way to agree that we're talking about the same thing."*
>
> — Qiaochu Yuan

In this chapter, we continue to use the symbol $\subseteq$ for the subset relation, and $\subset$ for the proper subset relation (as opposed to using $\subset$ for subset and $\subsetneq$ for proper subset). However, we occasionally use the symbol $\underset{\neq}{\subset}$ whenever it is important in a proof that the containment is proper, or to otherwise add emphasis when the inequality is important. This symbol is purposefully distinct from the symbol $\subsetneq$ used in the other convention.

6.1 Transfinite Iteration

We recall some definitions about the topology on $\mathbb{R}$:

1. A subset $U \subseteq \mathbb{R}$ is *open* if for every point $x \in U$, there exists $\varepsilon > 0$ such that $\mathbb{B}(x,\varepsilon) = (x-\varepsilon,x+\varepsilon) \subseteq U$.
2. A set $F \subseteq \mathbb{R}$ is *closed* if its complement $\mathbb{R} \setminus F$ is open.
3. Equivalently (in metric spaces), a set $F \subseteq \mathbb{R}$ is closed if and only if it contains the limit of every convergent sequence in F. That is, if $(x_n)_{n=1}^{\infty} \subseteq F$ converges to $x \in \mathbb{R}$, then $x \in F$.
4. A point $p \in F$ is *isolated* (in F) if there exists $\varepsilon > 0$ such that $\mathbb{B}(p,\varepsilon) \cap F = \{p\}$, or equivalently, $\mathbb{B}(p,\varepsilon) \cap \big(F \setminus \{p\}\big) = \varnothing$.

Example.

- Any interval with positive measure has no isolated points.
- The entire set $\mathbb{R}$ and the empty set $\varnothing$ has no isolated points.
- The set of rationals $\mathbb{Q}$ has no isolated points.
- The middle-third Cantor set has no isolated points.
- The point contained in a singleton set is isolated.
- Every point of $\mathbb{Z}$ is isolated (take any $\varepsilon < \frac{1}{2}$).
- The point 0 in the set $[-2,-1] \cup \{0\}$ is isolated (take any $\varepsilon < 1$).

△

We are interested in *removing* the isolated points. In the last example above, removing the isolated points yields the interval [0,1], which has no isolated points.

For any set $F \subseteq \mathbb{R}$, denote by $D(F)$ the *derived set* obtained by removing the isolated points from F, or equivalently, the set of all limit points of F. Note that if F is closed, the derived set $D(F)$ is also closed since isolated points, as singletons, of F are always open in F.

Example.

- $D\big([0,1]\big) = [0,1]$.
- $D\big([0,1] \cup \{2\}\big) = [0,1]$.
- $D\big(\{0\}\big) = \varnothing$.
- $D\big(\{\frac{1}{n} : n \in \mathbb{Z}^+\}\big) = \varnothing$.

△

The question is, "if we start with a closed set and remove all the isolated points, do we always get a set without isolated points?" Or more precisely, "given a closed set F, is the derived set $D(F)$ always free of isolated points?"

It turns out that the answer is "no" – in removing the isolated points, a point that was not isolated before may then become isolated in the resulting set.

For instance, define the set $X \subset \mathbb{R}$ by

$$X = [-2, -1] \cup \{0\} \cup \left\{\frac{1}{n} : n \in \mathbb{Z}^+\right\}$$

Nothing in the interval is isolated and 0 is not isolated since there are elements arbitrarily close to it, but all the elements $\frac{1}{n}$ are isolated (take $\varepsilon = \frac{n}{4}$ for each point $\frac{1}{n}$). Removing these points yields the derived set

$$D(X) = [-2, -1] \cup \{0\}$$

which has an isolated point, 0. The derived set of this set is then the interval $D\big(D(X)\big) = [-2, -1]$, and from that point onwards, applying the derivation operator leaves the set unchanged, as the set is now free from isolated points. Using superscripts to denote the number of iterations, we have the sequence

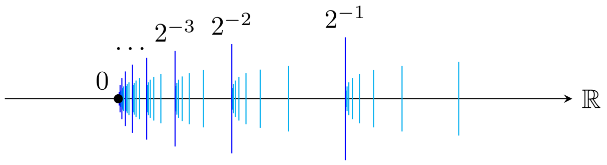
$$\begin{aligned}
D^0(X) &= X \\
D^1(X) &= [-2, -1] \cup \{0\} \\
D^2(X) &= [-2, -1] \\
D^3(X) &= [-2, -1]
\end{aligned}$$

so $D^2(X)$ is the first iteration where the set has no isolated points, after which the sequence stabilises to a fixed point.

In the previous example, X took 2 steps to stabilise because we had a sequence of isolated points that, when removed, produced a new isolated point. We can take inspiration from this to construct a set that takes another iteration to stablise by including sequences of isolated points that tend towards another sequence of isolated points. To avoid collisions, we use a geometric series rather than harmonic.

Define the set $Z \subset \mathbb{R}$ by

$$Z = \{0\} \cup \{2^{-n} : n \in \mathbb{Z}^+\} \cup \{2^{-n} + 2^{-n-m} : n,m \in \mathbb{Z}^+\}$$



The points of the form $2^{-n} + 2^{-n-m}$ are all isolated, and for each fixed n, the sequence $(2^{-n} + 2^{-n-m})$ tends to 2^{-n} as $m \to \infty$, so none of the points 2^{-n} are isolated. Also, the point 0 is not isolated, because the sequence (2^{-n}) tends towards it.

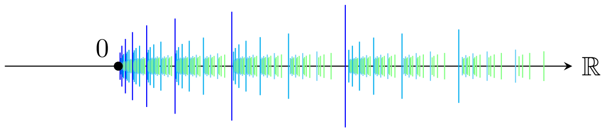
Thus, we have

$$\begin{aligned}
D^0(Z) &= \{0\} \cup \{2^{-n} : n \in \mathbb{Z}^+\} \cup \{2^{-n} + 2^{-n-m} : n,m \in \mathbb{Z}^+\} \\
D^1(Z) &= \{0\} \cup \{2^{-n} : n \in \mathbb{Z}^+\} \\
D^2(Z) &= \{0\} \\
D^3(Z) &= \varnothing \\
D^4(Z) &= \varnothing
\end{aligned}$$

So Z takes three iterations to stabilise.

By adding more and more sequences that converge to the sequences added in the previous step, this construction generalises, and it is possible to construct a set

$$E = \{0\} \cup \bigcup_{i \in \mathbb{Z}^+} \left\{ \sum_{j=1}^{i} 2^{-\sum_{k=1}^{j} n_k} : n_\alpha \in \mathbb{Z}^+ \right\}$$



such that $D^n(E)$ has isolated points for all natural n. This means that even if we remove the isolated points at all steps n, the set

$$D^\omega(E) := \bigcap_{n \in \mathbb{N}} D^n(E)$$

still has isolated points. (Note that ω is just notation right now.)

This is just a set, so it makes sense for us to apply the derivation operator again, which we may choose to denote by

$$D^{\omega+1}(E) := D\big(D^\omega(E)\big)$$

Again, this is still a set, so we may define

$$D^{\omega+2}(E) := D\big(D^{\omega+1}(E)\big)$$

and so on, until

$$D^{\omega+\omega}(E) := \bigcap_{\alpha=0,1,2,\ldots,\omega,\omega+1,\omega+2,\ldots} D^\alpha(E)$$

and we may suggestively define the notation $D^{2\cdot\omega}$ to represent $D^{\omega+\omega}$ more compactly, which suggest a further generalisation.

For any set S, we may then form the sequence of sets

$$D^0(S), D^1(S), \ldots, D^\omega(S), D^{\omega+1}(S), \ldots, D^{\overbrace{\omega+\omega}^{\omega\cdot 2}}(S), D^{\omega\cdot 2+1}(S), \ldots, D^{\omega\cdot 3}(S), \ldots, D^{\overbrace{\omega\cdot\omega}^{\omega^2}}(S), \ldots$$

The natural numbers can be used for two distinct purposes: to describe the size or *cardinality* of a set, or to describe the *position* of an element in a sequence, or more precisely, the *order-type* (sometimes called *ordinality*) of an ordered set.

The order-type of an ordered set is the first number not required to label the elements of the set. For instance, the set $\{a,b,c\}$ may be labelled by 0, 1, and 2, so it has ordinality 3. For finite sets, cardinality and order-type coincide.

However, the sequence above is too long for every element to be labelled by a natural number – after all, the natural numbers have all been exhausted by the time we reach $D^\omega(S)$ – so its order-type is greater than that of the naturals. On the other hand, the sequence is still countable, so its cardinality is the same as the naturals.

We still have not formally defined what any of these ω symbols mean – only the notation involving derivations that use them – but informally, they are the order-types of sets beyond the natural numbers.

6.2 The Set-Theoretic Universe

One way of specifying a (finite) set is to list out its members in curly brackets, e.g. $\{a,b,c,d\}$. The symbol $\in$ is called the *membership relation*, indicating that the object to the left is an "element of" or "member of" the set, so $a \in \{a,b,c,d\}$.

- Sets are unordered, so $\{b,a,c,d\} = \{a,b,c,d\}$;
- Elements are unique, so $\{a,a,a,b,b,c,d,d,d\} = \{a,b,c,d\}$.

The most basic set is the empty set, denoted by $\varnothing$, containing no elements. It may also be represented in the manner above, listing its elements as: $\{\}$.

The symbol $\subseteq$ is the *subset relation* between sets: $A \subseteq B$ if and only if every element in A is also an element in B.

Note that there are two ways for a set to be "in" another set: either as an element, or as a subset. To reduce ambiguity, whenever we say that A is "in" B, we mean $A \in B$, while the wording A is *contained* in B means $A \subseteq B$.

6.2.1 Atoms

Let A denote the collection of objects we want to talk about that are not themselves sets. For instance, the real number $\sqrt{2}$, the natural number 5, or the imaginary unit i, or anything else. Such an element is called an *atom* or *urelement.*

The goal is now to build a hierarchy of sets

$$V_0 \subseteq V_1 \subseteq V_2 \subseteq V_3 \subseteq \cdots$$

such that V_0 is the collection of all sets that can be formed from atoms. That is, an element of V_0 is a subset of A, so $V_0 = \mathcal{P}(A)$ is the power set of A.

Then, V_1 is the collection of all sets whose members are either atoms or sets in V_0. That is, an element of V_1 is a subset of $A \cup V_0$, so $V_1 = \mathcal{P}(A \cup V_0)$. A set containing only atoms is certainly a set containing only atoms *or* elements of V_1, so we also have $V_0 \subseteq V_1$.

We then recursively define

$$V_{n+1} \coloneqq \mathcal{P}(A \cup V_n)$$

and we have $V_n \subseteq V_{n+1}$ by induction.

The empty set is in V_0, so we have $\{\varnothing\} \in V_1$,$\{\{\varnothing\}\} \in V_2$,$\{\{\{\varnothing\}\}\} \in V_3, \ldots$, but the infinite set

$$\{\varnothing,\{\varnothing\},\{\{\varnothing\}\},\{\{\{\varnothing\}\}\},\ldots\}$$

is not in V_n for any natural n, since there will always be an element with at least $n+1$ nested brackets. To remedy this, we may take the infinite union

$$V_\omega := \bigcup_{n\in\mathbb{N}} V_n$$

(we still haven't defined ω, but the notation is coming in handy), which immediately extends to

$$\begin{aligned} V_{\omega+1} &:= \mathcal{P}(A \cup V_\omega) \\ V_{\omega+2} &:= \mathcal{P}(A \cup V_{\omega+1}) \\ &\vdots \end{aligned}$$

6.2.2 No Atoms

It turns out that atoms aren't really necessary.

Most higher mathematical objects are defined as sets equipped with certain operations on them, so as long as we can encode relations, and functions as sets, everything else should follow from there.

The previous hierarchy is also simpler without atoms. Because there are no atoms, the only set containing atoms is the empty set, so $V_0 = \{\varnothing\}$. Then, we have $V_1 = \mathcal{P}(V_0)$, and moer generally,

$$V_{n+1} := \mathcal{P}(V_n)$$

The ωth term, V_ω is the same as before, but the following sets follow the new pattern:

$$\begin{aligned} V_\omega &:= \bigcup_{n\in\mathbb{N}} V_n \\ V_{\omega+1} &:= \mathcal{P}(V_\omega) \\ V_{\omega+2} &:= \mathcal{P}(V_{\omega+1}) \\ &\vdots \end{aligned}$$

6.3 Unrestricted Comprehension

Rather than explicitly listing out the elements *extensionally*, we often define sets *intensionally* by specifying a property P and collecting all objects x that satisfy that property:

$$X = \{x : P(x)\}$$

However, we need to be careful with what we allow as the property P.

6.3.1 Frege's Natural Numbers

One intuitive implementation of the natural numbers as pure sets was given by Gottlob Frege. The idea is to define the number 1 as the set of all sets that have exactly one element. This may appear circular (1 vs "one"), but luckily, we can define the predicate $P(S)$ = "S has one element" in first order logic without appealing to a prior notion of "one":

$$P(S) := \exists x : \Big(x \in S \land \big(\forall y : (y \in S \to x = y)\big)\Big)$$

so the number 1 would be implemented as

$$[[1]] \coloneqq \big\{S : P(S)\big\}$$

Any finite number n can be similarly implemented as the set $[[n]]$ of all sets containing exactly n elements.

This construction seems reasonable. But let us now consider the set $\big\{[[1]]\big\}$. This set has exactly one element, namely $[[1]]$, so we have

$$[[1]] \in \big\{[[1]]\big\} \in [[1]]$$

There is, as yet, no reason that this circularity is incorrect, but it does seem somewhat suspect.

6.3.2 Universal Sets

Consider the *universal set* U that contains all sets:

$$U \coloneqq \{S : S \text{ is a set}\}$$

Since U is itself a set, it must be an element of U, so we have $U \in U$. This is again concerning.

6.3.3 Russell's Paradox

Consider the following set, proposed by Bertrand Russell:

$$R \coloneqq \{S : S \notin S\}$$

That is, R is the set of all sets that do not contain themselves. The problems then arise when we ask if $R \in R$ or not. If $R \in R$, then it is a set that does not contain itself, so we must have $R \notin R$. Conversely, if $R \notin R$, then by definition, $R \in R$.

6.3.4 Unrestricted Comprehension

In the previous examples, we obtained some questionable chains of membership, culminating with the self-contradictory membership of the Russell set. Informally, the problem is that these sets are "too large".

One method of resolving this problem is to switch to type theory. This approach has its own advantages, but the simple solution we will use is to introduce an axiom that restricts how sets are built.

So far, we have been using *unrestricted comprehension* to generate sets by collecting all objects that satisfy any given properties, but as we have seen, this is problematic. The first step is to only allow the collection of objects from an existing set. That is, we cannot form sets

$$\{x : P(x)\}$$

but only

$$\{x \in X : P(x)\}$$

where X is a known set. This already resolve some paradoxes – Russell's paradox included – but there is some ambiguity in what we mean by "property".

Let S be the set of the natural numbers that may be defined in less than 100 characters:

$$S = \{x \in \mathbb{N} : x \text{ can be defined in less than 100 characters}\}$$

So, for instance, "the third natural number", 3, is in S, as is "the 10000th prime number", 104 729, or "the numerator of the 15th coefficient of the Maclaurin series for tan", 689 005 380 505 609 448.

This set is huge. But it is finite: there are finitely many characters we can use, and as such, there are only finitely many strings with fewer than 100 characters. So, there is a smallest natural not in S.

The string,

"n is defined to be the smallest natural number that
cannot be defined in fewer than 100 characters."

that describes the number $n \notin S$ is 99 characters long. So, $n \in S$.

Despite restricting the set comprehension to only collect natural numbers, we still ran into a contradiction – we also need to restrict what qualifies as a "property". We replace "properties" with *formulae* in first order logic.

A formula may contain some or all of the following symbols:

- Logical symbols: $\wedge$, $\vee$, $\rightarrow$, $\leftrightarrow$, and $\neg$;
- Quantifiers: $\forall :$, $\exists :$ (the : are optional);
- Variable symbols: x_1,x_2,x_3, $\ldots$, or a,b,c,x,y,z,A,B,C, $\ldots$;
- Scoping symbols: (,);
- The equality symbol: $=$;
- The membership symbol: $\in$.

The syntax of valid formulae is defined recursively. Given a collection of valid symbols as above, the *atomic formulae* are as follows:

- The string $x = y$ is a valid formula for any variables x,y;
- The string $x \in y$ is a valid formula for any variables x,y.

and given two formulae φ and ψ, the following are all valid formulae:

- $(\varphi \wedge \psi)$;
- $(\varphi \vee \psi)$;
- $(\varphi \rightarrow \psi)$;
- $(\varphi \leftrightarrow \psi)$;
- $\neg\varphi$;
- $\forall x : \varphi$ for any variable x;
- $\exists x : \varphi$ for any variable x.

The inclusion of brackets ensures that every formula can be parsed unambiguously.

Note that we do not yet have the symbols $\neq$ and $\notin$. However, $x \neq y$ is just an abbreviation for the formula $\neg(x = y)$, and similarly, $x \notin y$ is an abbreviation of $\neg(x \in y)$.

We have also not yet defined the symbol $\subseteq$:

$$A \subseteq B \coloneqq \forall x(x \in A \rightarrow x \in B)$$

The notation of using memberships in quantifiers such as $\forall x \in X : P(x)$ or $\exists x \in X : P(x)$ can also be defined by:

$$\forall x \in X : P(x) \coloneqq \forall x\big(x \in X \rightarrow P(x)\big)$$
$$\exists x \in X : P(x) \coloneqq \exists x\big(x \in X \wedge P(x)\big)$$

(where $P(x)$ is some first order formula with free variable x).

Using these symbols, we can now begin to express some basic axioms for set theory.

6.4 The Axioms of ZF

6.4.1 Axiom of Extensionality

Axiom of Extensionality.
If two sets have exactly the same members, then they are equal:

$$\forall X \forall Y \big(\forall z (z \in X \leftrightarrow z \in Y) \to x = y\big)$$

Note that our theory concerns itself only with sets, so it matters not if we use upper or lowercase letters as variable symbols.

6.4.2 Axiom of The Empty Set

Axiom of the Empty Set.
There exists a set with no elements:

$$\exists E \forall x : x \notin E$$

We can prove a theorem with these two axioms:

Theorem 6.4.1. *There exists exactly one set with no members.*

Proof. By the axiom of the empty set, there exists at least one such set.

For uniqueness, suppose A and B are sets with no members. Then, for every x, the implication $x \in A \to x \in B$ holds vacuously, as does the reverse implication, so $A = B$ by the axiom of extensionality. ■

We call this set the *empty set*, with the theorem above justifying the wording "*the* empty set" over "*an* empty set".

6.4.3 Axiom of Pairing

Axiom of Pairing.
For any two sets u and v, there exists a set that contains exactly u and v as elements.

$$\forall u \forall v \exists X \forall x \big(x \in X \leftrightarrow (x = u \lor x = v)\big)$$

We denote the set obtained from pairing u and v by $\{u,v\}$, with uniqueness given by extensionality. If $u = v$, then we also denote this by $\{u\}$.

6.4.4 Axiom of Binary Union

Axiom of Binary Union.
For any two sets u and v, there is a set whose members are those sets that are members of u or of v:

$$\forall u \forall v \exists U \forall x : \big(x \in U \leftrightarrow (x \in u \lor x \in v)\big)$$

6.4.5 Axiom of the Power Set

Axiom of the Power Set.
For any set u, there is a set whose elements are exactly the subsets of u:

$$\forall u \exists P \forall s (s \subseteq u \leftrightarrow s \in P)$$

or omitting the abbreviation $\subseteq$,

$$\forall u \exists P \forall s \big(\forall x (x \in s \to x \in s) \leftrightarrow s \in P\big)$$

6.4.6 Free and Bound Variables

A *free variable*, also called a *parameter*, is a variable that is not *bound* by a preceding quantifier.

Example. In the formula

$$\forall x : x = y$$

the variable x is bound, while y is free. △

Quantifiers have a certain *scope* in which they bind their variables, so more accurately, we should talk about free and bound *instances* or *occurrences* of variables.

Example. Consider the formula

$$(\forall x : x \in y) \land (\exists y : y \in X)$$

Clearly, x is bound and X is free. However, y is free in the first clause and bound in the second. △

This is similar to variable binding in other areas of mathematics:

$$x + \int_0^1 x \, dx$$

Although the integral uses the symbol x, that instance of the symbol is bound, and doesn't really have anything to do with the free variable x outside of the integral. We could (and should) use a different label for that bound variable:

$$x + \int_0^1 t \, dt$$

Similarly, the symbol used to denote a bound variable may be changed freely in a formula of first order logic:

$$(\forall x : x \in y) \land (\exists z : z \in X)$$

(This is closely related to the notion of α-conversion in the lambda calculus.)

6.4.7 Truth Values

A formula that has no free variables is called a *sentence*. Assuming classical first order logic – specifically, the law of the excluded middle – any sentence has a truth value (though they may not be decidable in any particular theory c.f. Gödel's incompleteness theorems).

Formulae that are not sentences do not have truth values, because they contain free variables. A formula that is not a sentence may be true for all possible valuations of their parameters, such as

$$x = x$$

which is true for all x (such a formula is a *tautology*), but the formula itself does not have a truth value.

6.5 More Axioms

6.5.1 Axiom Schema of Specification

Most of the existence axioms so far have been of the form

$$\forall t_1 \forall t_2 \ldots \forall t_k \exists B \forall x \big(x \in B \leftrightarrow \varphi\big)$$

where φ is some meaningful statement about x and the sets $t_1, \ldots, t_k$. For instance, we have $k = 2$ with $t_1 = u$ and $t_2 = v$ in the pairing or axiom union, and $k = 0$ in the axiom of the empty set.

Note that for this formula to be meaningful, the statement φ cannot contain B as a variable, or else it's not a very useful definition of the set B. Also, to resolve Russell's paradox as before, we will only allow the collection of elements from some existing set. This leads to the next set of axioms:

Unlike the axioms we have seen so far, this is an axiom *schema* because it contains infinitely many axioms – one for each property φ.

Axiom Schema of Specification.
Let φ be any formula that does not contain the variable name B and has only bound variables, except for $x, t_1, \ldots, t_k$. Then, the following is an axiom:

$$\forall t_1 \forall t_2 \ldots \forall t_k \forall A \exists B \forall x \big(x \in B \leftrightarrow (x \in A \wedge \varphi)\big)$$

That is, for any property φ of x and any set A, there exists a set B that contains exactly the elements of A for which $\varphi(x)$ holds, and φ may depend on additional parameters $t_1, \ldots, t_k$.

This axiom schema is also known as *separation*, since it allows us to separate out the elements of a set that satisfy a property, or as *restricted comprehension*, since it constrains what sets can be constructed via set comprehension.

These axioms define new sets, which we denote as

$$B_{t_1,\ldots t_k,A} = \{x \in A : \varphi(x, t_1, \ldots, t_k)\}$$

or

$$B = \{x \in A : \varphi(x)\}$$

Theorem 6.5.1. *For any sets A and T, there is a unique set B whose elements are precisely those that are members of both A and T.*

This set is called the *intersection* of A and T, and is denoted (as usual) by $B = A \cap T$.

Proof. The axiom schema of specification contains the axiom

$$\forall t_1 \forall A \exists B \forall x \big(x \in B \leftrightarrow (x \in A \wedge x \in t_1)\big)$$

which implies the existence of at least one such set. For uniqueness, suppose X and Y are both intersections of A and T. Then, for each $x \in X$,

$$x \in X \leftrightarrow (x \in A \wedge x \in B) \leftrightarrow x \in Y$$

so $X = Y$ by extensionality. ■

We can also prove that Russell's paradox does not occur using these axioms:

Theorem 6.5.2. *Russell's set $R = \{x : x \notin x\}$ does not exist.*

Proof. Suppose R exists. If $R \in R$, then $R \notin R$; and if $R \notin R$, then $R \in R$. In either case, we have a contradiction. ■

Theorem 6.5.3. *There is no set of all sets. That is,*

$$\neg\exists U \forall x : x \in U$$

Proof. Suppose U is such a set, and let φ be the formula $x \notin x$. The formula φ does not contain U and has x free with no other bound variables, so specification yields the set

$$R = \{x \in U : x \notin x\}$$

That is,

$$x \in R \leftrightarrow (x \in U \land x \notin x)$$

Because U contains every set, $x \in U$ is a tautology, so

$$x \in R \leftrightarrow x \notin x$$

This implies that R is Russell's set, which does not exist by the previous theorem. ■

6.5.2 Axiom of Union

Using the axiom of binary union, we can form the union $A \cup B$ of two sets, and by repeating it, we can form the union of three or more sets as $(A \cup B) \cup C$. However, we cannot form arbitrary unions of infinitely many sets $u_1, u_2, \dots$

Because we attempting to axiomatise sets, we will require that any collection of sets we are attempting to take the union of is itself a set. That is, we wish to take the union of the members of a set $A = \{u_1, u_2, \dots\}$. Such a union would be a set B whose elements are exactly the members of the members of A.

Axiom of Union.
For any set A, there exists a set B whose members are precisely the members of the members of A:

$$\forall A \exists B \forall x \big(x \in B \leftrightarrow \exists y (y \in A \land x \in y)\big)$$

This set is unique through an argument identical to that for intersections. We denote this set by

$$B = \bigcup A$$

That is, $x \in \bigcup A$ if and only if there is a set $y \in A$ such that $x \in y$.

Note that $\bigcup$ is a unary operator, while the previous $\cup$ was binary. However, the new operator is stronger in that we can take the union of more sets.

Theorem 6.5.4. *The axiom of union and axiom of pairing imply the axiom of binary union.*

Proof. Pairing sets u and v gives $\{u,v\}$, and the union $\bigcup\{u,v\}$ is precisely $u \cup v$. ■

6.5.3 Arbitrary Intersections

We have already found the binary intersection using specification, but we would like to define the intersection of members of any set. Unlike for unions, a new axiom isn't required for this.

Theorem 6.5.5. *For any non-empty set A, there is a unique set B whose members are precisely those that are members of all members of A. That is,*

$$x \in B \leftrightarrow \forall y(y \in A \to x \in y)$$

Proof. Suppose A is non-empty and fix some set $w \in A$. Then, the set

$$\big\{x \in w : \forall y(y \in A \to x \in y)\big\}$$

exists by specification, and its members are precisely those that are members of every member of A. Uniqueness follows from extensionality. ■

Analogously to unions, we write $B = \bigcap A$ for this unary intersection. We can also express binary unions as the special case $x \cap y = \bigcap\{x,y\}$.

Theorem 6.5.6. *Let A be non-empty and let $A \subseteq B$. Then,*

$$\bigcap A \supseteq \bigcap B$$

Proof. Let $x \in \bigcap B$. By the definition of the intersection,

$$\begin{aligned} x \in \bigcap B &\leftrightarrow \forall y(y \in B \to x \in y) \\ &\to \forall y(y \in A \to x \in y) \\ &\leftrightarrow x \in \bigcap A \end{aligned}$$

so $\bigcap B \subseteq \bigcap A$ as required. ■

6.6 Ordered Pairs

One important basic mathematical structure is the *ordered pair*, written as (x,y) or $\langle x,y\rangle$, satisfying the characteristic property

$$\langle x,y\rangle = \langle a,b\rangle \leftrightarrow (x = a \land y = b)$$

The goal is to find a representation of this structure made from pure sets that satisfies the property above.

Let's first see two encodings that don't work:

1. $\langle x,y\rangle \coloneqq \{x,y\}$. For any sets x and y, we have,

$$\begin{aligned} \langle x,y\rangle &= \{x,y\} \\ &= \{y,x\} \\ &= \langle y,x\rangle \end{aligned}$$

2. $\langle x,y\rangle \coloneqq \big\{x,\{y\}\big\}$. For any sets x and y, we have,

$$\begin{aligned} \big\langle\{x\},y\big\rangle &= \big\{\{x\},\{y\}\big\} \\ &= \big\{\{y\},\{x\}\big\} \\ &= \big\langle\{y\},x\big\rangle \end{aligned}$$

In either encoding, if $x \neq y$, we have a contradiction.

The standard *Kuratowski construction* of the ordered pair is given by

$$\langle x,y \rangle \coloneqq \big\{\{x\},\{x,y\}\big\}$$

which exists by repeated applications of pairing.

Theorem 6.6.1. *The Kuratowski construction satisfies the characteristic property of the ordered pair. That is,*

$$\big\{\{x\},\{x,y\}\big\} = \big\{\{a\},\{a,b\}\big\} \leftrightarrow (x = a \wedge y = b)$$

Proof. Suppose that

$$\big\{\{x\},\{x,y\}\big\} = \big\{\{a\},\{a,b\}\big\}$$

We have $\{x,y\} \in \big\{\{a\},\{a,b\}\big\}$, so either

- (a) $\{x,y\} = \{a\}$ or
- (b) $\{x,y\} = \{a,b\}$

holds. We also have $\{x\} \in \big\{\{a\},\{a,b\}\big\}$, so either

- (c) $\{x\} = \{a\}$ or
- (d) $\{x\} = \{a,b\}$

holds.

If (a) holds, then $x = y = a$ and the equation reduces to

$$\big\{\{a\}\big\} = \big\{\{a\},\{a,b\}\big\}$$

so $x = y = a = b$. If (b) holds and

- if (c) also holds, we have $x = a$ and $\{x,y\} = \{x,b\}$. If $x = b$, then $x = y = a = b$. Otherwise, $y = b$.
- if (d) also holds, then $x = y = a = b$.

In all cases, $x = a$ and $y = b$.

The other direction is trivial. ■

Theorem 6.6.2.

- (i) *There is a formula with free variables x,y,z that is satisfied if and only if $z = \langle x,y \rangle$.*
- (ii) *There is a formula with free variable z that is satisfied if and only if z is an ordered pair.*
- (iii) *There is a formula with free variable x,z that is satisfied if and only if z is an ordered pair with x as the first coordinate.*
- (iv) *There is a formula with free variable y,z that is satisfied if and only if z is an ordered pair with y as the second coordinate.*

Proof.

(i)

$$\varphi(x,y,z) = \exists L \exists R \Big(\underbrace{(L \in z) \wedge (R \in z) \wedge \forall t \big(t \in z \to (t = L \vee t = R)\big)}_{z \text{ has precisely two elements, } L \text{ and } R}$$

$$\wedge \underbrace{(x \in L) \wedge \forall l (l \in L \to l = x)}_{L \text{ contains precisely } x} \wedge \underbrace{(x \in R) \wedge (y \in R) \wedge \forall r (r \in R \to r = x \vee r = y)}_{R \text{ contains precisely } x \text{ and } y} \Big)$$

The next three follow trivially as partial applications of this formula:

(ii) $\exists x \exists y : \varphi(x,y,z)$

(iii) $\exists y : \varphi(x,y,z)$

(iv) $\exists x : \varphi(x,y,z)$

■

6.6.1 Cartesian Product

Now we have ordered pairs, we can define cartesian products as follows:

$$A \times B \coloneqq \big\{ \langle a,b \rangle : a \in A \wedge b \in B \big\}$$

However, we have not yet proven the existence of this set. First, we identify a larger set that contains all such ordered pairs, before using specification to obtain the cartesian product.

Lemma 6.6.3. *If $a \in A$ and $b \in B$, then $\langle a,b \rangle \in \mathcal{P}\mathcal{P}(A \cup B)$.*

Proof.

$$\begin{aligned} a \in A \wedge b \in B &\to a \in A \cup B \wedge b \in A \cup B \\ &\leftrightarrow \{a\} \subseteq A \cup B \wedge \{a,b\} \subseteq A \cup B \\ &\leftrightarrow \{a\} \in \mathcal{P}(A \cup B) \wedge \{a,b\} \in \mathcal{P}(A \cup B) \\ &\leftrightarrow \big\{\{a\},\{a,b\}\big\} \in \mathcal{P}(A \cup B) \\ &\leftrightarrow \big\{\{a\},\{a,b\}\big\} \in \mathcal{P}\mathcal{P}(A \cup B) \\ &\leftrightarrow \langle a,b \rangle \in \mathcal{P}\mathcal{P}(A \cup B) \end{aligned}$$

■

Theorem 6.6.4. *The cartesian product of two sets is a set.*

Proof. Consider the formula

$$\psi(a,b,z) \equiv a \in A \wedge b \in B \wedge \varphi(a,b,z)$$

where φ is the formula from Theorem 6.6.2. Specification then yields the set

$$\big\{ z \in \mathcal{P}\mathcal{P}(A \cup B) : \psi(a,b,z) \big\}$$

which, by the lemma above, is $A \times B$. Uniqueness follows from extensionality. ■

The cartesian product is not commutative, as the pairs are ordered, and it is also not associative, as $(A \times B) \times C$ consists of pairs of the form $\langle \langle a,b \rangle, c \rangle$, while $A \times (B \times C)$ consists of pairs of the form $\langle a, \langle b,c \rangle \rangle$. They are, however, naturally isomorphic.

To reduce the number brackets required, the convention is that the product operator binds to the left. That is, $A \times B \times C \times D \times E$ should be read as $((((A \times B) \times C) \times D) \times E)$.

6.7 Relations and Functions

6.7.1 Relations

A *relation* R is a set of ordered pairs. If $\langle x,y\rangle \in R$, then we use infix notation and write xRy.

Given a relation R, we define the *domain*, *range*, and *field* of R as

$$\begin{aligned}\operatorname{dom}(R) &\coloneqq \{x \mid \exists y : xRy\}\\ \operatorname{ran}(R) &\coloneqq \{y \mid \exists x : xRy\}\\ \operatorname{field}(R) &\coloneqq \operatorname{dom}(R) \cup \operatorname{ran}(R)\end{aligned}$$

Lemma 6.7.1. *If $\langle x,y\rangle \in A$, then $x,y \in \bigcup\bigcup A$.*

Proof.

$$\begin{aligned}\langle x,y\rangle &\in A\\ \big\{\langle x,y\rangle\big\} &\subset A\\ \Big\{\{\{x\},\{x,y\}\}\Big\} &\subset A\\ \bigcup\Big\{\{\{x\},\{x,y\}\}\Big\} &\subset \bigcup A\\ \{\{x\},\{x,y\}\} &\subset \bigcup A\\ \bigcup\{\{x\},\{x,y\}\} &\subset \bigcup\bigcup A\\ \{x,y\} &\subset \bigcup\bigcup A\\ x,y &\in \bigcup\bigcup A\end{aligned}$$

■

Corollary 6.7.1.1. *The domain, range, and field of a relation are sets.*

Proof. By specification, the following are sets:

$$\begin{aligned}\operatorname{dom}(R) &= \Big\{x \in \bigcup\bigcup R \,\Big|\, \exists y : xRy\Big\}\\ \operatorname{ran}(R) &= \Big\{x \in \bigcup\bigcup R \,\Big|\, \exists x : xRy\Big\}\end{aligned}$$

so $\operatorname{field}(R) = \operatorname{dom}(R) \cup \operatorname{ran}(R)$ is a set by union. ■

6.7.2 Functions

In ordinary mathematics, we think of a function as a special kind of correspondence between pairs of object, where every given object x (the "input") is assigned exactly one corresponding object y (its "image") by the function. This means that each such object can be represented as an ordered pair $\langle x,y\rangle$.

A *function* F is a relation such that for every $x \in \operatorname{dom}(R)$ there exists a unique y with $\langle x,y\rangle \in F$. If $\langle x,y\rangle \in F$ for some function F, we write $y = F(x)$ to denote this.

The *inverse* of a relation R is the set

$$R^{-1} \coloneqq \big\{\langle x,y\rangle : yRx\big\}$$

and the *composition* of two relations R and T is the set

$$R \circ T \coloneqq \big\{\langle x,y\rangle : \exists z(xTz \wedge zRy)\big\}$$

Theorem 6.7.2. *If F and G are functions, then the composition $F \circ G$ is a function with domain*

$$\operatorname{dom}(F \circ G) = \{x \in \operatorname{dom}(G) : G(x) \in \operatorname{dom}(F)\}$$

Proof. Suppose $\langle x,y\rangle,\langle x,y'\rangle \in F \circ G$. Then, there exist t and t' such that

$$\begin{aligned} \langle x,t\rangle &\in G & \langle t,y\rangle &\in F \\ \langle x,t'\rangle &\in G & \langle t',y'\rangle &\in F \end{aligned}$$

Since G is a function, $t = t'$, so we have

$$\begin{aligned} \langle t,y\rangle &\in F \\ \langle t,y'\rangle &\in F \end{aligned}$$

and since F is a function, $y = y'$, so $\langle x,y\rangle = \langle x,y'\rangle$ and $F \circ G$ is a function. ■

6.7.3 Images

Given a relation R and a set X, the (R-)*image* of X under R is the set

$$R[X] \coloneqq \{y \mid \exists x \in X : \langle x,y\rangle \in R\}$$

and the (R-)*preimage* is the set

$$\begin{aligned} R[X] &\coloneqq \{y \mid \exists x \in X : \langle x,y\rangle \in R^{-1}\} \\ &= \{y \mid \exists x \in X : \langle y,x\rangle \in R\} \end{aligned}$$

Note that it is not necessary to have $X \subseteq \operatorname{dom}(R)$ when taking the R-image, or $X \subseteq \operatorname{ran}(R)$ when taking the R-preimage, since $R[X] = R[X \cap \operatorname{dom}(R)]$ and $R^{-1}[X] = R^{-1}[X \cap \operatorname{ran}(R)]$.

A function F is *injective* if for every y there is at most one x such that $\langle x,y\rangle \in F$.

Theorem 6.7.3. *A function F is injective if and only if F^{-1} is a function.*

Otherwise F^{-1} is a relation but not a function.

Theorem 6.7.4. *For any functions F and G, we have*

- $(F \circ G)(x) = F\big(G(x)\big)$*;*
- $(F \circ G)^{-1} = G^{-1} \circ F^{-1}$.

We write $f : A \to B$ to denote a function f with $\operatorname{dom}(f) = A$ and $\operatorname{ran}(f) \subseteq B$. A function f is *surjective* if $\operatorname{ran}(f) = B$.

Theorem 6.7.5.

- *For any $A \neq \varnothing$, there is no function $f : A \to \varnothing$;*
- *For any B, there is a unique function $f : \varnothing \to B$ given by $f = \varnothing$.*

6.7.4 Cantor's Diagonal Argument

Theorem (Cantor)**.** *For any set A, there is no surjection $f : A \to \mathcal{P}(A)$.*

Proof. Let $f : A \to \mathcal{P}(A)$ be a function, and define the set

$$S = \{a \in A : a \notin f(a)\}$$

which is a set by specification. Clearly, $S \subseteq A$, so $S \in \mathcal{P}(A)$.

Suppose $S \in \operatorname{ran}(f)$. Then, there exists $a_0 \in A$ such that $S = f(a_0)$. If $a_0 \in S$, then $a_0 \notin f(a_0) = S$, and if $a_0 \notin S$, then $a_0 \in f(a_0) = S$. This is a contradiction, so $S \notin \operatorname{ran}(f)$, and f is not surjective. ■

We define the *identity function* on a set A by

$$I_A \coloneqq \big\{\langle a,a\rangle : a \in A\big\}$$

That is, $I_A(a) = a$ for all $a \in A$.

Theorem 6.7.6. *If A is non-empty and there is an injective function $f : A \to B$, then there exists a surjective function $g : B \to A$ such that $g \circ f = I_A$.*

Proof. Let $a_0 \in A$. If $b \in \operatorname{ran}(f)$, then define $g(b)$ to be the unique $a \in A$ such that $f(a) = b$ (since f is injective), otherwise define $g(b) = a_0$. This function (set) exists because

$$g = f^{-1} \cup \big\{\langle b,a_0\rangle \in B \times A : b \notin \operatorname{ran}(f)\big\}$$

is the union of two sets, the latter of which exists by specification.

Clearly, g is surjective and we have $g \circ f = I_A$. ∎

Theorem 6.7.7. *If $f : A \to B$ is surjective, then there exists an injective function $g : B \to A$ such that $f \circ g = I_B$.*

The sets $f^{-1}\big[\{b\}\big]$ for $b \in B$ are non-empty and partition A.

For each b, select $a_b \in f^{-1}\big[\{b\}\big]$. Then, the map $g : B \to A$ defined by $b \mapsto a_b$ is injective, and clearly, $f \circ g = I_B$.

However, this proof is not yet valid with the axioms we have so far – we have yet to prove that the set constructed above exists. The problem is that we do not know if collecting an element a_b from infinitely many sets $f^{-1}\big[\{b\}\big]$ yields a set.

Axiom of Choice (first form).
For any relation R, there exists a function $F \subseteq R$ such that $\operatorname{dom}(F) = \operatorname{dom}(R)$.

That is, for each element $x \in \operatorname{dom}(R)$, the function F "chooses" exactly one y with $\langle x,y\rangle \in R$. That is, exactly one element $y \in R\big[\{x\}\big]$.

Proof of Theorem 6.7.7. (AC) Suppose $f : A \to B$ is surjective, and consider the relation f^{-1}. We have $\operatorname{dom}(f^{-1}) = B$ and $\operatorname{ran}(f^{-1}) = A$, so by the axiom of choice, there exists a function $F \subset f^{-1}$ with domain B. For every $b \in B$, we must have $\langle F(b),b\rangle \in f$, so $f\big(F(b)\big) = b$. That is, $f \circ F = I_B$. ∎

6.8 Constructing Numbers

6.8.1 Axiom of Infinity

Since we are attempting to embed all of mathematics into set theory, we should find sets that correspond to natural numbers, as well as a set that contains all natural numbers.

Zermelo proposed the following construction of the natural numbers:

- $0 = \varnothing$;
- $1 = \{\varnothing\}$;
- $2 = \{\{\varnothing\}\}$;
- $3 = \{\{\{\varnothing\}\}\}$;
- $n = \{n-1\}$.

However, the generally accepted (and in many senses better) convention given by von Neumann is as follows:

- $0 = \varnothing$;
- $1 = \{0\} = \{\varnothing\}$;
- $2 = \{0,1\} = \{\varnothing,\{\varnothing\}\}$;
- $3 = \{0,1,2\} = \{\varnothing,\{\varnothing\},\{\varnothing,\{\varnothing\}\}\}$;
- $4 = \{0,1,2,3\} = \{\varnothing,\{\varnothing\},\{\varnothing,\{\varnothing\}\},\{\varnothing,\{\varnothing\},\{\varnothing,\{\varnothing\}\}\}\}$;
- $n = \{0,1,\dots,n-1\}$.

One major advantage of this encoding is that the set representing n now has n elements (while in Zermelo's construction, every natural number n has exactly one element – $n-1$).

We also have

$$0 \in 1 \in 2 \in 3 \in \cdots$$

and

$$0 \subset 1 \subset 2 \subset 3 \subset \cdots$$

For a set x, we define its *successor* x^+ by $x^+ = x \cup \{x\}$.

Example.

- $0^+ = 0 \cup \{0\} = \{0\} = 1$;
- $1^+ = 1 \cup \{1\} = \{0,1\} = 2$;
- $2^+ = 2 \cup \{2\} = \{0,1,2\} = 3$;
- $3^+ = 3 \cup \{3\} = \{0,1,2,3\} = 4$;

so $0 = \varnothing$, $1 = \varnothing^+$, $2 = \varnothing^{++}$, $3 = \varnothing^{+++}$. △

It is now intuitively clear what natural numbers are in set theory, but without a further axiom, we cannot write a formula $\varphi(x)$ that identifies if x is a natural number or not, and we also cannot form the set of all natural numbers yet.

Lemma 6.8.1. *For all $m,n \in \omega$.*

(i) $0 \neq n^+$;

(ii) $m \in n \to m^+ \in n^+$;

(iii) $m^+ = n^+ \to m = n$.

(You may recognise some of these as Peano axioms.)

A set A is *inductive* if it contains the empty set and is closed under the successor operation. That is,

$$\forall x(x \in A \to x^+ \in A)$$

Axiom of Infinity.
There is an inductive set:

$$\exists A\big(\varnothing \in A \land \forall x(x \in A \to x^+ \in A)\big)$$

or omitting $\varnothing$ and x^+,

$$\exists A\big(\exists e(\forall z : z \notin e) \land e \in A \land \forall x(x \in A \to x \cup \{x\} \in A)\big)$$

The axiom of infinity gives the existence of inductive sets, but does not provide uniqueness. We are only interested in the "smallest" such set.

A *natural number* is a set that is a member of every inductive set.

Theorem (Existence of Natural Numbers).

(*i*) *There is a set ω whose elements are precisely the natural numbers.*

(*ii*) *The set ω is the unique set that is a subset of every inductive set.*

Proof. Let A be an inductive set given by the axiom of infinity. Then, by specification, the following is a set:

$$\begin{aligned}\omega &= \big\{a \in A : \forall S((\forall y : y \in S \to y^+ \in S) \to x \in S)\big\}\\ &= \{a \in A : x \text{ is an element of every inductive set}\}\end{aligned}$$

Clearly, its elements are precisely the natural numbers, and we also have $\omega \subseteq S$ for any inductive set S, with uniqueness given by extensionality. ■

We will use the notation $\mathbb{N} = \{0,1,2,\ldots\}$ to refer to the natural numbers when the encoding is irrelevant, and ω whenever von Neumann's convention is required.

Theorem (Induction Principle for ω). *Any inductive subset of ω coincides with ω.*

Proof. Clear from the definition of ω. ■

Theorem (Proof by Induction). *Suppose φ is a property of natural numbers such that $\varphi(0)$ holds and for every natural number x, $\varphi(x) \to \varphi(x^+)$. Then $\varphi(x)$ holds for all natural numbers x.*

Proof. Let $X = \{x \in \omega : \varphi(x)\} \subseteq \omega$. By assumption, $\varphi(0)$ holds, so $0 = \varnothing \in X$, and because $\varphi(x) \to \varphi(x^+)$, X is inductive. Thus, by the previous theorem, $X = \omega$. ■

Theorem 6.8.2. *Every element of ω is either 0 or the successor x^+ of a unique $x \in \omega$.*

Proof. Define the set

$$A = \{n \in \omega : n = \varnothing \vee \exists m \in \omega : n = m^+\}$$

Clearly, $0 = \varnothing$ is a member of A, and A is closed under the successor operation, so A is inductive, and hence $A = \omega$. Injectivity of the successor function (Lemma 6.8.1) implies uniqueness. ■

6.8.2 Ordering of ω

We define the *less than* relation $<_\omega$ on ω by

$$<_\omega := \big\{\langle m,n\rangle \in \omega \times \omega : m \in n\big\}$$

and the *less than or equal to* relation $\leq_\omega$ on ω by

$$\leq_\omega := \big\{\langle m,n\rangle \in \omega \times \omega : m \in n \vee m = n\big\}$$

If $\langle m,n\rangle \in <_\omega$, we write $m < n$, and similarly, if $\langle m,n\rangle \in \leq_\omega$, we write $m \leq n$.

Theorem 6.8.3. *The relation $\in$ linearly orders ω. That is, it is:*

(*i*) *irreflexive: $\forall n \in \omega : n \notin n$;*

(*ii*) *transitive: $\forall x,y,z \in \omega : (x \in y \wedge y \in z) \to x \in z$;*

(*iii*) *linear/total:* $\forall m,n \in \omega : m \in n \vee m = n \vee n \in m$.

Proof.

(*i*) Define the set

$$A = \{n \in \omega : n \notin n\}$$

We show A is inductive via induction on n.

The empty set has no elements, so $\varnothing \notin \varnothing$ holds, and $\varnothing \in A$.

Now assume that $n \in A$ and suppose for a contradiction that $n^+ \in n^+ = n \cup \{n\}$. Then, either $n^+ \in n$ or $n^+ = n$. In the former case, $n \in n \cup \{n\} = n^+ \in n$, so by transitivity (*ii*), $n \in n$, contradicting that $n \in A$. In the latter case, $n \in n \cup \{n\} = n^+ = n$, so $n \in n$, again contradicting that $n \in A$.

It follows that $n^+ \notin n^+$, so $n^+ \in A$. So, $A \subseteq \omega$ is inductive, giving $A = \omega$, and hence $\in$ is an irreflexive relation on ω.

(*ii*) Fix $x,y \in \omega$ and define the set

$$A = \{z \in \omega : x \in y \in z \to x \in z\}$$

We show A is inductive via induction on z.

If $z = \varnothing$, then the implication holds vacuously, so $\varnothing \in A$.

Now assume that $z \in A$, and suppose that $x \in y \in z^+$. As $y \in z^+ = z \cup \{z\}$, we either have $y \in z$ or $y = z$. If $y \in z$, then by the inductive hypothesis $x \in z$, and since $z \subseteq z^+$, we have $x \in z^+$, so $z^+ \in A$.

If $y = z$, then $x \in y$ gives $x \in z$. Again, $z \subseteq z^+$, so $x \in z^+$, and $z^+ \in A$.

So, $A \subseteq \omega$ is inductive, giving $A = \omega$, and hence $\in$ is a transitive relation on ω.

(*iii*) Define the set

$$B = \{n \in \omega : \varnothing = n \vee \varnothing \in n\}$$

We show B is inductive via induction on n.

If n is empty, then $n = \varnothing$ so $n \in B$.

Now assume that $n \in B$, so either $\varnothing \in n$ or $\varnothing = n$. In the former case, $\varnothing \in n \in n \cup \{n\} = n^+$, and in the latter case $\varnothing \in \varnothing \cup \{\varnothing\} = n^+$. In either case, $\varnothing \in n^+$ by transitivity (*ii*), so $B \subseteq \omega$ is inductive and hence $B = \omega$, so $\varnothing \in n \vee \varnothing = n$ holds for all $n \in \omega$.

Now define the set

$$A = \{m \in \omega : \forall n(m \in n \vee m = n \vee n \in m)\}$$

If $m = \varnothing$, then $m \in A$ by the above result.

Now assume $m \in A$. If $n \in m$, then $n \in m^+$ by transitivity. If $n = m$, then $n^+ = m^+$. If $m \in n$, then $m^+ \in n^+$, so either $m^+ \in n$ or $m^+ = n$. In all cases, one of the conditions hold, so $m^+ \in A$. So, $A \subseteq \omega$ is inductive, giving $A = \omega$, and hence $\in$ is total on ω.

■

6.8.3 Recursion

Theorem (Recursion on ω)**.** *Let X be a set and $x \in X$. Let $r : X \to X$ be a function. Then, there is a unique function $f : \omega \to X$ such that*

(*i*) $f(0) = x$;

(*ii*) $f(n^+) = r\big(f(n)\big)$.

Informally, given a set X, and an element $x \in X$, every function $r : X \to X$ generates a unique sequence of elements $(x_i)_{i=1}^{\infty} \subseteq X$ such that $x_0 = x$ and $x_{n+1} = r(x_n)$. The theorem above then says that the mapping that sends n to x_n defines a function $\omega \to X$.

That is, the following diagram commutes

$$\begin{tikzcd}
 & \mathbb{N} \arrow[r, "s"] \arrow[dd, "f"] & \mathbb{N} \arrow[dd, "f"] \\
\{*\} \arrow[ru, "0"] \arrow[rd, "x"'] & & \\
 & X \arrow[r, "r"'] & X
\end{tikzcd}$$

Proof sketch. Call a function v *acceptable* if the following four properties hold:

- $\mathrm{dom}(v) \subseteq \omega$;
- $\mathrm{ran}(v) \subseteq A$;
- $0 \in \mathrm{dom}(v) \to v(0) = a$;
- $n \in \omega \land n^+ \in \mathrm{dom}(v) \to n \in \mathrm{dom}(v) \land v(n^+) = F\big(v(n)\big)$.

The last property implies that if $n \in \mathrm{dom}(v)$, then $\{0,1,...,n\} \subseteq \mathrm{dom}(v)$. The empty set is also an accemptable function, as is the function $v = \big\{\langle 0,a\rangle\big\}$.

Let $\mathcal{K}$ be the collection of acceptable functions. Because any acceptable function is a function with domain contained in ω and range contained in A, all of its elements are ordered pairs in $\omega \times A$. That is, every acceptable function v is a subset of $\omega \times A$, or an element in $\mathcal{P}(\omega \times A)$. Thus, $\mathcal{K} \subseteq \mathcal{P}(\omega \times A)$ constitutes a set by specification.

Define

$$f \coloneqq \bigcup \mathcal{K}$$

That is, h is the relation formed from the union of all acceptable functions. This relation h satisfies $\langle n,y\rangle \in f \leftrightarrow \exists v \in \mathcal{K} : \langle n,y\rangle \in v$.

It can be proven that f is itself an acceptable function, and moreover, its domain is ω, thus satisfying the conclusions of the theorem. This function can also be shown to be unique by considering the set

$$T = \{n \in \omega : f_1(n) = f_2(n)\}$$

and proving that it is inductive. ■

6.8.4 Classes & Class-Functions

What happens if we try to iterate the power set operation? That is, does this theorem on recursion prove the existence of a function f on ω for which $f(0) = \varnothing$ and $f(n^+) = \mathcal{P}(f(n))$? The range of such a function would be the collection $\big\{\varnothing,\mathcal{P}(\varnothing),\mathcal{P}\mathcal{P}(\varnothing),\mathcal{P}\mathcal{P},\mathcal{P}(\varnothing),\ldots\big\}$.

With our current axioms, we cannot prove the existence of this function, nor of this set. The problem is that the power set operation is *not* a function, as it is not a set – the power set operation may be applied to any set, so its domain would be the collection of all sets, which we know is not a set.

A *class* is a collection of sets satisfying a formula φ:

$$H = \{x : \varphi(x)\}$$

This is similar to the earlier unrestricted comprehension, but now, these collections do not a priori constitute sets, being only classes unless proved otherwise.

Example. The class V defined by

$$V = \{x : x = x\}$$

is the class of all sets, called the *universe*. △

A class that is not a set is called a *proper class*. For instance, the universe class is a proper set. A class that happens to be a set is sometimes called a *small class*.

We can now define more precisely what kind of operation $x \mapsto \mathcal{P}(x)$ is.

A *class-function* is a class F whose elements are ordered pairs and for every set x there is exactly one set y such that $\langle x,y \rangle \in F$. So, a class-function is an operation "definable with a formula". A class-function can thus be regarded as an operation $V \to V$, but it is not a function since it is not a set.

A class (and therefore a class-function) is defined by a formula $\varphi(x)$, which one designated free variable x, and possibly other unlisted parameters. As we have defined it, whenever $\varphi(x)$ holds for a class-function, x must be an ordered pair. This is inconvenient.

Instead, we will talk about class-functions using formulae $\varphi(x,y)$ such that for every set x, there is exactly one other set y such that $\varphi(x,y)$ holds.

Theorem 6.8.4. *There is a formula that is satisfied if and only if $\varphi(x,y)$ defines a class function.*

Proof.

$$\underbrace{\forall x \exists y : \varphi(x,y)}_{\text{existence}} \wedge \underbrace{\forall x \forall y \forall y' \Big(\big(\varphi(x,y) \wedge \varphi(x,y')\big) \to y = y' \Big)}_{\text{uniqueness}}$$

■

6.8.5 Axiom Schema of Replacement

The axiom schema of specification can be rephrased as saying that the "intersection" of any class with a set is a set. More generally, one can safely assume that every sufficiently small class is a set, and proper sets are those that contain "too many" elements.

If we take the "image" of a set under a class-function, intuitively we do not get a "bigger" thing (i.e. a class) than the set we started with. This is the motivation behind the next axiom.

Axiom Schema of Replacement.
The image of a set under a class-function is a set; if φ is any formula that does not contain B, then:

$$\forall A \Bigg(\underbrace{\forall x \forall y \forall y' \Big(\big(x \in A \wedge \varphi(x,y) \wedge \varphi(x,y')\big) \to y = y' \Big)}_{\varphi \text{ is a class-function on at least } A} \to \underbrace{\exists B \forall y \Big(y \in B \leftrightarrow \exists x (x \in A \wedge \varphi(x,y)) \Big)}_{\text{there is a set } B \text{ consisting of } \varphi\text{-images of elements of } A} \Bigg)$$

That is, if φ represents a definable function f, A represents its class domain, and $f(x)$ is a set for every $x \in A$, then the image of f is a subset of some set B.

Using the axiom schema of replacement, we can prove a more general recursion theorem that does not presuppose the existence of the set A.

Theorem (Recursion on ω, Class Form)**.** *Let a be any set and let $\varphi(x,y)$ define a class-function. Then, there is a set A and a unique function $h : \omega \to A$ such that $h(0) = a$ and $\varphi\big(h(n),h(n^+)\big)$ for every $n \in \omega$.*

Proof sketch. Call a function n-acceptable if it is acceptable and has domain n. First, we prove that for each $n \in \omega$, there exists an n-acceptable function, and furthermore, that any n-acceptable function and m-acceptable function agree on $n \cap m = \min(n,m)$. This implies there is a unique n-acceptable function for each $n \in \omega$.

Then, let $\varphi(x,y)$ be the formula that if $x \in \omega$, then y is x-acceptable, and otherwise $y = \varnothing$. This defines a class-function by the uniqueness proved above, and the axiom of replacement for φ implies the existence of the set $\mathcal{K}$ of all acceptable functions. Then, as before, $h = \bigcup \mathcal{K}$ satisfies the conclusions of the theorem. ■

Corollary 6.8.4.1. *There is a function h with domain ω for which $h(0) = \varnothing$ and $h(n^+) = \mathcal{P}\big(h(n)\big)$ for all $n \in \omega$.*

Proof. Apply the class form of the recursion theorem for the formula $\varphi(x,y) \equiv y = \mathcal{P}(x)$; and let $a = \varnothing$. ■

Corollary 6.8.4.2. *There is a set $\big\{\varnothing,\mathcal{P}(\varnothing),\mathcal{P}\mathcal{P}(\varnothing),\mathcal{P}\mathcal{P}\mathcal{P}(\varnothing),\dots\big\}$.*

Proof. This is the range of the function of the previous corollary. ■

6.8.6 Addition and Multiplication on ω

Theorem 6.8.5 (Parametric Recursion on ω)**.** *Let $f_0 : A \to B$ and $u : B \times A \to A$ be functions. Then, there exists a unique function $f : A \times \omega \to B$ such that*

- *$f(a,0) = f_0(a)$ for all $a \in A$;*
- *$f(a,n^+) = u\big(a,f(a,n)\big)$ for all $n \in \omega$ and $a \in A$.*

Corollary 6.8.5.1. *There is a unique function $+ : \omega \times \omega \to \omega$ such that*

- *$+(m,0) = m$ for all $m \in \omega$;*
- *$+(m,n^+) = \big(+(m,n)\big)^+$ for all $m,n \in \omega$.*

Proof. Let $A = B = \omega$, and let $f_0 : \omega \to \omega$ be the identity function, and $u : \omega \times \omega \to \omega$ be defined by $(a,n) \mapsto n^+$ in the previous theorem. ■

We will write this function using infix notation like with relations. Intuitively, this theorem states that the equations $m + 0 = m$ and $m + (n + 1) = (m + n) + 1$ uniquely characterise addition.

Corollary 6.8.5.2. *There is a unique function $\cdot : \omega \times \omega \to \omega$ such that*

- *$m \cdot 0 = 0$ for all $m \in \omega$;*
- *$m \cdot n^+ = m + m \cdot n$ for all $m,n \in \omega$.*

Proof. Let $A = B = \omega$, $f_0 : \omega \to \omega$ be the constant zero function, and $u : \omega \times \omega \to \omega$ be defined by $(a,n) \mapsto a + n$ in the previous theorem. ■

Theorem 6.8.6 (Basic Properties of ω)**.** *The following general properties all hold:*

(*i*) $\forall a,b \in \omega : a +_\omega b = b +_\omega a$ *(commutativity of addition);*

(*ii*) $\forall a,b,c \in \omega : (a +_\omega b) +_\omega c = a +_\omega (b +_\omega c)$ *(associativity of addition);*

(*iii*) $a +_\omega 0_\omega = a$ *(existence of additive identity);*

(*iv*) $a + 1 = a^+$ *(equivalence of successor and addition);*

(*v*) $\forall a,b \in \omega : a \cdot_\omega b = b \cdot_\omega a$ *(commutativity of multiplication);*

(*vi*) $\forall a,b,c \in \omega : (a \cdot_\omega b) \cdot_\omega c = a \cdot_\omega (b \cdot_\omega c)$ *(associativity of multiplication);*

(*vii*) $\forall a,b,c \in \omega : a \cdot_\omega (b +_\omega c) = a \cdot_\omega b +_\omega a \cdot_\omega c$ *(distributivity of multiplication over addition);*

(*viii*) $\forall a \in \omega : a \cdot_\omega 1_\omega = a$ *(existence of multiplicative identity);*

(*ix*) $0_\omega \neq 1_\omega$ *(non-degeneracy);*

(*x*) $\forall a,b \in \omega : a \cdot_\omega b = 0_\omega \to (a = 0_\omega \vee b = 0_\omega)$ *(zero divisors).*

Proof of (*i*). Fix $a \in \omega$ and define the set

$$S = \{b \in \omega : a + b = b + a\}$$

If $b = 0$, then $a + b = b + a$ holds by property (*iii*), so $0 \in S$.

Now assume $b \in S$. Then, by the definition of addition,

$$\begin{aligned} a + b^+ &= (a + b)^+ \\ &= (b + a)^+ \\ &= b + a^+ \end{aligned}$$

so $b^+ \in S$, and $S \subseteq \omega$ is inductive, so $S = \omega$.

The proofs for the other properties are similar. ■

Theorem 6.8.7. *For any natural numbers m,n,p*

$$m \in n \quad \leftrightarrow \quad m + p \in n + p$$

and if $p \neq 0$*, then also*

$$m \in n \quad \leftrightarrow \quad m \cdot p \in n \cdot p$$

Theorem 6.8.8. *For any natural numbers m,n,p*

$$m + p \in n + p \quad \to \quad m = n$$

and if $p \neq 0$*, then also*

$$m \cdot p = n \cdot p \quad \to \quad m = n$$

Theorem 6.8.9. *ω is well-ordered by $<_\omega$. That is, every non-empty subset $A \subseteq \omega$ has an $<_\omega$-minimal element.*

6.8.7 Equivalence Relations

A relation $R \subseteq A \times A$ is an *equivalence relation* on A if it is:

- reflexive: $\forall x \in A : xRx$;
- symmetric: $\forall x,y \in A : xRy \leftrightarrow yRx$;
- transitive: $\forall x,y,z \in A : (xRy \wedge yRz) \to xRz$.

We define the *equivalence class* $[x]_R$ of $x \in A$ under R as the set

$$[x]_R \coloneqq \{t : xRt\}$$

Note that this is indeed a set by specification, as $[x]_R \subseteq \operatorname{ran}(R) = A$.

Theorem 6.8.10. *Two elements are equivalent under R if and only if their equivalence classes are equal.*

Proof. Suppose xRy, and let $a \in [x]_R$. Then, by definition, xRa, so by symmetry and transitivity, aRy, so $a \in [y]_R$, and hence $[x]_R \subseteq [y]_R$. Now, let $b \in [y]_R$. Then, by definition, yRb, so by transitivity, xRb, so $b \in [x]_R$, and hence $[y]_R \subseteq [x]_R$. So, $[x]_R = [y]_R$.

For the reverse implication, suppose $[x]_R = [y]_R$. By reflexivity, $y \in [y]_R$, and since $[x]_R = [y]_R$, we also have $y \in [x]_R$, so xRy, as required. ■

Theorem 6.8.11. *Equivalence classes partition A. That is, the union of all equivalence classes is A, and their pairwise intersections are empty.*

Proof. By reflexivity, $x \in [x]_R$ for all x, so the union of all equivalence classes must be A.

Let $x,y \in A$ be distinct, and suppose $[x]_R \cap [y]_R$ is non-empty. Let $a \in [x]_R \cap [y]_R$, so $a \in [x]_R$ and $a \in [y]_R$. Then, by definition, xRa and yRa, so by the previous theorem, $[x]_R = [a]_R = [y]_R$. ■

6.8.8 Integers

We can represent natural numbers and the operations of addition and multiplication as sets. The next goal is to find an encoding of the integers, then of the rationals.

A rational number p/q may be expressed as a pair of integers, p and q – but this representation is not unique. For instance, $p/q = 2p/2q$. So, the rationals are really an equivalence relation of these representations on pairs of integers.

We take the same approach for constructing the integers: just a rational is an equivalence class of quotients of integers, an integer can be expressed as an equivalence class of differences of naturals. For instance,

$$-3 = 0 - 3 = 1 - 4 = 2 - 5 = 3 - 6 = \cdots$$

Define $\sim$ to be the equivalence relation on $\omega \times \omega$ for which $\langle a,b \rangle \sim \langle x,y \rangle$ if and only if $a + y = b + x$. The *set of integers* $\mathbb{Z}$ is defined to be the set of equivalence classes

$$\mathbb{Z} \coloneqq \omega \times \omega \big/ \sim$$

For instance, the integer $2_{\mathbb{Z}}$ is the equivalence class

$$2_{\mathbb{Z}} = \big[\langle 2,0 \rangle\big] = \big\{\langle 2,0 \rangle, \langle 3,1 \rangle, \langle 4,2 \rangle, \langle 5,3 \rangle, \dots\big\}$$

while the integer $-3_{\mathbb{Z}}$ is the equivalence class

$$-3_{\mathbb{Z}} = \big[\langle 0,3 \rangle\big] = \big\{\langle 0,3 \rangle, \langle 1,4 \rangle, \langle 2,5 \rangle, \langle 3,6 \rangle, \dots\big\}$$

Note that in this construction, ω is not a subset of $\mathbb{Z}$, and that

$$0_\omega = \varnothing \neq \big\{\langle n,n\rangle : n \in \omega\big\} = 0_\mathbb{Z}$$

How should we define addition? Informally, we have

$$(a-b)+(x-y)=(a+x)-(b+y)$$

Theorem (Addition on $\mathbb{Z}$)**.** *There is a unique function $+_\mathbb{Z} : \mathbb{Z}\times\mathbb{Z}\to\mathbb{Z}$ such that*

$$\big[\langle a,b\rangle\big] +_\mathbb{Z} \big[\langle x,y\rangle\big] = \big[\langle a +_\omega x, b +_\omega y\rangle\big]$$

Proof. We check that this operation is well-defined. Suppose $\langle a,b\rangle \sim \langle a',b'\rangle$ and $\langle x,y\rangle \sim \langle x',y'\rangle$, so $a+b'=b+a'$ and $x+y'=y+x'$. Adding these together, we have $a+b'+x+y'=b+a'+y+x'$. As $+_\omega$ is commutative, we have $(a+x)+(b'+y')=(b+y)+(a'+x')$, so

$$\langle a+x,b+y\rangle \sim \langle a'+x',b'+y'\rangle$$

as required. ■

For multiplication, informally, we have

$$(a-b)(x-y)=(ax+by)-(ay+bx)$$

Theorem (Multiplication on $\mathbb{Z}$)**.** *There is a unique function $\cdot_\mathbb{Z} : \mathbb{Z}\times\mathbb{Z}\to\mathbb{Z}$ such that*

$$\big[\langle a,b\rangle\big] \cdot_\mathbb{Z} \big[\langle x,y\rangle\big] = \big[\langle ax+by, ay+bx\rangle\big]$$

The ring unit and zero are then given by

$$\begin{aligned} 0_\mathbb{Z} &= \big[\langle 0,0\rangle\big] \\ 1_\mathbb{Z} &= \big[\langle 1,0\rangle\big] \end{aligned}$$

Theorem 6.8.12 (Basic Properties of $\mathbb{Z}$)**.** *Replacing $(+_\omega, \cdot_\omega, 0_\omega, 1_\omega)$ by $(+_\mathbb{Z}, \cdot_\mathbb{Z}, 0_\mathbb{Z}, 1_\mathbb{Z})$, the same results as in Theorem 6.8.6 for ω hold for $\mathbb{Z}$, with the addition of*

(xi) $\forall a \in \mathbb{Z} : \exists b \in \mathbb{Z} : a +_\mathbb{Z} b = 0_\mathbb{Z}$ *(existence of additive inverses).*

Although ω is not a true subset of $\mathbb{Z}$, there is a natural embedding $E : \omega \to \mathbb{Z}$ defined by $E(n) = \big[\langle n,0\rangle\big]$ such that $E(\omega)$ behaves like ω:

$$\begin{aligned} E(m +_\omega n) &= E(m) +_\mathbb{Z} E(n); \\ E(m \cdot_\omega n) &= E(m) \cdot_\mathbb{Z} E(n). \\ E(0_\omega) &= 0_\mathbb{Z}; \\ E(1_\omega) &= 1_\mathbb{Z}. \end{aligned}$$

(That is, E is a semiring homomorphism.) We also have

$$\langle m,n\rangle = E(m) - E(n)$$

6.8.9 Rationals

As mentioned previously, rationals can be expressed as the quotient of two integers (non-zero, in the case of the divisor). For instance,

$$\frac{1}{2} = \frac{-1}{-2} = \frac{2}{4} = \frac{-2}{-4} = \frac{3}{6} = \frac{-3}{-6} = \cdots$$

Let $\mathbb{Z}' := \mathbb{Z} \setminus \{0_{\mathbb{Z}}\}$ be the set of non-zero integers and define the equivalence relation $\sim$ on $\mathbb{Z} \times \mathbb{Z}'$ for which $\langle a,b\rangle \sim \langle x,y\rangle$ if and only if $a \cdot y = b \cdot x$. The *set of rationals* $\mathbb{Q}$ is the set of equivalence classes

$$\mathbb{Q} := \mathbb{Z} \times \mathbb{Z}' /_{\sim}$$

Informally, we have

$$\frac{a}{b} + \frac{x}{y} = \frac{ay + bx}{by}$$

so,

Theorem 6.8.13 (Addition on $\mathbb{Q}$)**.** *There is a unique function $+_{\mathbb{Q}} : \mathbb{Q} \times \mathbb{Q} \to \mathbb{Q}$ such that*

$$\big[\langle a,b\rangle\big] +_{\mathbb{Q}} \big[\langle x,y\rangle\big] = \big[\langle ay +_{\mathbb{Z}} bx, by\rangle\big]$$

for all $a,b \in \mathbb{Z}$ and $x,y \in \mathbb{Z}'$.

Similarly, for multiplication, we should have

$$\frac{a}{b} \cdot \frac{x}{y} = \frac{ax}{by}$$

so

Theorem 6.8.14 (Multiplication on $\mathbb{Q}$)**.** *There is a unique function $\cdot_{\mathbb{Q}} : \mathbb{Q} \times \mathbb{Q} \to \mathbb{Q}$ such that*

$$\big[\langle a,b\rangle\big] \cdot_{\mathbb{Q}} \big[\langle x,y\rangle\big] = \big[\langle ax, by\rangle\big]$$

for all $a,b \in \mathbb{Z}$ and $x,y \in \mathbb{Z}'$.

The unit and zero are then given by

$$\begin{aligned} 0_{\mathbb{Q}} &= \big[\langle 0,1\rangle\big] \\ 1_{\mathbb{Q}} &= \big[\langle 1,1\rangle\big] \end{aligned}$$

Again, $\mathbb{Z}$ is not a true subset of $\mathbb{Q}$, but there is a natural embedding $E : \mathbb{Z} \to \mathbb{Q}$ defined by $E(a) = \big[\langle a,1_{\mathbb{Z}}\rangle\big]$ such that $E(\mathbb{Z})$ behaves like $\mathbb{Z}$:

$$\begin{aligned} E(a +_{\mathbb{Z}} b) &= E(a) +_{\mathbb{Q}} E(b); \\ E(a \cdot_{\mathbb{Z}} b) &= E(a) \cdot_{\mathbb{Q}} E(b); \\ E(0_{\mathbb{Z}}) &= 0_{\mathbb{Q}}; \\ E(1_{\mathbb{Z}}) &= 1_{\mathbb{Q}}. \end{aligned}$$

(That is, E is a semiring homomorphism.)

Theorem 6.8.15 (Basic Properties of $\mathbb{Q}$)**.** *Replacing the relevant operations and constants by $(+_{\mathbb{Q}}, \cdot_{\mathbb{Q}}, 0_{\mathbb{Q}}, 1_{\mathbb{Q}})$, the same results as in Theorem 6.8.6 and Theorem 6.8.12 for ω and $\mathbb{Z}$ hold for $\mathbb{Q}$, with the addition of*

(xii) $\forall a \in \mathbb{Q} : \exists b \in \mathbb{Q} : a \cdot_{\mathbb{Q}} b = 1_{\mathbb{Q}}$ (existence of multiplicative inverses).

6.8.10 Real Numbers

There are many possible approaches to construct the real numbers.

- Decimal expansions: every real number can be expressed as an integer and an infinite sequence of digits (a function $\omega \to \{0,...,9\}$).

  Disadvantages: multiplication is messy to define; also, real numbers may have two distinct decimal expansions, so we need to use equivalence classes.

- Equivalence classes of Cauchy sequences of rationals.

  Advantage: defining addition and multiplication is very easy.

- Dedekind cuts.

  Advantages: defining addition is easy, as is proving the existence of suprema; also no equivalence classes needed.

We take the third approach here.

A *Dedekind cut* is a subset $X \subset \mathbb{Q}$ such that

(*i*) $\varnothing \neq X \neq \mathbb{Q}$;

(*ii*) X is *downward closed* – that is,

$$q \in X \wedge r < q \to r \in X$$

(*iii*) X has no largest member – that is,

$$\neg \exists m \in X : \forall x \in X : x < m$$

(A Dedekind cut may also alternatively be defined to be the pair $\langle X, \mathbb{Q} \setminus X \rangle$, but X alone completely determines the pair, so no information is lost by just considering the first component.)

The *set of real numbers* $\mathbb{R}$ is the set of all Dedekind cuts.

We define the relation $<_{\mathbb{R}}$ on $\mathbb{R}$ by $x <_{\mathbb{R}} y$ if and only if $x \subset y$, and the relation $\leq_{\mathbb{R}}$ by $x \leq_{\mathbb{R}} y$ if and only if $x \subseteq y$.

Theorem 6.8.16. *The relation $<_{\mathbb{R}}$ is a linear ordering on $\mathbb{R}$.*

Proof. Irreflexivity and transitivity follows from that of the strict subset relation. Obviously, at most one of

$$x \subset y, \qquad x = y, \qquad y \subset x$$

can hold. To show at least one holds, suppose the first two cases fail, so $x \nsubseteq y$.

Since $x \nsubseteq y$, there exists a rational $r \in x \setminus y$. Now, let $q \in y$. If $q \geq r$, then $r \in y$, as y is downward closed, but $r \notin y$ by definition, so $q < r$. Since x is downward closed, $q \in x$. So, $y \subset x$. ■

6.8.10.1 Bounds

- A number $u \in \mathbb{R}$ is an *upper bound* of a set $A \subseteq \mathbb{R}$ if $a \leq_{\mathbb{R}} u$ for all $a \in A$.
- The set $A \subseteq \mathbb{R}$ is *bounded from above* if there exists an upper bound of A.
- A *least upper bound* is an upper bound less than any other upper bound.

Theorem 6.8.17. *Any non-empty subset of $\mathbb{R}$ that is bounded from above has a least upper bound.*

Proof. Let $A \subseteq \mathbb{R}$ be non-empty and bounded from above. We claim that $\bigcup A$ is a Dedekind cut, and is the least upper bound of A.

Since A is a collection of Dedekind cuts, which are sets of rational numbers, we have $\bigcup A \subseteq \mathbb{Q}$. Since A is non-empty, $\bigcup A \neq \varnothing$, and since A is bounded above, there exists an upper bound, say u, so $u+1 \notin A$, and $A \neq \mathbb{Q}$.

Let $a \in \bigcup A$ and $r \in \mathbb{Q}$ such that $r < a$. Because $a \in \bigcup A$, there exists a Dedekind cut $x \in A$ such that $a \in x$. Because x is a Dedekind cut, it is downwards closed, so $r \in x$, and hence $r \in A$. So $\bigcup A$ is downward closed.

Now, let $m \in \bigcup A$, so there exists $x \in A$ such that $m \in x$. If m is a largest element of $\bigcup A$, then it would also be the largest element of x. But x is a Dedekind cut, which has no largest element.

Hence, $\bigcup A$ is a Dedekind cut. For all $x \in A$, we have $x \subseteq \bigcup A$ (that is, $x \leq_{\mathbb{R}} \bigcup A$), so $\bigcup A$ is an upper bound of A. Now, let z be any upper bound of A, so $x \leq_{\mathbb{R}} z$ $(x \subseteq z)$ for every $x \in A$. Then, $\bigcup A \subseteq z$ $(\bigcup A \leq_{\mathbb{R}} z)$, so $\bigcup A$ is the least upper bound. ■

For any $x, y \in \mathbb{R}$, we define the set

$$x +_{\mathbb{R}} y = \{p + q \in \mathbb{Q} : q \in x, r \in y\}$$

This coincides with our usual idea of addition, since, for example, if

$$\begin{aligned} x &= \{q \in \mathbb{Q} : q < 1\} \\ y &= \{q \in \mathbb{Q} : q < 3\} \end{aligned}$$

then

$$\begin{aligned} x +_{\mathbb{R}} y &= \{a + b \in \mathbb{Q} : a \in x, b \in y\} \\ &= \{a + b \in \mathbb{Q} : a < 1, b < 3\} \\ &= \{q \in \mathbb{Q} : q < 4\} \end{aligned}$$

Lemma 6.8.18. *If $x, y \in \mathbb{R}$, then $x +_{\mathbb{R}} y \in \mathbb{R}$.*

Proof. Clearly, $x +_{\mathbb{R}} y \subseteq \mathbb{Q}$. If $a, b \in \mathbb{Q}$ such that $a \notin x$ and $b \notin y$, then for every $p \in x$ and $q \in y$, $p < a$ and $q < b$, so every element $(p+q) \in x +_{\mathbb{R}} y$ is less than $a + b$, so $a + b \notin x +_{\mathbb{R}} y$, giving $x +_{\mathbb{R}} y \neq \mathbb{Q}$.

Let $a < (p + q) \in x +_{\mathbb{R}} y$. Then, $a - q < p$, so $(a - q) \in x$ by downward-closedness of x as a Dedekind cut. Then, $a = (a - q) + q \in x +_{\mathbb{R}} y$, so $x +_{\mathbb{R}} y$ is downward closed.

Suppose $p + q \in x +_{\mathbb{R}} y$ is the largest element. As x is a Dedekind cut, p is not the largest element in x, so there exists a larger element $p < p' \in x$. Then, $p + q < p' + q \in x +_{\mathbb{R}} y$, contradicting that $p + q$ was the largest.

Thus, $x +_{\mathbb{R}} y$ is a Dedekind cut. ■

Theorem 6.8.19 (Basic Properties of $\mathbb{R}$)**.** *Replacing the relevant operations and constants by $(+_{\mathbb{R}}, \cdot_{\mathbb{R}}, 0_{\mathbb{R}}, 1_{\mathbb{R}})$, the same results as in Theorem 6.8.6, Theorem 6.8.12, and Theorem 6.8.15 for ω, $\mathbb{Z}$, and $\mathbb{Q}$ hold for $\mathbb{R}$.*

Again, $\mathbb{Q}$ is not a true subset of $\mathbb{R}$, but there is a natural embedding $E : \mathbb{Q} \to \mathbb{R}$ defined by $E(r) = \{q \in Q : q < r\}$ such that $E(\mathbb{Q})$ behaves like $\mathbb{Q}$:

$$E(a +_{\mathbb{Q}} b) = E(a) +_{\mathbb{R}} E(b);$$

$$\begin{aligned} E(a \cdot_{\mathbb{Q}} b) &= E(a) \cdot_{\mathbb{R}} E(b); \\ E(a) <_{\mathbb{R}} E(b) &\leftrightarrow a <_{\mathbb{Q}} b; \\ E(0_{\mathbb{Q}}) &= 0_{\mathbb{R}}; \\ E(1_{\mathbb{Q}}) &= 1_{\mathbb{R}}. \end{aligned}$$

(That is, E is a order semiring homomorphism.)

6.8.11 Complex Numbers

A complex number consists of a real part, and an imaginary part, which is just a real number scaling the imaginary unit. As such, complex numbers can easily be represented using ordered pairs: $\mathbb{C} = \mathbb{R} \times \mathbb{R}$, with addition and multiplication defined as usual:

$$\begin{aligned} \langle a,b\rangle +_{\mathbb{C}} \langle c,d\rangle &= \langle a+c,b+d\rangle \\ \langle a,b\rangle \cdot_{\mathbb{C}} \langle c,d\rangle &= \langle ac-bd,ac+bd\rangle \end{aligned}$$

Again, $\mathbb{R}$ is not a true subset of $\mathbb{C}$ – for instance, $1_{\mathbb{C}} = \langle 1_{\mathbb{R}},0_{\mathbb{R}}\rangle$ – but, as usual, there is a natural embedding that lifts one to the other.

6.9 Cardinality

Informally, the cardinality of a set means the "size" of that set, in the sense of how many elements it has.

Two sets A and B are *equinumerous* if there exists a bijection between A and B, and we denote this relation by $A \sim B$.

Equinumerosity is an "equivalence relation" since,

- For all A, $A \sim A$ via the identity on A;
- For all A,B, $A \sim B \to B \sim A$, as a bijection between A and B is also a bijection between B and A;
- For all A,B,C, $(A \sim B \land B \sim C) \to A \sim B$ via composition.

However, equinumerosity is not a relation in the sense that it is not a set: it is a proper class, consisting of pairs of sets that are equinumerous.

We say that two sets A and B have the *same cardinality*, written as $|A| = |B|$, if they are equinumerous.

For comparing these cardinalities, we have several options. We could say that a non-empty set A is "smaller than" (or is "at most as large as") another set B if:

(*i*) A is equinumerous to a subset of B. That is, $A \sim B_0 \subseteq B$;

(*ii*) There exists an injection $A \rightarrowtail B$;

(*iii*) There exists an surjection $A \twoheadrightarrow B$.

The first two are equivalent, and they also imply the third, but the third only implies the first two with the axiom of choice.

We say that the cardinality of A is at most the cardinality of B, written as $|A| \leq |B|$ if there is an injective function from A to B.

Theorem 6.9.1. *If* $|A| \leq |B|$ *and* $|B| \leq |C|$*, then* $|A| \leq |C|$.

Proof. Let $f : A \to B$ and $g : B \to C$ be injective. Then, $g \circ f : A \to C$ is injective. ■

Theorem 6.9.2 (Cantor-Bernstein)**.** *Let A and B be sets. If there is an injection $f : A \to B$ and an injection $g : B \to A$, then there is a bijection between A and B. That is, if $|A| \leq |B|$ and $|B| \leq |A|$, then $|A| = |B|$.*

Proof (König, 1906). Without loss of generality that A and B are disjoint (any elements in the intersection can be paired with their copy in the other set, and hence ignored).

For any $a \in A$, we may consider its orbit,

$$\cdots \to f^{-1}\big(g^{-1}(a)\big) \to g^{-1}(a) \to a \to f(a) \to g\big(f(a)\big) \to \cdots$$

This sequence may terminate at some point to the left, if f^{-1} or g^{-1} is not defined.

Because f and g are injective, each $a \in A$ and $b \in B$ is in exactly one such sequence, since if an element $a \in A$ occurs in two sequences, the following and preceding elements are just functions applied to that element, so the two sequences must agree. Therefore, the sequences partition the disjoint union of A and B, so it is sufficient to give a bijection between the elements of A and B in each sequence separately.

Call a sequence A*-terminating* if it terminates to the left because g^{-1} cannot be taken at a certain point. That is, the sequence begins with an element in A. Define B*-terminating* sequences similarly.

Then, for an A-terminating sequence, f is a bijection between its A-elements and its B-elements, and similarly, for a B-terminating sequence, g is a bijection between its A-elements and its B-elements. For a doubly infinite or cyclic sequence, both f and g provide bijections.

To explicitly give a bijection $A \to B$, define the set $A_0 = A \setminus g(B)$. Note that every A-terminating sequence starts with an element in A_0, or else g^{-1} would be defined for its starting element.

Then, recursively define $A_n = g(f(A_{n-1}))$, and define A' to be their union:

$$\begin{aligned} A' &= \bigcup_{n\in\mathbb{N}} A_n \\ &= A_0 \cup g(f(A_0)) \cup g(f(g(f(A_0)))) \cup \cdots \end{aligned}$$

That is, A' is the orbit of A_0 under $g \circ f$.

Then,

$$h(x) := \begin{cases} f(x) & x \in A' \\ g^{-1}(x) & x \notin A' \end{cases}$$

is a bijection $A \to B$. ■

Corollary 6.9.2.1. $[0,1] \sim [0,1)$. *That is, the closed unit interval* $[0,1]$ *is equinumerous with the half-open unit interval* $[0,1)$.

Proof. The functions $f : [0,1] \to [0,1)$ and $g : [0,1) \to [0,1]$ defined by $f(x) = x/2$ and $g(x) = x$ are injections. Then, we have,

$$\begin{aligned} A_0 &= [0,1] \setminus g\big([0,1)\big) \\ &= [0,1] \setminus [0,1) \\ &= \{1\} \\ A_n &= \left\{\frac{1}{2^n}\right\} \\ A' &= \bigcup_{n\in\mathbb{N}} A_n \end{aligned}$$

$$= \left\{ \frac{1}{2^n} : n \in \mathbb{N} \right\}$$

So

$$h(x) := \begin{cases} \frac{x}{2} & x \in A' \\ x & x \notin A' \end{cases}$$

is a bijection [0,1] → [0,1). ■

We recall Cantor's theorem and state a new corollary related to cardinality.

Theorem (Cantor)**.** *For any set A, there is no surjection $f : A \to \mathcal{P}(A)$.*

Corollary 6.9.2.2 (Cantor)**.** *For every set A, we have $|A| < |\mathcal{P}(A)|$.*

Proof. The function $f : A \to \mathcal{P}(A)$ defined by $a \mapsto \{a\}$ is an injection, so $|A| \leq |\mathcal{P}(A)|$, but Cantor's theorem states that there is no surjection $A \to \mathcal{P}(A)$, which implies that no bijection $A \to \mathcal{P}(A)$ may exist, so $|A| \neq |\mathcal{P}(A)|$. ■

For any sets A and B, is it true that at least one of $|A| \leq |B|$ and $|B| \leq |A|$ holds? This is surprisingly non-trivial, and is in fact equivalent to the axiom of choice. We prove this later.

6.9.1 Finite Sets

How do we determine if a set is finite or not? We cannot directly write a first order formula that states that a set X has finitely many elements. Fortunately, in our definition of the natural numbers, each set $n \in \omega$ is defined to be the set $\{0,1,\ldots,n-1\}$, so it "has n elements".

A set is *finite* if it is equinumerous to a natural number. This can be written as a formula for any set X as

$$\begin{aligned} &\exists n \exists f \Big(\underbrace{n \in \omega \land f \subseteq X \times n}_{f \text{ is a function } X \to n} \\ &\quad \land \Big(\underbrace{\forall x \in X : \forall y \in X : \forall a \in n\big((\langle x,a\rangle \in f \land \langle y,a\rangle \in f) \to (\langle x,a\rangle = \langle y,a\rangle \to x = y)\big)\Big)}_{f \text{ is injective}} \\ &\quad \land \Big(\underbrace{\forall m : m \in n \to \exists x(x \in X \land \langle x,m\rangle \in f)}_{f \text{ is surjective}}\Big)\Big) \end{aligned}$$

A finite set X *has cardinality* n or *has* n *elements* if there is a bijection from X to n, and we write $|X| = n$ in this case.

Theorem (Pigeonhole Principle)**.** *No natural number is equinumerous to a proper subset of itself.*

Lemma 6.9.3. *Let X be finite. Then, there exists a unique $n \in \omega$ such that $|X| = n$.*

Proof. By the definition of finiteness, there exists at least one such natural number. For uniqueness, suppose $|X| = n$ and $|X| = m$, so there exist bijections $f : X \to n$ and $g : X \to n$. Suppose further that $m < n$. Then $g \circ f^{-1}$ is a bijection $n \to m$, so n is equinumerous to a proper subset of itself, contradicting the pigeonhole principle. ■

Corollary 6.9.3.1. *No finite set is equinumerous to a proper subset of itself.*

Proof. Let X be finite, so there exists a unique $n \in \omega$ such that $|X| = n$, so there is a bijection $f : X \to n$. Let $Y \subset X$ be a proper subset, and suppose that $X \sim Y$ (X and Y are equinumerous), so there exists a bijection $g : X \to Y$. Then, $f \circ g \circ f^{-1}$ is a bijection $n \to f(Y) \subsetneq n$, contradicting the pigeonhole principle. ∎

A set is *infinite* if it is not finite. Note that if X is finite and $Y \sim X$, then Y is also finite. Similarly, if X is infinite, and $Y \sim X$, then Y is also infinite.

Theorem 6.9.4. *ω is infinite.*

Proof. Let $s : \omega \to \omega$ be the function $s(n) = n^+$. Then, $\operatorname{ran}(s) = \omega \setminus \{0\} \subset \omega$, so ω is equinumerous to a proper subset of itself, so ω is not finite. ∎

6.9.1.1 Dedekind Finiteness

There are other possible notions of finiteness.

A set X is *Dedekind finite* if no proper subset of X is equinumerous to X.

The pigeonhole principle implies that every finite set is Dedekind finite, but is the converse true? Yes, but this direction requires the axiom of choice.

Theorem 6.9.5.

(*i*) *Finite sets are Dedekind finite.*

(*ii*) *(AC) Dedekind finite sets are finite.*

Proof.

(*i*) Follows from the pigeonhole principle.

(*ii*) (*Proof sketch.*) Let A be an infinite set, and let $a_0 \in A$. Then, choose $a_1 \in A\backslash\{a_0\}$, $a_2 \in A\backslash\{a_0,a_1\}$, and so on. Since A is infinite, this process can continue forever.

Now define a function $f : A \to A$ such that $f(a_n) = a_{n+1}$, and $f(a) = a$ for any $a \notin \{a_n\}_{n\in\mathbb{N}}$. Then, f is injective, but not surjective as $a_0 \notin \operatorname{ran}(f)$, so A is equinumerous to $A \setminus \{a_0\} \subset A$.

∎

6.9.2 Countability

A set X is *countable* if there is an injective function $f : X \to \omega$.

Clearly, finite sets are countable.

Because injections imply surjective inverses (Theorem 6.7.6), equivalently, a set X is countable if is empty or if there is a surjection from ω to X. We have a converse theorem (Theorem 6.7.7), so the previous implication is actually biconditional, but this theorem requires the axiom of choice.

However, it turns out that we can prove this result for ω without the axiom of choice.

Theorem 6.9.6. *A set X is countable if and only if there is a surjection $g : \omega \to X$, or if X is empty.*

Proof. Suppose X is countable, so there exists an injective function $f : X \to \omega$, so by Theorem 6.7.6, there exists a surjection $g : \omega \to X$, or X is empty.

If X is empty, then the unique empty function $(X = \varnothing) \to \omega$ is vacuously injective, so X is countable. Suppose otherwise that there is a surjection $g : \omega \to X$. Define $f : X \to \omega$ by

$$f(x) = \min\Big(g^{-1}[\{x\}]\Big) \subseteq \omega$$

The set $g^{-1}\big[\{x\}\big]$ is non-empty as g is surjective, and the minimum element exists as ω is well-ordered (Theorem 6.8.9). The preimages are also disjoint for distinct inputs (or else anything in the intersection has multiple images under g), so f is injective, as required. ■

The existence of such a surjection is often phrased as "the elements of X can be listed in a sequence", since a sequence in X is just a function $\omega \to X$.

Example. The sets $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{N} \times \mathbb{N}$ are countable.

By the above theorem, it is enough to enumerate the elements of each set in a sequence:

- $\mathbb{N}$: $0,1,2,3,\ldots$
- $\mathbb{Z}$: $0,1,-1,2,-2,3,-3,\ldots$
- $\mathbb{Q}$: A listing of all the positive rationals was famously given by Cantor:

$$\begin{array}{cccccccccc}
\frac{1}{1} & \to & \frac{1}{2} & & \frac{1}{3} & \to & \frac{1}{4} & & \frac{1}{5} & \to \\
 & \swarrow & & \nearrow & & \swarrow & & \nearrow & & \cdots \\
\frac{2}{1} & & \frac{2}{2} & & \frac{2}{3} & & \frac{2}{4} & & \frac{2}{5} & \cdots \\
\downarrow & \nearrow & & \swarrow & & \nearrow & & & & \cdots \\
\frac{3}{1} & & \frac{3}{2} & & \frac{3}{3} & & \frac{3}{4} & & \frac{3}{5} & \cdots \\
 & \swarrow & & \nearrow & & & & & & \cdots \\
\frac{4}{1} & & \frac{4}{2} & & \frac{4}{3} & & \frac{4}{4} & & \frac{4}{5} & \cdots \\
\downarrow & \nearrow & & & & & & & & \cdots \\
\frac{5}{1} & & \frac{5}{2} & & \frac{5}{3} & & \frac{5}{4} & & \frac{5}{5} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}$$

Cantor's zig-zag argument

This lists all the positive rationals, but to include the negative rationals, we can interleave them like with the integers (and also add 0 to the beginning of the list).

- $\mathbb{N} \times \mathbb{N}$: Like the grid above, but listing (p,q) instead of $\frac{p}{q}$.

△

To distinguish between finite sets and infinite countable sets, we call a set *countably infinite* if it is infinite and countable.

Lemma 6.9.7. *Let $A \subseteq \omega$ be infinite. Then, there is a bijection $f : A \to \omega$*

Proof sketch. Intuitively, we match the smallest element of A with 0, the second smallest with 1, and so on. This does not require the axiom of choice, since ω (and hence A) is well-ordered, so no arbitrary choice functions are required, and the recursion theorem then gives the required function. ■

Theorem 6.9.8. *A set X is countably infinite if and only if there is a bijection between X and ω.*

Proof. Suppose there is a bijection $f : X \to \omega$, so $X \sim \omega$. Since ω is infinite, X is infinite, and since f is injective, X is countable.

Now suppose that X is countably infinite. Since X is countable, there is an injection $f : X \to \omega$. Consider $A = f(X)$. Since $f : X \to A$ is a bijection, $X \sim A$, and since X is infinite, $A \subseteq \omega$ is infinite, so there is a bijection $h : A \to \omega$ by the previous lemma. The composition $h \circ f$ is then a bijection between X and ω. ■

If X is countably infinite, then we say that its cardinality is $|X| = \aleph_0$.

Corollary 6.9.8.1. *If X and Y are countably infinite, then $X \sim Y$. That is, there exists a bijection XY.*

Proof. By the above theorem, $X \sim \omega$ and $Y \sim \omega$, so $X \sim Y$. ■

Theorem 6.9.9. *If A and B are countable, then*

(*i*) *$A \cup B$ is countable;*

(*ii*) *$A \times B$ is countable.*

Proof.

(*i*) If A or B are empty, then the union is just one of the sets, which is countable, so suppose otherwise that A and B are both non-empty.

Enumerate A as $a_1,a_2,\ldots$ and B as $b_1,b_2,\ldots$, possibly listing each element multiple times. Then, $A \cup B$ may be enumerated as $a_1,b_1,a_2,b_2,\ldots$.

(*ii*) If A or B are empty, then the product is empty, which is countable, so suppose otherwise that A and B are both non-empty.

As A and B are countable, there exist surjections $f : \mathbb{N} \to A$ and $g : \mathbb{N} \to B$. Define $h : \mathbb{N} \times \mathbb{N} \to A \times B$ by

$$h\big(\langle a,b\rangle\big) = \big\langle f(a),g(b)\big\rangle$$

■

By induction, the union and product of finitely many countable sets is countable.

Is it true that the union of *countable many* countable sets is countable? Yes, but surprisingly, this requires the axiom of choice.

Theorem 6.9.10. *(AC) The union of countably many countable sets is countable. That is, if X is countable and $A \in X$ is countable for all A, then $\bigcup X$ is countable.*

Proof sketch. On a grid, enumerate each A_n on a horizontal line. Then, Cantor's zig-zag argument applies.

Slightly more formally, without loss of generality assume that $\varnothing \notin X$, and fix a surjection $f : \mathbb{N} \to X$. Each $A = f(n)$ is a non-empty countable set, so choose a surjection $g_n : \mathbb{N} \to A$, and let $h : \mathbb{N} \times \mathbb{N} \to \bigcup X$ be defined by

$$h(n,m) = g_n(m)$$

This function is a surjection as f and g_n are surjections.

Now, let $i : \mathbb{N} \to \mathbb{N} \times \mathbb{N}$ be a surjection. Then $h \circ i : \mathbb{N} \to \bigcup X$ is a surjection, so $\bigcup X$ is countable. ■

In more detail, the axiom of choice is required when choosing g_n:

Define R to be the set of ordered pairs $\langle n,g \rangle$ $n \in \mathbb{N}$ such that g is a surjection $\mathbb{N} \to f(n)$. Because $f(n)$ is countable for all n, such a surjection always exists, so every n is in the domain of R, so $\text{dom}(R) = \mathbb{N}$.

By the first version of the axiom of choice, there is a function $G \subseteq R$ with the same domain, $\mathbb{N}$. That is, for each n, there is a unique g_n such that $\langle n,g_n \rangle \in G$. This gives the surjection from $\mathbb{N}$ to $f(n)$, and the proof from this point is identical to the one above.

A set is *uncountable* if it is not countable, and we write $|X| > \aleph_0$.

6.9.3 Continuum

In set theory, *the continuum* refers to the size of $\mathbb{R}$.

Theorem 6.9.11 (Cantor's Diagonal Argument)**.** $\mathbb{R}$ *is not countable.*

Proof. This is Cantor's original diagonal argument.

Every real number can be uniquely expressed in base 10 as a series

$$d_0.d_1d_2d_3\ldots = n + \sum_{i=0}^{\infty} \frac{d_i}{10^i}$$

where d_0 is an integer, and d_i, $i > 0$, is a digit from 0 to 9, and the sequence (d_i) is not eventually all 9s.

Suppose that $\mathbb{R}$ is countable, so there is a list containing all real numbers

$$d^i_0.d^i_1d^i_2d^i_3\ldots$$

Let $r = d'_0.d'_1d'_2d'_3\ldots$ where

$$d'_i = \begin{cases} 1 & d_i = 0 \\ 0 & d_i \neq 0 \end{cases}$$

Then r is a real number not listed, as it differs from the ith real number at the ith digit. ■

Or, as a table, the list of real numbers is given by:

i	d_0	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	$\cdots$
0	0	1	2	3	4	5	6	7	8	9	$\cdots$
1	1	4	1	4	2	1	3	5	6	2	$\cdots$
2	3	1	4	1	5	9	2	6	5	3	$\cdots$
3	1	3	7	0	3	5	9	9	0	8	$\cdots$
4	1	6	1	8	0	3	3	9	8	8	$\cdots$
5	0	1	1	2	3	5	8	3	1	4	$\cdots$
6	0	1	4	2	8	5	7	1	4	2	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

and the new real number can be generated by examining the diagonal entries of the table, giving this proof its name:

i	d_0	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	$\cdots$
0	0	1	2	3	4	5	6	7	8	9	$\cdots$
1	1	4	1	4	2	1	3	5	6	2	$\cdots$
2	3	1	4	1	5	9	2	6	5	3	$\cdots$
3	1	3	7	0	3	5	9	9	0	8	$\cdots$
4	1	6	1	8	0	3	3	9	8	8	$\cdots$
5	0	1	1	2	3	5	8	3	1	4	$\cdots$
6	0	1	4	2	8	5	7	1	4	2	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$
?	1	0	0	1	1	0	$\cdots$				

If $A \sim \mathbb{R}$, then we say that A has cardinality of the continuum, and we write $|A| = \mathfrak{c}$. Cantor's theorem above can then be written as $|\mathbb{R}| = \mathfrak{c} > \aleph_0$.

Theorem 6.9.12. $\mathbb{R} \sim (0{,}1)$. *That is, there is a bijection between* $\mathbb{R}$ *and the open unit interval* $(0{,}1)$.

Proof. The logistic function $\sigma : \mathbb{R} \to (0{,}1)$

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

with inverse $f(x) = -\ln(\frac{1}{x-1})$ provides the required bijection. ■

Lemma 6.9.13. $[0{,}1) \sim (0{,}1)$. *That is, the half-open unit interval* $[0{,}1]$ *is equinumerous with the open unit interval* $(0{,}1)$.

Proof. We have previously proved that $[0{,}1] \sim [0{,}1)$ (Corollary 6.9.2.1). The bijection here is similar:

$$f(x) = \begin{cases} \frac{1}{2} & x = 0 \\ \frac{1}{2^{n+1}} & x = \frac{1}{2^n} \\ x & \text{otherwise} \end{cases}$$

■

So, we have $[0{,}1] \sim [0{,}1) \sim (0{,}1) \sim \mathbb{R}$, so these sets all have the same cardinality. Combining with scaling transformations, this implies that all non-trivial intervals of $\mathbb{R}$ are equinumerous to $\mathbb{R}$ and thus have the cardinality of the continuum.

Corollary 6.9.13.1. *If* $S \subseteq \mathbb{R}$ *contains an open interval, then* $|S| = \mathfrak{c}$.

Proof. A suitable scaling map map injects $(0{,}1)$ into the open interval in S, so there is an injection from $\mathbb{R} \sim (0{,}1)$ into S. Also, there is an injection from S into $\mathbb{R}$ given by the inclusion map. The Cantor-Bernstein theorem then implies $|S| = |\mathbb{R}| = \mathfrak{c}$. ■

Theorem 6.9.14. $[0{,}1] \sim [0{,}1]^2$

Corollary 6.9.14.1. $|\mathbb{R}^2| = \mathfrak{c}$

So far, we know the cardinalities $0,1,2,\dots,\aleph_0,\mathfrak{c}$, and Cantor's theorem implies the existence of infinitely many infinite cardinalities given by iterated power sets:

$$|\mathbb{N}| < |\mathcal{P}(\mathbb{N})| < |\mathcal{P}\mathcal{P}(\mathbb{N})| < |\mathcal{P}\mathcal{P}\mathcal{P}(\mathbb{N})| < \dots$$

Where does $\mathfrak{c}$ lie in this infinite chain?

Theorem 6.9.15. $|\mathcal{P}(\mathbb{N})| = |\mathbb{R}|$

Proof. For each subset $A \subseteq \mathbb{N}$, define x_A to be the real number whose decimal expansion is

$$0.d_1d_2d_3\dots$$

where

$$d_n = \begin{cases} 1 & n \in A \\ 0 & n \notin A \end{cases}$$

The function $A \mapsto x_A$ injects $\mathcal{P}(\mathbb{N})$ into $\mathbb{R}$, so $|\mathcal{P}(\mathbb{N})| \leq |\mathbb{R}|$.

Now, given $x \in (0,1) \subset \mathbb{R}$, write x in its unique binary expansion

$$0.b_0b_1b_2\dots$$

where the sequence (b_i) is not eventually all 1s, and define the set

$$A_x \coloneqq \{n \in \mathbb{N} : x_n = 1\}$$

Then, $x \mapsto A_x$ injects $(0,1) \sim \mathbb{R}$ into $\mathcal{P}(\mathbb{N})$, so $|\mathbb{R}| = |(0,1)| \leq |\mathbb{P}(\mathbb{N})|$. ■

6.9.3.1 Transcedental Numbers

A real number is *algebraic* if it is the root of a polynomial with integer (or rational) coefficients. A real number is *transcendental* if it is not algebraic.

It is very difficult to construct explicit examples of transcendental numbers, but surprisingly, most real numbers are transcendental:

Theorem 6.9.16. *The set of algebraic numbers is countable.*

Proof. A polynomial $p \in \mathbb{Z}[x]$ of degree n has the form

$$p(x) = a_nx^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0$$

with $a_n,\dots,a_0 \in \mathbb{Z}$. Define the *height* of p to be the natural number

$$n + \sum_{i=0}^{n} |a_i|$$

Note that there are only finitely many polynomials with any given fixed height, since there are finitely many integer partitions of any natural number. Every non-zero non-constant polynomial also has at most n (i.e. finitely many) roots. So, for each height $h \in \mathbb{N}$, there are finitely many roots of polynomials of height h.

Therefore, these roots may be enumerated by enumerating the roots of polynomials of height 1, then height 2, etc. ■

Corollary 6.9.16.1. *There exist transcendental numbers. Furthermore, the set of transcendental numbers is not countable.*

Proof. Let A be the set of algebraic numbers and $T = \mathbb{R} \setminus A$ be the set of transcendental numbers. If T were countable, then $A \cup T = \mathbb{R}$ is countable as the union of two countable sets. But $\mathbb{R}$ is not countable. ■

6.9.4 Cardinal Arithmetic

If κ and λ are cardinalities, then we define $\kappa + \lambda$ to be the cardinality of any set $S = A \cup B$ where $|A| = \kappa$, $|B| = \lambda$, and $A \cap B = \varnothing$.

For natural numbers (i.e. elements of ω), this definition is equivalent to the previous definition of addition, $+_\omega$.

If κ and λ are cardinalities, then we define $\kappa \cdot \lambda$ to be the cardinality of any set $S = A \times B$ where $|A| = \kappa$, $|B| = \lambda$.

Example.

- $1 + \aleph_0 = \aleph_0$, given by $f : \{-1\} \cup \mathbb{N} \to \mathbb{N} : x \mapsto x + 1$.
- $\aleph_0 + \aleph_0 = \aleph_0$: $\mathbb{N} \sim E := \{n : \exists k \in \mathbb{N} : n = 2k\}$ and $\mathbb{N} \sim O := \{n : \exists k \in \mathbb{N} : n = 2k + 1\}$ via $x \mapsto 2x$ and $x \mapsto 2x + 1$, respectively, and $E \cup O = \mathbb{N}$.
- $2 \cdot \aleph_0 = \aleph_0$: $\mathbb{N} \times \mathbb{N}$ is countably infinite, and hence bijects to ω.
- $\mathfrak{c} \cdot \mathfrak{c} = \mathfrak{c}$.
- $\aleph_0 \cdot \mathfrak{c} = \mathfrak{c}$.

△

Theorem 6.9.17 (Basic Properties of Cardinal Arithmetic)**.** *The following all hold:*

(i) $\forall \kappa, \lambda : \kappa + \lambda = \lambda + \kappa$ *(commutativity of cardinal addition);*

(ii) $\forall \kappa, \lambda, \mu : (\kappa + \lambda) + \mu = \kappa + (\lambda + \mu)$ *(associativity of cardinal addition);*

(iii) $\forall \kappa : \kappa + 0 = \kappa$ *(existence of additive identity);*

(iv) $\forall \kappa, \lambda : \kappa \cdot \lambda = \lambda \cdot \kappa$ *(commutativity of multiplication);*

(v) $\forall \kappa, \lambda, \mu : (\kappa \cdot \lambda) \cdot \mu = \kappa \cdot (\lambda \cdot \mu)$ *(associativity of multiplication);*

(vi) $\forall \kappa : \kappa \cdot 1 = \kappa$ *(existence of multiplicative identity);*

(vii) $\forall \kappa, \lambda, \mu : \kappa \cdot (\lambda + \mu) = \kappa \cdot \lambda + \kappa \cdot \mu$ *(distributivity of multiplication over addition);*

(viii) $\forall \kappa, \kappa', \lambda, \lambda$ *if* $\kappa \leq \kappa'$ *and* $\lambda \leq \lambda'$*, then* $\kappa + \lambda \leq \kappa' + \lambda'$ *and* $\kappa \cdot \lambda \leq \kappa' \cdot \lambda'$ *(weakly monotone);*

Proof. All trivial from basic properties of unions and products. For instance, distributivity holds because

$$A \times (B \cup C) = (A \times B) \cup (A \times C)$$

■

These cardinal operations are not, however, *strictly* monotone, as $1 + \aleph_0 = 2 + \aleph_0$, and $1 \cdot \aleph_0 = 2 \cdot \aleph_0$.

Next, we define cardinal exponentiation.

Let A and B be sets. We denote the set of functions $B \to A$ by ${}^B A$ or A^B. That is,

$${}^B A = A^B := \{f : f \text{ is a function from } B \text{ to } A\}$$

Note that for finite sets, A^B has $|A|^{|B|}$ many elements.

If κ and λ are cardinalities, then we define κ^λ to be the cardinality of A^B where $|A| = \kappa$ and $|B| = \lambda$.

Theorem 6.9.18 (Basic Properties of Cardinal Exponentiation)**.** *For all cardinalities* κ, λ, μ*, the following hold:*

(i) $\kappa^{\lambda + \mu} = \kappa^\lambda \cdot \kappa^\mu$*;*

(ii) $(\kappa \cdot \lambda)^\mu = \kappa^\mu \cdot \lambda^\mu$;

(iii) $(\kappa^\lambda)^\mu = \kappa^{\lambda \cdot \mu}$.

Lemma 6.9.19. *For any cardinality κ, $2^\kappa = |\mathcal{P}(S)|$ for any set S with cardinality $|S| = \kappa$.*

Proof. Let S have cardinality κ. The set $\{0,1\}$ has cardinality 2, so $2^\kappa = |\{0,1\}^S|$ by definition. Let $f : S \to \{0,1\}$, and define the set

$$A_f \coloneqq \{s \in S : f(s) = 1\}$$

Then, $f \mapsto A_f$ is a bijection from $\{0,1\}^S$ to $\mathcal{P}(S)$. ■

Corollary 6.9.19.1. *For any cardinality κ, $\kappa < 2^\kappa$*

Proof. Apply Cantor's theorem to the previous lemma. ■

Corollary 6.9.19.2. $2^{\aleph_0} = \mathfrak{c}$

Proof. By the previous lemma,

$$2^{\aleph_0} = |\{0,1\}^{\mathbb{N}}| = |\mathcal{P}(\mathbb{N})| = |\mathbb{R}| = \mathfrak{c}$$ ■

Corollary 6.9.19.3. $\mathfrak{c}^{\aleph_0} = \mathfrak{c}$

Proof.

$$\mathfrak{c}^{\aleph_0} = (2^{\aleph_0})^{\aleph_0} = 2^{\aleph_0 \cdot \aleph_0} = 2^{\aleph_0} = \mathfrak{c}$$ ■

Corollary 6.9.19.4. *The set of sequences of real numbers has cardinality equal to that of the continuum. That is, $|\mathbb{R}^{\mathbb{N}}| = |\mathbb{R}|$.*

Proof. By the previous corollary,

$$|\mathbb{R}^{\mathbb{N}}| = \mathfrak{c}^{\aleph_0} = \mathfrak{c} = |\mathbb{R}|$$ ■

6.10 Axiom of Choice

6.10.1 Equivalent Formulations

We recall the first form of the axiom of choice:

> **Axiom of Choice (first form).**
> For any relation R, there exists a function $F \subseteq R$ such that $\operatorname{dom}(F) = \operatorname{dom}(R)$.

We used this to prove that the existence of a surjection $B \to A$ implies the existence of an injection $A \to B$.

We now present another more intuitive form of the axiom of choice. Informally, given a collection of non-empty sets, there is a function that chooses one element from each set.

> **Axiom of Choice (second form).**
> Let $\mathcal{S}$ be a set with $\varnothing \notin \mathcal{S}$. Then, there is a *choice function* for $\mathcal{S}$. That is, a function $\sigma : \mathcal{S} \to \bigcup \mathcal{S}$ such that $\sigma(A) \in A$ for all $A \in \mathcal{S}$.

Example. Let $\mathcal{S} \subseteq \mathcal{P}(\mathbb{N}) \setminus \{\varnothing\}$ be a collection of non-empty sets of natural numbers. A choice function is a function $\sigma : \mathcal{S} \to \bigcup \mathcal{S} = \mathbb{N}$ that sends each set $A \in \mathcal{S}$ to an element $\sigma(A) = a \in A$.

One possible choice function is given by selecting the smallest number in each set. Even if $\mathcal{S}$ contains infinitely many sets, or even all possible sets of natural numbers, it is always possible to choose the smallest element from each set to produce a new set, because $\mathbb{N}$ is well-ordered.

In this case, we don't have to invoke the axiom of choice because we can construct this function explicitly

$$\sigma = \left\{\langle A,n\rangle \in \mathcal{P}(\omega) \times \omega : A \in \mathcal{S}, n \in A, \forall m \in \omega : m <_\omega n \to m \notin A\right\}$$

△

However, for some sets, a choice function is not known. For instance, the set of all non-empty subsets of the real numbers does not admit such a choice function. In this case, the axiom of choice must be invoked.

As Russell once remarked, "*The Axiom of Choice is necessary to select a set from an infinite number of pairs of socks, but not an infinite number of pairs of shoes.*"

Axiom of Choice (third form).
Let $\mathcal{A}$ be a set of pairwise disjoint non-empty sets. Then, there exists a set C who has as a member exactly one element from each member of $\mathcal{A}$. That is, for each $B \in \mathcal{A}$, $|C \cap B| = 1$.

6.10.1.1 Infinite Cartesian Products

Let I be an set (the *indexing set*) and let H be a function whose domain includes I. Then, for each $i \in I$, we have a set $H(i)$. We define the *indexed product* of the $H(i)$ as

$$\bigtimes_{i \in I} H(i) \coloneqq \{f : f \text{ is a function with domain } I \text{ and } f(i) \in H(i) \text{ for all } i \in I\}$$

Note that, up to natural isomorphism, this agrees with our earlier definition of an iterated cartesian product. For instance, a binary cartesian product $X_1 \times X_2$ is the set of pairs $\langle x_1,x_2\rangle$ with $x_1 \in X_1$ and $x_2 \in X_2$, but such a pair can be naturally identified with a function $x : \{1,2\} \to X_1 \cup X_2$, with $x(1) = x_1 \in X_1$ and $x(2) = x_2 \in X_2$.

This definition of an indexed product, however, makes sense even if the indexing set is not finite, or even countable.

Clearly, if any of the sets $H(i)$ is empty, then there are no such functions, and the entire product is empty. However, is it true that the product of non-empty sets is non-empty? This is again the axiom of choice:

Axiom of Choice (fourth form).
For any indexing set I and function H with domain I, if $H(i) \neq \varnothing$ for all $i \in I$, then

$$\bigtimes_{i \in I} H(i) \neq \varnothing$$

Theorem 6.10.1. *In the presence of the other ZF axioms, the first, second, third, and fourth forms of the axiom of choice are all equivalent.*

Proof of $(1) \leftrightarrow (2)$. Assume the first form holds. Let $\mathcal{S}$ be a set with $\varnothing \notin \mathcal{S}$, and define the relation

$$R = \left\{\langle x,y\rangle \in \mathcal{S} \times \bigcup \mathcal{S} : y \in x\right\}$$

Let $F \subseteq R$ be a function such that $\mathrm{dom}(F) = \mathrm{dom}(R)$, given by the first form. Then, for every $x \in \mathcal{S}$, $y = F(x) \in x$, so F is a choice function for $\mathcal{S}$.

Now assume the second form holds. The idea is that an arbitrary relation is like a multi-valued function, and a choice function given by the second form can choose for every x in the domain exactly one corresponding y.

Let R be any relation. Then, $R \subset \mathrm{dom}(R) \times \mathrm{ran}(R)$. Let

$$\mathcal{S} = \mathcal{P}\big(\mathrm{ran}(R)\big) \setminus \{\varnothing\}$$

Let $\sigma : \mathcal{S} \to \bigcup\mathcal{S}$ be a choice function given by the second form. That is, for every $A \in \mathcal{S}$ (so $\varnothing \neq A \subseteq \mathrm{ran}(R)$), we have $\sigma(A) \in A$. For $x \in \mathrm{dom}(R)$, let

$$F(x) = \sigma\Big(\big\{y \in \mathrm{ran}(R) : \langle x,y\rangle \in R\big\}\Big)$$

This defines a function $F : \mathrm{dom}(R) \to \mathrm{ran}(R)$ with $F \subseteq R$. ■

6.10.2 Partial Orders

A relation $R \subseteq X \times X$ is a (*weak* or *non-strict*) *partial order* on X if it satisfies, for all $x,y,z \in X$:

(i) reflexivity: xRx;

(ii) transitivity: $(xRy \wedge yRz) \to xRz$;

(iii) antisymmetry: $(xRy \wedge yRx) \to x = y$.

The usual notation for a weak partial order R is $\preceq_R$ or just $\preceq$.

A relation $R \subseteq X \times X$ is a *strict partial order* on X if it satisfies, for all $x,y,z \in X$:

(i) irreflexivity: $\neg xRx$;

(ii) transitivity: $(xRy \wedge yRz) \to xRz$;

(iii) asymmetry: $xRy \to \neg yRx$;

Note that asymmetry is implied by irreflexivity and transitivity, and may optionally be omitted from this definition.

The usual notation for a strict partial order R is $\prec_R$ or just $\prec$. If $a \prec b$, then we say that *a precedes b* or that *b covers a*.

The pair $(X, \prec)$ (where $\prec$ is any partial order, weak or strict depending on context) is then called a *partially ordered set* or a *poset*.

Weak and strict partial orders are essentially the same notions, differing only by the diagonal elements $\langle x,x\rangle$: given a strict partial order, adding these pairs into it yields a corresponding weak partial order, and conversely, removing these pairs from a weak partial order yields a strict partial order.

Note that not all elements in a poset may be comparable under the ordering. If every pair of elements *are* comparable, then the ordering is *total*.

Example. Consider the set of all subsets of $\mathbb{N}$ with at most three elements ordered by inclusion, $\subseteq$. The sets $\{0\}$ and $\{1\}$ are incomparable under this relation because neither $\{0\} \subseteq \{1\}$ nor $\{1\} \subseteq \{0\}$ holds. △

In a poset $(X, \preceq)$, an element $x \in X$ is *maximal* if for all $y \in X$, $x \preceq y$ only if $y = x$, or equivalently, x is maximal if there does not exist any y such that $x \prec y$.

A partial ordering may have any number of maximal elements, including none. For instance, the integers have no maximal element, while the set [0,1] has one maximal element, and a set with k mutually incomparable elements has k maximal elements.

This notion is distinct from that of a *maximum* element, which is an element $x \in X$ such that $y \preceq x$ for all $y \in X$. Clearly, a maximum element is maximal, and if it exists, it is unique.

Informally, a maximal element is an element that is not less than any other element, while a maximum element is an element that is greater than every other element.

Example. Consider the set of all subsets of $\mathbb{N}$ with at most three elements ordered by inclusion, $\subseteq$. The set $\{0,1,2\}$ is maximal because it is not a subset of any other set apart from itself, but it is not a maximum, because, for example, it is not a superset of $\{3\}$. △

6.10.2.1 Zorn's Lemma

Let $\preceq$ be a weak partial order on a set Z.

A *chain* is a subset $C \subseteq Z$ such that $\preceq$ is total on C. That is, every pair of elements in C are comparable under $\preceq$:

$$\forall c_1,c_2 \in C : c_1 \preceq c_2 \vee c_2 \preceq c_1$$

Clearly, every subset of a chain is itself a chain.

An element $x \in Z$ is an *upper bound* of a chain C if $c \preceq x$ for all $c \in C$.

Zorn's Lemma.
Let $(Z, \preceq)$ be a poset, and suppose that every chain $C \subseteq Z$ has an upper bound. Then, Z has a maximal element.

Although called a lemma, we normally treat this as an axiom, for it is equivalent to the axiom of choice.

Theorem 6.10.2. *In the presence of the other ZF axioms, Zorn's lemma is equivalent to the axiom of choice.*

Proof, forward direction only. We prove the first form of the axiom of choice, assuming Zorn's lemma.

Let R be any relation and define

$$Z = \{f \subseteq R : f \text{ is a function}\}$$

Z is partially ordered by inclusion, $\subseteq$. (Recall that $f \subseteq g$ if and only if g extends the function f).

Let $C \subseteq Z$ be a chain. We claim that $\bigcup C \in Z$, and that $\bigcup C$ is an upper bound of C. As the union of functions in R, $\bigcup C$ is a subset of R, and since $\bigcup C$ is the union of relations, it is itself a relation.

Suppose that $\langle x,y\rangle,\langle x,y'\rangle \in \bigcup C$. Then, there are functions $f,f' \in C$ such that $\langle x,y\rangle \in f$ and $\langle x,y'\rangle \in f'$. Since C is a chain, every function is comparable. Without loss of generality, suppose $f \subseteq f'$. Then, both pairs lie within f', so $y = y'$. Thus, $\bigcup C$ is a function, so it is in Z.

Now, for any $f \in C$, we have $f \subseteq \bigcup C$ by the definition of a union, so $\bigcup C$ is an upper bound for C.

Zorn's lemma then says that $(Z, \subseteq)$ has a maximal element F.

As $F \subseteq R$, $\operatorname{dom}(F) \subseteq \operatorname{dom}(R)$. Suppose for a contradiction that $\operatorname{dom}(F) \neq \operatorname{dom}(R)$, so there exists $x_0 \in \operatorname{dom}(R) \setminus \operatorname{dom}(F)$. Let y_0 be such that $\langle x_0,y_0\rangle \in R$, and define

$$F' = F \cup \{\langle x_0,y_0\rangle\}$$

Clearly, $F' \subseteq R$ is a function, so $F' \in Z$. But then, $F \subsetneqq F'$, contradicting that F is maximal in Z, so $\operatorname{dom}(R) = \operatorname{dom}(F)$, as required. ■

Theorem 6.10.3. *(AC) Every vector space has a basis.*

Proof. Let V be a vector space over a field K, and define the set

$$Z = \{S \subseteq V : S \text{ is linearly independent over } K\}$$

Consider the partial order on Z given by $\subseteq$. The empty set is linearly independent, and $\varnothing \in Z$, so $Z \neq \varnothing$.

Let $C \subseteq Z$ be a chain. Clearly, $\bigcup C \subseteq V$.

Suppose

$$\sum_{i=1}^{n} k_i v_i = 0$$

for some vectors $v_1, \ldots, v_n \in \bigcup C$ and scalars $k_1, \ldots, k_n \in K$. Since $v_i \in \bigcup C$, there are $S_i \in C$ with $v_i \in S_i$, and since C is a chain, the S_i also form a chain. Without loss of generality, suppose the ordering is as follows:

$$S_1 \subseteq S_2 \subseteq \cdots \subseteq S_n$$

so $v_1, \ldots, v_n \in S_n$. Since $S_n \in Z$ is linearly independent, $k_1 = \cdots = k_n = 0$, so $\bigcup C$ is linearly independent, and hence $\bigcup C \in Z$. $\bigcup C$ is also an upper bound for C since $S \subseteq \bigcup C$ for all $S \in C$.

By Zorn's lemma, there is a maximal element $S \in Z$. We claim S is a basis for V.

Since $S \in Z$, S is linearly independent. If $S = V$, then we are done. Otherwise, let $u \in V \setminus S$. Since S is maximal in Z, $S \cup \{u\}$, is not linearly independent, so

$$k_0 u + \sum_{i=1}^{n} k_i v_i = 0$$

for some vectors $v_1, \ldots, v_n \in S$ and scalars $k_0, \ldots, k_n \in K$ not all equal to 0. If $k_0 = 0$, then $\sum_{i=1}^{n} k_i v_i = 0$ with $k_1, \ldots, k_n$ not all zero, contradicting that S is linearly independent. So, $k_0 \neq 0$, and hence

$$u = -\frac{1}{k_0}(k_1 v_1 + \cdots + k_n v_n)$$

is in the linear span of S, so S is a basis for V. ■

Corollary 6.10.3.1. *$\mathbb{R}$ as a vector space over $\mathbb{Q}$ has a basis. That is, there is a set $H \subset \mathbb{R}$ such that every $x \in \mathbb{R}$ can be expressed as a unique linear combination*

$$x = \sum_{i=1}^{n} q_i x_i$$

of vectors $x_1, \ldots x_n \in H$ and scalars $q_1, \ldots q_n \in \mathbb{Q}$.

Such a basis is called a *Hamel basis.*

6.10.3 Cardinal Comparability

Recall that we write $|A| \leq |B|$ if there exists an injection $A \to B$. Is this ordering total on the class of all cardinals?

Cardinal Comparability.
For any sets A and B, we have $|A| \leq |B|$ or $|B| \leq |A|$. That is, there is an injective function $A \to B$ or there is an injective function $B \to A$.
Equivalently, for any two cardinals κ and λ, we have $\kappa \leq \lambda$ or $\lambda \leq \kappa$.

It turns out that cardinal comparability is again equivalent to the axiom of choice.

6.10.4 Absorption Law

So far, we have seen that cardinal arithmetic for finite cardinalities agrees with arithmetic on ω. That is, if $|A| = n$ (i.e. there exists a bijection between A and $n \in \omega$) and $|B| = m$ and A and B are disjoint, then $n + m = |A \cup B| = n +_\omega m$, where the addition on the left is cardinal addition. Similarly, if $|A| = n$ and $|B| = m$, then $n \cdot m = |A \times B| = n \cdot_\omega m$.

Theorem 6.10.4. *(AC) For every infinite cardinality κ, $\kappa \cdot \kappa = \kappa$. That is, for every infinite set X, there is a bijection $X \to X \times X$.*

Corollary (Absorption Law)**.** *(AC) If κ or λ is infinite, then*

$$\kappa + \lambda = \max(\kappa,\lambda)$$

If one is infinite and the other is non-zero, then

$$\kappa \cdot \lambda = \max(\kappa,\lambda)$$

Proof. Suppose without loss of generality that $\kappa \geq \lambda$. Then,

$$\kappa \leq \kappa + \lambda \leq \kappa + \kappa \leq \kappa \cdot 2 \leq \kappa \cdot \kappa = \kappa$$

so $\kappa + \lambda = \kappa = \max(\kappa,\lambda)$. If additionally $\lambda \neq 0$, then,

$$\kappa \leq \kappa \cdot \lambda \leq \kappa \cdot \kappa = \kappa$$

so $\kappa \cdot \lambda = \kappa = \max(\kappa,\lambda)$. ■

6.11 Well-Ordered Sets

6.11.1 Linearly Ordered Sets

A binary relation $<_X$ on a set X is a (*weak* or *non-strict*) *total* or *linear order* if:

- For all $a,b \in X$, exactly one of $a <_X b$, $b <_X a$, and $a = b$ holds (trichotomy);
- For all $a,b,c \in X$, if $a <_X b$ and $b <_X c$, then $a <_X c$ (transitivity).

or equivalently, if $<_X$ is a partial order in which every pair of elements is comparable or equal.

Example. The *double line* is the set $X = \mathbb{R} \times \{0,1\}$ equipped with the *lexicographical ordering* $<_{\text{lex}}$ where $\langle r,i \rangle <_X \langle s,j \rangle$ if and only if $r < s$ or $r = s$ and $i < j$. △

Two ordered sets $(X, <_X)$ and $(Y, <_Y)$ are *order-isomorphic*, written as $(X, <_X) \cong (Y, <_Y)$ if there is a bijection $f : X \to Y$ that compatible with the ordering. That is, $a <_X b$ if and only if $f(a) <_Y f(b)$.

Every finite totally ordered set is order-isomorphic to a subset of the natural numbers, and similarly, every countable totally ordered set is order-isomorphic to a subset of the rationals.

6.11.2 Well-Ordered Sets

A total ordering $(X, <_X)$ is *well-ordered* if every non-empty subset $S \subseteq X$ has a minimal element. That is, for every $S \subseteq X$, there exists $s \in S$ such that $s \leq_X t$ for all $t \in S$.

Example.

- $\mathbb{N}$ is well-ordered.
- $\mathbb{Z}$ with its usual numerical ordering $<_\mathbb{Z}$ is not well-ordered.

- $\mathbb{Z}$ with the ordering $a \prec b$ if and only if $|a| < |b|$ or if $a = |b|$,

  $$0 \prec 1 \prec -1 \prec 2 \prec -2 \prec 3 \prec -3 \prec \cdots$$

  is well-ordered. This ordering is order-isomorphic to $(\mathbb{N}, <_{\mathbb{N}})$.
- $\{-\frac{1}{n} : n \in \mathbb{Z}^+\} \cup \mathbb{N}$ is well-ordered.
- $\mathbb{Z}$ with the ordering $a \prec b$ if and only if $0 \leq a < b$, $a \geq 0 > b$ or if $0 \geq a > b$,

  $$0 \prec 1 \prec 2 \prec 3 \prec \cdots \prec -1 \prec -2 \prec -3 \prec \cdots$$

  is well-ordered. This ordering is order-isomorphic to $\{-\frac{1}{n} : n \in \mathbb{Z}^+\} \cup \mathbb{N}$.
- $\mathbb{R}$ is not well-ordered.
- $(\mathbb{N} \times \mathbb{N}, <_{\text{lex}})$ is well-ordered.

△

Theorem 6.11.1. *(AC) A total ordering $(X, <_X)$ is well-ordered if and only if there is no strictly decreasing infinite sequence $x_0 > x_1 > x_2 > \cdots$ in X.*

Proof. If such a sequence exists, take $S = \{x_0, x_1, \ldots\}$. Then, S has no minimal element. Conversely, if S is non-empty and has no minimal element, then choose some element $x_0 \in S$. Because x_0 is not minimal, there exist elements $x_1 \in S$ less than x_0. Choose one such element and add it to the sequence. This element is again not minimal, and so on. ■

Theorem 6.11.2. *Suppose $(X, <)$ is a well-order. Then, if $x \in X$ is not maximal, then there is a unique element $x^+ \in X$ such that $x < x^+$ and there is no element $y \in X$ with $x < y < x^+$.*

Proof. Let x^+ be the minimal element of the non-empty subset $\{y \in X : y > x\}$. ■

Theorem (Fundamental Lemma)**.** *Let $(X, <)$ be a well-ordering, and let $f : X \to X$ be order preserving. That is, $x < y \to f(x) < f(y)$ for all $x, y \in X$. Then, $f(x) \geq x$ for all $x \in X$.*

Proof. Suppose the set

$$S = \{x \in X : f(x) < x\}$$

of counterexamples is non-empty. Since X is well-ordered, $S \subseteq X$ has a minimal element x_0. By definition of S, we have $f(x_0) = x_1 < x_0$, and since f is order preserving, we also have $f(x_1) < x_1$, so $x_1 \in S$. But $x_1 < x_0$, contradicting that x_0 is minimal. ■

Corollary 6.11.2.1. *If $(X, <_X) \cong (Y, <_Y)$ are isomorphic well-ordered sets, then the order-isomorphism $X \to Y$ is unique.*

Proof. Suppose $f, g : X \to Y$ are order-isomorphisms. Then, $f^{-1} \circ g : X \to X$ is an order-isomorphism, and by the fundamental lemma, $(f^{-1} \circ g)(x) = f^{-1}\big(g(x)\big) \geq x$ for all x. So, $g(x) \leq f(x)$ for all x. Similarly, by considering $g^{-1} \circ f : X \to X$, we obtain that $f(x) \leq g(x)$ for all x, so $f = g$. ■

6.11.3 Trichotomy Theorem for Well-Ordered Sets

Let $(X, <)$ be a well-ordering, and let $a \in X$. We define the *initial segment determined by the element* a to be the set

$$X \restriction a = \{x \in X : x < a\}$$

This set is well-ordered by the ordering inherited from X.

Example.

- For $(\mathbb{N}, <_{\mathbb{N}})$, $\mathbb{N} \restriction a = \{0{,}1, \ldots, a-1\}$ equipped with the usual ordering $0 < 1 < \cdots < a-1$.
- For $\mathbb{Z}$ in the unusual ordering $0{,}1{,}2{,}3, \ldots, -1, -2, -3, -4, \ldots$, the initial segment $\mathbb{Z} \restriction -1$ is simply $\mathbb{N}$ in its usual ordering.

△

Theorem 6.11.3. *If X is well-ordered, and $x_0 \in X$, then $X \ncong X \restriction x_0$.*

Proof. Suppose $f : X \to X \restriction x_0$ is an order-isomorphism. Then, the same map is equivalently a order homomorphism $f : X \to X$, but with $f(x_0) < x_0$, contradicting the fundamental lemma. ■

Theorem (Trichotomy). *Let $(X, <_X)$ and $(Y, <_Y)$ be well-ordered sets. Then, exactly one of the following holds:*

- *$X \cong Y$;*
- *There exists $x_0 \in X$ such that $X \restriction x_0 \cong Y$;*
- *There exists $y_0 \in Y$ such that $Y \restriction x_0 \cong X$;*

6.11.4 Well-Ordering Principle

Well-Ordering Principle (Cantor).
Every set is well-orderable. That is, given any set X, there is a relation $<$ on X such that $(X, <)$ is well-ordered.

The well-ordering principle is equivalent to the axiom of choice.

Recall that every finite totally ordered set is well-ordered. We also have that $\omega = \mathbb{N}$ with natural ordering is well-ordered, but $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$, with their natural ordering, are not well-ordered.

We can prove that $\mathbb{Z}$ and $\mathbb{Q}$ can be well-ordered without invoking the axiom of choice: both sets are countable, so take any bijection $f : \mathbb{Z} \to \omega$ or $f : \mathbb{Q} \to \omega$, and define the ordering $<$ on $\mathbb{Z}$ or $\mathbb{Q}$ by

$$a < b \quad \leftrightarrow \quad f(a) <_\omega f(b)$$

Since $(\omega, <_\omega)$ is well-ordered, $\mathbb{Z}$ or $\mathbb{Q}$ equipped with this ordering $<$ will also be well-ordered. However, this method does not work on $\mathbb{R}$ since it is uncountable. Here, the well-ordering principle must be used.

6.11.5 Order-types

Informally, the cardinality of a set describes the "size" of that set. The *order-type* instead describes the "length" of a well-ordered set.

Let $(X, <_X)$ and $(Y, <_Y)$ be well-orderings. We say that

$$\mathrm{type}(X, <_X) = \mathrm{type}(Y, <_Y)$$

if $(X, <_X) \cong (Y, <_Y)$.

We seem to have very good notation for the order-type of many well-ordered sets:

- $\omega = \text{type}(\mathbb{N})$;
- $n = \text{type}\big(\{0,1,\ldots,n-1\}\big)$;
- $0 = \text{type}(\varnothing)$.

Let α and β be order-types of well-ordered sets. Fix well-ordered sets $(X, <_X)$ and $(Y, <_Y)$ with order-type α and β, respectively.

- We write $\alpha = \beta$ if $X \cong Y$;
- We write $\alpha > \beta$ if there exists $x_0 \in X$ such that $X \restriction x_0 \cong Y$;
- We write $\alpha < \beta$ if there exists $y_0 \in Y$ such that $X \cong Y \restriction y_0 \cong$.

Recall that exactly one of these three options holds by the trichotomy theorem.

6.11.6 Ordinal Arithmetic

We have seen that cardinalities admit a sensible notion of arithmetic. What about order-types?

Let $\alpha = \text{type}(X, <_X)$ and $\beta = \text{type}(Y, <_Y)$.

- Suppose that X and Y are disjoint. The union $X \cup Y$ is well-ordered by the relation $<$ where $a < b$ if $(a \in X$ and $b \in Y)$ or $(a,b \in X$ and $a <_X b)$ or $(a,b \in Y$ and $a <_Y b)$.
- The product $X \times Y$ can also be well-ordered by the *anti-lexicographic ordering* with $\langle x,y\rangle <_{\text{anti-lex}} \langle x',y'\rangle$ if and only if $y <_Y y'$, or if $y = y'$ and $x <_X x'$.

Then, we define $\alpha + \beta = \text{type}(X \cup Y, <)$ and $\alpha \times \beta = \text{type}(X \times Y, <_{\text{anti-lex}})$.

Recall the labels used in Cantor's transfinite iteration from §6.1:

$$0,1,\ldots,\omega,\omega+1,\ldots,\underbrace{\omega+\omega}_{\omega\cdot 2},\ldots,\underbrace{\omega+\omega+\omega}_{\omega\cdot 3},\ldots,\underbrace{\omega\cdot\omega}_{\omega^2},\ldots$$

Formally, this notation lists out the order-types in increasing order.

6.12 Transfinite Induction

Induction works on $\mathbb{N}$ precisely because $\mathbb{N}$ is well-ordered. It turns out that there is nothing special about $\mathbb{N}$, and induction works equally as well on *any* well-ordered set.

6.12.1 Induction on $\mathbb{N}$

To prove that a property φ holds for all natural numbers, the usual way is to show that $\varphi(0)$ holds, then to prove that for every $n \in \mathbb{N}$ if $\varphi(n)$ holds, then $\varphi(n+1)$ holds.

Theorem (Induction on $\mathbb{N}$)**.** *Let φ be a property of natural numbers. If $\varphi(0)$ holds, and for every $n \in \mathbb{N}$, $\varphi(n)$ implies $\varphi(n+1)$, then for every $n \in \mathbb{N}$, $\varphi(n)$ holds:*

$$\Big(\varphi(0) \wedge \forall n \in N\big(\varphi(n) \to \varphi(n+1)\big)\Big) \to \forall n \in \mathbb{N} : \varphi(n)$$

This version of induction does not generalise easily to arbitrary well-ordered sets, so first we rephrase this.

Theorem (Strong Induction on $\mathbb{N}$)**.** *Let φ be a property of natural numbers, and assume that φ satisfies the following property:*

If for every $n \in \mathbb{N}$, if $\varphi(m)$ holds for all $m < n$, then $\varphi(n)$ holds.

Then, for every $n \in \mathbb{N}$, $\varphi(n)$ holds:

$$\Big(\forall n \in \mathbb{N} : \big(\forall m \in \mathbb{N} : m < n \to \varphi(m)\big) \to \varphi(n)\Big) \to \forall n \in \mathbb{N} : \varphi(n)$$

It may seem that we are missing the base case $\varphi(0)$ in our assumptions, but for $n = 0$, the assumption becomes the following: if $\varphi(m)$ holds for all naturals $m < 0$, then $\varphi(0)$ holds. The conditional clause is vacuously true, since there are no naturals $m < 0$.

Proof. Let $\psi(n)$ be the statement "if $m < n$, then $\varphi(m)$ holds". Then, $\psi(0)$ holds as there is no $m < 0$.

Now assume $\psi(n)$ holds, so $\varphi(m)$ holds for all $m < n$ and therefore $\varphi(n)$ holds by the hypothesis of the theorem. Thus, $\varphi(m)$ holds for all $m < n + 1$, which is the statement of $\psi(n + 1)$. Then, standard induction for ψ gives that $\psi(n)$ holds for all $n \in \mathbb{N}$, and therefore $\varphi(n)$ holds for all $n \in \mathbb{N}$. ■

6.12.2 Transfinite Induction on Well-Ordered Sets

Strong induction on $\mathbb{N}$ generalises nicely to arbitrary well-ordered sets.

Theorem (Transfinite Induction)**.** *Let $(X, <)$ be a well-ordered set, and let φ be a property of elements of X. That is, for each $x \in X$, $\varphi(x)$ is either true or false.*

Assume that for every $x \in X$, the following holds:

If $\varphi(y)$ holds for every $y < x$ then $\varphi(x)$ holds.

Then, for every $x \in X$, $\varphi(x)$ holds:

$$\Big(\forall x \in X \Big(\forall y \in X \big(y < x \to \varphi(y)\big)\Big)\Big) \to \forall x \in X : \varphi(x)$$

Proof. Define the set

$$S = \big\{x \in X : \neg\varphi(x)\big\}$$

Suppose for a contradiction that S is non-empty. Then, since X is well-ordered, $S \subseteq X$ has a minimal element x_0. Then, for every $y < x_0$, $y \notin S$ so $\varphi(y)$ holds. By the assumption, this implies $\varphi(x_0)$ holds, so $x_0 \notin S$. ■

6.13 Ordinals as Sets

6.13.1 Concept Evolution

Recall the evolution of the concept of natural numbers. Early in life, we become familiar with the abstract concept of "three", common to all collections of three objects. Then, we learnt how to manipulate 3, and other numbers.

Then, in set theory, we identified the number 3 with a particular set, $\big\{\varnothing,\{\varnothing\},\{\varnothing,\{\varnothing\}\}\big\}$. This choice is arbitrary, but it has some nice properties, and, importantly, it is a set, so we can continue on our quest to embed all of mathematics into set theory.

One may notice that we have not completed this final step for cardinalities – or ordinals, for that matter. We have looked at all those sets that can be mapped bijectively to $\mathbb{N}$, and called them countably infinite.

Then, we introduced the notation $\aleph_0$, common to all these sets. Then, we defined some arithmetic operations to manipulate these cardinalities.

But, we haven't yet talked about what kind of object $\aleph_0$ is, nor any other cardinality κ.

This isn't really a problem, as mathematics is not interested in the internal composition of its objects – only about their external structure and how they behave. This, we have defined, even for cardinalities. We know what $\kappa = \lambda$ or $\kappa < \lambda$ means; and we have defined $\kappa + \lambda$, $\kappa \cdot \lambda$, κ^λ, etc.

However, our goal was to embed all of mathematics into set theory – so this should include cardinalities too. The task is to present a unique set, called a *cardinal*, for each cardinality.

It turns out that it is more convenient if we do this for order-types first. In this case, the representing sets are called *ordinals*.

6.13.2 Mapping Order-Types to Sets

We wish to define the *ordinals*: sets equipped with a natural well-ordering such that each order-type of well-ordered sets corresponds to exactly one ordinal.

We can think of this as a "map" that assigns the abstract concept of order-types to well-ordered sets. However, the order-types are just classes of well-ordered sets, so we are really looking for a class-function that sends well-ordered sets to representing well-ordered sets, such that order-isomorphic well-ordered sets have the same image.

Consider the three-element set $X = \{a,b,c\}$ with ordering $a < b < c$, and let E be a function with domain X that satisfies the equation

$$E(x) = \big\{E(y) : y <_X x\big\}$$

for all $x \in X$.

Then,

$$\begin{aligned}
E(x_0) &= \big\{E(y) : y <_X a\big\} = \varnothing = 0 \in \omega \\
E(x_1) &= \big\{E(y) : y <_X b\big\} = \big\{E(a)\big\} = \{\varnothing\} = 1 \in \omega \\
E(x_2) &= \big\{E(y) : y <_X c\big\} = \big\{E(a),E(b)\big\} = \big\{\varnothing,\{\varnothing\}\big\} = 2 \in \omega
\end{aligned}$$

More generally, for any finite well-ordered set X, the function E as defined above maps the smallest element of X to $\varnothing = 0$, the second smallest to $\{0\} = 1$, the next smallest to $\{0,1\}$, and so on; and furthermore, E gives an order-isomorphism between X and the set $n = \{0,1,2,\ldots,n-1\}$.

Note that if $x <_X x'$ for some $x,x' \in X$, then $E(x) \in E(x')$. This recursive property allows us to define E for all well-ordered sets.

Lemma (Epsilon-image of a Well-Ordered Set)**.** *Let $(X, <_X)$ be a well-ordered set. Then, there is a unique function E with domain X such that for all $x \in X$,*

$$\begin{aligned}
E(x) &= \big\{E(y) : y \leq_X x\big\} \\
&= E\big[\{y \in X : y <_X x\}\big] \\
&= E[X \restriction x]
\end{aligned}$$

Theorem (Transfinite Recursion for Well-Ordered Sets)**.** *For any formula $\varphi(f,y)$, the following is a theorem:*

Suppose that for any function f, there is a unique y such that $\varphi(f,y)$ holds, and let $(X, <_X)$ be a well-ordered set. Then, there exists a unique function F with domain X such that

$$\varphi(F|_{X\restriction x},F(x))$$

holds for all $x \in X$.

This map E has some very useful properties.

Lemma (Epsilon-images II)**.** *The function E defined in the previous lemma satisfies the following properties:*

- *Whenever $x <_X x'$, we have $E(x) \in E(x')$, and also $E(x) \subsetneq E(x')$;*
- *E is injective;*
- *E is an order-isomorphism between $(X, <_X)$ and $E[X]$ with its well-ordering given by the $\in$ relation.*

Lemma 6.13.1. *Let $(X, <_X) \cong (Y, <_Y)$ be order-isomorphic well-ordered sets. Then, their epsilon-images $E_X[X]$ and $E_Y[Y]$ coincide.*

Proof. This is a proof by transfinite induction. Let $f : X \to Y$ be an order-isomorphism, and define the set

$$S \coloneqq \{x \in X : E_X(x) \neq E_Y\big(f(x)\big)\}$$

Suppose for a contradiction that S is non-empty. Let x_0 be the smallest element of S ($S \subseteq X$, so S is well-ordered and x_0 exists). By definition of S, we have $E_X(x) = E_Y\big(f(x)\big)$ for all $x < x_0$. Then,

$$\begin{aligned} E_X(x_0) &= \{E_X(x) : x <_X x_0\} \\ &= \{E_Y\big(f(x)\big) : x <_X x_0\} \\ &= \{E_Y\big(y\big) : y <_X f(x_0)\} \\ &= E_Y\big(f(x_0)\big) \end{aligned}$$

contradicting that $x_0 \in S$. ■

6.13.3 Ordinals

A set α is an *ordinal* if

- it is well-ordered by the $\in$ relation;
- whenever $y \in x \in \alpha$, we have $y \in \alpha$.

The second property is equivalent to $x \in \alpha \to x \subseteq \alpha$ (α is *transitive*).

Note that by iterating the second property, we also have that if $x_3 \in x_2 \in x_1 \in \alpha$, then $x_3 \in x_2 \in \alpha$, and $x_3 \in \alpha$. Informally, α contains as elements "all the things it references".

Theorem 6.13.2.

(i) *The epsilon-image of well-ordered sets are ordinals.*

(ii) *Every ordinal is the epsilon-image of a well-ordered set. Namely, each ordinal is the epsilon-image of itself.*

(iii) *Ordinals represent order-types of well-ordered sets.*

Proof.

(i) This follows from Lemma 6.13.1. Let $(X, <_X)$ be a well-ordering. Suppose $a \in b \in E[X]$, so $a = E(x)$ and $b = E(x')$ for some $x < x'$. Then $E(x) \in E[X]$.

(ii) Let α be an ordinal. Then, by definition,

$$E(x) = \{E(y) : y \in x\}$$

We claim that $E(x) = x$ for every $x \in \alpha$. Suppose otherwise, and let x_0 be the $\in$-minimal element in α for which $E(x_0) \neq x_0$. Then,

$$\begin{aligned} E(x_0) &= \big\{E(y) : y \in x_0\big\} \\ &= \big\{y : y \in x_0\big\} \\ &= x_0 \end{aligned}$$

This is a contradiction, so $E(x) = x$ for all $x \in \alpha$, and $E[\alpha] = \alpha$.

(iii) The previous two statements combined with Lemma 6.13.1 is precisely this result.

■

Theorem (Properties of Ordinals)**.**

(i) *Any element of an ordinal is itself an ordinal.*

(ii) *For any ordinals α,β,γ, if $\alpha \in \beta \in \gamma$, then $\alpha \in \gamma$.*

(iii) *For any ordinal α, we have $\alpha \notin \alpha$.*

(iv) *For any two ordinals α,β, exactly one of the $\alpha \in \beta$, $\alpha = \beta$, and $\beta \in \alpha$ holds (trichotomy).*

(v) *Any non-empty set S of ordinals has a least element. That is, there exists an ordinal $\delta \in S$ such that $\delta \in \alpha$ for all $\alpha \in S$ distinct from δ.*

(vi) *Every ordinal is the set of ordinals smaller than it: $\alpha = \{\beta : \beta < \alpha\}$, where $\beta < \alpha \equiv \beta \in \alpha$.*

(vii) *For each ordinal α, the set*

$$\alpha + 1 \coloneqq \alpha^+ \coloneqq \alpha \cup \{\alpha\}$$

is also an ordinal, and is the smallest ordinal larger than α:

$$\alpha^+ = \min\{\beta : \beta > \alpha\}$$

(viii) *If A is any set of ordinals, then $\bigcup A$ is also an ordinal, and is the least upper bound of A.*

An infinite ordinal is called a *limit ordinal* if it is not the successor of any ordinal.

Example. ω is a limit ordinal, while ω^+ and ω^{++} are successor ordinals. △

Example. The following sets are ordinals:

- $0 = \varnothing$;
- $1 = \{\varnothing\}$;
- $2 = \big\{\varnothing,\{\varnothing\}\big\}$;
- $n = \{0,1,2,\dots,n-1\}$ for every $n \in \omega$;
- $\omega = \{0,1,2,\dots\}$;
- $\omega + 1 = \omega^+ = \omega \cup \{\omega\} = \{0,1,2,\dots,\omega\}$;
- $\omega + 2 = \omega^{++} = \omega^+ \cup \{\omega^+\} = \{0,1,2,\dots,\omega,\omega+1\}$;
- $\omega + n = \big\{0,1,2,\dots,\omega,\omega+1,\dots,\omega+(n-1)\big\}$;
- $\omega \cdot 2 = \omega + \omega = \{0,1,2,\dots,\omega,\omega+1,\omega+2,\dots\}$;

△

An ordinal α is *countable* if α is a countable set. Every ordinal in the example above is countable.

However, what is the order-type of an uncountable well-ordered set? Such sets exist, by the well-ordering principle, so it must have an order type.

We define the ordinal

$$\omega_1 := \text{the smallest uncountable ordinal}$$

Since any subset of the ordinals has a minimal element, such a smallest ordinal exists.

So, α is a countable ordinal if and only if $\alpha < \omega_1$. That is, if $\alpha \in \omega_1$. Since ω_1 is uncountable, there are uncountably many countable ordinals. This also means that not every countable ordinal can be explicitly expressed like those above.

Theorem (Burali-Forti)**.** *There is no set to which every ordinal number belongs.*

Proof. The class of ordinals is well-ordered by $\in$, and every element of an ordinal is an ordinal, so if the class of ordinals was a set, it would be an ordinal itself. But then it would be a member of itself, and no ordinal has this property. ■

6.13.4 Cardinals

Let A be a set. Then, its *cardinal* $\kappa = |A|$ is the smallest ordinal that is equinumerous to A.

Note that cardinals are always limit ordinals, because for any infinite ordinal α, α and $\alpha + 1 = \alpha \cup \{\alpha\}$ are equinumerous.

We show that $|A|$ is well-defined, and that this definition agrees that equinumerous sets have the same cardinals. By the well-ordering principle, A can be well-ordered by an ordering $<_A$. Let β be be the ordinal representing the order type of $(A, <_A)$. So, $\beta = E[A]$, and the epsilon function is a bijection, so $\beta \sim A$. Now, consider the set of ordinals $\{\gamma : \gamma \leq \beta\}$. This has a minimal ordinal equinumerous to A; this minimal ordinal is the cardinal $|A|$.

Now, let $f : A \to B$ be a bijection. Clearly, the cardinal defined for A is still the smallest ordinal equinumerous to B, so $|A| = |B|$.

Example. The smallest countably infinite ordinal is ω, so $\omega = \aleph_0 = |\mathbb{N}| = |\mathbb{Z}|$.

Note that it is still sensible to retain the differing notations ω and $\aleph_0$ despite them being the same set, because cardinal and ordinal arithmetic are distinct. That is, $\omega + 1 \neq \aleph_0 + 1$. Also, $\omega = \aleph_0 = |\omega + 1|$ △

Let $\aleph_1$ denote the cardinality of ω_1. Actually, $\aleph_1 = \omega_1$ by the definition above. Then, $\aleph_1$ is the smallest uncountable cardinality. We similarly define $\aleph_2$ to be the smallest cardinal larger than $\aleph_1$, and so on. Note that we did not know until now that there is a smallest uncountable cardinality.

6.14 Applications

6.14.1 Transfinite Recursion

Theorem 6.14.1. *It is possible to draw disjoint letters* `T` *in* $\mathbb{R}^2$ *above every rational point of the x-axis.*

(A letter `T` above a real number x with height $h > 0$ and width $w > 0$ is a union of two line segments: the vertical line segment connecting $(x,0)$ to (x,h), and the horizontal line segment connecting $(x - \frac{w}{2},h)$ to $(x + \frac{w}{2},h)$.)

Constructive proof. For the rational number $\frac{p}{q}$ with $p \neq 0$, $q > 0$, and p,q coprime, draw a letter `T` with width $\frac{1}{2q^2}$ and height $\frac{1}{q}$.

Let $\frac{p}{q}$ and $\frac{p'}{q'}$ be two rationals in simplest form. If $q' = q$, then the horizontal distance of these rational numbers is at least $\frac{1}{q}$, so the letters T are clearly disjoint. Otherise, suppose without loss of generality that $q' < q$. Then,

$$\left|\frac{p}{q} - \frac{p'}{q'}\right| = \left|\frac{pq' - p'q}{qq'}\right| \geq \frac{1}{qq'} \geq \frac{1}{q^2}$$

so the horizontal segment of the letter T for $\frac{p}{q}$ does not reach the vertical segment of the letter for $\frac{p'}{q'}$, and since the height of the letter T for $\frac{p'}{q'}$ is greater than the height of the letter T for $\frac{p}{q}$, the letters are indeed disjoint. ■

Proof by recursion. Enumerate the rationals as x_0,x_1,x_2,.... First draw a letter T above x_0. Then draw a letter T x_1, disjoint from the one above x_0. Then draw a letter T above x_2, disjoint from the previously drawn letters. And so on.

At each step, there are always finitely many letters T drawn, so we can always draw the next one disjoint from all previous ones. ■

Theorem 6.14.2. *(AC)* $\mathbb{R}^3$ *is a union of disjoint circles of unit radius.*

Proof. Let $\{p_\alpha : \alpha < \mathfrak{c}\}$ enumerate the points of $\mathbb{R}^3$. That is, well-order $\mathbb{R}^3$ so that its order type is the smallest possible ordinal with cardinality of the continuum; or equivalently, fix a bijection from the cardinal/ordinal $\mathfrak{c}$ to $\mathbb{R}^3$.

For each ordinal $\alpha < \mathfrak{c}$, we will choose a circle C_α to cover the point p_α unless p_α was already covered by previous circles $\{C_\beta : \beta < \alpha\}$, in which case, we set $C_\alpha = \varnothing$.

With transfinite recursion, define sets C_α such that for each $\alpha < \mathfrak{c}$,

- $\bigcup_{\beta \leq \alpha} C_\beta$ contains the point p_α,
- C_α is disjoint from the set $\bigcup_{\beta < \alpha} C_\beta$.

Once this is done, the non-empty sets C_α are pairwise disjoint unit circles whose union is $\mathbb{R}^3$.

Assume that for some $\alpha < \mathfrak{c}$, the sets C_β for $\beta < \alpha$ are already defined. From this, we will construct C_α to finish the result by strong induction.

If $p_\alpha \in \bigcup_{\beta<\alpha} C_\beta$, then define $C_\alpha := \varnothing$. Otherwise, $p_\alpha \notin \bigcup_{\beta<\alpha} C_\beta$. because $\mathfrak{c}$ is the minimal ordinal with cardinality continuum, the ordinal $\alpha = \{\beta : \beta < \alpha\}$ has cardinality less than continuum, so at this point of the construction, we have less than continuum many circles.

Each circle C_β lies in a plane, H_β. There are continuum many planes through p, and we have less than continuum many circles constructed, so there exists a plane H distinct from the other planes H_β. This plane H can intersect each circle C_β in at most two points, or otherwise it would contain the whole circle. So, H intersects $\bigcup_\beta C_\beta$ in less than continuum many points. Denote this set of intersections as $S \subset H$, with cardinality $\kappa < \mathfrak{c}$. There are continuum many circles in H passing through p, and each point of S disqualifies only two such circles. So, there exists a circle through p in H disjoint from S, and therefore disjoint from all the circles C_β. ■

6.14.2 Exactly Two Points on Every Line

Theorem 6.14.3. *(AC) There exists a subset A of the plane $\mathbb{R}^2$ that intersects every straight line in exactly two points.*

It is a major unsolved problem in fractal geometry whether there is a *Borel* set A with this property.

Proof. Let L be the set of all straight lines in $\mathbb{R}^2$. Then $|L| = \mathfrak{c}$. Enumerate the lines as

$$L = \{\ell_\alpha : \alpha < \mathfrak{c}\}$$

We construct sets $A_\alpha \subset \mathbb{R}^2$ by induction and recursion such that, for all $\alpha < \mathfrak{c}$,

- A_α has at most two points;
- $\bigcup_{\beta\leq\alpha} A_\beta$ does not have three collinear points;
- $\bigcup_{\beta\leq\alpha} A_\beta$ has exactly two points of the line ℓ_α, and thus exactly two points of any line ℓ_β for $\beta \leq \alpha$.

Once we have this set, the set

$$A = \bigcup_{\alpha<\mathfrak{c}} A_\alpha$$

has the desired properties. It has at least two points on every line ℓ_α, and it cannot have three on any line, as then there would be an ordinal α such that the set

$$\bigcup_{\beta\leq\alpha} A_\beta$$

already contains those three collinear points.

To begin the construction, consider the first line ℓ_0. Choose two arbitrary points $p_0,q_0 \in \ell_0$ and let $A_0 = \{p_0,q_0\}$. Now consider the line ℓ_1. If A_0 already contains a point on ℓ_1, then let $A_1 = \{p_1\}$ for some other point p_1 on ℓ_1. Otherwise, choose two arbitrary points p_1,q_1 from $\ell_1 \setminus \ell_0$ and let $A = \{p_1,q_1\}$ (we disallow the point of intersection if it exists, or there would be three points on ℓ_0 already).

Let $1 \leq \alpha < \mathfrak{c}$ be arbitrary and assume that the sets A_β for $\beta < \alpha$ are already defined.

The set $S = \bigcup_{\beta<\alpha} A_\beta$ has cardinality $|S| \leq \alpha < \mathfrak{c}$ by the absorption law as it is the union of α many sets of cardinality at most 2.

By the inductive hypothesis, the set S can contain at most two points of the line ℓ_α. If $|S \cap \ell_\alpha| = 2$, then let $A_\alpha = \varnothing$ and we are done. If $|S \cap \ell_\alpha|$ is one or two, then we choose one or two new point(s) on ℓ_α, respectively, and define A_α to be the set containing these new points. We have to show that these new points do not create three collinear points anywhere.

Consider the set M of lines that contain (at least) two points of S. Then, $|M| \leq |S \times S| = |S|^2 = |S| < \mathfrak{c}$, so the line ℓ_α contains less than continuum many points that are incident to a line of M. Apart from these, we can choose any other points on ℓ_α for A_α. ■

6.14.3 Ultrafilters

A set $\mathcal{U} \subset \mathcal{P}(\mathbb{N})$ is a *filter* if it satisfies the following properties:

- $\mathbb{N} \in \mathcal{U}$;
- $\varnothing \notin \mathcal{U}$;
- (monotone) if $A \in \mathcal{U}$ and $A \subseteq B \subseteq \mathbb{N}$, then $B \in \mathcal{U}$;
- (closed under intersection) if $A,B \in \mathcal{U}$, then $A \cap B \in \mathcal{U}$;

A filter is an *ultrafilter* if it additionally satisfies

- (maximal) for every $A \subseteq \mathbb{N}$, either $A \in \mathcal{U}$ or $\mathbb{N} \setminus A \in \mathcal{U}$.

Example. Let $\mathcal{U}$ be the family of sets $A \subseteq \mathbb{N}$ that contain 5. Then $\mathcal{U}$ is an ultrafilter. △

An ultrafilter $\mathcal{U}$ is *principal* if there is no $n_0 \in \mathbb{N}$ such that every $A \in \mathcal{U}$ satisfies $n_0 \in A$, and is *free* otherwise. Equivalently, $\mathcal{U}$ is free if it does not contain any finite set.

Theorem 6.14.4. *There is a free ultrafilter $\mathcal{U} \subset \mathcal{P}(\mathbb{N})$.*

Proof. Let $\mathcal{F}_0$ be the filter that contains those sets A whose complement $\mathbb{N} \setminus A$ is finite, and let Z be the set of those filters that contain $\mathcal{F}_0$. Then, Z forms a partial order for the relation $\subseteq$ and furthermore satisfies the hypotheses of Zorn's lemma. Then, any maximal filter $\mathcal{U}$ in Z is a free ultrafilter. (If it wasn't an ultrafilter, because say, $C \in \mathcal{U}$ and $\mathbb{N} \setminus C \notin \mathcal{U}$, then

$$\mathcal{U}' = \big\{(C \cap A) \cup T : A \in \mathcal{U}, T \subseteq \mathbb{N} \setminus C\big\}$$

is a larger filter than $\mathcal{U}$.) ■

6.14.3.1 Ultraproducts and the Hyperreals

Using ultrafilters to construct ultraproducts of structures is ubiquitous in model theory. We will only define the hyperreals here, which are the ultraproduct of $\mathbb{N}$ many copies of $\mathbb{R}$.

Let $(X_n)_{n\in\mathbb{N}}$ be non-empty sets, and let $\mathcal{U}$ be a free ultrafilter on $\mathbb{N}$. Then, their *ultraproduct*

$$\prod_{\mathcal{U}} X_n$$

consists of the equivalence classes of the standard product

$$\prod_{n\in\mathbb{N}} X_n$$

under the equivalence relation $\sim_{\mathcal{U}}$ where $(x_i) \sim_{\mathcal{U}} (y_i)$ if and only if $\{i \in \mathbb{N} : x_i = y_i\} \in \mathcal{U}$.

Notice that if, for example, $\mathcal{U}$ is the principal ultrafilter of sets that contain 5, then the ultraproduct is just the set X_5.

The set $\mathbb{R}^*$ is the ultraproduct of the field of real numbers $\mathbb{R}$ under any free ultrafilter $\mathcal{U}$. That is,

$$\mathbb{R}^* = \prod_{\mathcal{U}} \mathbb{R} = \mathbb{R}^{\mathbb{N}}/{\sim_{\mathcal{U}}}$$

consists of the equivalence classes of real sequences $(x_n)_{n=0}^{\infty}$ under the equivalence relation $\sim_{\mathcal{U}}$ where $(x_i) \sim_{\mathcal{U}} (y_i)$ if and only if $\{i \in \mathbb{N} : x_i = y_i\} \in \mathcal{U}$.

Since the complement of finite sets are always in $\mathcal{U}$, changing finitely many entries of a sequence does not change its equivalence class.

There is also a natural embedding $E : \mathbb{R} \to \mathbb{R}^*$ defined by

$$E(r) = [r,r,r,r,\ldots]$$

and we can also extend the ordering relation $<$ from $\mathbb{R}$ to $\mathbb{R}^*$. We have

$$[x_0,x_1,x_2,x_3\ldots] < [y_0,y_1,y_2,y_3,\ldots]$$

if $\{n \in \mathbb{N} : x_n < y_n\} \in \mathcal{U}$. This ordering is a total ordering on $\mathbb{R}^*$.

There are hyperreals greater than any real number. For instance,

$$[1,2,3,4,5,\ldots] > E(r)$$

for any $r \in \mathbb{R}$. There are also infinitesimal hyperreals, smaller than any non-zero real. For instance, the hyperreal

$$\varepsilon = [1,\tfrac{1}{2},\tfrac{1}{3},\tfrac{1}{4},\tfrac{1}{5},\ldots]$$

satisfies $E(0) < \varepsilon < E(r)$ for any $r > 0$.

Addition and multiplication also extend naturally to the hyperreals by being applied pointwise, and these extensions satisfy their usual properties. For instance, the multiplicative inverse of $[0,1,2,3,4,\ldots]$ is $[0,1,\frac{1}{2},\frac{1}{3},\frac{1}{4},\ldots]$, and their product is $[0,1,1,1,1,\ldots] = [1,1,1,1,1,\ldots]$.

Theorem 6.14.5. $\mathbb{R}^*$ *is a field.*

Conversely, the standard product $\prod_{n\in\mathbb{N}} \mathbb{R}$ is not a field: while addition and multiplication can be defined, not every non-zero sequence has a multiplicative inverse. (It is still a commutative ring, though with no obvious linear ordering.)

Theorem 6.14.6 (Łoś' Theorem, Fundamental Theorem of Ultraproducts)**.**

*(i) (*General Statement*) A first-order formula is true in the ultraproduct $\prod_{\mathcal{U}} X_n$ if and only if the set of indices n for which the formula holds in X_n is in $\mathcal{U}$.*

*(ii) (*Transfer Theorem for the Hyperreals*) Let φ be a first-order formula expressible in the language of the real numbers (e.g. using $0,1,+\,,-\,,\cdot\,,/,<$, etc.). Then, φ holds for the hyperreals $\mathbb{R}^*$ if and only if it holds for the reals $\mathbb{R}$.*

On the other hand, $\mathbb{R}$ and $\mathbb{R}^*$ really are distinct fields. The hyperreals have many interesting properties that are not first-order definable in the language of the reals. For instance, $\mathbb{R}$ is *Archimedian* (Theorem 11.4.1), but $\mathbb{R}^*$ is not: the element $\omega = [1,2,3,4,\ldots]$ is larger than any of $[1,1,\ldots]$, $[1,1,\ldots]+[1,1,\ldots]$, $[1,1,\ldots] + [1,1,\ldots] = [1,1,\ldots]$, and so on.

6.14.4 Continuum Hypothesis

Is it true that a set $A \subset \mathbb{R}$ is either countable or there is a bijection $A \to \mathbb{R}$? That is, is it true that there is no cardinal κ with $\aleph_0 < \kappa < 2^{\aleph_0}$? Or equivalently, is it true that $\mathfrak{c} = \omega_1$?

This last statement is the *continuum hypothesis*, and it has been shown to be independent from the axioms of ZFC set theory. That is, there are models of set theory in which CH holds, and models in which it fails.

6.14.5 Borel sets, σ-algebras, and ω_1

Let G be a finitely generated group with generators $g_1,\ldots,g_r$. For $g \in G$, define $|g|$ (unrelated to the order of the element) to be the minimal $k \in \mathbb{N}$ such that g can be written as the product of k many generators or their inverses.

Letting $G_0 = \{1\}$ and defining for each $1 \leq k \in \mathbb{N}$ by recursion

$$G_k = \bigcup_{i=1,2,\ldots,r} g_i G_{k-1} \cup \bigcup_{i=1,2,\ldots,r} g_i^{-1} G_{k-1}$$

we see that $|g|$ is the minimal k such that $g \in G_k$.

The point is that generating a group is a process of length ω. Generating a σ-algebra takes considerably longer.

Let X be a set. A collection $\mathcal{F} \subseteq \mathcal{P}(X)$ is a *σ-algebra* if it is closed under countable unions, countable intersections, and complementation. That is,

- If $A_1,A_2,\ldots \in \mathcal{F}$, then $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}$ and $\bigcap_{n\in\mathbb{N}} A_n \in \mathcal{F}$;
- If $A \in \mathcal{F}$, then $X \setminus A \in \mathcal{F}$;

A *Borel set* is any subset of a topological space X that can be formed by taking countable unions, countable intersections, and complements of open sets (or equivalently, closed sets) in X.

For any topological space X, the collection of all Borel sets on X forms a σ-algebra called the *Borel algebra* of X. Equivalently, the Borel algebra on X is the smallest σ-algebra containing all open sets (or equivalently, all closed sets).

We consider the Borel algebra on $\mathbb{R}$.

We define the *rank* of a Borel set recursively. Open and closed sets have rank 0 and form the set $\mathcal{B}_0$. Then, for $1 \leq \alpha < \omega_1$, a Borel set has rank *at most* α if it is the union of intersection of countably many Borel sets of rank *less than* α (or a complement of such a set). That is, $\mathcal{B}_\alpha$ – the family of Borel sets of rank at most α – consists of those sets that can be obtained as the union of intersection of countably many sets from

$$\bigcup_{\beta<\alpha} \mathcal{B}_\beta$$

or a complement of such a set.* We say that a Borel set B has rank α if α is the minimal ordinal such that $B \in \mathcal{B}_\alpha$.

Theorem 6.14.7.

$$\mathcal{B} = \bigcup_{\alpha<\omega_1} \mathcal{B}_\alpha$$

That is, every Borel set has a rank that is an ordinal less than ω_1.

Proof. Since every open (and every closed) set is Borel, we have $\mathcal{B} \supseteq \mathcal{B}_0$. Suppose that

$$\mathcal{B} \not\supseteq \bigcup_{\alpha<\omega_1} \mathcal{B}_\alpha$$

Then, there is a minimal α such that $\mathcal{B} \not\supseteq \mathcal{B}_\alpha$. Then,

$$\mathcal{B} \supseteq \bigcup_{\beta<\alpha} \mathcal{B}_\beta$$

and since every set in $\mathcal{B}_\alpha$ can be obtained as a union, intersection, or complement of countably many sets from the right side, and $\mathcal{B}$ is a σ-algebra, we must have that $\mathcal{B} \supseteq \mathcal{B}_\alpha$, which is a contradiction. So,

$$\mathcal{B} \supseteq \bigcup_{\alpha<\omega_1} \mathcal{B}_\alpha$$

For the other direction, all we have to show is that

$$\mathcal{B}' \coloneqq \bigcup_{\alpha<\omega_1} \mathcal{B}_\alpha$$

is a σ-algebra. If $A \in \mathcal{B}_\alpha$, then $\mathbb{R} \setminus A \in \mathcal{B}_\alpha$, so the complement is clear. Now suppose that $A_1, A_2, A_3, \ldots \in \mathcal{B}'$, so $A_i \in \mathcal{B}_{\alpha_i}$ for some ordinals $\alpha_i < \omega_1$.

Let $\alpha = \sup_i \alpha_i = \bigcup_i \alpha_i$. Then, α is the countable union of countable sets, so $\alpha < \omega_1$, and $\bigcup_i A_i$ and $\bigcap_i A_i$ have rank at most $\alpha + 1$, and are thus elements of $\mathcal{B}_{\alpha+1}$. This shows that $\mathcal{B}'$ is a σ-algebra. ■

This *Borel hierarchy* is usually split into not one, but two well-ordered sequences of length ω_1. To simplify, we have merged these two sequences into one here. The Borel hierarchy can also be used to prove that there are only continuum many Borel sets on $\mathbb{R}$.

* Actually, since $\mathcal{B}_0$ is closed under complementation, this last part of the sentence can be omitted and we get the same family of sets.

6.14.6 Cantor-Bendixson Theorem

In §6.1, we iteratively described the derived sets of closed subsets of $\mathbb{R}$. With ordinals, we can define this transfinite sequence more formally.

Recall that $D(A)$ is the set of points in A that are not isolated.

Let $A \subset \mathbb{R}$ be closed. Let $D^0(A) = A$, and define for any ordinal α

$$D^{\alpha+1}(A) = D\big(D^{\alpha}(A)\big)$$

For limit ordinals α, let

$$D^{\alpha}(A) = \bigcap_{\beta<\alpha} D^{\beta}(A)$$

When $D^{\alpha+1}(A) = D^{\alpha}(A)$, then $D^{\alpha}(A)$ has no isolated points, and obviously for any $\beta > \alpha$, we have $D^{\beta}(A) = D^{\alpha}(A)$. Any ordinal $\alpha < \omega_1$ can be the smallest ordinal when this happens first.

However, this must happen at some point strictly before ω_1. This follows from the next theorem.

A set is *perfect* if it is closed and has no isolated points.

Theorem (Cantor-Bendixson)**.** *Every closed set $A \subset \mathbb{R}$ can be uniquely written as the union of a perfect set and a countable set.*

6.15 Axiom of Regularity

We have proved that for every natural number n, we have $n \notin n$. However, there is an additional axiom that implies that no set is an element of itself.

Axiom of Regularity.
Every non-empty set X has an element x such that $x \cap X = \varnothing$. That is:

$$\forall X\big(X \neq \varnothing \to \exists x : x \in X \wedge x \cap X = \varnothing\big)$$

That is, x does not contain any element of X. This axiom is often stated as, "*every non-empty set has an $\in$-minimal element.*"

Theorem 6.15.1 (Corollaries of Regularity)**.**

(i) *For every set x, $x \notin x$.*

(ii) *There are no sets a and b such that $a \in b$ and $b \in a$.*

Proof.

(i) Suppose $x \in x$. Let $X = \{x\}$ (this is a set by the axiom of pairing). Since x is the only element of X, we must have $x \cap X = \varnothing$ by the axiom of regularity. However, $x \in x$ and $x \in X$, so $x \in x \cap X \neq \varnothing$.

(ii) Suppose $a \in b$ and $b \in a$. Let $X = \{a,b\}$. By regularity, either $a \cap X = \varnothing$ or $b \cap X = \varnothing$. However, neither hold, since $a \in b \cap X$ and $b \in a \cap X$.

■

The axiom of regularity is equivalent to the statement that there is no infinite sequence of sets, $x_0,x_1,x_2,\ldots$ such that

$$x_0 \ni x_1 \ni x_2 \ni \cdots$$

Since a sequence is just a function from ω, more precisely, there is no function f with domain ω such that

$$f(0) \ni f(1) \ni f(2) \ni \cdots$$

Theorem 6.15.2. *The axiom of regularity is equivalent to the statement that there are no functions f with domain ω such that $f(0) \ni f(1) \ni f(2) \ni \cdots$.*

Proof. Let $X = \operatorname{ran}(f)$. By regularity, $x \cap X = \varnothing$ for some $x \in X$. But $x = f(n)$ for some $n \in \omega$, so $f(n+1) \in x \cap X \neq \varnothing$.

The other direction requires the axiom of choice. Let X be a non-empty set, and let $x_0 \in X$. If $x_0 \cap X = \varnothing$, then take $x = x_0$. Otherwise, choose some $x_1 \in x_0 \cap X$ and check if $x_1 \cap X = \varnothing$. If so, take $x = x_1$. Otherwise, choose some $x_2 \in x_1 \cap X$. And so on. ■

6.15.1 Cumulative Hierarchy and Rank

Recall the hierarchy of sets described in §6.2.1. We did not end up needing atoms, so we restate the atomless version of the hierarchy of sets more formally here.

Informally, we defined V_0 to be a certain set, then recursively defined $V_1 = \mathcal{P}(V_0)$, $V_2 = \mathcal{P}(V_1)$, and so on. We formalise the "and so on" part using ordinals.

Let

$$V_0 = \varnothing$$

(This is slightly different from the introduction, where we instead had $V_0 = \{\varnothing\}$.) Then, for each ordinal α, define

$$V_{\alpha+1} = \mathcal{P}(V_\alpha)$$

For limit ordinals α, define

$$V_\alpha = \bigcup_{\beta<\alpha} V_\beta$$

(Proving, by a version of transfinite recursion, that this is well-defined is quite some work.)

Theorem 6.15.3. *The axiom of regularity is equivalent to the statement that every set appears in the Cumulative Hierarchy. That is, for every set A, there is an ordinal α such that $A \in V_\alpha$.*

The *rank* of a set A is defined as

$$\operatorname{rank}(A) = \min\{\beta : A \in V_\beta + 1\}$$

The previous theorem implies that every set has a rank.

Notice that the smallest ordinal β for which $A \in V_\beta$ can never be a limit ordinal, as the set V_β for a limit ordinal β is just the union of V_γ for $\gamma < \beta$. This is why the definition is given with the minimal β for which $A \in V_{\beta+1}$. This way, for every ordinal α, there is a set with rank α; in fact, the ordinal α has rank α.

6.16 Condensed List of ZFC Axioms

Axiom of Extensionality.
If two sets have exactly the same members, then they are equal:

$$\forall X \forall Y \big(\forall z (z \in X \leftrightarrow z \in Y) \to x = y\big)$$

Axiom of the Empty Set.
There exists a set with no elements:

$$\exists E \forall x : x \notin E$$

Axiom of Pairing.
For any two sets u and v, there exists a set that contains exactly u and v as elements.

$$\forall u \forall v \exists X \forall x \big(x \in X \leftrightarrow (x = u \lor x = v)\big)$$

Axiom of the Power Set.
For any set u, there is a set whose elements are exactly the subsets of u:

$$\forall u \exists P \forall s (s \subseteq u \leftrightarrow s \in P)$$

or omitting the abbreviation $\subseteq$,

$$\forall u \exists P \forall s \big(\forall x (x \in s \to x \in s) \leftrightarrow s \in P\big)$$

Axiom Schema of Specification.
Let φ be any formula that does not contain the variable name B and has only bound variables, except for $x, t_1, \ldots, t_k$. Then, the following is an axiom:

$$\forall t_1 \forall t_2 \ldots \forall t_k \forall A \exists B \forall x \big(x \in B \leftrightarrow (x \in A \land \varphi)\big)$$

That is, for any property φ of x and any set A, there exists a set B that contains exactly the elements of A for which $\varphi(x)$ holds, and φ may depend on additional parameters $t_1, \ldots, t_k$.

Axiom of Union.
For any set A, there exists a set B whose members are precisely the members of the members of A:

$$\forall A \exists B \forall x \big(x \in B \leftrightarrow \exists y (y \in A \land x \in y)\big)$$

Axiom of Infinity.
There is an inductive set:

$$\exists A \big(\varnothing \in A \land \forall x (x \in A \to x^+ \in A)\big)$$

or omitting $\varnothing$ and x^+,

$$\exists A \big(\exists e (\forall z : z \notin e) \land e \in A \land \forall x (x \in A \to x \cup \{x\} \in A)\big)$$

Axiom Schema of Replacement.
The image of a set under a class-function is a set; if φ is any formula that does not contain B, then:

$$\forall A \Bigg(\underbrace{\forall x \forall y \forall y' \Big(\big(x \in A \land \varphi(x,y) \land \varphi(x,y')\big) \to y = y' \Big)}_{\varphi \text{ is a class-function on at least } A} \to \underbrace{\exists B \forall y \Big(y \in B \leftrightarrow \exists x (x \in A \land \varphi(x,y)) \Big)}_{\text{there is a set } B \text{ consisting of } \varphi\text{-images of elements of } A} \Bigg)$$

Axiom of Regularity.
Every non-empty set X has an element x such that $x \cap X = \varnothing$. That is:

$$\forall X \big(X \neq \varnothing \to \exists x : x \in X \wedge x \cap X = \varnothing \big)$$

Axiom of Choice (first form).
For any relation R, there exists a function $F \subseteq R$ such that $\operatorname{dom}(F) = \operatorname{dom}(R)$.

Axiom of Choice (second form).
Let $\mathcal{S}$ be a set with $\varnothing \notin \mathcal{S}$. Then, there is a *choice function* for $\mathcal{S}$. That is, a function $\sigma : \mathcal{S} \to \bigcup \mathcal{S}$ such that $\sigma(A) \in A$ for all $A \in \mathcal{S}$.

Axiom of Choice (third form).
Let $\mathcal{A}$ be a set of pairwise disjoint non-empty sets. Then, there exists a set C who has as a member exactly one element from each member of $\mathcal{A}$. That is, for each $B \in \mathcal{A}$, $|C \cap B| = 1$.

Axiom of Choice (fourth form).
For any indexing set I and function H with domain I, if $H(i) \neq \varnothing$ for all $i \in I$, then

$$\mathop{\times}_{i \in I} H(i) \neq \varnothing$$

Well-Ordering Principle (Cantor)**.** (Equivalent to Choice)
Every set is well-orderable. That is, given any set X, there is a relation $<$ on X such that $(X, <)$ is well-ordered.

Cardinal Comparability. (Equivalent to Choice)
For any sets A and B, we have $|A| \leq |B|$ or $|B| \leq |A|$. That is, there is an injective function $A \to B$ or there is an injective function $B \to A$.
Equivalently, for any two cardinals κ and λ, we have $\kappa \leq \lambda$ or $\lambda \leq \kappa$.

Zorn's Lemma. (Equivalent to Choice)
Let $(Z, \preceq)$ be a poset, and suppose that every chain $C \subseteq Z$ has an upper bound. Then, Z has a maximal element.

Chapter 7

Combinatorics I

> *"If I walk randomly on a grid, never visiting any square twice, placing a marble every N steps, on average how many marbles will be in the longest line after $N \times K$ steps? Somehow the answer is important in like three unrelated fields."*
>
> — Cueball (xkcd 2529), Randall Munroe

There is a substantial amount of disagreement about the scope of combinatorics, but very broadly speaking, combinatorics is the branch of mathematics that deals with counting and discrete structures. In any case, combinatorics is well known for the sheer breadth of the problems it handles.

Most topics in mathematics have a generally clear goal, such as the prime number theorem in number theory, and if such a goal doesn't exist, there is usually a narrow focus, as in group theory. Combinatorics is not like that – it is often described as a collection of unrelated problems, some of which we think have solutions. Many problems in combinatorics end up being embedded into many other disparate areas of maths, and there isn't a general theory for many combinatorial problems. Combinatorics is about solving problems – it is about techniques rather than specific results.

Because we will often be using sets purely to count things, we introduce the reasonably standard notation of writing $[n]$ to denote the set $\{1,2,\ldots,n\}$. Note that $[n]$ is sometimes taken to include 0 in other applications, but in this chapter, we will take $[n]$ to exclude 0.

7.1 Introduction

The *multiplication principle* states that if you have n options for one item, and m for a second item, then the total number of ways to choose both items is $n \times m$. In general, if there are k items to be chosen, with n_k options for the kth item, then there are $n_1 \times n_2 \times \cdots \times n_k$ ways to choose the k items. We call these *combinations* of the items. Combinations do not care about the ordering of the items in question.

In contrast, say you are arranging 5 different books on a shelf, each with a different colour, and all snugly fitting together. How many ways are there to arrange the books on the shelf?

We call these different orderings *permutations*, so an equivalent question is to ask how many permutations of 5 objects there are.

For the first position on the shelf, we have 5 options to pick from – any book could go in the first place. Then, for the next space, we have 4 options to pick from, then for the next, 3. We multiply all of these together, giving $5 \times 4 \times 3 \times 2 \times 1 = 120$ ways of arranging the 5 books.

Particularly in combinatorics, multiplying sequences of integers occurs often enough that we have notation for this called the *factorial*: $n! \coloneqq n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$. This function comes up in a lot of places outside of combinatorics, but here, we restrict the factorial function to take natural inputs only. One thing of note is that $0! = 1$, which corresponds to the idea that if you have zero objects, there is only one permutation – having zero objects.

Another point to remember, is that there is only one *combination* of books here. Regardless of how we arrange them, we will end up with the same 5 books on the shelf – combinations do not care about order.

Now, what if we had 8 different books, but still only 5 spaces on the shelf? Then, we have 8 options for the first position, then 7 for the second,..., down to 4 for the fifth space. We can compactly write this as $\frac{8!}{3!}$.

$$\text{Total permutations} = \frac{(\text{Number of items we can choose from})!}{(\text{Number of items we \underline{can't} choose})!}$$

Or, if we have n items and k spaces, these are called k-permutations of n,[*] and are denoted with $P(n,k)$, nPk, or P^n_k, and

$$P^n_k = \frac{n!}{(n-k)!}$$

Now, say we have a shelf with 3 spaces, and 7 different books. How many ways to choose 3 books out of 7 do we have?

We know there are $P^7_3 = \frac{7!}{(7-3)!} = 210$ permutations. Each set of 3 objects can be arranged in $3! = 6$ ways, so we've counted each combination exactly 6 times, so the number of combinations is $\frac{7!}{4!3!} = 35$.

$$\text{Total combinations} = \frac{(\text{Number of items we can choose from})!}{(\text{Number of items we \underline{can} choose})! \times (\text{Number of items we \underline{can't} choose})!}$$

We call these k-combinations of n objects, denoted $C(n,k)$, C^n_k, nCk,, or $\binom{n}{k}$.

$$C^n_k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Now, we have 5 books again, with 5 spaces on the shelf, but this time, 3 of the books are indistinguishable. How many permutations do we have now?

We still start with the $5! = 120$ permutations from before, but now, every permutation has set of matching versions with the indistinguishable books swapped around. Since those permutations are indistinguishable, we divide them out, giving $\frac{5!}{3!} = 20$ total permutations.

In general, for permutations with repeated elements,[†] we start as if the items are all distinguishable, then divide out by the repeated elements swapping with each other:

$$\text{Total permutations with repeated objects} = \frac{(\text{Total number of objects})!}{(\text{Group 1})! \times (\text{Group 2})! \times (\text{Group 3})! \cdots}$$

Example. How many distinct anagrams of the word "MISSISSIPPI" are there?

The word is 11 letters long, and we have 4 groups of letters, M, I, S and P with multiplicities 1, 4, 4, and 2, respectively, so there are,

$$\frac{11!}{1!4!4!2!} = 34\,650$$

such anagrams. △

[*] A "permutation" in more general contexts is a bijection from a set to itself, and should use all elements of the set, so this is a slightly different notion of a permutation.

[†] Since sets cannot contain duplicates of elements, we consider structures called "multisets" instead, and these permutations are technically called *multiset permutations*.

Of course, when rolling 10 D6 dice, there are 6 possible values for the first die, 6 for the second, 6 for the third, and so on, giving 6^{10} possible permutations, but how many possible *combinations* are there when rolling 10 D6 dice? (More properly, we call these "k-multicombinations" or "k-multisubsets" of n.)

We let x_i denote the number of dice in with value i per multisubset. The question is then to find the number of non-negative integer solutions[*] to $x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 10$.

We use a method called the *stars and bars*. A solution to the equation can be represented with x_1 symbols called *stars*, followed by a separator called a *bar*, then x_2 more stars, another bar, and so on. For example, the combination of dice values, 1112234446 (order doesn't matter) would be represented as $\star\ \star\ \star \,|\, \star\ \star \,|\, \star \,|\, \star\ \star\ \star \,||\, \star$.

In general, any valid solution to $\sum_{i=1}^{k} x_i = n$ is represented by n stars and $k-1$ separating bars, so the number of solutions is the number of permutations of the n stars and $k-1$ bars, which we know from above is,

$$\text{Number of combinations of } n \text{ objects with } k \text{ choices each} = \frac{(n+k-1)!}{n!(k-1)!} = \binom{n+k-1}{k-1}$$

so there are $\binom{6+10-1}{6-1} = 5\,005$ different combinations you could get from rolling 10 D6 dice.

These are all examples of simple problems in enumerative combinatorics – problems where we want to count some objects, or count how many ways we can do something. We will systematically classify these problems later, but first, we give a common presentation of these types of problems.

7.1.1 Balls and Boxes

We met a few different situations so far where we choose a selection of objects, subject to some constraints. One common framing for these kinds of problems is to count the number of ways to place balls into boxes, with various restrictions:

- Suppose we have k distinguishable, or *labelled*, balls (e.g., they have the integers $1,\ldots,k$ written on them), and n labelled boxes. How many ways are there to distribute the balls into the boxes?
- (Indistinguishable balls/boxes) What if the balls are now indistinguishable, or *unlabelled*, but the boxes are labelled? What about the reverse situation? What if neither balls nor boxes are labelled?
- (Restricted capacities) What if each box can only hold at most 1 ball? Or 2 balls? Or N balls?
- (Semi-distinguished balls/boxes) Suppose we have k_1 red balls, k_2 green balls,…, and/or n_1 blue boxes, n_2 yellow boxes,…
- (General case) Suppose we have k_1 balls of colour c_1, k_2 balls of colour c_2,…, k_r balls of colour c_r; and n_1 boxes of colour d_1 with capacities $N_{1,1},\ldots,N_{1,n_1}$, and n_2 boxes of colour d_2 with capacities $N_{2,1},\ldots,N_{2,n_2},\ldots$, and n_s boxes of colour d_s, with capacities $N_{s,1},\ldots,N_{s,n_s}$. How many ways are there to distribute the balls into the boxes?

Even more generally, we can impose *minimum* capacities for each box (each box must receive at least 1 ball, or 2 balls, or even more generally, $M_{1,1},\ldots,M_{1,n_1},M_{2,1},\ldots,M_{s,n_s}$ balls).

The general case is rather difficult and we will not be tackling it here, but the main point is that many other problems can be rephrased as to be about balls and boxes.

For instance, how many sequences of length k are there whose entries are $1,\ldots,n$? This problem is equivalent to placing k labelled balls into n labelled boxes; there are n^k sequences/distributions in both problems.

[*] Such equations are called *Diophantine* equations.

If we have two counting problems that seem to be "the same" like this, sometimes we can show a relationship between them by matching up the objects being counted. How can you match up the sequences with the ball arrangements in the previous problem?

Such a matching is called a *bijection*.

7.2 Bijective Proofs

In the introduction, we met a couple of different formulae, such as,

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

If we were just given this definition arbitrarily, it would not be clear that this function always returns an integer. After all, we have some large integer on the numerator being divided by another large integer on the denominator. However, it must be an integer, because the number of ways to choose k objects from n must be a whole number!

It is surprisingly common in combinatorics to have an expression that we suspect is an integer like this, and indeed, one way to prove that such an expression is an integer is to find a counting problem it is the answer to.

Lemma (Pascal)**.**

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$$

We can prove this in two ways. The first is to expand the expression, and prove the equation algebraically:

Proof.

$$\begin{aligned}
\binom{n}{k} &= \frac{n!}{k!(n-k)!} \\
&= \frac{(n-1)!n}{k!(n-k)!} \\
&= \frac{(n-1)!\big((n-k)+k\big)}{k!(n-k)!} \\
&= \frac{(n-1)!(n-k)}{k!(n-k)!} + \frac{(n-1)!k}{k!(n-k)!} \\
&= \frac{(n-1)!}{k!(n-k-1)!} + \frac{(n-1)!}{(k-1)!(n-k)!} \\
&= \frac{(n-1)!}{k!\big((n-1)-k\big)!} + \frac{(n-1)!}{(k-1)!\big((n-1)-(k-1)\big)!} \\
&= \binom{n-1}{k} + \binom{n-1}{k-1}
\end{aligned}$$

■

While this proof is valid, it isn't very elegant, and moreover, it doesn't really provide any insight as to why the formula is true. A better proof is to show that the left and right sides of the equation count the same things in possibly different ways:

Proof. $\binom{n}{k}$ is the number of ways to select a subset of k elements from the set $[n]$. For each subset, we could choose to include the element 1, then pick $k-1$ elements from the remaining $n-1$, or we could choose to exclude the element 1, and pick k elements from the remaining $n-1$, so there are $\binom{n}{k-1}+\binom{n-1}{k-1}$ subsets of size k. Because these count the same thing, we have

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$$

as required. ■

Such a proof is called a proof by *double counting*, because we show two expressions are equal by showing that they are two ways of counting the same thing.

Another related way to show that two expressions are equal is to find sets of things the two expressions count, then exhibit a bijection between the sets. These proofs are called *bijective* or *combinatorial* proofs. (A proof by double counting is then a special case of a combinatorial proof where the two sets coincide.)

These will be our main techniques for solving enumerative combinatorial problems.

Here is another result that can be proved with a combinatorial approach:

Theorem 7.2.1. *The following equation holds for all $n \geq 0$:*

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n$$

We give both the algebraic and bijective proofs.

Proof. Let $P(n)$ be the statement that $\sum_{k=0}^{n} \binom{n}{k} = 2^n$. We induct on n. $P(0)$ holds as $\binom{0}{0} = \frac{0!}{0!0!} = 2^0$. Suppose $P(n)$ holds for an arbitrary fixed $n \geq 0$. Then,

$$\begin{aligned}
\sum_{k=0}^{n+1} \binom{n+1}{k} &= \binom{n+1}{n+1} + \sum_{k=0}^{n} \binom{n+1}{k} \\
&= 1 + \sum_{k=0}^{n} \left(\binom{n}{k} + \binom{n}{k-1} \right) \\
&= 1 + \sum_{k=0}^{n} \binom{n}{k} + \sum_{k=0}^{n} \binom{n}{k-1} \\
&= 1 + 2^n + \sum_{k=0}^{n} \binom{n}{k-1} \\
&= 2^n + 1 + \sum_{k=-1}^{n-1} \binom{n}{k} \\
&= 2^n + \binom{n}{n} + \binom{n}{-1} + \sum_{k=0}^{n-1} \binom{n}{k} \\
&= 2^n + \binom{n}{n} + \sum_{k=0}^{n-1} \binom{n}{k} \\
&= 2^n + \sum_{k=0}^{n} \binom{n}{k} \\
&= 2^n + 2^n
\end{aligned}$$

$$= 2^{n+1}$$

so $P(n)$ holds for all $n \geq 0$ by induction. ■

Proof. Alternatively, let $[n] = \{1,2,\ldots,n\}$. How many subsets of $[n]$ are there? There are $\binom{n}{k}$ subsets of cardinality k, so we can add these up for $k = 0$ up to $k = n$, so there are $\sum_{k=0}^{n} \binom{n}{k}$ subsets of $[n]$.

Another way to generate these subsets is to choose to include or exclude each element $1,\ldots,n$ in each subset. This gives n binary choices for each subset, so there are 2^n possible subsets. These two methods count the same thing, so they must be equal, and hence

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n$$

as required. ■

As can be seen, bijective proofs are often more informative and concise (but not necessarily easier to come up with!), so we will omit the algebraic proofs from this point forward.

We give a few more results about the choose function.

Theorem (Symmetry)**.** *The following equation holds for all n and k:*

$$\binom{n}{k} = \binom{n}{n-k}$$

Proof. Let R have n elements, and let $S \subseteq R$ be a subset with k elements. Consider the mapping $\phi(S) = R \setminus S$, noting that $|R \setminus S| = |R| - |S| = n - k$.

$\phi\big(\phi(S)\big) = R \setminus (R \setminus S) = S$, so ϕ is an involution and is hence bijective; that is, for every subset of cardinality k, there is a unique subset of cardinality $n - k$ obtained by complementation. There are $\binom{n}{k}$ of the former, and $\binom{n}{n-k}$ of the latter, and the bijection says these quantities are equal. ■

Theorem (Binomial Theorem)**.** *Let n be a positive integer. Then,*

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

Proof. Specifying a term in the expansion of $(a+b)(a+b)\cdots(a+b)$ consists of choosing, for each factor, either the a or b, and there are $\binom{n}{k}$ ways to choose k as and $n-k$ bs. ■

This theorem also explains the alternative name of $\binom{n}{k}$ as the *binomial coefficient* function. Letting $a = b = 1$ in the above also yields another bijective proof of Theorem 7.2.1.

We can write out the binomial coefficients in a triangle:

	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$
$n=0$	1									
$n=1$	1	1								
$n=2$	1	2	1							
$n=3$	1	3	3	1						
$n=4$	1	4	6	4	1					
$n=5$	1	5	10	10	5	1				
$n=6$	1	6	15	20	15	6	1			
$n=7$	1	7	21	35	35	21	7	1		
$n=8$	1	8	28	56	70	56	28	8	1	
$n=9$	1	9	37	84	126	126	84	37	9	1

This triangle is called *Pascal's triangle*, and we can quickly fill in the table using Pascal's identity from before – each entry in the table is the sum of the two entries above it (this is more clear if the triangle is drawn centred).

Note that the nth row also gives the binomial coefficients for an expansion of degree n – for instance, $(a+b)^4 = 1a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + 1b^4$, and the 4th row is 1,4,6,4,1.

There are many patterns and properties of this triangle. For instance, the column $k=1$ consists of all the natural numbers, column $k=2$ consists if the triangular numbers, column $k=3$ of the tetrahedral numbers, and column $k=i$ of the i-simplex numbers.

A lower triangular matrix containing Pascal's triangle can also be obtained by taking the exponential (§33.9.3.2) of the matrix which has the natural numbers along the first subdiagonal, and zero elsewhere:

$$\exp\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 1 & 3 & 3 & 1 & 0 \\ 1 & 4 & 6 & 4 & 1 \end{bmatrix}$$

7.2.1 Multinomial Coefficients

We have previously derived a formula for counting the number of ways to arrange objects with indistinguishable subgroups. We can reframe this in terms of balls and boxes as follows:

If we have n labelled boxes, and n balls with k_1 labelled with colour c_1, k_2 labelled with colour c_2, …, and k_r labelled with colour c_r, where $\sum_{i=1}^{r} k_i = n$, then the number of distinct ways of distributing the balls among the boxes is

$$\frac{n!}{k_1!k_2!\cdots k_r!} = n!\prod_{i=1}^{r}\frac{1}{k_i!}$$

We give this quantity a symbol in analogy with binomial coefficients:

$$\binom{n}{k_1,k_2,\ldots,k_r}$$

and we call this a *multinomial coefficient.*

Theorem (Multinomial Theorem). *Let n be a positive integer. Then,*

$$\left(\sum_{i=1}^{r} a_r\right)^n = \sum_{\sum_{i=1}^{r} k_i = n} \binom{n}{k_1,\ldots,k_r} \prod_{i=1}^{r} a_i^{k_i}$$

7.2.2 Inclusion-Exclusion Principle

A common problem in combinatorics and discrete probability is to find the cardinality of the union of two sets. If we have a set A with n elements, and B with m elements, it could be that A and B are not disjoint, so their union has fewer than $n+m$ elements. We have, however, double counted exactly the elements in their *intersection.*

Theorem (Binary Inclusion-Exclusion)**.** *Let A and B be sets. Then,*

$$|A \cup B| = |A| + |B| - |A \cap B|$$

Proof.

$$\begin{aligned} |A \cup B| &= |A \cup (B \setminus A)| \\ &= |A| + |B \setminus A| \qquad (1) \\ |B| &= |(B \setminus A) \cup (A \cap B)| \\ &= |B \setminus A| + |A \cap B| \qquad (2) \end{aligned}$$

Combining (1) and (2) gives the result. ■

In general, to find the cardinality of the union of n sets, we include the cardinality of the sets, exclude the cardinalities of the pairwise intersections, include the cardinalities of the 3-wise intersections, exclude 4-wise, and continue up to n.

Theorem (Inclusion-Exclusion)**.**

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{\varnothing \neq J \subseteq [n]} (-1)^{|J|+1} \left|\bigcap_{j \in J} A_j\right|$$

Proof. Let $J \subseteq \{1,\dots,r\}$ be the set of A_is containing element x. Then x is counted as $+1$ in term I for all $I \subseteq J$ with an odd number of elements, and x is counted as -1 in term I for all non-empty I with an even number of elements. J has an equal number of odd-cardinality and even-cardinality subsets; if J has odd cardinality, this is easy, otherwise this follows by complementation. Removing the empty set, there is one extra odd-cardinality subset, so x is counted overall as $+1$. ■

Example. A *derangement* is a permutation with no fixed points: no element appears in its original position. How many derangements are there of n objects?

For some fixed n and all $1 \leq k \leq n$, let S_k be the set of permutations of n objects that fix the kth object. The number of derangements is then the total number of permutations, minus the union of these sets, $n! - |\bigcup_{i=1}^{n} S_i|$

Any intersection of a collection of i of these sets then fixes i objects and contains $(n-i)!$ permutations, and there are $\binom{n}{i}$ such collections, so,

$$\begin{aligned} \left|\bigcup_{i=1}^{n} S_i\right| &= \sum_{\varnothing \neq J \subseteq [n]} (-1)^{|J|+1} \left|\bigcap_{j \in J} S_j\right| \\ &= \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} (n-i)! \\ &= \sum_{i=1}^{n} (-1)^{i+1} \frac{n!}{i!(n-i)!} (n-i)! \end{aligned}$$

$$= n! \sum_{i=1}^{n} \frac{(-1)^{i+1}}{i!}$$

So the number of derangements is $n! - n! \sum_{i=1}^{n} \frac{(-1)^{i+1}}{i!} = n! \sum_{i=1}^{n} \frac{(-1)^{i}}{i!}$.[*] △

See also §47.1.2 for an application of the inclusion-exclusion principle in probability.

7.2.3 Twelvefold Way

Recall that a function $f : A \to B$ is:

- *injective* if $f(a_1) = f(a_2)$ implies $a_1 = a_2$;
- *surjective* if for every $b \in B$, there exists a $a \in A$ such that $f(a) = b$;
- *bijective* if it is both injective and surjective.

Note that it is only possible for f to be injective if A has at most as many elements as B; that is, if $|A| \leq |B|$. Similarly, it is only possible for f to be surjective if A has at least as many elements as B; that is, if $|A| \geq |B|$. It is also only possible for f to be bijective if $|A| = |B|$.[†]

A function $f : N \to X$ can be considered from the perspective of N or X, with each giving different combinatorial interpretations:

- The function f *labels* elements of X by elements of N;
- The function f *selects* or *chooses* an element of X for each element of N;
- The function f *groups* the elements of N together that are mapped to the same element of X.

When viewing f as a labelling of the elements of N, we may think of ordering N in sequence, and successively applying labels from X to each element. A requirement that f is injective then means that labels can only be used once; yielding sequences of labels *without repetition*.

If we instead view f as an unordered selection of elements of X, then injectivity gives a similar restriction: the selection must involve k distinct elements of X, so it is a subset of X with k elements – this is exactly an k-combination from before. If we do not require injectivity, the same element of X may occur multiple times in the selection, so we obtain a k-multicombination of X.

More generally, imagine a set X of objects numbered from 1 to x, from which we choose n (by means of a function f), yielding an ordered list of objects. Applying various restrictions to f yield a variety of combinatorial problems:

- Any f – after selecting an item, we replace it and are free to select it again.
- Injective f – after selecting an item, we remove it, so we obtain n distinct items. Note that if $n \geq x$, no such lists can be chosen.
- Surjective f – after selecting an item, we may replace it and select it again, but we require that every item is selected at least once. Note that if $n \leq x$, no such lists can be chosen.

We can also reorder or relabel the lists before counting them:

- Distinct – do not modify the lists.

[*] You might notice that this is the taylor polynomial for e, so another expression for the number of derangements is $\left[\frac{n!}{e}\right]$, where $[x]$ is the nearest integer to x.

[†] This is somewhat circular, because cardinality in set theory is defined by bijections; we write $|K| = |N|$ if there exists a bijection $f : A \to B$, and the more colloquial notion of cardinality as "number of elements" follows from forming a bijection to an initial subset of $\mathbb{N}$. We will leave these subtleties to the chapter on set theory, and not dwell on these matters here; for us, cardinality will just informally be the number of elements of a set.

- S_n orbits* – before counting, sort the list by item number. This has the effect of treat all permutations as the same list. e.g. the lists (1,2,3), (2,3,1), and (3,1,2) would be considered as the list (1,2,3).
- S_x orbits – before counting, renumber the items in the order they were selected, repeating if an item was selected multiple times. e.g. the lists (3,5,3), (3,2,3), (5,8,5), and (8,2,8) would all be relabelled to the list (1,2,1), while (2,2,4), (3,3,2), (9,9,6) would be relabelled to (1,1,2).
- $S_n \times S_x$ orbits – two lists are considered the same if they can be reordered or relabelled as above, and produce the same result. e.g. (2,9,2) and (3,4,4) can both be reordered into (2,2,9) and (4,4,3), then relabelled into the same list (1,1,2).

These different ways of modifying the function and lists yields twelve kinds of enumerative problems. This classification of these problems is called the *twelvefold way*:

f-class	Any f	Injective f	Surjective f
Distinct f	n-sequence in X	n-permutation of X	composition of N with x subsets
S_n orbits $f \circ S_n$	n-multisubset of X	n-subset of X	composition of N with x terms
S_x orbits $S_x \circ f$	partition of N into $\leq x$ subsets	partition of N into $\leq x$ elements	partition of N into x subsets
$S_n \times S_x$ orbits $S_x \circ f \circ S_n$	partition of n into $\leq x$ parts	partition of n into $\leq x$ parts 1	partition of n into x parts

Each of the classes has a formula, some of which we have already seen, and some we will explore later:

f-class	Any f	Injective f	Surjective f
Distinct f	x^n	$x^{\underline{n}}$	$x!S(n,x)$
S_n orbits $f \circ S_n$	$\binom{x+n-1}{n}$	$\binom{x}{n}$	$\binom{n-1}{n-x}$
S_x orbits $S_x \circ f$	$\sum_{k=0}^{x} S(n,k)$	$[n \leq x]$	$S(n,k)$
$S_n \times S_x$ orbits $S_x \circ f \circ S_n$	$p_x(n+x)$	$[n \leq x]$	$p_x(n)$

[*] *Orbits* here refer to the same orbits as in group theory. Enumerative combinatorics is closely related to group actions – in particular, by Burnside's lemma and the orbit-stabiliser theorem.

We can also state these classes in terms of boxes and balls:

f-class	Any f	Injective f	Surjective f
Balls and Boxes labelled	How many ways to place n labelled balls into x labelled boxes?	How many ways to place n labelled balls into x labelled boxes, with every box receiving at most one ball?	How many ways to place n labelled balls into x labelled boxes, with every box receiving at least one ball?
Balls unlabelled, Boxes labelled	How many ways to place n unlabelled balls into x labelled boxes?	How many ways to place n unlabelled balls into x labelled boxes, with every box receiving at most one ball?	How many ways to place n unlabelled balls into x labelled boxes, with every box receiving at least one ball?
Balls labelled, Boxes unlabelled	How many ways to place n labelled balls into x unlabelled boxes?	How many ways to place n labelled balls into x unlabelled boxes, with every box receiving at most one ball?	How many ways to place n labelled balls into x unlabelled boxes, with every box receiving at least one ball?
Balls and Boxes unlabelled	How many ways to place n unlabelled balls into x unlabelled boxes?	How many ways to place n unlabelled balls into x unlabelled boxes, with every box receiving at most one ball?	How many ways to place n unlabelled balls into x unlabelled boxes, with every box receiving at least one ball?

We will spend the rest of this section exploring more of these classes.

7.2.4 Stars and Bars

Earlier, we asked the question of how many *combinations* are possible when rolling n dice each with k faces (our example had $n = 10$ and $k = 6$). Here are some equivalent formulations:

- How many ways can we place n unlabelled balls into k labelled boxes?
- How many combinations of n numbers with repetition from $[k]$ are there?
- How many degree-n monomials in k variables are there?
- How many non-negative integer solutions are there to $\sum_{i=1}^{n} x_i = k$?
- How many ways are there to arrange n stars and $k - 1$ separating bars?

Example. $k = 3$, $n = 3$:

	[3][0][0] [0][1][2]	[2][1][0] [0][0][3]	[1][2][0] [1][0][2]	[0][3][0] [2][0][1]	[0][2][1] [1][1][1]
	111 233	112 333	122 331	222 311	223 123
	x^3 yz^2	x^2y z^3	xy^2 xz^2	y^3 x^2z	y^2z xyz
x_1,x_2,x_3 =	3,0,0 0,1,2	2,1,0 0,0,3	1,2,0 1,0,2	0,3,0 2,0,1	0,2,1 1,1,1
	$\star\star\star\vert\vert$ $\vert\star\vert\star\star$	$\star\star\vert\star\vert$ $\vert\vert\star\star\star$	$\star\vert\star\star\vert$ $\star\vert\vert\star\star$	$\vert\star\star\star\vert$ $\star\star\vert\vert\star$	$\vert\star\star\vert\star$ $\star\vert\star\vert\star$

△

We previously showed that the dice problem is equivalent to the integer solution problem by constructing a sum $\sum_{i=1}^{n} x_i = k$, where x_i is the number of dice with value i (and we have n dice, so the x_i sum to k). Each solution can then be represented as x_i many stars, followed by a separating bar, then x_2 many stars, etc. which gives the neat and tidy formula

$$\binom{n+k-1}{k-1}$$

for all the these problems.

Exercise. Construct an explicit bijection from the original dice problem to these other formulations.

Looking at the balls and boxes formulation, we can count these distributions in a different way. We could first place $0 \leq i \leq n$ balls in the first box, leaving $n - i$ balls to distribute among the $k - 1$ remaining boxes. Applying the formula above with double counting, we have,

$$\binom{n+k-1}{k-1} = \sum_{i=0}^{n} \binom{n+k-2-i}{k-2}$$

Reindexing the variables by $n + k - 2 \to n$ and $k - 2 \to k$, we obtain:

Theorem 7.2.2.

$$\binom{n+1}{k+1} = \sum_{i=0}^{n+k-2} \binom{n-i}{k} = \sum_{i=0}^{n} \binom{i}{k}$$

This identity is sometimes called the *hockey-stick identity* because of the way the relevant terms are arranged on Pascal's triangle.

7.2.5 Set Partitions and Stirling Numbers

Let $k \leq n$. How many ways are there to place n labelled balls into k labelled boxes, if every box must get a ball?

Example. $n = 4$, $k = 2$. Each box must have at least one ball, so we can either place 3 balls in one box, and 1 in the other, or place 2 in each. In the former case, there are $\binom{4}{1} = 4$ ways to pick the ball that is to be on its own, and 2 places to put it, giving 8 total ways. In the latter case, there are $\binom{4}{2} = 6$ ways to choose which two balls to place into the first box (which also fully determines the balls placed into the other box). This gives a total of 14 ways. △

This is the same as counting surjective functions $f : [n] \to [k]$ – the requirement that every box receives a ball is equivalent to every element of the codomain receiving a non-empty preimage under f.

If the boxes are unlabelled, then the answer will simply be lower by a factor of $k!$ – because every box receives a ball, we can simply remove permutations of the boxes.

This unlabelled box form is also equivalent to counting *set partitions* – the number of ways to *partition* $[n]$ into exactly k non-empty subsets. These numbers are written as $S(n,k)$ or

$$\begin{Bmatrix} n \\ k \end{Bmatrix}$$

and are called *Stirling numbers of the second kind.* We also define $S(0,k)$ to be 1 if $k = 0$, and 0 otherwise.

These numbers turn out to be somewhat more complicated to calculate. It is easier to count surjections – $k!S(n,k)$ – then to divide by $k!$. In fact, it is easier to count the functions that are not surjections – $k^n - k!S(n,k)$ – and apply the inclusion-exclusion principle. But how do we transform this into a problem about set unions?

Let A_i be the set of functions $[n] \to [k]$ such that i is not in the image of the function (no ball goes into box i). Then, $|A_i| = (k-1)^n$, since these are just functions from $[n]$ to a set of cardinality $k - 1$ (or in terms of balls and boxes, we have n balls, each with $k - 1$ choices of destination). The intersection $A_i \cap A_j$ is then the set of functions such that no ball goes into box i *nor* j, so $|A_i \cap A_j| = (k-2)^n$. Similarly, $\left|\bigcap_{i \in I} A_i\right| = \big(k - |I|\big)^n$ for all $I \subseteq [k]$. Then,

$$\begin{aligned} k!S(n,k) &= \sum_{\varnothing \neq I \subseteq [k]} (-1)^{|I|-1}(k - |I|)^n \\ &= \sum_{i=1}^{k} (-1)^{i-1}(k-i)^n \binom{k}{i} \end{aligned}$$

and hence

$$S(n,k) = \frac{1}{k!} \sum_{i=0}^{k} (-1)^i (k-i)^n \binom{k}{i}$$

We can write out the Stirling numbers in a triangle:

	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$
$n=0$	1									
$n=1$	1	1								
$n=2$	1	3	1							
$n=3$	1	7	6	1						
$n=4$	1	15	25	10	1					
$n=5$	1	31	90	65	15	1				
$n=6$	1	63	301	350	140	21	1			
$n=7$	1	127	966	1701	1050	266	28	1		
$n=8$	1	255	3025	7770	6951	2646	462	36	1	
$n=9$	1	511	9330	34105	42525	22827	5880	750	45	1

Just like with Pascal's triangle, we might see some patterns in this triangle:

Theorem 7.2.3.

$$S(n,k) = S(n-1,k-1) + kS(n-1,k)$$

Proof. Consider the first element $1 \in [n]$. If 1 is in a part by itself, then there are $S(n-1,k-1)$ ways to partion the rest of the set into $k-1$ non-empty parts. Otherwise, 1 is in a subset with other elements. There are $S(n-1,k)$ ways to partition the elements that are not 1 into k parts, and we then have k choices of subset to place 1 into. ■

Theorem 7.2.4.

$$S(n,n-1) = \binom{n}{2}$$

Proof. Dividing $[n]$ into $n-1$ subsets is equivalent to picking a single subset of cardinality 2, with every other element necessarily forming a singleton part. ■

Theorem 7.2.5.

$$6S(n,3) + 6S(n,2) + 3S(n,1) = 3^n$$

Proof. There are 3^n functions $f : [n] \to [3]$, since there are n inputs and 3 choices of outputs. We can also count these functions according to the cardinalities of their images. $S(n,1)$ of these functions map every input to the same output, and there are $P(3,1) = 3$ ways to do this; $S(n,2)$ of these functions map the inputs to 2 distinct outputs, and there are $P(3,2)$ permutations of the output; and $S(n,3)$ of these functions are surjective, and there are $P(3,3)$ ways to permute the outputs. This covers all possible cases, so

$$6S(n,3) + 6S(n,2) + 3S(n,1) = 3^n$$

as required. ■

For $n \geq 0$, we define the *Bell numbers* B_n to be the number of ways to partition $[n]$ into any number of parts. The Bell numbers are given by the formula:

$$B_n = \sum_{k=0}^{n} S(n,k)$$

and the first few Bell numbers are as follows:

$$1, 1, 2, 5, 15, 52, 203, 877, 4140, 21147, 115975, 678570, 4213597, 27644437, 190899322, \ldots$$

Theorem 7.2.6.

$$B_n = \sum_{k=0}^{n-1} \binom{n-1}{k} B_k$$

Proof. In each partition of $[n]$, removing the part that contains n yields a partition of $0 \leq k < n$ items, and there are $\binom{n-1}{n}$ possible combinations for the k remaining items, with B_k ways to partition each one, so summing over all k yields B_n. ■

7.2.6 Integer Partitions

Let $k \leq n$. How many ways are there to divide n unlabelled balls into k unlabelled boxes, if every box must get a ball? In the previous section, the balls were labelled, but here, all that matters is the total number of balls in each box. This is equivalent to counting the number of ways to write the integer n as a sum of k positive integers, and we write the answer to this problem as $p_k(n)$. These are called *integer partitions of n into k parts* (not to be confused with the *set partitions* explored in the previous section).

Example. $p_3(10) = 8$; there are 8 partitions of 10 into 3 parts:

$$\begin{array}{cccc} 8+1+1 & 7+2+1 & 6+3+1 & 6+2+2 \\ 5+4+1 & 5+3+2 & 4+4+2 & 4+3+3 \end{array}$$

It is standard to write these sums in descending order of summands. △

If we modify the problem to allow any number of boxes, or equivalently, to write n as a sum of any number of positive integers, we obtain the number of *integer partitions* of n, written as $p(n)$. Note that $p(0) = 1$, because the unique partition of 0 is the empty sum, consisting of no parts.

Example. $p(10) = 42$; there are 42 partitions of 10. △

We also define $p_k^{\leq a}(n)$ to be the number of partitions of n into k parts of size at most a, and similarly, $p^{\leq a}(n)$ to be the number of partitions of n into any number of parts of size at most a.

There is no known closed-form expression for $p(n)$, nor any of these other variants. However, we do have some relations that relate them together:

Theorem 7.2.7.

$$p(n) = \sum_{i=1}^{n} p^{\leq i}(n-i)$$

$$p_k(n) = \sum_{i=1}^{n-k+1} p_{k-1}^{\leq i}(n-i)$$

Again, we can write these values as a triangle:

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$
$n=1$	1									
$n=2$	1	1								
$n=3$	1	1	1							
$n=4$	1	2	1	1						
$n=5$	1	2	2	1	1					
$n=6$	1	3	3	2	1	1				
$n=7$	1	3	4	3	2	1	1			
$n=8$	1	4	5	5	3	2	1	1		
$n=9$	1	4	7	6	5	3	2	1	1	
$n=10$	1	5	8	9	7	5	3	2	1	1

And the first few values of $p(n)$:

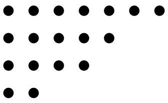
$$1, 2, 3, 5, 7, 11, 15, 22, 30, 42, 56, 77, 101, 135, 176, 231, 297, 385, 490, 627, \ldots$$

While we don't have a closed-form expression for $p(n)$, we do have an interesting asymptotic relationship:

Theorem 7.2.8 (Hardy-Ramanujan)**.**

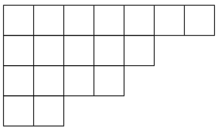
$$p(n) \sim \frac{1}{4\sqrt{3}} \exp\left(\pi\sqrt{\frac{2n}{3}}\right)$$

Here is one way of visualising a partition: draw a horizontal line of dots for each part in the partition, then stack them on top of each other. For example, here is the picture of the partition $18 = 7+5+4+2$:



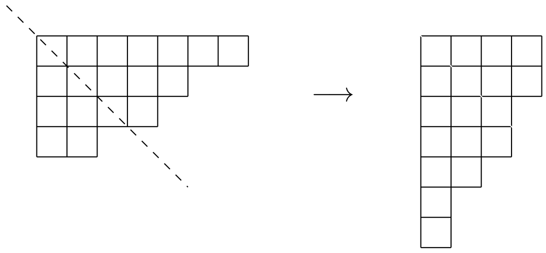
This is called a *Ferrers diagram.*

Instead of drawing dots, we could alternatively draw a line of *boxes*:



This is called a *Young diagram.* These seem very similar to Ferrers diagrams, and they are often also called as such, but this small change turns out to be very useful. Unlike a Ferrers diagram, the boxes in a Young diagram can be filled with various numbers or objects to form structures called *Young tableaux.*

Let λ be a partition of n. The *conjugate partition* $\lambda^\top$ of λ is the partition corresponding to the reflection of the Young diagram of λ over the diagonal. For instance,



so the conjugate of $7+5+4+2$ is given by $4+4+3+3+2+1+1$. This symmetry immediately implies many results about partitions.

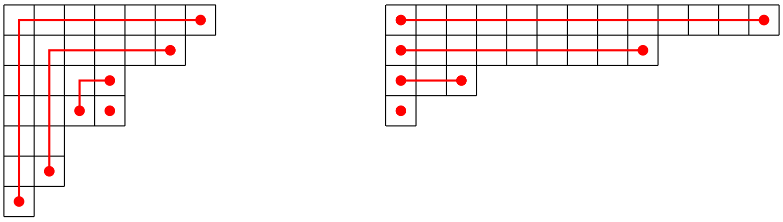
Theorem 7.2.9. *The number of partitions of n into at most k parts is equal to the number of partitions of n into parts of size at most k. That is,*

$$\sum_{i=0}^{k} p_i(n) = p^{\leq k}(n)$$

Proof. The two sets of partitions are in bijection by conjugation. ■

Theorem 7.2.10. *The number of partitions of n into distinct odd parts is equal to the number of self-conjugate partitions of n.*

Consider the following partitions of 26 into $7+6+4+4+4+2+2+1$, which is self conjugate, and and $13+9+3+1$, which consists of distinct odd parts:



By “folding” and “unfolding” along the midpoint of the marked lines, we can transform these partitions into each other. This motivates the strategy for our proof.

Proof. Given a Young diagram of a partition λ, let (i,j) denote the box in the ith row and jth column. We define the *hook length* or *hook number* $h_\lambda(i,j)$ of a box (i,j) to be the number of boxes below and to the right of (i,j), including the box itself. For example, here is a Young tableau filled with the hook lengths of each box:

13	11	8	7	4	3	1
11	9	6	5	2	1	
8	6	3	2			
7	5	2	1			
4	2					
3	1					
1						

Let γ be a self-conjugate partition of n. Because γ is self-conjugate, every hook number along the diagonal must be odd, as the boxes of the form (i,i) must have an equal number of boxes below and to the right of them; adding the box itself then yields an odd number. Hook numbers must also necessarily be decreasing, so the diagonals are also distinct.

Thus, any self-conjugate partition γ corresponds to a partition $\lambda : (\lambda_1 + \lambda_2 + \cdots + \lambda_k)$ of n into distinct odd parts λ_i via the mapping $\lambda_i = h_\gamma(i,i)$ for each $1 \leq i \leq k$. ■

Theorem 7.2.11. *The number of Young tableaux of shape λ, denoted by f^λ is given by*

$$f^\lambda = \frac{n!}{\prod_{i,j} h_\lambda(i,j)}$$

7.2.7 Generating Functions

A *generating function* is a method of encoding information in the *coefficients* of a power series, rather than in the variables of an equation. These encodings lie within the intersection of combinatorics and analysis, allowing us to apply analytic methods to combinatorial problems. This will be more apparent later, but first, we show how to reframe and solve enumerative problems in terms of generating functions.

One basic example of a generating function is as follows. To describe the outcomes when rolling a die, we could encode the number of ways to roll k as the coefficient of x^k; for a standard $D6$, we would have

$$f(x) = x^1 + x^2 + x^3 + x^4 + x^5 + x^6$$

noting that we cannot roll less than 1, nor greater than 6, so those coefficients are zero and the power series is in fact a finite polynomial. In doing this, we have encoded the sample space of a dice roll in a single algebraic object.

Note that x is an *indeterminate*; just a formal symbol that we do not attach any meaning to. More generally, when encoding infinite sequences of data, the resulting power series is also formal, meaning that we do not concern ourselves with issues of convergence, since we do not care about the value of the indeterminate itself.

You'd might ask why we bother having an indeterminate at all, rather than just storing the sequence of coefficients as a pure sequence. However, the utility of this seemingly extraneous addition will become apparent when combining generating functions together:

Example. How many ways can we make £1 out of pennies, 2 pence pieces, and 5 pence pieces?

Classically, you could model this situation with

$$x_1 + 10x_2 + 20x_3 = 100$$

where x_1 is the number of pennies, x_2 is the number of 2 pence pieces, and x_3 is the number of 5 pence pieces. Crucially, the information is contained within the variables themselves.

Let's see how to model this with generating functions.

How many ways can we make the value k with just pennies? We can create a value of 0 with no pennies, 1 with one penny, 2 with two pennies... and so on, and in each case, there is a single way of doing so. (Note that we cannot, however, create negative values.)

We model this by:

$$f_1(x) = x^0 + x^1 + x^2 + x^3 \cdots$$

that is, the coefficient of x^k is the number of ways to create k. We could also terminate this power series at x^{100} if we wanted to, since we are only interested in sums up to £1, but we will later see that the infinite series is actually easier to work with than the finite polynomial.

Considering only 2 pence pieces then similarly gives:

$$f_2(x) = x^0 + x^2 + x^4 + x^6 \cdots$$

and only 5 pence pieces yields the possibilities:

$$f_5(x) = x^0 + x^5 + x^{10} + x^{15} \cdots$$

Now, how many ways could we make the value k with 2 pence pieces *or* 5 pence pieces?

Suppose we make the value $a \leq k$ with 2 pence pieces, and then make the value $b = k - a$ with 5 pence pieces. Then, the number of ways to make k with this particular decomposition is the number of ways to make a with 2 pence pieces *multiplied* by the number of ways to make b with 5 pence pieces. The total number of ways to make k is then given by the sum over all decompositions $a + b = k$.

Consider what happens if we multiply the two power series together. In particular, what generates the coefficient of x^k in the expansion?

This term x^k can be generated by multiplying some x^a in f_2 and some x^b in f_5 with $a + b = k$. In particular, the coefficients of x^a and x^b are multiplied together in this process; and the final coefficient of x^k after collecting terms is given by the sum over all decompositions.

That is, the coefficient of x^k in the product $f_2 f_5$ is precisely the number of ways to make k with 2 and 5 pence pieces.

The generating function for our original problem is thus the product of the three polynomials:

$$g(x) = (x^0 + x^1 + x^2 + x^3 \cdots)(x^0 + x^2 + x^4 + x^6 \cdots)(x^0 + x^5 + x^{10} + x^{15} \cdots)$$

and the solution to the original problem would then be given by the coefficient of x^{100} in $g(x)$. △

Example. How many integer solutions are there to $a + b + c + d = 28$, subject to

$$\begin{gathered} 0 \leq a \leq 10 \\ b \geq 12 \\ 1 \leq c,d \leq 9 \end{gathered}$$

$$g(x) = (1 + x^1 + x^2 + \cdots + x^{10})(x^{12} + x^{13} + x^{14} \cdots + x^{28})(x^1 + x^2 + x^3 + \cdots + x^9)^2$$

The total number of solutions is then given by $[x^{28}]g(x)$. △

Example. I have a large bag of sweets of various colours: red, orange, yellow, green, blue, and violet. How many ways can I select 25 of these sweets such that I have at least two of every colour, an even number of green sweets, at least 7 blue sweets, and no more than 5 violet sweets?

- Red $(x^2 + x^3 + \cdots + x^{25})$
- Orange $(x^2 + x^3 + \cdots + x^{25})$
- Yellow $(x^2 + x^3 + \cdots + x^{25})$
- Green $(x^2 + x^4 + \cdots + x^{24})$
- Blue $(x^7 + x^8 + \cdots + x^{25})$
- Violet $(x^2 + x^3 + x^4 + x^5)$

$$g(x) = (x^2 + x^3 + \cdots + x^{25})^3(x^2 + x^4 + \cdots + x^{24})(x^7 + x^8 + \cdots + x^{25})(x^2 + x^3 + x^4 + x^5)$$

The solution is then given by $[x^{25}]g(x)$. △

Often, people have trouble with enumerative combinatorics because they can think of multiple ways of counting a problem, but run into numerical difficulties, and each method yields a different solution. For instance, in the above, one might attempt to count the combinations of all of the sweets, then subtract some invalid combinations; or alternatively, one could build up the collection of valid combinations one criterion at a time, or one colour at a time, or one sweet at a time.

Generating functions allow us to transform abstract counting problems into concrete questions about power series, i.e. "*what is the kth coefficient of some polynomial $g(x)$?*"

7.2.7.1 The Extended Binomial Theorems

We have seen various examples of how to represent problems as generating functions, and how to extract their solutions in terms of coefficients, but how do we actually *compute* these coefficients *efficiently* without having to expand everything out?

First, recall the binomial theorem:

Theorem (Binomial Theorem)**.**

$$(x + y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

This allows us to efficiently compute coefficients in at least binomials:

Example. What is $[x^4](3x + 2)^9$?

$$\begin{aligned}[x^4](3x+2)^9 &= \binom{9}{4}(3x)^4(2)^{9-4} \\ &= \left(\frac{9!}{5!4!}\right)(81x^4)(32) \\ &= 326\,592x^4\end{aligned}$$

(Because these can become extremely large, we will often choose to leave our coefficients in terms of choose functions/unexpanded products, rather than a numerical result, as shown here.) △

However, this isn't of much use in the problems shown in the previous section.

We can rearrange the choose function as follows:

$$\begin{aligned}\binom{n}{k} &= \frac{n!}{k!(n-k)!} \\ &= \frac{n(n-1)(n-2)\cdots(n-k+1)(n-k)!}{k!(n-k!)} \\ &= \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}\end{aligned}$$

Because n is no longer being input into a factorial, we can now plug in more exotic inputs than just natural numbers for n. (k still has to be a natural number, however.) This is the *extended choose function/extended binomial coefficients.*

Example. Compute $\binom{5/3}{3}$.

$$\begin{aligned}\binom{5/3}{3} &= \frac{\frac{5}{3}(\frac{5}{3}-1)(\frac{5}{3}-2)}{3!} \\ &= \frac{(\frac{5}{3})(\frac{2}{3})(-\frac{1}{3})}{3!} \\ &= -\frac{5}{81}\end{aligned}$$

△

Here is an important special case: what if we choose from a negative integer?

Theorem (Negative Binomial Coefficients)**.**

$$\binom{-n}{k} = (-1)^k\binom{n+k-1}{k}$$

Proof.

$$\begin{aligned}\binom{-n}{k} &= \frac{-n(-n-1)(-n-2)\cdots(-n-k+1)}{k!} \\ &= (-1)^k\frac{n(n+1)(n+2)\cdots(n+k-1)}{k!} \\ &= (-1)^k\frac{(n+k-1)!}{k!(n-1)!} \\ &= (-1)^k\binom{n+k-1}{k}\end{aligned}$$

■

Example. Compute $\binom{-6}{3}$.

$$\begin{aligned}\binom{-6}{3} &= (-1)^3\binom{6+3-1}{3} \\ &= -\binom{8}{3}\end{aligned}$$

△

Example. Compute $[x^5]\frac{1}{(1-x)^3}$.

$$\begin{aligned}[x^5]\frac{1}{(1-x)^3} &= [x^5](1-x)^{-3} \\ &= [x^5]\binom{-3}{5}(-x)^5 \\ &= [x^5](-1)^5\binom{3+5-1}{5}(-x)^5 \\ &= [x^5](-1)^5\binom{3+5-1}{5}(-1)^5x^5 \\ &= [x^5]\binom{7}{5}x^5 \\ &= \binom{7}{5}\end{aligned}$$

△

As seen above, we can now compute the coefficients of binomials with negative powers. Combining this with the standard sum of geometric series,

$$1 + x + x^2 + x^3 + \cdots = \frac{1}{1-x}$$

we can now efficiently compute coefficients of any generating function.

Example. How many ways can we arrange 20 balls in 5 boxes if each box must have at least 2 balls?

$$\begin{aligned}[x^{20}](x^2+x^3+x^4+\cdots)^5 &= [x^{20}]x^{10}(1+x+x^2+\cdots)^5 \\ &= [x^{10}](1+x+x^2+\cdots)^5 \\ &= [x^{10}]\left(\frac{1}{1-x}\right)^5 \\ &= [x^{10}](1-x)^{-5} \\ &= [x^{10}]\binom{-5}{10}(-x)^{10} \\ &= [x^{10}](-1)^{10}\binom{5+10-1}{10}(-1)^{10}x^{10}\end{aligned}$$

$$
\begin{aligned}
&= [x^{10}]\binom{14}{10}x^{10} \\
&= \binom{14}{10}
\end{aligned}
$$

△

Example. How many ways can we arrange 20 balls in 5 boxes if each box must have at least 2 balls, but no more than 10 balls?

$$
\begin{aligned}
[x^{20}](x^2 + x^3 + x^4 + \cdots + x^{10})^5 &= [x^{20}]x^{10}(1 + x + x^2 + \cdots + x^8)^5 \\
&= [x^{10}](1 + x + x^2 + \cdots + x^8)^5 \\
&= [x^{10}]\left(\frac{1 - x^9}{1 - x}\right)^5 \\
&= [x^{10}](1 - x^9)^5(1 - x)^{-5}
\end{aligned}
$$

We now perform a discrete convolution of the two brackets; we need 10 total copies of x from the two brackets. In the left bracket, we may choose between 1 and x^9 in each of the 5 factors:

- If we had 0 copies of x^9 from the left bracket, then we would need 10 copies of x from the right bracket:
$$\binom{5}{0}(-x^9)^0 1^5 \times \binom{-5}{10}x^{10} = \binom{5}{0}\binom{-5}{10}x^{10}$$
- If we had 1 copy of x^9 from the left bracket, then we would need 1 copies of x from the right bracket:
$$\binom{5}{1}(-x^9)^1 1^4 \times \binom{-5}{1}(-x)^1 = \binom{5}{1}\binom{-5}{1}x^{10}$$
- If we had 2 or more copies, then we would already exceed 10, so we are done.

$$[x^{10}](1 - x^9)^5(1 - x)^{-5} = \binom{5}{0}\binom{-5}{10} + \binom{5}{1}\binom{-5}{1}$$

△

7.2.7.2 A Pair of Dice

Note the total upon rolling a standard pair of 6 sided dice. There is one way of obtaining 2, two ways of obtaining 3, etc.

- Find all pairs of 6 sided dice such that:
  - The probability of obtaining each sum matches that of a standard pair of dice.
  - Each face has a natural value.

Note that the two dice may be distinct.

We can represent a standard die with the following generating function;

$$\begin{aligned} f(x) &= \sum_{i=1}^{6} x^i \\ &= x^1 + x^2 + x^3 + x^4 + x^5 + x^6 \end{aligned}$$

The sum of a pair of dice is then represented by:

$$\begin{aligned} \big(f(x)\big)^2 &= \left(\sum_{i=1}^{6} x^i\right)^2 \\ &= x^2 + 2x^3 + 3x^4 + 4x^5 + 5x^6 + 6x^7 + 5x^8 + 4x^9 + 3x^{10} + 2x^{11} + x^{12} \end{aligned}$$

So, for any other pair of dice to return the same probability distribution on the sums must also multiply to this same polynomial. The task is now to factor this polynonial into two factors, f_1 and f_2.

First, note, using the standard factorisation of $x^n - 1$, we can factor $f(x)$ to

$$\begin{aligned} f(x) &= \sum_{i=1}^{6} x^i \\ &= x\sum_{i=0}^{5} x^i \\ &= x(1 + x + x^2 + x^3 + x^4 + x^5) \\ &= x\left(\frac{x^6 - 1}{x - 1}\right) \\ &= x\left(\frac{(x^3 - 1)(x^3 + 1)}{x - 1}\right) \\ &= x\left(\frac{x^3 - 1}{x - 1}\right)(x^3 + 1) \\ &= x\left(x^2 + x + 1\right)(x^3 + 1) \\ &= x(x^2 + x + 1)(x + 1)(x^2 - x + 1) \end{aligned}$$

so,

$$\big(f(x)\big)^2 = x^2(x^2 + x + 1)^2(x + 1)^2(x^2 - x + 1)^2$$

The two additional requirements constrain the factorisation further:

- Each face requiring a natural value means that each factor has a minimum degree of 1.
- Each die has 6 faces, so $f_1(1) = f_2(1) = 6$.

The first constraint implies that f_1 and f_2 each take one copy of the x factor, as every other factor contains a constant term. For the second, consider evaluating the factors of $f(x)$ at 1:

$$\begin{aligned} (\lambda x.x)(1) &= 1 \\ (\lambda x.x^2 + x + 1)(1) &= 3 \\ (\lambda x.x + 1)(1) &= 2 \\ (\lambda x.x^2 - x + 1)(1) &= 1 \end{aligned}$$

This implies that f_1 and f_2 each take one copy of (x^2+x+1) and $(x+1)$, or otherwise a product of 6 is impossible.

This only leaves the two (x^2-x+1) factors to be distributed between f_1 and f_2. Giving one copy to each just gives the factorisation $\big(f(x)\big)^2 = f(x)f(x)$, which just gives back the standard set of dice.

Otherwise, we can give both copies to the same factor, and without loss of generality, let this factor be f_1.

$$\begin{aligned} f_1(x) &= x(x+1)(x^2+x+1)(x^2-x+1)^2 \\ &= x+x^3+x^4+x^5+x^6+x^8 \\ f_2(x) &= x(x+1)(x^2+x+1) \\ &= x+2x^2+2x^3+x^4 \end{aligned}$$

So, the only other pair of dice satisfying these requirements are dice with face values {1,2,2,3,3,4} and {1,3,4,5,6,8}.

7.2.7.3 Discrete Fourier Transform

Example. Find the number of subsets of $\{1,\ldots,2000\}$ whose element-sum is divisible by 5.

Consider the following polynomial:

$$f(x) = (1+x^1)(1+x^2)(1+x^3)\cdots(1+x^{2000})$$

A term in its expansion amounts to choosing 2000 binary choices, selecting either 1 or x^i from each bracket; this precisely corresponds to selecting a subset of $\{1,\ldots,2000\}$, with $\{a_1,a_2\ldots,a_m\}$ corresponding to the term $x^{a_1}x^{a_2}\cdots x^{a_m} = x^{a_1+a_2+\cdots+a_m}$.

Hence, we are looking for the sum of the coefficients of the terms x^{5k} in $f(x)$:

$$S = \sum_{k=0}^{400}[x^{5k}]f(x)$$

Let us rewrite f in its expanded form:

$$f(x) = \sum_{n=0}^{N} c_n x^n$$

We evaluate this polynomial using the factored form above to deduce information about the coefficients in the expanded form. For instance,

$$\begin{aligned} f(0) &= (1+0^1)(1+0^2)\cdots(1+0^{2000}) \\ &= 1 \end{aligned}$$

so we know $c_0 = 1$, as every other term in the expanded form vanishes. This is not very informative, as this simply tells us that there is a unique subset of $\{1,\ldots,2000\}$ with sum 0, i.e. the empty set.

Evaluating at 1 yields

$$\begin{aligned} f(1) &= (1+1^1)(1+1^2)\cdots(1+1^{2000}) \\ &= 2^{2000} \\ &= c_0+c_1+c_2+c_3+\cdots+c_N \end{aligned}$$

Again, this just tells us that there are 2^{2000} subsets of $\{1,\dots,2000\}$.

Now, what if we evaluate at -1?

$$\begin{aligned} f(-1) &= (1+(-1)^1)(1+(-1)^2)\cdots(1+(-1)^{2000}) \\ &= 0 \\ &= c_0 - c_1 + c_2 - c_3 + \cdots + c_N \end{aligned}$$

That is, there are the same number of subsets that have an even sum as there are subsets that have an odd sum. This is not immediately obvious, and even if it is reasonable that this might be true due to the symmetry of the situation, it is not clear how one would prove such a thing. However, this result just appears automatically by evaluating this generating function at a single point.

Now, consider the sum of the previous two evaluations:

$$\begin{aligned} f(1) + f(-1) &= (c_0 c_1 + c_2 + c_3 + \cdots + c_N) - (c_0 - c_1 + c_2 - c_3 + \cdots + c_N) \\ &= 2c_0 + 2c_2 + 2c_4 + \cdots + 2c_N \end{aligned}$$

Dividing by 2 yields

$$2^{1999} = \frac{1}{2}\big(f(1) + f(-1)\big) = c_0 + c_2 + c_4 + \cdots + c_N = S = \sum_{k=0}^{1000} [x^{2k}] f(x)$$

This is almost what we want; we have extracted all of the multiples of 2, when we want the multiples of 5.

Now, despite the fact that this started as a combinatorial problem involving discrete sets of natural numbers, we now consider evaluating f at *complex* inputs.

Firstly, we note that 1 and -1 are the 2nd roots of unity: the oscillatory behaviour of -1 has period 2, and in particular, for terms that do not align precisely to the period, the two roots are distinct and opposing, thus removing the term via *destructive* interference in the final sum. Conversely, for terms that do align to the period, the roots are equal, and *constructively* interfere, adding up to a coefficient of 2. (This is perhaps clearer in the following working, where the period is larger.)

Inspired by this, we will evaluate f at the complex 5th roots of unity, $\zeta^k = e^{2\pi i k/5}$, and sum over all the evaluations:

$$f(\zeta^0) + f(\zeta^1) + f(\zeta^2) + f(\zeta^3) + f(\zeta^4)$$

First, note that

$$\zeta^0 + \zeta^1 + \zeta^2 + \zeta^3 + \zeta^4 = 0$$

This is an easy exercise in geometry, or by simple calculation.

Squaring each term and reducing the exponents modulo 5, it can be easily verified that this just permutes the roots amongst themselves:

$$\begin{aligned} \zeta^0 + \zeta^2 + \zeta^4 + \zeta^6 + \zeta^8 &= \zeta^0 + \zeta^2 + \zeta^4 + \zeta^1 + \zeta^3 \\ &= 0 \end{aligned}$$

This can also easily be verified geometrically, as squaring each root simply doubles their arguments, rotating each root to another root.

Similarly,

$$\zeta^0 + \zeta^3 + \zeta^6 + \zeta^9 + \zeta^{12} = \zeta^0 + \zeta^3 + \zeta^1 + \zeta^4 + \zeta^2$$

$$= 0$$

$$\begin{aligned} \zeta^0 + \zeta^4 + \zeta^8 + \zeta^{12} + \zeta^{16} &= \zeta^0 + \zeta^4 + \zeta^3 + \zeta^2 + \zeta^1 \\ &= 0 \end{aligned}$$

However, by construction, taking 5th powers collapses each root simultaneously to 1:

$$\begin{aligned} \zeta^0 + \zeta^5 + \zeta^{10} + \zeta^{15} + \zeta^{20} &= 1 + 1^1 + 1^2 + 1^3 + 1^4 \\ &= 5 \end{aligned}$$

Thus, using the expanded form of f, we have:

$$\begin{array}{ccccccccccccccccccccc}
f(\zeta^0) & = & c_0 & + & c_1\zeta^0 & + & c_2\zeta^0 & + & c_3\zeta^0 & + & c_4\zeta^0 & + & c_5\zeta^0 & + & c_6\zeta^0 & + & c_7\zeta^0 & + & c_8\zeta^0 & + & \cdots \\
+ & & & & & & & & & & & & & & & & & & & & \\
f(\zeta^1) & = & c_0 & + & c_1\zeta^1 & + & c_2\zeta^2 & + & c_3\zeta^3 & + & c_4\zeta^4 & + & c_5\zeta^5 & + & c_6\zeta^6 & + & c_7\zeta^7 & + & c_8\zeta^8 & + & \cdots \\
+ & & & & & & & & & & & & & & & & & & & & \\
f(\zeta^2) & = & c_0 & + & c_1\zeta^2 & + & c_2\zeta^4 & + & c_3\zeta^6 & + & c_4\zeta^8 & + & c_5\zeta^{10} & + & c_6\zeta^{12} & + & c_7\zeta^{14} & + & c_8\zeta^{16} & + & \cdots \\
+ & & & & & & & & & & & & & & & & & & & & \\
f(\zeta^3) & = & c_0 & + & c_1\zeta^3 & + & c_2\zeta^6 & + & c_3\zeta^9 & + & c_4\zeta^{12} & + & c_5\zeta^{15} & + & c_6\zeta^{18} & + & c_7\zeta^{21} & + & c_8\zeta^{24} & + & \cdots \\
+ & & & & & & & & & & & & & & & & & & & & \\
f(\zeta^4) & = & c_0 & + & c_1\zeta^4 & + & c_2\zeta^8 & + & c_3\zeta^{12} & + & c_4\zeta^{16} & + & c_5\zeta^{20} & + & c_6\zeta^{24} & + & c_7\zeta^{28} & + & c_8\zeta^{32} & + & \cdots \\
 & & \| & & \| & & \| & & \| & & \| & & \| & & \| & & \| & & \| & & \\
 & & 5c_0 & + & 0 & + & 0 & + & 0 & + & 0 & + & 5c_5 & + & 0 & + & 0 & + & 0 & + & \cdots
\end{array}$$

so we have that

$$\frac{1}{5}\big(f(\zeta^0) + f(\zeta^1) + f(\zeta^2) + f(\zeta^3) + f(\zeta^4)\big) = \sum_{k=0}^{N} [x^{5k}]f(x)$$

as desired.

Now to actually compute these values, we have one last trick. Because the ζ^k are the roots of $g(x) = x^5 - 1$, we can factorise this polynomial as:

$$g(x) = x^5 - 1 = (x - \zeta^0)(x - \zeta^1)(x - \zeta^3)(x - \zeta^4)(x - \zeta^5)$$

Evaluating this polynomial at -1 in the original form, we have

$$g(-1) = (-1)^5 - 1 = -2$$

while in the factorised form, we have

$$\begin{aligned} g(-1) &= (-1 - \zeta^0)(-1 - \zeta^1)(-1 - \zeta^3)(-1 - \zeta^4)(-1 - \zeta^5) \\ &= (-1)^5\big((1 + \zeta)(1 + \zeta^2)(1 + \zeta^3)(1 + \zeta^4)(1 + \zeta^5)\big) \\ &= -\big((1 + \zeta)(1 + \zeta^2)(1 + \zeta^3)(1 + \zeta^4)(1 + \zeta^5)\big) \end{aligned}$$

so

$$(1 + \zeta)(1 + \zeta^2)(1 + \zeta^3)(1 + \zeta^4)(1 + \zeta^5) = 2$$

and hence,

$$f(\zeta) = (1 + \zeta)(1 + \zeta^2)(1 + \zeta^3)(1 + \zeta^4)(1 + \zeta^5)(1 + \zeta^6) \cdots (1 + \zeta^{2000})$$

$$\begin{aligned}&= (1+\zeta)(1+\zeta^2)(1+\zeta^3)(1+\zeta^4)(1+\zeta^5)(1+\zeta^1)\cdots(1+\zeta)\\&= \big((1+\zeta)(1+\zeta^2)(1+\zeta^3)(1+\zeta^4)(1+\zeta^5)\big)^{400}\\&= 2^{400}\end{aligned}$$

Similarly, $f(\zeta^k) = 2^{400}$ for $k = 2,3,4$, while $k = 5$ yields $f(\zeta^5) = f(1) = 2^{2000}$. Thus,

$$\begin{aligned}S = \frac{1}{5}\big(f(\zeta^0) + f(\zeta^1) + f(\zeta^2) + f(\zeta^3) + f(\zeta^4)\big) &= \frac{1}{5}\big(2^{2000} + 4 \times 2^{400}\big)\\&= \frac{1}{5}(2^{2000} + 2^{402})\end{aligned}$$

△

7.2.8 More Generating Functions

We've seen several infinite sequences and triangles of special numbers, some of which do not have explicit closed formulae. This is not unusual, and we will often stumble into new sequences that we don't yet understand.

An infinite sequence of numbers is rather unwieldy, especially if there isn't a formula for the nth term. For example, suppose we have two sequences, perhaps originating from distinct counting problems, and we suspect they are equal, but can't find a bijective proof. We could compute terms from the two sequences, and check that they agree, but we cannot prove that two infinite sequences are equal in this way.

We can also use generating functions to this end: to represent infinite sequences as more tangible finite-appearing functions. This generalises even to triangles, or other higher-dimensional analogues of sequences.

Recall that the Bell numbers are given by

$$B_n = \sum_{k=0}^{n} S(n,k)$$

The first few of these numbers are:

$$1, 1, 2, 5, 15, 52, 203, 877, 4140, 21147, \dots$$

If we play about with these numbers for a while, we might think to write them as the coefficients of a power series:

$$\frac{1}{0!} + \frac{1x}{1!} + \frac{2x^2}{2!} + \frac{5x^3}{3!} + \frac{15x^4}{4!} + \frac{52x^5}{5!} + \cdots$$

You would be justified in asking why we have done this, since there still seems to be very little pattern in the coefficients – especially with the messy factorials thrown in as divisors – but it turns out that this power series is the Maclaurin series of e^{e^x-1}. This expression is much easier to remember!

This should be *a*) hugely surprising, and *b*) evocative of the power of these representations.

Unlike the previous section where we hand-built our generating functions to solve a specific problem in enumeration, it is not immediately clear where this correspondence between the Bell numbers and the Maclaurin series of e^{e^x-1} comes from.

Let $(a_n)_{n=1}^{\infty}$ be a sequence of numbers. The *formal power series*

$$\sum_{n=0}^{\infty} a_n x^n$$

is called the *ordinary generating function* of (a_n), and the formal power series

$$\sum_{n=0}^{\infty} a_n \frac{x^n}{n!}$$

is called the *exponential generating function* of (a_n).

Formal power series here means that x is treated purely as an indeterminate variable and we are not concerned with the radius of convergence of the series, but we can still manipulate these expressions just like any other ordinary power series.

One application of generating functions is in solving recurrence relations.

Example. Let G_n denote the number of sequences of length n of elements of [3] whose consecutive entries differ by at most 1. The first few values are as follows: $G_0 = 1$, $G_1 = 3$, $G_2 = 7$, $G_3 = 17$. Compute a closed-form expression for G_n.

Let A_n denote the number of sequences of length n that begin with 1 (and by symmetry, with 3), and let B_n denote the number of sequences of length n that begin with 2, so $G_n = 2A_n + B_n$.

If a sequence begins with 1, then the sequence with the first element removed begins with either a 1, or a 2, so $A_n = A_{n-1} + B_{n-1}$. If a sequence begins with 2, then the truncated sequence could begin with any digit, so $B_n = 2A_{n-1} + B_{n-1} = G_{n-1}$. Then,

$$\begin{aligned} G_n &= 2A_n + B_n \\ &= 2(A_{n-1} + B_{n-1}) + 2A_{n-1} + B_{n-1} \\ &= 2(2A_{n-1} + B_{n-1}) + B_{n-1} \\ &= 2G_{n-1} + G_{n-2} \end{aligned}$$

Then, we use the ordinary generating function of G_n as follows:

$$\begin{aligned} \sum_{n=0}^{\infty} G_n x^n &= G_0 + G_1 x + \sum_{n=0}^{\infty} G_{n+2} x^{n+2} \\ &= 1 + 3x + \sum_{n=0}^{\infty} (2G_{n+1} + G_n) x^{n+2} \\ &= 1 + 3x + 2x \sum_{n=0}^{\infty} G_{n+1} x^{n+1} + x^2 \sum_{n=0}^{\infty} G_n x^n \\ &= 1 + 3x + (2xG_0x^0 - 2xG_0x^0) + 2x \sum_{n=0}^{\infty} G_{n+1} x^{n+1} + x^2 \sum_{n=0}^{\infty} G_n x^n \\ &= 1 + 3x - 2x + \left(2xG_0x^0 + 2x \sum_{n=0}^{\infty} G_{n+1} x^{n+1} \right) + x^2 \sum_{n=0}^{\infty} G_n x^n \\ &= 1 + x + 2x \sum_{n=0}^{\infty} G_n x^n + x^2 \sum_{n=0}^{\infty} G_n x^n \\ &= 1 + x + (2x + x^2) \sum_{n=0}^{\infty} G_n x^n \\ &= \frac{1+x}{1-2x-x^2} \end{aligned}$$

Performing partial fraction decomposition, and using geometric series, we have:

$$\begin{aligned}
\sum_{n=0}^{\infty} G_n x^n &= \frac{1+x}{(-x+\sqrt{2}-1)(x+\sqrt{2}+1)} \\
&= \frac{1}{2}\left(\frac{1}{-x+\sqrt{2}-1}\right) - \frac{1}{2}\left(\frac{1}{x+\sqrt{2}+1}\right) \\
&= \frac{1}{2}\left(\frac{1}{\sqrt{2}-1-x}\right) - \frac{1}{2}\left(\frac{1}{\sqrt{2}+1+x}\right) \\
&= \frac{1}{2(\sqrt{2}-1)}\left(\frac{1}{1-\left(\frac{1}{\sqrt{2}-1}\right)x}\right) - \frac{1}{2(\sqrt{2}+1)}\left(\frac{1}{1-\left(-\frac{1}{\sqrt{2}+1}\right)x}\right) \\
&= \frac{1}{2(\sqrt{2}-1)}\sum_{n=0}^{\infty}\left(\frac{1}{\sqrt{2}-1}\right)^n x^n - \frac{1}{2(\sqrt{2}+1)}\sum_{n=0}^{\infty}\left(-\frac{1}{\sqrt{2}+1}\right)^n x^n \\
&= \sum_{n=0}^{\infty}\left(\frac{1}{2(\sqrt{2}-1)}\left(\frac{1}{\sqrt{2}-1}\right)^n - \frac{1}{2(\sqrt{2}+1)}\left(-\frac{1}{\sqrt{2}+1}\right)^n\right) x^n \\
&= \sum_{n=0}^{\infty}\frac{1}{2}\left(\left(\frac{1}{\sqrt{2}-1}\right)^{n+1} + \left(-\frac{1}{\sqrt{2}+1}\right)^{n+1}\right) x^n \\
&= \sum_{n=0}^{\infty}\frac{1}{2}\left(\left(\sqrt{2}+1\right)^{n+1} + \left(-\sqrt{2}+1\right)^{n+1}\right) x^n
\end{aligned}$$

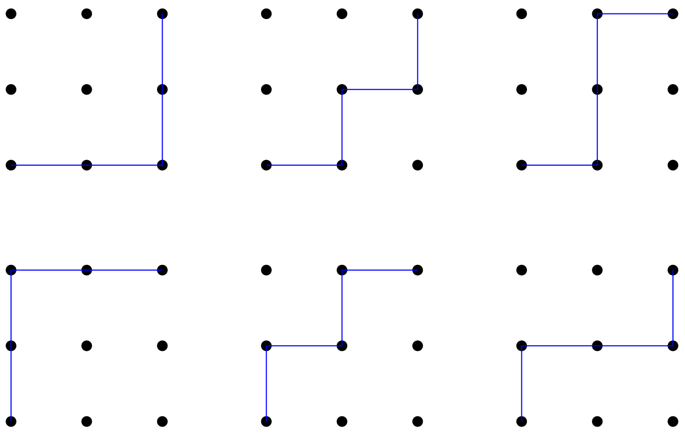
So,

$$G_n = \frac{1}{2}\left(\left(\sqrt{2}+1\right)^{n+1} + \left(-\sqrt{2}+1\right)^{n+1}\right)$$

△

This may not seem very impressive given that there are much simpler techniques for solving linear recurrence relations, but generating functions generalise to non-linear, and higher dimensional recurrence relations.

Example. A *north-east lattice walk* is a path in the 2D Cartesian plane consisting of integer-length steps in the positive x (*north*) or positive y (*east*) direction. For instance, here are all the north-east lattice walks from (0,0) to (2,2):



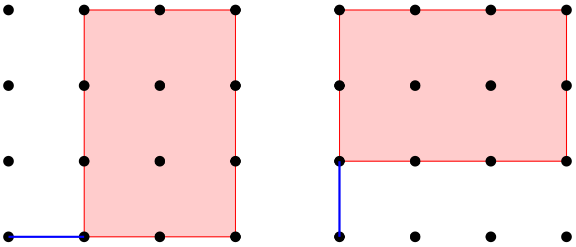
How many north-east lattice walks are there from (0,0) to (n,m)?

Let $N(n,m)$ denote the number of paths from (0,0) to (n,m).

If $n = 0$ and $m = 0$, then there is only the single trivial empty walk so $N(0,0) = 1$. If exactly one of n and m are zero, then the lattice is a straight line and again, only one trivial walk exists, so $N(n,0) = N(0,m) = 1$.

We also note that the problem is symmetric, in the the north and east directions are arbitrary, so $N(a,b) = N(b,a)$ for all a,b.

At (0,0), walking east on a (n,m) grid leaves a $(n-1,m)$ grid remaining to be walked, and similarly, walking north leaves a $(n,m-1)$ grid.



The 4×4 grid reduces to a 3×4 grid with an east step, and similarly to a 4×3 grid with a north step.

So, we have,

$$N(a,b) = N(a-1,b) + N(a,b-1)$$

with boundary conditions,

$$N(k,0) = N(0,k) = 1$$

We solve this recurrence relation with generating functions:

$$\begin{aligned}
\sum_{i,j\geq 0} c_{i,j}x^iy^j &= c_{0,0} + \sum_{i=1}^{\infty} c_{i,0}x^i + \sum_{j=1}^{\infty} c_{0,j}y^j + \sum_{i,j\geq 1} c_{i,j}x^iy^j \\
&= 1 + \sum_{i=1}^{\infty} x^i + \sum_{j=1}^{\infty} y^j + \sum_{i,j\geq 1} c_{i-1,j}x^iy^j + \sum_{i,j\geq 1} c_{i,j-1}x^iy^j \\
&= 1 + \frac{x}{1-x} + \frac{y}{1-y} + x\sum_{i,j\geq 1} c_{i-1,j}x^{i-1}y^j + y\sum_{i,j\geq 1} c_{i,j-1}x^iy^{j-1} \\
&= 1 + \frac{x}{1-x} + \frac{y}{1-y} + x\sum_{i\geq 0,j\geq 1} c_{i,j}x^iy^j + y\sum_{i\geq 1,j\geq 0} c_{i,j}x^iy^j \\
&= 1 + \frac{x}{1-x} + \frac{y}{1-y} + x\left(\sum_{i,j\geq 0} c_{i,j}x^iy^j - \sum_{j=0}^{\infty} c_{0,j}y^j\right) + y\left(\sum_{i,j\geq 0} c_{i,j}x^iy^j - \sum_{i=0}^{\infty} c_{i,0}x^i\right) \\
&= 1 + \frac{x}{1-x} + \frac{y}{1-y} + x\left(\sum_{i,j\geq 0} c_{i,j}x^iy^j - \frac{1}{1-x}\right) + y\left(\sum_{i,j\geq 0} c_{i,j}x^iy^j - \frac{1}{1-y}\right) \\
&= 1 + x\sum_{i,j\geq 0} c_{i,j}x^iy^j + y\sum_{i,j\geq 0} c_{i,j}x^iy^j \\
&= 1 + (x+y)\sum_{i,j\geq 0} c_{i,j}x^iy^j \\
&= \frac{1}{1-x-y}
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{1-(x+y)} \\
&= \sum_{n=0}^{\infty}(x+y)^n \\
&= \sum_{n=0}^{\infty}\sum_{k=0}^{n}\binom{n}{k}x^{n-k}y^k
\end{aligned}$$

Letting $n = i + j$ and $k = j$, we have,

$$= \sum_{i,j\geq 0}\binom{i+j}{j}x^i y^j$$

so,

$$c_{i,j} = \binom{i+j}{j}$$

△

This particular problem happens to have another elegant combinatorial solution that does not involve generating functions at all, and the simplicity of the final form we obtained is rather suggestive of this fact.

Exercise. Construct this alternative proof.

In the previous problem, we actually managed to find a simple form for the generating function for some of the binomial coefficients:

$$\sum_{n,k\geq 0}\binom{n+k}{k}x^n y^k = \frac{1}{1-x-y}$$

Let us fix k and just work with the one dimensional sequence $\binom{k}{k}, \binom{k+1}{k}, \ldots$. Define

$$A(x) = \sum_{n=0}^{\infty}\binom{k+n}{k}x^n$$

Expanding the binomial coefficient, we have,

$$A(x) = \sum_{n=0}^{\infty}\frac{(k+n)!}{k!n!}x^n$$

So, taking the derivative, we obtain:

$$\begin{aligned}
A'(x) &= \sum_{n=1}^{\infty}\frac{(k+n)!}{k!(n-1)!}x^{n-1} \\
(1-x)A'(x) &= (1-x)\sum_{n=1}^{\infty}\frac{(k+n)!}{k!(n-1)!}x^{n-1} \\
(1-x)A'(x) &= \sum_{n=1}^{\infty}\frac{(k+n)!}{k!(n-1)!}x^{n-1} - \sum_{n=1}^{\infty}\frac{(k+n)!}{k!(n-1)!}x^n \\
&= \sum_{n=1}^{\infty}(k+n)\binom{k+n-1}{k}x^{n-1} - \sum_{n=1}^{\infty}n\binom{k+n}{n}x^n
\end{aligned}$$

$$\begin{aligned}
&= \sum_{n=0}^{\infty}(k+n+1)\binom{k+n}{k}x^n - \sum_{n=0}^{\infty} n\binom{k+n}{n}x^n \\
&= (k+1)A(x) + \sum_{n=0}^{\infty} n\binom{k+n}{k}x^n - \sum_{n=0}^{\infty} n\binom{k+n}{n}x^n \\
&= (k+1)A(x)
\end{aligned}$$

giving the differential equation

$$(1-x)A'(x) = (k+1)A(x)$$

Rearranging, we have,

$$\begin{aligned}
\frac{A'(x)}{A(x)} &= \frac{k+1}{1-x} \\
\int \frac{A'(x)}{A(x)}\,dx &= \int \frac{k+1}{1-x}\,dx \\
\ln\big(A(x)\big) &= (k+1)\ln(1-x) + C \\
A(x) &= \frac{C}{(1-x)^{k+1}}
\end{aligned}$$

We also have the initial condition $A(0) = \binom{k}{k} = 1$, which gives $C = 1$.

We can also show this formula with an alternative combinatorial proof. Recall that $\binom{k+n}{k}$ is the number of ways to place n unlabelled balls into $k+1$ labelled boxes – or equivalently, the number of degree-n monomials in $k+1$ variables. Consider the product

$$\prod_{j=0}^{k+1}\sum_{i=0}^{\infty} x_j^i = (1 + x_1 + x_1^2 + \cdots)(1 + x_2 + x_2^2 + \cdots)\cdots(1 + x_{k+1} + x_{k+1}^2 + \cdots)$$

In the expansion of this product, every monomial in $k+1$ variables appears exactly once. Setting all variables $x_1, \ldots, x_{k+1}$ equal to x thus gives $A(x)$, and the product reduces to

$$\prod_{j=0}^{k+1}\sum_{i=0}^{\infty} x^i = \left(\sum_{i=0}^{\infty} x^i\right)^{k+1} = \frac{1}{(1-x)^{k+1}}$$

Earlier, we proved that the Bell numbers satisfy

$$B_n = \sum_{k=0}^{n-1}\binom{n-1}{k}$$

so we can attempt to write the generating function for the Bell numbers. Define

$$A(x) = \sum_{n=0}^{\infty} B_n \frac{x^n}{n!}$$

Then,

$$A'(x) = \sum_{n=0}^{\infty} B_n \frac{x^{n-1}}{(n-1)!}$$

$$\begin{aligned}
&= \sum_{n=0}^{\infty}\sum_{k=0}^{n-1}\binom{n-1}{k}B_k\frac{x^{n-1}}{(n-1)!}\\
&= \sum_{k=0}^{\infty}\sum_{n=k+1}^{\infty}\binom{n-1}{k}B_k\frac{x^{n-1}}{(n-1)!}\\
&= \sum_{k=0}^{\infty}\sum_{n=0}^{\infty}\binom{k+n}{k}B_k\frac{x^{k+n}}{(k+n)!}\\
&= \sum_{k=0}^{\infty}B_k\frac{x^k}{k!}\sum_{n=0}^{\infty}\frac{x^n}{n!}\\
&= A(x)e^x
\end{aligned}$$

Which yields

$$\begin{aligned}
\frac{A'(x)}{A(x)} &= e^x\\
\ln\big(A(x)\big) &= e^x + CA(x) = e^{e^x+C}
\end{aligned}$$

We also have the initial condition $A(0) = B_0 = 1$, which gives $C = 1$, thus proving:

Theorem 7.2.12. *The generating function for the Bell numbers is given by*

$$A(x) = e^{e^x-1}$$

Let us explore another sequence of numbers. Earlier, we encountered the partition numbers $p(n)$ which count the number of ways to write n as a sum of non-negative integers.

What is the coefficient of x^n in the following expression?

$$\prod_{j=0}^{k}\sum_{i=0}^{\infty}x^{ij} = (1+x+x^2+\cdots)(1+x^2+x^4+\cdots)\cdots(1+x^k+x^{2k}+\cdots)$$

The answer is equal to the number of ways to write n as a sum of the form $a_1 + 2a_2 + 3a_3 + \cdots + ka_k$. Such sums are in bijection with the partitions of n into parts that are at most k, with the bijection being that a_i is the number of copies of i in the partition. That is,

$$\sum_{n=0}^{\infty}p_{\leq k}(n)x^n = \prod_{i=1}^{k}\frac{1}{1-x^i}$$

Removing the limit on k, we obtain:

$$\sum_{n=0}^{\infty}p(n)x^n = \prod_{i=1}^{\infty}\frac{1}{1-x^i}$$

These generating functions are exceptionally clean, but despite that, $p(x)$ and all of its variants still do not admit closed formulae.

We can apply similar reasoning as the above to obtain more generating functions. For instance, the coefficient of x^n in

$$\prod_{j\text{ odd}}^{\infty}\sum_{i=0}^{\infty}x^{ij} = (1+x+x^2+\cdots)(1+x^3+x^6+\cdots)(1+x^5+x^{10}+\cdots)\cdots$$

is the number of partitions of n into odd parts. Similarly, the coefficient of x^n in

$$\prod_{i=1}^{\infty}(1+x^i)$$

is the number of ways partitions of n into a sum of distinct parts.

These two restrictions of partitioning n don't seem to be related, but in fact:

Theorem (Euler)**.** *The number of partitions of n into odd parts is equal to the number of partitions of n into distinct parts.*

Proof. Using $(1+x)^i = \frac{1-x^{2i}}{1-x^i}$, we have,

$$\begin{aligned}
\sum_{n=0}^{\infty} p_{\text{distinct}}(n)x^n &= \prod_{i=1}^{\infty}(1+x^i) \\
&= \prod_{i=1}^{\infty}\frac{1-x^{2i}}{1-x^i} \\
&= \frac{(1-x^2)(1-x^4)(1-x^6)\cdots}{(1-x)(1-x^2)(1-x^3)\cdots} \\
&= \frac{1}{(1-x)(1-x^3)(1-x^5)\cdots} \\
&= \frac{1}{1-x}\cdot\frac{1}{1-x^3}\cdot\frac{1}{1-x^5}\cdots \\
&= (1+x+x^2+\cdots)(1+x^3+x^6+\cdots)(1+x^5+x^{10}+\cdots)\cdots \\
&= \prod_{j\text{ odd}}^{\infty}\sum_{i=0}^{\infty}x^{ij} \\
&= \sum_{n=0}^{\infty} p_{\text{odd}}(n)x^n
\end{aligned}$$

■

We can also give a combinatorial proof of this result:

Proof. Given a partition of n into odd parts, we count the number of times each odd number occurs, so

$$n = 1a_1 + 3a_3 + 5a_5 + \cdots$$

and we can write a_i as a sum of powers of two:

$$n = 1(2^{b_{1,1}} + 2^{b_{1,2}} + \cdots) + 3(2^{b_{3,1}} + 2^{b_{3,2}} + \cdots) + 5(2^{b_{5,1}} + 2^{b_{5,2}} + \cdots) + \cdots$$

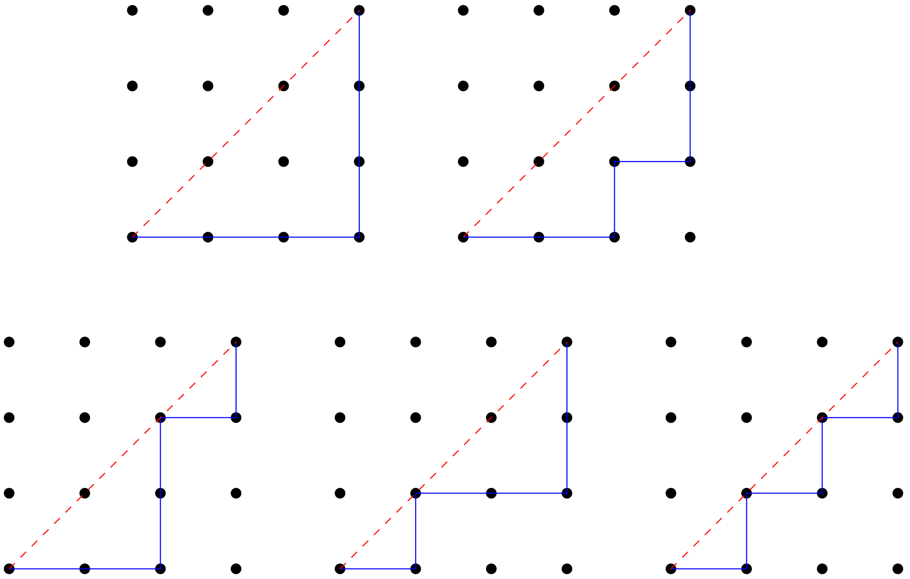
Note that the binary representation of a number is unique. Then, because each of the coefficients is an odd number, every term in the expansion is distinct, so this yields a unique partition of n into distinct parts. This operation also works in reverse, so there is a bijection between the number of partitions of n into distinct parts and the number of partitions of n into odd parts. ■

7.2.9 Catalan Numbers

In the previous problem, we explored the number of north-east lattice paths from (0,0) to (n,m). That is, the number of ways to travel from (0,0) to (n,m) on the Cartesian plane using only steps (1,0) and (0,1). Note that such a path must have length $n + m$, consisting of n steps east, and m steps north.

How many lattice paths are there from (0,0) to (n,n) if we add the restriction that they never strictly cross above the diagonal line $y = x$ from (0,0) to (n,n)?

Example. For $n = 3$, there are 5 such paths:



△

These paths are called *Dyck paths*, and they are counted by the *Catalan numbers*, C_n, and the first few are:

$$1, 1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796, 58786, 208012, 742900, 2674440, \ldots$$

The Catalan numbers describe a surprising amount of other things. For instance:

(i) How many valid strings of length $2n$ of opening and closing brackets are there such that the brackets are correctly matched? For example, ()() and (()) are valid strings of length 4, but ((() and))() are not valid. (These are called *Dyck words* or *Dyck strings*.)

(ii) How many ways are there to split a convex $(n + 2)$-sided polygon into triangles by connecting corners together such that the lines formed do not intersect? (This is called a *triangulation* of the polygon)

(iii) Pick $2n$ points on a circle. How many ways are there to pair them up into n non-intersecting chords?

(iv) How many non-decreasing sequences $(a_i)_{i=1}^{n} \subseteq \mathbb{Z}$ are there such that $a_1 \leq 1$, and all partial sums are non-negative?

(v) How many rooted ordered trees are there on $n + 1$ nodes?

(vi) Draw $n+1$ points on the x-axis of the Cartesian plane. How many ways are there to connect these points with n arcs such that the arcs are all above the x-axis such that the arcs do not intersect, the arcs exit each node in the same direction, and the graph formed is a tree?

(vii) Now, allow intersections in the above, but we do not allow any arc to lie strictly below another arc. How many ways are there now?

(viii) How many permutations of $[n]$ are there such that, when written in cycle notation, the permutation does not contain a 3-cycle?

(ix) How many ways can you fill in a $2 \times n$ grid of squares with the numbers 1 to $2n$ such that each row and each column is increasing?

(x) How many non-decreasing sequences $(a_i)_{i=1}^{n} \subseteq \mathbb{N}$ are there, such that $a_i \leq i$?

(xi) How many non-decreasing sequences $(a_i)_{i=1}^{n} \subseteq \mathbb{N}$ are there, such that $1 \leq a_i \leq 2i$?

(xii) How many permutations of the multiset $\{1,1,2,2,3,3,...,n,n\}$ are there, such that the first occurrences of 1 to n are in increasing order, and there is no subsequence of the form $abab$?

Exercise. Find bijections between these different formulations.

There is a book, simply titled *Catalan Numbers*, by Richard P. Stanley, which famously describes 214 different counting problems whose solutions are all the Catalan numbers.

Let us find a formula for C_n.

First, we have the following recurrence relation for C_n:

Theorem 7.2.13.

$$C_{n+1} = \sum_{i=0}^{n} C_i C_{n-i}$$

Proof. Given a Dyck path P from (0,0) to $(n+1,n+1)$, let $(i+1,i+1)$ denote its first point of intersection of P with the diagonal, after the point (0,0), so $0 \leq i \leq n+1$.

We can split P into two smaller Dyck paths: P_1 from (0,0) to $(i+1,i+1)$, and P_2 from $(i+1,i+1)$ to $(n+1,n+1)$. Then, P_2 is a Dyck path of length $2(n-i)$, so there are C_{n-i} many choices for P_2.

P_1 is also a Dyck path, but with the additional property that it never intersects the diagonal line between (0,0) and $(i+1,i+1)$ – or equivalently, it never strictly crosses the subdiagonal line $y = x - 1$, apart from the first and last steps, which necessarily connect to (0,0) and $(i+1,i+1)$. Deleting these segments yields a Dyck path from (1,0) to $(i+1,i)$, which is a Dyck path of length $2i$, so there are C_i many choices for P_1. ■

There are $\binom{2n}{n}$ total paths from (0,0) to (n,n), of which C_n are Dyck paths. The goal is now to count how many bad paths do strictly cross the diagonal.

Suppose a path touches the diagonal $y = x + 1$ for the first time at a point $(i,i+1)$. Reflecting the remainder of the path over this line yields a new path ending at $(n-1,n+1)$. We claim that this gives a bijection between bad paths from (0,0) to (n,n), and all paths from (0,0) to $(n-1,n+1)$. We describe the inverse operation.

Every path from (0,0) to $(n-1,n+1)$ must cross the diagonal $y = x + 1$ since it starts below it and ends above it, so we can reflect about the first point of contact between the path and $y = x + 1$. We are guaranteed to get a bad path from (0,0) to (n,n), as the resulting path must touch $y = x + 1$.

This bijection shows that there are $\binom{2n}{n-1}$ bad paths, and thus,

$$\begin{aligned} C_n &= \binom{2n}{n} - \binom{2n}{n-1} \\ &= \frac{2n!}{n!n!} - \frac{2n!}{(n-1)!(n+1)!} \end{aligned}$$

$$
\begin{aligned}
&= \frac{2n!(n+1-n)}{n!(n+1)!} \\
&= \frac{2n!}{n!(n+1)!} \\
&= \frac{1}{n+1}\binom{2n}{n}
\end{aligned}
$$

7.2.10 Pigeonhole Principle

The *pigeonhole principle* states that if n elements (often described as pigeons) are partitioned into m non-empty sets (pigeonholes), with $n > m$, then at least set must contain more than one element. For example, if we have three balls to be placed into two boxes, then one of the boxes must contain more than one ball. This seemingly obvious fact can be used to prove seemingly unexpected results.

Example. Let nine points be placed inside a square of side length 1, with no three points lying on the same line. Prove that it is always possible to select 3 points that form a triangle with an area of at most $\frac{1}{8}$.

Divide the unit square into 4 subregions of area $\frac{1}{4}$; for simplicity, and without loss of generality, let these regions be squares of side length $\frac{1}{2}$.

As there are 9 points, and 4 squares, there will always be at least one square containing at least 3 points by the pigeonhole principle (note: a point that lies on the edge of the square can be considered to be contained within that square). Selecting these three points within the square to be the vertices of a triangle, the entire triangle must be fully contained within that square.

The largest area it can be is half the area of the square. As the square has area $\frac{1}{4}$, it follows that the area of the triangle is at most $\frac{1}{8}$, as required. △

Example. Prove that, for any 5 points placed on a sphere, at least one hemisphere will contain 4 of the points.

Pick any pair of points. These points describe a great circle on the sphere, which divides the remaining 3 points into 2 hemispheres. By the pigeonhole principle, at least 2 of these points lie in the same hemisphere, and, including the points on the great circle, this hemisphere will contain at least 4 points. △

Example. Let $X \subset [200]$ be a subset with 101 elements. Show that there exist distinct elements x,$y \in X$ such that x divides y.

Every integer can be written in the form $2^k \cdot a$, where k is a non-negative integer, and a is odd. If we do this for any number less than 200, a must be one of the 100 odd numbers $\{1,3,5,\dots,199\}$. By the pigeonhole principle, at least two of the 101 integers in X share the same a value, say, $2^r \cdot a$ and $2^s \cdot a$, with $r \neq s$. If $r < s$, then the first divides the second; otherwise, the second divides the first. △

Example. Let $X \subset [80]$ be a subset with 10 elements. Show that there exist two disjoint subsets of X whose elements sum to the same number.

The largest possible sum of the elements of X is $80 + 79 + 78 + \cdots + 71 < 80 \cdot 10 < 1000$. There are $2^{10} - 1 = 1023$ non-empty subsets of X, so by the pigeonhole principle, there are at least two non-empty subsets A and B such that the sum of the elements in A equals the sum of the elements in B. This also implies that A and B are not proper subsets of each other, so we can remove their intersection from both sets, yielding the disjoint subsets $A \setminus B$ and $B \setminus A$ with the same sums. △

Example. Given a set of 16 distinct positive integers that are at most 100, prove there is a subset of four integers a,b,c,d such that $a+b=c+d$.

Let $(a_i)_{i=1}^{16}$ denote the 16 numbers. We consider the differences between pairs of these integers, noting that there are $\binom{16}{2}=120$ such pairs. For convenience, let (a,b) denote a pair such that $a>b$.

If we have two distinct pairs (a_{i_1},a_{i_2}) and (a_{i_3},a_{i_4}) such that $a_{i_1}-a_{i_2}=a_{i_3}-a_{i_4}$, then we have the quadruplet $(a,b,c,d)=(a_{i_1},a_{i_4},a_{i_2},a_{i_3})$, unless $a_{i_2}=a_{i_3}$.

We say that x is *bad* for the pair of pairs (a_{j_1},x) and (x,a_{j_2}) if $a_{j_1}-x=x-a_{j_2}$, or equivalently, if $2x=a_{j_1}+a_{j_2}$. Note that if x is bad for (at least) two distinct pairs, we are done; if x is bad for (a_{i_1},x), (x,a_{i_2}) and (a_{i_3},x), (x,a_{i_4}), then $a_{i_1}+a_{i_2}=2x=a_{i_3}+a_{i_4}$.

Now, suppose each of the a_i is bad for at most one pair of pairs of numbers. For each such pair, remove one pair of numbers, so there are no bad numbers remaining. Then, we still have at least $120-16=104$ pairs of numbers remaining. The difference of the numbers in each remaining pair ranges from 1 to 99, so by the pigeonhole principle, some of these differences have the same value. △

Some variants on the pigeonhole principle are as follows:

(i) If n balls are placed into k boxes, and $k\nmid n$, then at least one box contains strictly greater than $\frac{n}{k}$ balls.

(ii) If infinitely many balls are placed into finitely many boxes, then at least one box contains infinitely many balls.

(iii) If uncountably many balls are placed into countably many boxes, then at least one box contains uncountably many balls.

7.3 Exercises

1. How many ordered quadruples (x_1,x_2,x_3,x_4) of positive odd integers are there such that $x_1 + x_2 + x_3 + x_4 = 98$?
2. Prove that $\binom{2n}{n}$ is even with a combinatorial argument.
3. How many rational numbers between 0 and 1 have the property that when written in simplest form, the product of the numerator and denominator is 20! (twenty factorial)?
4. Each square of a 1998 × 2002 chessboard contains either a 0 or a 1 such that the total number of squares containing 1 is odd in each row and in each column. Prove that the number of white squares containing 1 is even.
5. How many subsets $X \subset [18]$ of cardinality 5 have the property that every pair of numbers differ by at least 2?
6. How many rooted binary trees are there on n vertices?
7. Show that there is a bijection between the set of partitions of a set X, and the set of equivalence relations on X.

7.3.1 Solutions

1. Because the integers are odd, we may write each in the form $x_i = 2y_i - 1$. Then,

$$98 = \sum_{i=1}^{4}(2y_i - 1)$$
$$98 = 2\left(\sum_{i=1}^{4} y_i\right) - 4$$
$$51 = \sum_{i=1}^{4} y_i$$

   so we are looking for the number of ordered quadruples of integers whose sum is 51. This can be done with stars and bars: each such quadruple corresponds to 51 stars split into 4 parts by 3 separating bars, so there are

$$\binom{50}{3} = 19600$$

   ways to insert 3 bars into the 50 spaces.

2. Suppose we need to pick n objects from $2n$ total. Given any selection, the complement of that selection is also a valid selection, so the total number of selections is even.
3. The numerator and denominator must be relatively prime, so their prime factorisations must be disjoint. There are eight prime factors of 20 – namely, 2,3,5,7,11,13,17, and 19. Each prime factor may go to the numerator and denominator, so there are $2^8 = 256$ fractions whose products are 20!.

   These fractions can be paired up as reciprocals (i.e. swap where every prime factor goes), each containing exactly one fraction less than 1. Thus, there are 128 fractions less than 1 whose numerator and denominator multiply to 20!.

4. Let (i,j) denote the position of the unit square in the ith row and jth column, and let $a_{i,j}$ denote the number within that square.

The sum of the numbers in the 999 odd rows:

$$R = \sum_{i=1}^{999} \sum_{j=1}^{2002} a_{2i-1,j}$$

is odd, as it is the sum of 999 odd numbers. Similarly, the sum of all the numbers in the even columns:

$$C = \sum_{i=1}^{1001} \sum_{j=1}^{1998} a_{2i,j}$$

is odd, as it is the sum of 1001 odd numbers. Consider the set B of black squares in the even columns, and let $S(B)$ denote the sum of the numbers in squares in B.

The numbers in each of the squares in B appears precisely once in the sum R, and once in the sum C. Finally, note that each of the numbers in the white squares appears exactly once in the sum $R+C$. Thus, the sum of the numbers in all the white squares is $R+C-2S(B)$, which is even, so the number of white squares containing 1 is even.

5. Let $a_1 < a_2 < a_3 < a_4 < a_5$ be the five chosen numbers. Consider the numbers

$$(b_1,b_2,b_3,b_4,b_5) = (a_1,a_2-1,a_3-2,a_4-3,a_5-4)$$

Then, the b_i are five distinct numbers from the first fourteen positive integers.

Conversely, given any five distinct numbers $b_1 < b_2 < b_3 < b_4 < b_5$, we can reconstruct

$$(a_1,a_2,a_3,a_4,a_5) = (b_1,b_2+1,b_3+2,b_4+3,b_5+4)$$

to obtain five numbers such that every pair of numbers differ by at least 2.

Thus, there is a bijection between the set of 5-tuples of numbers satisfying the required conditions, and the set of 5-tuples of distinct numbers from the first fourteen positive integers. Therefore, there are $\binom{14}{5} = 2002$ possible subsets.

6. Denote the number of rooted binary trees on n vertices by C_n. $C_0 = 1$, as there is only the empty tree, and $C_1 = 1$ as there is only the trivial graph.

On 0 or 1 vertices, there is only 1 such tree. Then, given any two rooted binary trees on a and b vertices, we may construct a new binary tree on $a+b+1$ vertices by adding a new vertex, and connecting it to the root node of the two given trees. So, C_n satisfies the recurrence relation:

$$C_{n+1} = \sum_{i=0}^{n} C_i C_{n-i}$$

with boundary conditions $C_0 = C_1 = 1$, which describes the Catalan numbers.

7. Let $\sim$ be an equivalence relation. By reflexivity, $x \in [x]$ for all $x \in X$, so the union of the equivalence classes is X. Now, let $x,y \in X$ be distinct, and suppose $[x] \cap [y]$ is non-empty. Let $a \in [x] \cap [y]$. Then, by definition, $a \sim x$ and $a \sim y$, so $[x] = [a] = [y]$. So, the equivalence classes form a partition of X.

Conversely, given a partition $\{A_i\}_{i \in I}$ of X, define an relation $\sim$ on X such that $x \sim y$ if and only if x,y are in the same part. This is clearly reflexive, as every element is in the same part as itself, and also symmetric, because if $x,y \in A_i$, then $y,x \in A_i$. Let $x \sim y$, so $x,y \in A_i$, and $y \sim z$, so $y,z \in A_i$. Then, $x,z \in A_i$, so $x \sim z$. It follows that $\sim$ is an equivalence relation.

7.4 Extremal Combinatorics

7.5 Graph Theory

A *graph* G is represented by V, a set of *vertices* or *nodes*, and E, a set of pairs of vertices, called *edges* or *arcs*, and we write $G = (V,E)$. If we are using multiple graphs at once, we can refer to the vertex (edge) set of a graph G by writing $V(G)$ $(E(G))$, but when the context is clear, we will often write things like $G \cup v$ to mean the graph formed by adding the vertex v to the graph G.

If the edge pairs are ordered, the graph is *directed* or *oriented*, and can also be referred to as a *digraph.*

A vertex and an edge are *incident* if the vertex is at either end of the edge. The *degree*, *valency* or *order* of a vertex is the number of edges incident to it. The *indegree* and *outdegree* of a vertex of a digraph is the number of edges pointing into and out from the vertex. A vertex of degree 1 is called a *leaf.* If every vertex of a graph have the same degree k, then the graph is said to be k*-regular.*

7.5.1 Vertex Covers

7.5.2 Edge Covers

7.5.3 Bipartite Graphs

7.5.3.1 Matchings

7.5.4 Chromatic Numbers

7.5.5 Eulerian Graphs

7.5.6 Hamiltonian Cycles

7.5.7 Cayley's Tree Enumeration Theorem

7.5.8 Hall's Theorem

7.5.9 Turán's Theorem

7.5.10 Ramsey's Theorem

Chapter 8

Combinatorics II

"Every hard problem in mathematics has something to do with combinatorics."

— Lennart Carleson

8.1 Projective Planes and Latin Squares

8.1.1 Projective Planes

There is a deep connection between algebra and geometry, but once we move beyond linear algebra, there is a certain inconvenience in the vector spaces in which we do geometry, which stems primarily from an asymmetry between points and lines. Any two points are incident to a single line, but it is not true that any two lines are incident to a single point: they could be parallel, or in higher dimensions, skew.

To resolve this imbalance, we add a point "at infinity" which some lines may be incident to in *projective geometry*.

Consider the vector space K^3 over a field K. Removing the origin, we define an equivalence relation on $K^3 \setminus \{\mathbf{0}_K\}$ by $(a,b,c) \sim (x,y,z)$ if there exists $0 \neq \lambda \in K$ such that $(a,b,c) = \lambda(x,y,z)$. That is, vectors are equivalent up to scaling.

The *projective plane over K*, denoted $K\mathbb{P}^2$ is then the set of equivalence classes of non-zero vectors in K^3. Equivalently, the points of $K\mathbb{P}^2$ may be viewed as the lines through the origin in K^3.

Example. Consider the case $K = \mathbb{R}$, giving the *real projective plane* $\mathbb{R}\mathbb{P}^2$. Each line through the origin in $\mathbb{R}^3$ intersects the unit sphere at two antipodal points, so we can also view the set of points of $\mathbb{R}\mathbb{P}^2$ as the surface of the sphere with antipodal points identified.

Intuitively, a line in $\mathbb{R}\mathbb{P}^2$ is then just a great circle on the sphere, also with antipodal points identified. Such a great circle also be viewed as the intersection of a plane through the origin with the sphere, which can be characterised by the normal vector $(\lambda,\mu,\nu) \neq \mathbf{0}_K$. The great circle is then the (equivalence class of the) set of points (x,y,z) satisfying

$$\lambda x + \mu y + \nu z = 0$$

where λ, μ, and ν are elements of K that are not all zero. △

We will be studying the discrete analogue of these spaces in which our projective planes have only finitely many points.

If a field K is finite with q elements, then $K^3 \setminus \{\mathbf{0}_K\}$ has $q^3 - 1$ elements. Each equivalence class has $q - 1$ elements, as there are $q - 1$ non-zero elements of K to scale by. So, there are

$$\frac{q^3 - 1}{q - 1} = q^2 + q + 1$$

elements, or *points*, in $K\mathbb{P}^2$.

A *line* in $K\mathbb{P}^2$ is then the set of points (x,y,z) satisfying

$$\lambda x + \mu y + \nu z = 0$$

where λ, μ, and ν are elements of K that are not all zero. That is, a line is the set of points orthogonal to a point (λ,μ,ν).

Note that this is well-defined due to the bilinearity of the dot product.

Example. Consider the case $K = \mathbb{Z}_2$, the field of two elements $\{0,1\}$ in which $1 + 1 = 0$. In this case, the equivalence classes in $\mathbb{Z}_2\mathbb{P}^2$ are singletons, so the points in the projective planes are given by the seven non-zero vectors

$$(0,0,1), \quad (0,1,0), \quad (0,1,1), \quad (1,0,0), \quad (1,0,1), \quad (1,1,0), \quad (1,1,1)$$

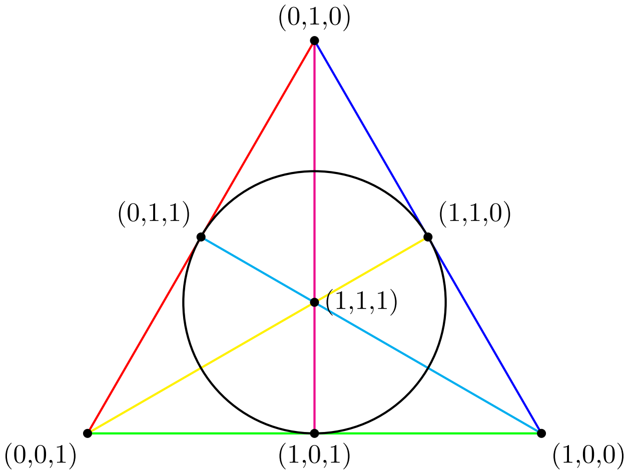
Each line in $\mathbb{Z}_2\mathbb{P}^2$ contains 3 points: for instance, the line represented by $(1,0,0)$ (i.e. $1x + 0y + 0z = 0$) consists of the three points

$$(0,0,1), \quad (0,1,0), \quad (0,1,1)$$

and the line represented by $(1,1,1)$ (i.e. $1x + 1y + 1z = 0$) consists of the three points

$$(0,1,1), \quad (1,0,1), \quad (1,1,0)$$

This projective plane is also called the *Fano plane*, and its points are often drawn arranged in a triangle:



△

As we would expect, any two distinct points determine a unique line connecting them.

Lemma 8.1.1 (Points in the Projective Plane)**.** *Given any two distinct points in $K\mathbb{P}^2$, there is exactly one line incident to both of them.*

Proof. Let the points be (represented by) (a,b,c) and (x,y,z) and form the cross product

$$\begin{aligned}(\lambda,\mu,\nu) &\coloneqq (a,b,c) \times (x,y,z)\\ &= (bz-cy,cx-az,ay-bx)\end{aligned}$$

The cross product is also zero if and only if the two starting vectors are linearly dependent, but by assumption, (a,b,c) and (x,y,z) are representatives of distinct points and are therefore linearly independent, so (λ,μ,ν) is non-zero and defines a line.

By construction, the cross product is orthogonal to (a,b,c) and (x,y,z) (or otherwise, this can be checked by hand), so both the points are incident to the line defined by (λ,μ,ν).

For uniqueness, suppose (λ',μ',ν') defines a line incident to (a,b,c) and (x,y,z):

$$\lambda' a + \mu' b + \nu' c = 0 \tag{1}$$

$$\lambda' x + \mu' y + \nu' z = 0 \tag{2}$$

From the above, we have that $bz-cy$, $cx-az$, and $ay-bx$ are not all zero, so without loss of generality, suppose that $\phi \coloneqq bz-cy \neq 0$. Then, multiplying (1) by z and (2) by c, and subtracting (2) from (1), we have:

$$\begin{aligned}c(\lambda' x + \mu' y + \nu' z) - z(\lambda' a + \mu' b + \nu' c) &= 0 - 0\\ \lambda' cx + \mu' cy - \lambda' az - \mu' bz &= 0\\ \lambda'(cx-az) + \mu'(cy-bz) &= 0\\ \lambda'(cx-az) &= \mu'(bz-cy)\\ \lambda'(cx-az) &= \mu'\phi\end{aligned}$$

and similarly,

$$\lambda'(ay-bx) = \nu'\phi$$

so

$$\begin{aligned}\lambda' &= (\phi^{-1}\lambda')(bz-cy)\\ \mu' &= (\phi^{-1}\mu')(cx-az)\\ \nu' &= (\phi^{-1}\nu')(ay-bx)\end{aligned}$$

so $(\lambda',\mu',\nu') \sim (\lambda,\mu,\nu)$, and the line is unique. ■

The preceding proof shows that if (a,b,c) and (x,y,z) are non-equivalent elements of $K^3 \setminus \{\mathbf{0}\}$, then there is a unique equivalence class of elements $(\lambda,\mu,\nu) \in K^3 \setminus \{\mathbf{0}\}$ satisfying

$$\lambda a + \mu b + \nu c = \lambda x + \mu y + \nu z = 0$$

However, we may also interpret (a,b,c) and (x,y,z) as (equivalence classes of) lines and $\big[(\lambda,\mu,\nu)\big]$ as a point, so this also shows that any pair of distinct lines are incident at a single point.

Lemma 8.1.2 (Lines in the Projective Plane). *Given any two distinct lines in $K\mathbb{P}^2$, there is exactly one point incident to both of them.*

You will notice that this lemma is precisely the same as the previous, only with the words "point" and "line" interchanged. This is not a coincidence: points and lines in projective planes are *dual* in the sense that any result about points and lines in projective planes will still hold true if the two are interchanged.

This is also the reasoning for the choice of wording "incident to" for describing the relation between points and lines, rather than saying "two lines meet at a point" or "two points lie on a line", since this makes the dualisation process easier.

8.1.2 Finite Projective Planes

Based on the previous algebraic construction, we define a combinatorial object.

A *finite projective plane* (*FPP*) is a finite set P of *points*, and a set $L \subseteq \mathcal{P}(P)$ of *lines* satisfying:

(i) Every pair of points are incident to exactly one common line;

(ii) Every pair of points are incident to exactly one common point;

(iii) There are four points, no three of which belong to a single line.

The last condition is there only to rule out certain degenerate cases which lack the desired symmetries we like to work with.

Lemma 8.1.3 (Point-Line Matching)**.** *Let* (P,L) *be an FPP,* $\ell \in L$ *be a line, and* $p \in P$ *be a point not incident to* ℓ*. Then, the number of points incident to* ℓ *is equal to the number of lines incident to* p*.*

Proof. By axiom (i), each point on ℓ is incident to exactly one line through p; and by axiom (ii), each line through p is incident to exactly one point on ℓ. ■

Theorem 8.1.4 (FPP Structure)**.** *Let* (P,L) *be an FPP. Then, there is a number* q *such that:*

(i) *Each line is incident to* $q+1$ *points;*

(ii) *Each point is incident to* $q+1$ *lines;*

(iii) *There are* q^2+q+1 *points;*

(iv) *There are* q^2+q+1 *lines.*

The number q is then called the *order* of the FPP.

Proof.

(i) It suffices to show that any two lines are incident to the same number of points (and call this $q+1$). Suppose ℓ and ℓ' are two lines. By Lemma 8.1.2, it is sufficient to find a point p not on either line, since each line would then be incident to as many points as there are lines incident to p.

Consider the 4 points p_1,p_2,p_3,p_4 guaranteed by axiom (iii). If one is in neither line, we are done. Otherwise, all four are on ℓ or ℓ', and by axiom (iii), there must be two on each line, say, $p_1,p_2 \in \ell$ and $p_3,p_4 \in \ell'$. Now, consider the lines ℓ_{13} and ℓ_{24} connecting p_1 to p_3 and p_2 to p_4, respectively. These lines meet at a point p.

If $p \in \ell$, then p_1 and p are points common to both ℓ and ℓ_{13}, so $\ell = \ell_{13}$ by uniqueness, and p_1, p_2, and p_3 all lie on the line $\ell = \ell_{13}$, contradicting axiom (iii). Similarly, $p \notin \ell'$.

(ii) Let p be a point. Again, by Lemma 8.1.1, it suffices to find a line not incident to p. Consider the 4 points p_1,p_2,p_3,p_4 guaranteed by axiom (iii), and without loss of generality, suppose $p \neq p_1$. Then, the line connecting p_1 and p_2 and the line connecting p_1 and p_3 cannot simultaneously contain p since they already both contain p_1.

(iii) Let p be a point and consider the $q+1$ lines incident to it. Every pair of these lines intersect only at p, so each contains q points other than p, and the lines jointly cover the plane. So the total number of points is $(q+1)q+1 = q^2+q+1$.

(iv) Each point is incident to $q+1$ lines and every line is incident to $q+1$ points, so the number of lines must be equal to the number of points.

■

8.1.3 Latin Squares

A *Latin square* is an $n \times n$ array of n distinct symbols such that every symbol appears in every row and every column.

Example.

$$\begin{array}{cc} \text{A} & \text{B} \\ \text{B} & \text{A} \end{array} \qquad \begin{array}{ccc} \text{A} & \text{B} & \text{C} \\ \text{B} & \text{C} & \text{A} \\ \text{C} & \text{A} & \text{B} \end{array}$$

△

Example. Any Cayley table forms a Latin square. For instance, C_4 yields:

	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

△

Two $n \times n$ Latin squares $A = (a_{ij})$ and $B = (b_{ij})$ are *orthogonal* if the n^2 pairs (a_{ij},b_{ij}) cover all possible pairs.

Example. If we pair

$$\begin{array}{ccc} \text{A} & \text{B} & \text{C} \\ \text{B} & \text{C} & \text{A} \\ \text{C} & \text{A} & \text{B} \end{array} \qquad \text{and} \qquad \begin{array}{ccc} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 2 & 3 & 1 \end{array}$$

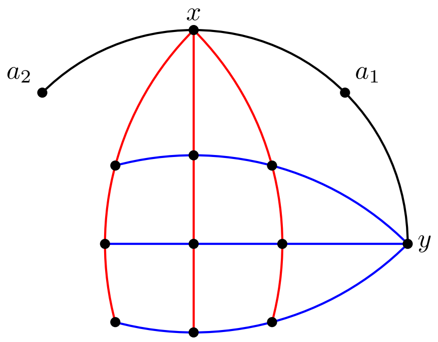
we obtain

$$\begin{array}{ccc} (\text{A},1) & (\text{B},2) & (\text{C},3) \\ (\text{B},3) & (\text{C},1) & (\text{A},2) \\ (\text{C},2) & (\text{A},3) & (\text{B},1) \end{array}$$

All 9 possible pairs are present, so these Latin squares are orthogonal. △

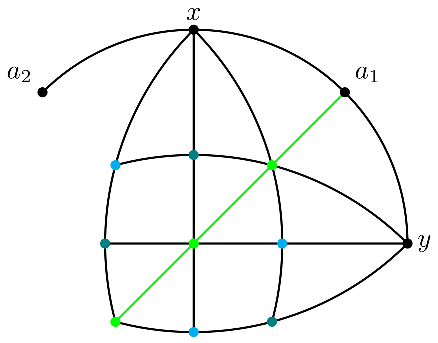
Let (P,L) be an FPP, and choose two points, say $x,y \in P$. They are connected by a line ℓ incident to $q+1-2 = q-1$ other points $a_1,\dots,a_{q-1}$, and $P \setminus \ell$ contains q^2 other points not on this line.

The point x is incident to q other lines, each disjointly incident to q points, so the x-lines hence partition these q^2 other points. Similarly, the point y is incident to q lines, each disjointly incident to q points, so the y-lines also partition these q^2 points. Also, each x-line meets all other y-lines, so the q^2 intersections have a Cartesian product structure and form a grid.



a_1 also lies on q other lines that also partition the grid. Also, each of these q lines meets each x-line and each y-line, so each line through a_1 is incident to q points within the grid; one in each row, and one in each column.

So, these q lines through a_1 generate a Latin square on the grid by labelling the points on the first line by a_{11}, points on the second line a_{12}, etc.



So in the example above, a_1 generates:

$$\begin{array}{ccc} a_{12} & a_{13} & a_{11} \\ a_{13} & a_{11} & a_{12} \\ a_{11} & a_{12} & a_{13} \end{array}$$

while the point a_2 generates:

$$\begin{array}{ccc} a_{21} & a_{23} & a_{22} \\ a_{22} & a_{21} & a_{23} \\ a_{23} & a_{22} & a_{21} \end{array}$$

The $q-1$ points on the line connecting x and y thus generate $q-1$ Latin squares.

Theorem 8.1.5. *The $q-1$ Latin squares generated in this way are pairwise orthogonal.*

Proof. Without loss of generality, consider the Latin squares generated by the points a_1 and a_2. These points each lie on q other lines corresponding to the symbols a_{1j} and a_{2j} used in their Latin squares, and each a_1-line meets each a_2-line at one of the q^2 points of the grid. So, every possible pair of symbols appears when the grids are merged. ■

Theorem 8.1.6. *There is an FPP of order $q > 1$ if and only if there are $q-1$ pairwise orthogonal $q \times q$ Latin squares.*

8.2 Error-Correcting Codes

8.2.1 Introduction

Suppose Alice wishes to send Bob a message encoded in binary across an unreliable or *noisy* channel. That is, some bits in the message may be flipped during transmission.

One simple protocol to resist noise is for Alice to send each bit of the message repeatedly, say, ten times, then for Bob to take the most frequent received bit in each block of ten received bits to be the intended bit.

Message Bit	Code Bit
0	0000000000
1	1111111111

For instance, if Bob receives the string "1011011111 0100011000", he can be fairly confident that the original message was "1 0".

This replacement procedure constitutes an *error-correcting code.* The idea is that only certain strings of ten bits are valid or *admissible* strings, also called *codewords*, and that these admissible strings are selected to be very distinct from each other to minimise the chance that one is converted into by noise.

However, this code is not very *efficient*, because the rate of transmission decreases by a factor of ten when using this code.

Consider the similar repetition code:

Message Bit	Code Bit
0	00
1	11

This code detects one bit errors: if the string 01 is received, Bob will know there has been at least one bit flip. However, this code cannot detect two bit flips, as the intended message 00 could be converted into the admissible string 11 with two bit flips. The rate of this code is also $1/2$.

If we use this code to send two bits of information, we have the encodings:

Message Bit	Code Bit
00	0000
01	0011
10	1100
11	1111

Again, this code is only safe against single bit flips. However, using three bits, we can achieve the same resilience against noise:

Message Bit	Code Bit
00	000
01	011
10	110
11	101

Any pair of these strings differ in two places, so again, two bits have to be corrupted to change one admissible string into another in this code. However, this code is also faster than the previous scheme, having rate 2/3.

Let the *alphabet* A be a set of symbols called *letters.* A *code* is a subset $C \subseteq A^n$, where n is the *length* of the code, and the elements of C are called *codewords.* If a code uses n bits of codewords to send k bits of plaintext, then the code has *rate* k/n.

Example. In the examples above, we have been working over the alphabet $A = \{0,1\}$, and we call such codes *binary.* △

An *encoding* is a bijection $e : W \to C$ from the set W of words in the plain text, to the code C.

Example. The table above describes an encoding from $W = \{0,1\}^2$ to a code $C = \{000,011,110,101\} \subseteq \{0,1\}^3$. △

For our purposes, it will not matter how this encoding is selected; all that is relevant is how "well-separated" the codewords are.

To quantify this separation, we define the *Hamming distance* of two codewords as the number of positions in which they differ. The Hamming distance forms a metric on any set of strings.

Example. The codewords 110 and 011 differ in the first and third positions, so they are Hamming distance $d(110,011) = 2$ apart. △

Note that the minimum separation $\min_{X \neq Y \in C} d(X,Y)$ of a code determines the maximum number of bit flips it can detect, as any number of bit flips exceeding this number could then potentially turn one codeword into another valid codeword. If for a code C, this minimum separation is

$$D := \min_{X \neq Y \in C} d(X,Y)$$

then we say that C is *D-separated.*

8.2.2 Block Codes

We will only be considering *block codes*, where the message is divided into blocks of fixed length k, each of which can be encoded without reference to any of the other blocks. That is, we will take the set of words W to be the set $\{0,1\}^k$ of all possible binary strings of length k.

To *decode* a message encoded with a block code, we break the received transmission into blocks of length n, where n is the length of the code used. If a block is a codeword (i.e. admissible), then we assume that this block was correctly transmitted. Otherwise, we find the closest codeword to the received block, and interpret that as the intended codeword. If the code is well designed, this closest codeword should be unique.

Lemma 8.2.1. *If a block code is $(2r+1)$-separated, then it can correct r bit flips.*

Proof. An invalid block x can be at distance at most r from its nearest codeword y.

$$d(x,y) \leq r$$

Then, for any other codeword z, the reverse triangle inequality gives:

$$\begin{aligned}\left|d(z,y) - d(y,x)\right| &\leq d(x,z)\\ \left|(2r+1) - r\right| &\leq d(x,z)\\ r+1 &\leq d(x,z)\end{aligned}$$

so the codeword closest to x is unique. ■

Given a binary block code C, any encoding gives a bijection $W \to C$, we have $|C| = |W| = 2^k$, so the rate of a binary block code is simple to compute as:

$$\frac{\log_2|C|}{n} = \frac{k}{n}$$

In a binary code, we can also interpret the alphabet $\{0,1\}$ as the finite field $\mathbb{Z}_2 := \mathbb{Z}/2\mathbb{Z}$ of characteristic 2. If the codewords have length n, then they can be interpreted as elements of the vector space $\mathbb{Z}_2^n$.

In the code of length 3 above, we had the vectors

$$000, \qquad 011, \qquad 110, \qquad 101$$

These elements form a linear subspace of $\mathbb{Z}_2^3$, and we call this code a *linear code.*

The aim is to find codes in which every pair of codewords are far from each other in the Hamming metric. Linear codes have a simple feature that makes this easier to achieve:

Lemma 8.2.2. *Suppose $C \subseteq \mathbb{Z}_2^n$ is a linear code such that every element of C other than $\mathbf{0}$ contains at least D coordinates equal to 1. Then, C is D-separated. That is,*

$$\forall(X \neq Y \in C) : d(X,Y) \geq D$$

Proof. Suppose u and v are distinct codewords with $d(u,v) = r < d$. That is, they differ only in $r < d$ coordinates. Then, $u \oplus v$ is also a codeword, as C is linear. This codeword has a 1 only in the positions where u and v disagree, since vector addition in $\mathbb{Z}_2^n$ is componentwise exclusive disjunction, so $u \oplus v$ has $r < d$ coordinates equal to 1, contradicting the construction of C. ■

It is also very simple to determine the rate of a linear code; if C is a linear subspace of $\mathbb{Z}_2^n$ of dimension k, then it has 2^k elements, so the rate is just

$$\frac{\dim C}{n} = \frac{k}{n}$$

Example. Consider the code

$$C = \{000,011,110,101\} \subset \mathbb{Z}_n^3$$

The minimum number of 1s in a non-zero codeword is 2, so C is 2-separated. C is also a 2-dimensional subspace of $\mathbb{Z}_2^3$, so the rate is $\frac{2}{3}$. △

So, the goal is to look for k-dimensional subspaces of $\mathbb{Z}_2^n$ whose non-zero elements contain a large number of 1s to get good separation. We also want k to be large to get a high rate. Since we want this subspace to have large dimension, it is usually easier to define it using $n-k$ linear equations, rather than using a basis of size k.

Example. The code above consists of the elements $(x_1,x_2,x_3) \in \mathbb{Z}_2^3$ which contain an even number of 1s, so they can be specified to be the elements satisfying

$$x_1 + x_2 + x_3 = 0$$

In other words, this code is defined by a *parity check.* △

8.2.3 Hamming Codes

In this section, we describe an efficient 3-separated binary code based on parity checks.

Let r be a positive integer and $n = 2^r - 1$ be the length of the code. The code will be an $(n-r)$-dimensional subspace of $\mathbb{Z}_2^n$, so we need r linear equations to specify the codewords, and the rate will be almost 1:

$$\text{rate}(C) = \frac{\log_2(n)}{n} = \frac{n-r}{n} = 1 - \frac{r}{n} = 1 - \frac{r}{2^r - 1} \approx 1$$

for large r.

We arrange the linear equations into an $r \times n$ matrix B, so the code will be given by

$$C = \{x \in \mathbb{Z}_2^n : Bx = 0\}$$

The *Hamming code* of length n is given by the matrix with columns consisting of all binary numbers from 1 to n.

Example. For $r = 3$ and $n = 2^r - 1 = 7$, then B is given by

$$B = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

△

Note that $(1,1,1,0,\ldots,0) \in C$ regardless of r, since the first 3 columns of B have 0 below the top two rows. So, Hamming codes are at most 3-separated.

Lemma 8.2.3. *Hamming codes are precisely* 3*-separated.*

Proof. We have already shown that Hamming codes are at most 3-separated. If two words are less than distance 3 apart, then they must either be distance 1 or distance 2 apart. In the former case, both would fail to satisfy a parity check involving the different bit, and in the latter case, then there would be a non-codeword in betweem them, which would decode as either one of them. ■

As with any block code, to decode a message encoded with a Hamming code, we break the received code bits into blocks of length n, and check if the block is a codeword. If it is, then we assume that this block was correctly transmitted; otherwise, we find the closest codeword to the received block, and interpret that as the intended codeword.

We show that this closest codeword is at distance at most 1 for any possible received block, and furthermore, that this closest codeword is unique.

Theorem 8.2.4. *Let C be a Hamming code of length $n = 2^r - 1$. Then, for any invalid block $x \in \mathbb{Z}_2^n \setminus C$, there is a unique element y of C at distance $d(x,y) = 1$ from x.*

Proof. Suppose $x \in \mathbb{Z}_2^n \setminus C$ is not a codeword, so $Bx \neq \mathbf{0}$. Consider the vector

$$u := Bx \in \mathbb{Z}_2^r \setminus \{\mathbf{0}\}$$

By construction, the columns of B contain all possible non-zero vectors, so u must coincide with some column of B, say, the mth column B_m representing the binary number m.

Now, consider the vector

$$\tilde{x} := x \oplus \mathbf{e}_m$$

where $\mathbf{e}_m$ is the standard mth basis vector, so $\tilde{x}$ differs from x only in the mth coordinate. In particular, $d(x,\tilde{x}) = 1$.

Multiplying $\tilde{x}$ by B, we have

$$\begin{aligned} B\tilde{x} &= Bx \oplus B\mathbf{e}_m \\ &= u \oplus B_m \\ &= B_m \oplus B_m \\ &= 0 \end{aligned}$$

and we see that $\tilde{x}$ is a codeword at distance 1 from x, as required.

For uniqueness, suppose u and v are distinct codewords at distance 1 from x. Then, by the triangle inequality, we have $d(u,v) \leq d(u,x) + d(x,v) = 2$, contradicting that Hamming codes are 3-separated. ■

Example. Decode the received string 1010011.

The first bit parity checks the positions whose unit digit is 1. That is, the 1st, 3rd, 4th, and 7th bits. We have $1 + 1 + 0 + 1 = 1$, so an error has occured somewhere within in these odd digits.

The second bit parity checks the positions whose 2s digit is 1. That is, the 2nd, 3rd, 6th, and 7th bits. We have $0 + 1 + 1 + 1 = 1$, so an error has occured somewhere within these digits.

The fourth bit parity checks the positions whose 4s digit is 1. That is, the 4th, 5th, 6th, and 7th bits. We have $0 + 0 + 1 + 1 = 0$, so no error has occured within these digits.

From this, we deduce that the 3rd digit has been flipped, so the closest codeword we correct to is 1000011. △

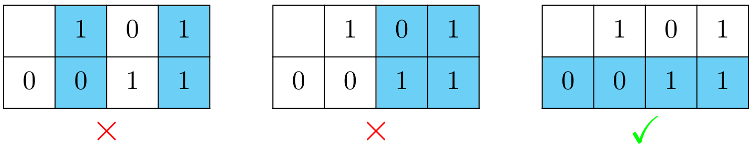
This process is somewhat involved, so we present a graphical method to quickly decide which positions to parity check:

Example. Decode the received string 1010011.

Arrange the string into a grid as follows, skipping the first entry:

	1	0	1
0	0	1	1

Now, perform parity checks within each of the highlighted regions:

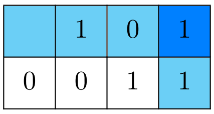


	1	0	1
0	0	1	1

	1	0	1
0	0	1	1

	1	0	1
0	0	1	1

The error is in the first two regions, so it must be in the 4th column. The error is not in the last region, so it must be in the first row.



	1	0	1
0	0	1	1

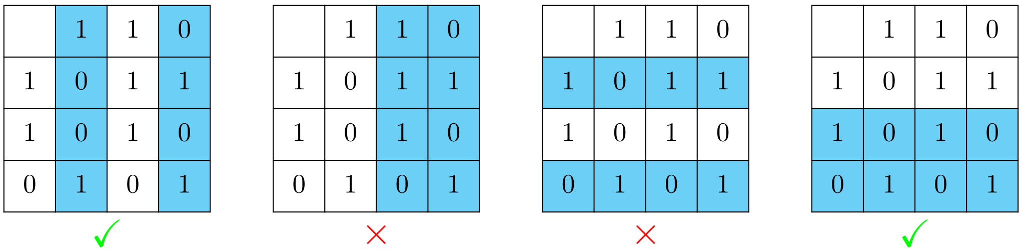
We deduce that the error is in the 3rd bit, so the closest codeword we correct to is 1000011. △

Example. Decode the received string 110 1011 1010 0101.

Arrange the string into a grid as follows, skipping the first entry:

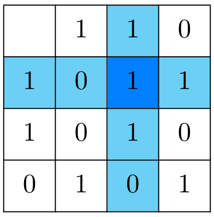
	1	1	0
1	0	1	1
1	0	1	0
0	1	0	1

Now, perform parity checks within each of the highlighted regions:



	1	1	0
1	0	1	1
1	0	1	0
0	1	0	1

The error is in the second region, but not the first, so the error must be in the third column; the error is in the third region, but not the fourth, so the error is in the second row:



	1	1	0
1	0	1	1
1	0	1	0
0	1	0	1

So, we correct the received block to 110 1001 1010 0101. △

Theorem 8.2.5 (Sphere-packing Bound). *Let C be a $(2r+1)$-separated binary code of length n. Then,*

$$|C| \sum_{i=0}^{r} \binom{n}{i} \leq 2^n$$

Proof. In Lemma 8.2.1, we showed that if a block code is $(2r+1)$-separated, then it can correct r bit flips, since every string that is distance at most r from a codeword x is closer to x than any other codeword. In other words, the balls of radius r centred on each codeword are all disjoint.

How many strings are contained in each ball of radius r?

We count these strings based on their distance from the centre x. If we change 0 bits from x, then we just get the string x; if we change 1 bit, then there are $n = \binom{n}{1}$ many strings at distance 1 from x; if we change 2 bits, then there are $\binom{n}{2}$ many strings at distance 2 from x; and so on, so the balls each contain

$$1 + \binom{n}{1} + \binom{n}{2} + \binom{n}{3} + \cdots + \binom{n}{r} = \sum_{i=0}^{r} \binom{n}{i}$$

strings. Since these balls are all disjoint, the total number of elements contained in all of these balls is $|C|$ times this sum, and there are 2^n possible strings, so

$$|C| \sum_{i=0}^{r} \binom{n}{i} \leq 2^n$$

as required. ■

A code that attains this bound is called a *perfect code.*

Lemma 8.2.6 (Domain of Codewords)**.** *Let C be the Hamming code of length $n = 2^r - 1$. Then, each codeword is the closest codeword to n other elements of $\mathbb{Z}_2^n$.*

Proof. There are n possible bits to flip in each codeword. ■

Theorem 8.2.7. *Hamming codes are perfect codes.*

Proof. By the previous lemma, each codeword is closest to 2^r many possible bit strings: itself, and $n = 2^r - 1$ others adjacent to it. Also, there are $|C| = 2^{n-r}$ many codewords, so

$$\begin{aligned} |C| \cdot \left|B_r\right| &= 2^{n-r}2^r \\ &= 2^n \end{aligned}$$

■

8.2.4 Shannon's Theorem

Suppose we have a binary communication channel which flips bits with probability p. We define the *Shannon capacity* to be

$$R \coloneqq 1 + p\log_2(p) + (1-p)\log_2(1-p)$$

Theorem 8.2.8 (Shannon's Limit)**.** *Using a binary communication channel which flips bits with probability p, there exists a code C with rate almost R and almost perfect accuracy.*

That is, for all $\varepsilon > 0$, there exists a code C with rate

$$\text{rate}(C) \geq 1 + p\log_2(p) + (1-p)\log_2(1-p) - \varepsilon$$

and such that the probability of decoding a codeword incorrectly is less than ε. Moreover, subject to any accuracy constraint, rates greater than R are not achievable.

Proof sketch. Choose a large value of n for the length of a block code and let F be a random variable measuring the number of bits flipped out of a message of length n. F has expectation $\mathbb{E}(F) = np$ and standard deviation $\sigma = \sqrt{npq}$ (where $q = 1-p$). Notably, for large n, $\sigma \ll \mathbb{E}(F)$, so we will almost never have more than np bits flipped. Let $d \coloneqq n(p+\varepsilon)$.

The first idea one might have is to find a $(2d+1)$-separated code with the given rate, but it turns out that this is extremely difficult to do.

Instead, choose M codewords from $\mathbb{Z}_2^n$ uniformly and independently to form a code C. Now, suppose a codeword S is sent using this scheme, and is received as the string S'. There are two things that could go wrong during decoding:

A: More than d bits are flipped.

B: There is an incorrect codeword $Y \neq S$ at distance $d(S',Y) \leq d$ from S'.

The first case is rare by our choice of d, since $\sigma \ll \mathbb{E}(F) < d$. The second case occurs whenever one of the $M-1$ codewords X in $C \setminus \{S\}$ is within distance d of S'

$$\mathbb{P}(B) = (M-1)\frac{\#\text{ of strings } X \text{ with } d(S',X) \leq d}{2^n}$$

As before, the number of strings at distance r from S' is given by $\binom{n}{r}$, so the numerator is given by the sum

$$= (M-1)\frac{\sum_{i=0}^{d}\binom{n}{i}}{2^n}$$

If d is not too large, then the sum of binomial coefficients is approximately the last summand, so $\mathbb{P}(B) \approx (M-1)\frac{\binom{n}{d}}{2^n}$. So, we need $(M-1)\binom{n}{d}$ to be small compared to 2^n:

$$(M-1)\binom{n}{d} = \alpha 2^n$$

for some small constant $\alpha > 0$.

The rate of C is then given by

$$\begin{aligned}
\text{rate}(C) &:= \frac{\log_2(M)}{n} \\
&\approx \frac{1}{n}\log_2\left(\frac{2^n\alpha}{\binom{n}{d}}\right) \\
&= \frac{1}{n}\Big(\log_2(2^n) + \log_2(\alpha) - \log_2\tbinom{n}{d}\Big) \\
&= \frac{1}{n}\Big(n + \log_2(\alpha) - \log_2\tbinom{n}{d}\Big) \\
&= 1 + \frac{1}{n}\log_2(\alpha) - \frac{1}{n}\log_2\tbinom{n}{d}
\end{aligned}$$

$\frac{1}{n}\log_2(\alpha)$ is very small even for small α, so,

$$\begin{aligned}
&\approx 1 - \frac{1}{n}\log_2\tbinom{n}{d} \\
&= 1 - \frac{1}{n}\log_2\left(\frac{n!}{d!(n-d)!}\right) \\
&\approx 1 - \frac{1}{n}\log_2\left(\frac{n!}{(np)!\big(n(1-p)\big)!}\right)
\end{aligned}$$

By Stirling's formula and discarding sublinear factors,

$$\begin{aligned}
&\approx 1 - \frac{1}{n}\log_2\left(\frac{n^n}{(np)^{np}\big(n(1-p)\big)^{n(1-p)}}\right) \\
&= 1 - \frac{1}{n}\log_2\left(\frac{n^n}{n^{np}p^{np}n^{n(1-p)}(1-p)^{n(1-p)}}\right) \\
&= 1 - \frac{1}{n}\log_2\left(\frac{n^n}{n^n p^{np}(1-p)^{n(1-p)}}\right) \\
&= 1 - \frac{1}{n}\log_2\left(\frac{1}{p^{np}(1-p)^{n(1-p)}}\right) \\
&= 1 - \frac{1}{n}\log_2\left(\left(\frac{1}{p^p(1-p)^{(1-p)}}\right)^n\right) \\
&= 1 - \log_2\left(\frac{1}{p^p(1-p)^{(1-p)}}\right) \\
&= 1 + \log_2\big(p^p(1-p)^{(1-p)}\big)
\end{aligned}$$

$$= 1 + p\log_2(p) + (1-p)\log_2(1-p)$$

■

The point is that, if we fix the rate, then the number d of bit flips we can correct is proportional to n, and if we fix the probability p, then the expected number of bits flipped is np, also proportional to n. However, the standard deviation $\sigma = \sqrt{npq}$ grows slower than proportionally to n, so as n increases, the chance that $n(p+\varepsilon)$ bits are flipped decreases to 0.

8.3 Discrete Geometry

A set $C \subseteq \mathbb{R}^d$ is *convex* if for all pairs $x,y \in C$, we have $\lambda x + (1-\lambda)y \in C$ for all $\lambda \in [0,1]$. That is, the line segment connecting x to y is also contained in C.

Example. The unit ball of any normed space is convex. △

Lemma 8.3.1. *The arbitrary intersection of convex sets is convex.*

Proof. Let $\{S_i\}_{i=1}^n$ be a family of convex sets. Then, for any $x,y \in \bigcap_{i=1}^n S_i$ we have $x,y \in S_i$ for all i, and all the S_i are convex, so $\lambda x + (1-\lambda)y \in S_i$ for all i, so $\lambda x + (1-\lambda)y \in \bigcap_{i=1}^n S_i$ and $\bigcap_{i=1}^n S_i$ is convex. ■

The expression $\lambda x + (1-\lambda)y$ is called a *convex combination* of x and y. More generally, the convex combination of a collection of points $x_1,\ldots,x_m \in \mathbb{R}^d$ is a point of the form

$$\sum_{i=0}^{m} \lambda_i x_i$$

where $\lambda_i \geq 0$, and $\sum_{i=0}^m \lambda_i = 1$.

We write $\mathrm{cc}(E)$ to denote the set of all convex combinations of a set E.

Lemma 8.3.2. *The convex combinations operator is idempotent:*

$$\mathrm{cc}\big(\mathrm{cc}(E)\big) = \mathrm{cc}(E)$$

Lemma 8.3.3. *A set E is convex if and only if it it contains all of its convex combinations:*

$$E = \mathrm{cc}(E)$$

Proof. For the forward direction, we induct on m. For $m = 1$, a convex combination of points in E is just a point in E. Now suppose E contains all convex combinations of at most m of its points. Then, we can reduce

$$\begin{aligned}
\sum_{i=0}^{m+1} \lambda_i x_i &= \lambda_{m+1}x_{m+1} + \sum_{i=1}^{m} \lambda_i x_i \\
&= \lambda_{m+1}x_{m+1} + \left(\sum_{i=1}^{m} \lambda_i\right)\left(\sum_{i=1}^{m} \frac{\lambda_i}{\sum_{j=1}^m \lambda_j} x_i\right) \\
&= \lambda_{m+1}x_{m+1} + (1-\lambda_{m+1})\left(\sum_{i=1}^{m} \frac{\lambda_i}{\sum_{j=1}^m \lambda_j} x_i\right)
\end{aligned}$$

The sum on the right is a convex combination, and by the inductive hypothesis, this is an element of E. Then, by convexity of E, the whole expression is a point in C.

For the reverse direction, if E contains all convex combinations of its points, then it contains all convex combinations of two of its points, which is the definition of convexity. ■

Given a set $E \subseteq \mathbb{R}^d$, we define the *convex hull* of E to be the intersection of all convex sets containing E.

$$\operatorname{conv}(E) \coloneqq \bigcap_{\substack{C \supseteq E \\ C \text{ convex}}} C$$

As the intersection of convex sets, the convex hull is convex. The convex hull also satisfies:

(*i*) $E \subseteq \operatorname{conv}(E)$, since every set in the intersection contains E;

(*ii*) If C is convex and $E \subseteq C$, then $\operatorname{conv}(E) \subseteq C$, since the intersection is the minimal element of the poset of convex sets containing E.

Theorem 8.3.4. *For any $E \subseteq \mathbb{R}^d$,*

$$\operatorname{conv}(E) = \operatorname{cc}(E)$$

Proof. Note that $\operatorname{cc}(E)$ is convex and that $E \subseteq \operatorname{cc}(E)$ since every point in E is a convex combination of itself. So, by property (*ii*) of convex hulls, $\operatorname{conv}(E) \subseteq \operatorname{cc}(E)$.

Any point of $\operatorname{cc}(E)$, i.e. a convex combination of points in E, is also a convex combination of points in $\operatorname{conv}(E)$ since $E \subseteq \operatorname{conv}(E)$. Because $\operatorname{conv}(E)$ is convex, it contains all of these convex combinations, so $\operatorname{cc}(E) \subseteq \operatorname{conv}(E)$. ■

8.3.1 Separation

We will be concerned almost entirely with convex sets that are closed, and in almost all cases, they will also be bounded and hence compact (by Heine-Borel).

Lemma 8.3.5. *Every linear functional $\phi : \mathbb{R}^d \to \mathbb{R}$ is of the form*

$$x \mapsto \langle x,y \rangle$$

where y is some fixed non-zero vector in $\mathbb{R}^d$.

Proof. Let ϕ be a linear functional. Define $y_i \coloneqq \phi(\mathbf{e}_i)$ for each $0 \le i \le d$, where $\mathbf{e}_i$ is the ith standard basis vector. Then, if $x = \sum_{i=1}^d x_i \mathbf{e}_i$, we have

$$\begin{aligned}\phi(x) &= \phi\left(\sum_{i=1}^d x_i \mathbf{e}_i\right) \\ &= \sum_{i=1}^d x_i \phi(\mathbf{e}_i) \\ &= \sum_{i=1}^d x_i y_i \\ &= \langle x,y \rangle\end{aligned}$$

■

We define a *hyperplane* in $\mathbb{R}^d$ to be a set of the form

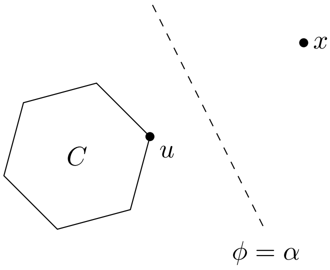
$$\Pi = \{x \in \mathbb{R}^d : \phi(x) = \alpha\}$$

for some non-zero linear functional ϕ and constant $\alpha \in \mathbb{R}$. Equivalently, it is an affine subspace of codimension 1.

Theorem 8.3.6 (Separation Principle I)**.** *If $C \subseteq \mathbb{R}^d$ is compact and convex, and $x \in \mathbb{R}^d \setminus C$, then there is a hyperplane separating x from C. That is, there exists a linear functional $\phi : \mathbb{R}^d \to \mathbb{R}$ and a number α such that*

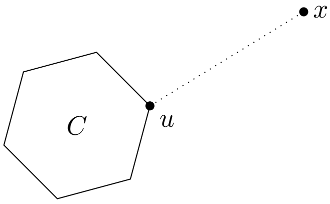
- *$\phi(x) > \alpha$;*
- *$\phi(c) < \alpha$ for all $c \in C$.*

Example.



△

Proof. Consider the function $C \to \mathbb{R}$ defined by $c \mapsto \|x - c\|$ that returns the distance from a point in C to the point x. This function is continuous, and so has a minimum on C. That is, there is a closest point u of C to x. Since $x \notin C$, $u \neq x$, so $\|x - c\| > 0$.



Now, define $\phi : \mathbb{R}^d \to \mathbb{R}$ by $y \mapsto \langle x - u, y\rangle$. We have

$$\begin{aligned}\phi(x) - \phi(u) &= \langle x - u, x\rangle - \langle x - u, u\rangle \\ &= \langle x - u, x - u\rangle \\ &= \|x - u\|^2 \\ &> 0\end{aligned}$$

so $\phi(u) < \phi(x)$. Let α satisfy $\phi(u) < \alpha < \phi(x)$. To complete the proof, it remains to show that if $c \in C$, then $\phi(c) \leq \phi(u)$.

Suppose $c \in C$, but $\phi(c) > \phi(u)$. Consider the convex combination $p := \delta c + (1 - \delta)u$. Then,

$$\begin{aligned}\|x - p\|^2 &= \|x - \delta c - (1 - \delta)u\|^2 \\ &= \|(x - u) - \delta(c - u)\|^2 \\ &= \|x - u\|^2 - 2\delta\langle x - u, c - u\rangle + \delta^2\|c - u\|^2 \\ &= \|x - u\|^2 - 2\delta\phi(c - u) + \delta^2\|c - u\|^2\end{aligned}$$

By assumption, $\phi(c - u) < 0$, and for small δ, $\delta^2\|c - u\|^2 \ll 2\delta\langle x - u, c - u\rangle$, so,

$$< \|x - u\|^2$$

so p is closer to x than u, contradicting the construction of u. ■

A *half-space* of $\mathbb{R}^d$ is a set

$$H = \{x \in \mathbb{R}^d : \phi(x) \leq \alpha\}$$

for some non-zero linear functional ϕ and constant $\alpha \in \mathbb{R}$.

Corollary 8.3.6.1. *If $C \subset \mathbb{R}^d$ is a compact convex set, then C can be expressed as an intersection of half-spaces.*

Proof. For each point $x \in \mathbb{R}^d \setminus C$, there is a half-space containing C and not x, so the intersection of all half-spaces containing C will exclude all points $x \in \mathbb{R}^d \setminus C$ and hence is equal to C. ■

Given a set C, a *supporting hyperplane* of C is a hyperplane H that contains a boundary point x of C, but does not intersect the interior of C. Or equivalently, the (non-zero) linear functional ϕ given by the orthogonal vector of the hyperplane satisfies $\phi(c) \leq \phi(x)$ for all $c \in C$.

Theorem 8.3.7 (Supporting Hyperplanes)**.** *If $C \subset \mathbb{R}^d$ is compact and convex, and $x \in \partial C$, then there is a hyperplane supporting C at x. That is, there is a non-zero linear functional $\phi : \mathbb{R}^d \to \mathbb{R}$ such that $\phi(c) \leq \phi(x)$ for all $c \in C$.*

Proof. Let $(x_i)_{i=1}^{\infty} \subseteq \mathbb{R}^d \setminus C$ be a sequence of points converging to x. By the separation principle, there exists, for each i, a linear functional ϕ_i defined by

$$u \mapsto \langle u,v_i \rangle$$

and a number α_i such that

- $\phi_i(x_i) > \alpha$;
- $\phi_i(c) < \alpha$ for all $c \in C$.

Without loss of generality, suppose each v_i is a unit vector, and hence that they have an accumulation point v which is also a unit vector. Passing to a subsequence and re-indexing, assume that $(v_i) \to v$.

For each i, we have

$$\langle x,v_i \rangle < \alpha_i < \langle x_i,v_i \rangle$$

Taking limits as $i \to \infty$, the inner products converge to $\langle x,v \rangle$, so $(a_i) \to \langle x,v \rangle$. Then, for each $c \in C$, we have $\langle c,v_i \rangle \leq \alpha_i$, so taking limits, we have $\langle c,v \rangle \leq \langle x,v \rangle$, as required. ■

We can relax the hypotheses of the separation principle by only requiring that C is closed and not compact:

Theorem 8.3.8 (Separation Principle II)**.** *If $C \subseteq \mathbb{R}^d$ is closed and convex, and $x \in \mathbb{R}^d \setminus C$, then there is a hyperplane separating x from C. That is, there exists a linear functional $\phi : \mathbb{R}^d \to \mathbb{R}$ and a number α such that*

- *$\phi(x) > \alpha$;*
- *$\phi(c) < \alpha$ for each $c \in C$.*

Proof. Let $c \in C$ and define $R := \|x - c\|$. Now, consider the intersection of C with the ball $B_R(x)$ of radius R centred on x. This is a compact set, so it has a point u closest to x. The proof from this point onwards is then identical to the first form of the theorem. ■

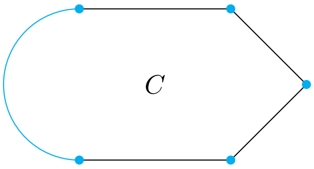
8.3.2 Extrema

Given a convex set $C \subseteq \mathbb{R}^d$, an *extreme point* of C is a point $c \in C$ not in the interior of any line segment contained in C. That is, if c is an extreme point and $x,y \in C$ satisfy

$$c = \lambda x + (1-\lambda)y$$

with $\lambda \in (0,1)$, then $x = y = c$.

Example. In the following figure, the blue points are extreme points.



△

Lemma 8.3.9 (Extreme Points of Faces). *Let H be a supporting hyperplane to the compact convex set C. Then,*

(*i*) *$H \cap C$ is compact and convex;*

(*ii*) *Every extreme point of $H \cap C$ is an extreme point of C.*

Proof.

(*i*) The intersection of a compact and closed set is compact, and the intersection of convex sets is convex, so $H \cap C$ is compact and convex.

(*ii*) Now, suppose x is an extreme point of $H \cap C$, but not an extreme point of C, so it is in the interior of a line segment in C. Because x is extreme in $H \cap C$, this line segment cannot be contained in $H \cap C$, so the segment must have endpoints on either side of H. But, this is impossible, since C is on one side of H.

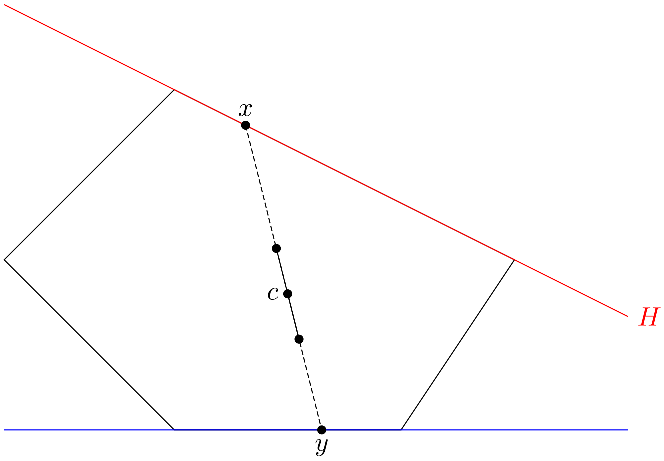
■

Theorem 8.3.10 (Extreme Point Theorem). *Let $C \subseteq \mathbb{R}^d$ be compact and convex, and let E be the set of its extreme points. Then,*

$$C = \operatorname{conv}(E) = \operatorname{cc}(E)$$

Proof. We induct on d. For $d = 1$, this is trivial.

We already know that $\operatorname{conv}(E) \subseteq C$, so we show the other inclusion. Let $c \in C$. If c is an extreme point, then there is nothing to prove. So, supposing otherwise, c lies on a line segment in C, which we may extend in each direction until it intersects the boundary of C, at points, say, x and y.



Let H be a supporting hyperplane to C at x. Then, $H \cap C$ is a compact convex set by part (i) of the previous lemma, and it has codimension at least 1, so by the strong inductive hypothesis, $H \cap C$ is the convex hull of its extreme points, so

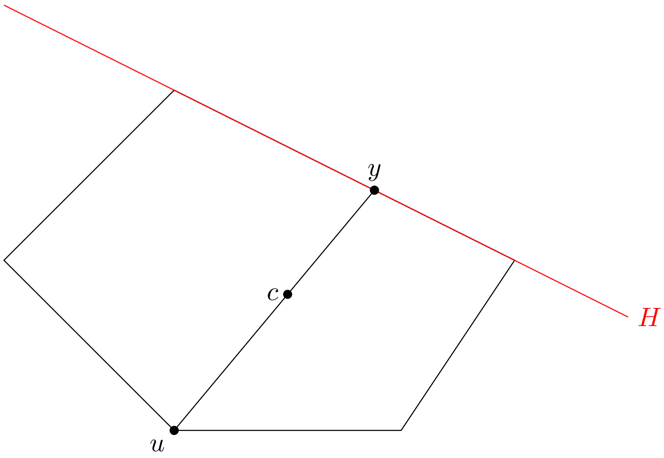
$$x \in H \cap C = \operatorname{conv}\bigl(E_{H \cap C}\bigr)$$

By part (ii) of the previous lemma, $E_{H \cap C} \subseteq E$, so $x \in \operatorname{conv}(E)$. Through an identical argument, we have $y \in \operatorname{conv}(E)$. Then, c lies on the line segment connecting x and y, so we also have $c \in \operatorname{conv}(E)$, as required. ■

Theorem 8.3.11 (Caratheodory)**.** *Each point of a compact convex set $C \subset \mathbb{R}^d$ is a convex combination of at most $d+1$ of its extreme points.*

Proof. We induct on d. For $d = 1$, this is trivial.

Let $c \in C$ and choose an extreme point u of C. Consider the line passing through c and u. This line intersects the boundary of C at u on one side of c, and at a point y on the other.



Now, let H be a supporting hyperplane to C at y. By the strong inductive hypothesis, y is a convex combination of at most d extreme points in $H \cap C$, so c is a convex combination of these at most d points, and an extra point u. ■

8.3.3 Polyhedra and Polytopes

Here are two important constructions of convex sets in $\mathbb{R}^d$:

- A *polyhedron* is a bounded intersection of a finite set of half-spaces.
- A *polytope* is the convex hull of a finite set E.

Theorem 8.3.12 (Polyhedra are Polytopes)**.** *Every polyhedron $C \subset \mathbb{R}^d$ is a polytope.*

Proof. Let

$$C \coloneqq \bigcap_{i=1}^{n} S_i$$

be the bounded intersection of half-spaces S_i bounded by hyperplanes H_i. Let E be the set of extreme points of C.

We claim that every extreme point of C is the intersection of at most d hyperplanes. That is, for all $e \in E$, there exists $I \subseteq [n]$ with $|I| \leq d$ such that

$$\{e\} = \bigcap_{i \in I} H_i$$

This would imply that there are at most $\binom{n}{d}$ (i.e. finitely many) extreme points, so $C = \operatorname{conv}(E)$ would be a polytope.

To prove the claim, we induct on d. For $d = 1$, this is trivial.

Let $e \in E$. If it were in the interior of all the half-spaces, then it would be in the interior of the intersection, C, in which case, e is not extreme. So, e must be on one of the hyperplanes, say H_1.

Note that $H_1 \cap C = \bigcap_{i=1}^{n}(H_1 \cap S_i)$, and each $H_1 \cap S_i$ is a half-space, so $H_1 \cap C$ is a polyhedron of dimension at most $d - 1$. Moreover, e is extreme in $H_1 \cap C$, since if it were not, any line witnessing this would also witness this in C, contradicting that $e \in E$.

So, by the strong inductive hypothesis, in $H_1 \cap C$, we have

$$\{e\} = \bigcap_{i \in I'} (H_1 \cap H_i)$$

where $|I'| \leq d - 1$. So, in C, we have

$$\begin{aligned}\{e\} &= H_1 \cap \bigcap_{i \in I'} H_i \\ &= \bigcap_{i \in I' \cup \{1\}} H_i\end{aligned}$$

so e is the intersection of at most $|I' \cup \{1\}| \leq (d - 1) + 1 = d$ hyperplanes, completing the induction. This proves the claim, and the result follows. ■

8.3.4 Polars

Given a compact convex set $C \subseteq \mathbb{R}^d$, we define its *polar* to be the set

$$C^\circ \coloneqq \{y \in \mathbb{R}^d : \forall x \in C, \langle x, y \rangle \leq 1\}$$

Under very weak conditions, polarity gives a bijection between C and C°:

Lemma 8.3.13. *If $C \subseteq \mathbb{R}^d$ is a compact convex set containing* $\mathbf{0}$*, then polarity is an involution:*

$$C^{\circ\circ} = C$$

Proof. By definition of a polar, for all $x \in C$ and $y \in C^\circ$, we have $\langle x,y \rangle = \langle y,x \rangle \leq 1$. By symmetry of the inner product, we have $\langle y,x \rangle \leq 1$, and

$$C^{\circ\circ} = \{x' \in \mathbb{R}^d : \forall y \in C^\circ, \langle y,x' \rangle \leq 1\}$$

so $C \subseteq C^{\circ\circ}$.

Now, suppose $C^{\circ\circ} \not\subseteq C$, so there exists $x \in C^{\circ\circ} \setminus C$ satisfying $\langle x,y \rangle \leq 1$ for all $y \in C^\circ$.

By the separation principle, there exists a linear functional $\phi(v) = \langle v,u \rangle$ for some fixed u and a constant α such that $\phi(x) > \alpha$ and $\phi(c) < \alpha$ for all $c \in C$.

Since $\mathbf{0} \in C$, $\alpha > \phi(0) = \langle 0,u \rangle > 0$, so by rescaling the orthogonal vector u to $u' := \frac{1}{\alpha}u$, we may assume that the constant is $\alpha' = 1$, so $\phi(c) = \langle c,u' \rangle < \alpha' = 1$ for all $c \in C$, and hence $u' \in C^\circ$. But then, $\phi(x) = \langle x,u' \rangle > \alpha' = 1$, so $x \notin C^{\circ\circ}$, contradicting our choice of x. ■

Lemma 8.3.14 (Polytope Polars)**.** *If $C = \operatorname{conv}\big(\{x_i\}_{i=1}^m\big)$, then*

$$C^\circ = \{y \in \mathbb{R}^d : \forall i, \langle x_i,y \rangle \leq 1\}$$

That is, we only have to check that $\langle x,y \rangle \leq 1$ for the vertices x_i, and not every point $x \in C$.

Proof. Define

$$C' := \{y \in \mathbb{R}^d : \forall i, \langle x_i,y \rangle \leq 1\}$$

Any $y \in C^\circ$ satisfies $\langle x_i,y \rangle \leq 1$ for all x_i, since $x_i \in C$, so $C^\circ \subseteq C$.

For the reverse inclusion, let $y \in C'$. Then, any $x \in C = \operatorname{conv}\big(\{x_i\}_{i=1}^m\big)$ is a convex combination

$$x = \sum_{i=1}^m \lambda_i x_i$$

so

$$\begin{aligned}\langle x,y \rangle &= \sum_{i=1}^m \lambda_i \langle x_i,y \rangle \\ &\leq \sum_{i=1}^m \lambda_i \\ &= 1\end{aligned}$$

so $y \in C^\circ$. Since y was arbitrary, we have $C' \subseteq C^\circ$. ■

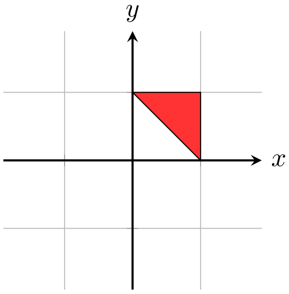
This lemma allows us to interpret the vectors of C as facets of C°. More precisely, notice that for any fixed $i \leq m$, the set

$$\{y \in \mathbb{R}^d : \langle x_i,y \rangle \leq 1\}$$

is a half-space (i.e. with orthogonal vector x_i and $\alpha = 1$), so this lemma equivalently says that C° is the intersection of these m half-spaces. From this, we deduce:

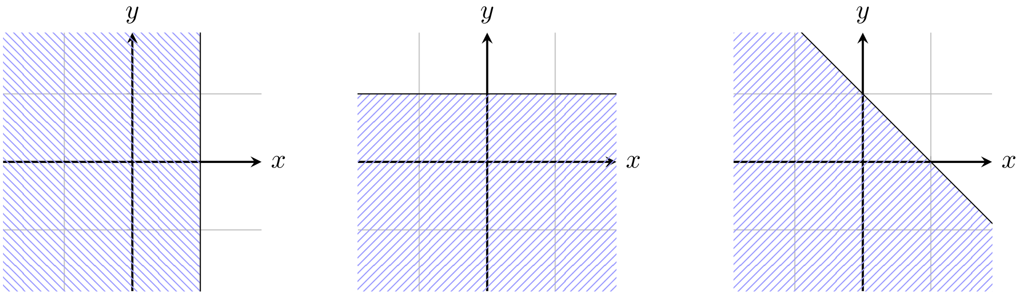
Corollary 8.3.14.1. *If C is a polytope, then C° is an intersection of half-spaces, and is hence a polyhedron if it is bounded.*

Example. Let $C = \operatorname{conv}\big(\{(1,0),(0,1),(1,1)\}\big)$:



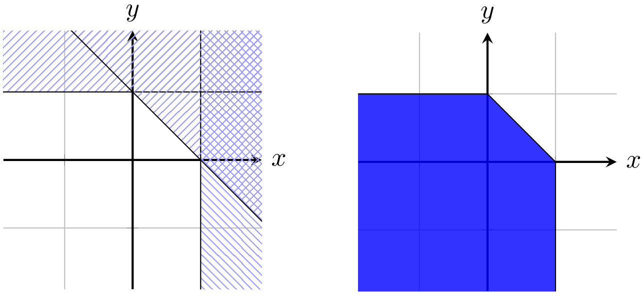
By the previous lemma, C° is the intersection of half-spaces

$$\begin{aligned}\big\{(x,y) : \big\langle(1,0),(x,y)\big\rangle \leq 1\big\} &= \big\{(x,y) : x \leq 1\big\}\\ \big\{(x,y) : \big\langle(0,1),(x,y)\big\rangle \leq 1\big\} &= \big\{(x,y) : y \leq 1\big\}\\ \big\{(x,y) : \big\langle(1,1),(x,y)\big\rangle \leq 1\big\} &= \big\{(x,y) : x + y \leq 1\big\}\end{aligned}$$



$S_1 = \big\{(x,y) : x \leq 1\big\}$ $S_2 = \big\{(x,y) : y \leq 1\big\}$ $S_3 = \big\{(x,y) : x + y \leq 1\big\}$

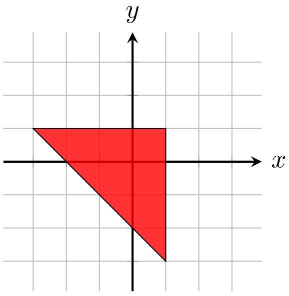
Shading the unwanted region so the intersection is easier to see, we have:



$S_1 \cap S_2 \cap S_3$ C°

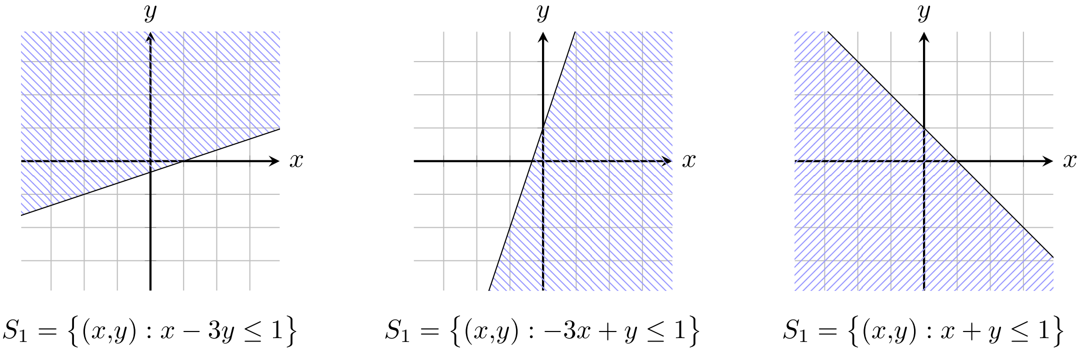
Note that in this case, C does not contain $\mathbf{0}$, and C° is unbounded, and hence not a polygon. △

Example. Let $C = \operatorname{conv}\big(\{(1,-3),(-3,1),(1,1)\}\big)$:



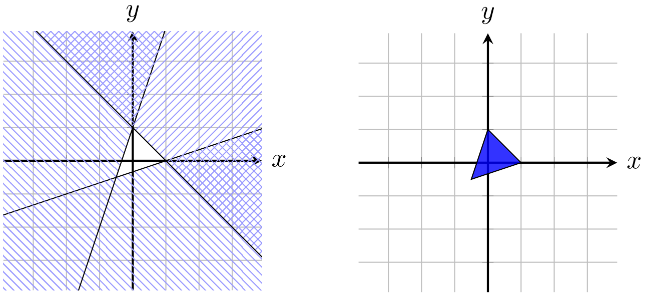
By the previous lemma, C° is the intersection of half-spaces

$$\big\{(x,y) : \big\langle(1,-3),(x,y)\big\rangle \le 1\big\} = \big\{(x,y) : x - 3y \le 1\big\}$$
$$\big\{(x,y) : \big\langle(-3,1),(x,y)\big\rangle \le 1\big\} = \big\{(x,y) : -3x + y \le 1\big\}$$
$$\big\{(x,y) : \big\langle(1,1),(x,y)\big\rangle \le 1\big\} = \big\{(x,y) : x + y \le 1\big\}$$



$S_1 = \big\{(x,y) : x - 3y \le 1\big\}$ $S_1 = \big\{(x,y) : -3x + y \le 1\big\}$ $S_1 = \big\{(x,y) : x + y \le 1\big\}$

Shading the unwanted region so the intersection is easier to see, we have:



$S_1 \cap S_2 \cap S_3$ C°

This time, $\mathbf{0} \in C$, so $C = C^{\circ\circ}$ and C° is a polytope whose vertices correspond to the facets of C. So, another way to find the vertices of the polar is to find the lines in which the facets lie.

In this case, we have the lines $1x + 0y = 1$, $0x + 1y = 1$, and $1x + 1y = -2$, which rearranges to $-\frac{1}{2}x - \frac{1}{2}y = 1$, so the vertices are $(1,0)$, $(0,1)$, and $(-\frac{1}{2}, -\frac{1}{2})$. △

Lemma 8.3.15 (Inversion)**.** *If C and D are convex sets with $D \subseteq C$, then $C^\circ \subseteq D^\circ$.*

Proof. If $y \in C^\circ$, then $\langle x,y\rangle \le 1$ for all $x \in C$. We have $D \subseteq C$, so y also satisfies $\langle x,y\rangle \le 1$ for all $x \in D \subseteq C$, so $y \in D^\circ$. ■

Theorem 8.3.16 (Polytopes are Polyhedra)**.** *Every polytope $C \subset \mathbb{R}^d$ is a polyhedron.*

Proof. By translating if necessary, we may assume that C contains $\mathbf{0}$.

The strategy is to prove that the polar C° is bounded, and is hence a polyhedron. We have previously proved that polyhedra are polytopes, so C° is also a polytope. So, we may repeat this argument with C° replacing C, giving that $C^{\circ\circ}$ is also a polyhedron. Then, $C = C^{\circ\circ}$ is a polyhedron, as required.

We induct on d. If $d = 1$, this is trivial as polyhedra and polytopes are both just intervals.

Let C be the convex hull of finitely many points. If C is contained in a hyperplane, then the result immediately follows from the inductive hypothesis, so suppose otherwise.

Pick a point $u \in C$. Since C is d-dimensional, the set

$$\{v - u : v \in C\}$$

spans $\mathbb{R}^d$. Pick a basis $(v_i - u)_{i=1}^d$ consisting of vectors of this form.

The convex hull of the points $u, v_1, \dots, v_d$ has non-empty interior since it contains a ball of radius $r > 0$ say around the barycentre

$$p := \frac{1}{d+1}(u + v_1 + \dots + v_d)$$

By translating the polytope C, suppose that $p = \mathbf{0}$ and that C is the convex hull of $x_1, \dots, x_m$.

We claim that C° is bounded.

Suppose that $y \in C^\circ$ has norm $k > 0$ and define the point $x := \frac{r}{k}y$. Then, x has norm

$$\|x\| = \left\|\frac{r}{k}y\right\| = \left|\frac{r}{k}\right| \|y\| = \frac{r}{k}k = r$$

so $x \in C$. So, by the definition of a polar, y must satisfy

$$\begin{aligned} \langle x, y\rangle &\le 1 \\ \langle \tfrac{r}{k}y, y\rangle &\le 1 \\ \frac{r}{k}\langle y, y\rangle &\le 1 \\ \frac{r}{k}\|y\|^2 &\le 1 \\ \frac{r}{k}k^2 &\le 1 \\ k &\le \frac{1}{r} \end{aligned}$$

So, $C^\circ \subseteq \mathbb{B}_{1/r}$ and is hence bounded. So, C° is a polyhedron.

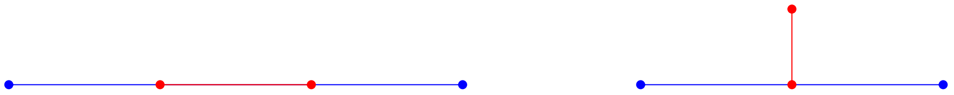
Repeating this argument with C° replacing C, we have that $C^{\circ\circ} = C$ is a polyhedron, as required. ■

Along with Theorem 8.3.12, we have proved that polytopes and polyhedra in $\mathbb{R}^d$ are equivalent.

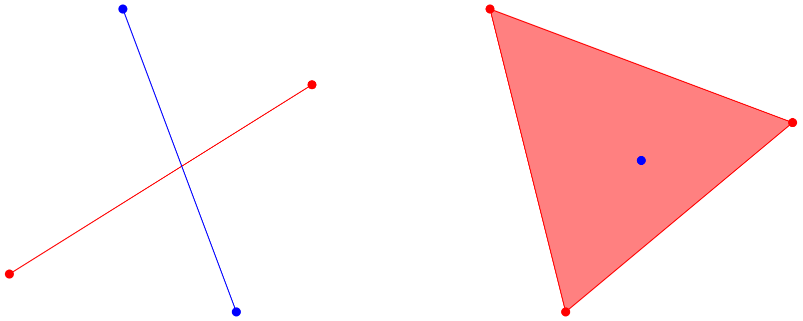
8.3.5 Radon's Lemma and Helly's Theorem

Lemma 8.3.17 (Radon)**.** *Let $X \subseteq \mathbb{R}^d$ have cardinality $d + 2$. Then, there exists a partition of X into two subsets whose convex hulls have non-empty intersection.*

Example. Consider $d = 2$, with 4 points in the plane. If at least 3 are colinear, then we can place the outermost points in one part, and the remaining points in the other part:



Otherwise, the points could form a convex quadrilateral, in which case, the diagonals intersect; alternatively, they could form a triangle, with a point inside:



△

Proof. Let $I = \{1, \ldots, d+2\}$. Let the points be $(x_i)_{i \in I} \subseteq \mathbb{R}^d$. Adjoin an extra coordinate to each x_i, and set the coordinate equal to 1 to obtain $d+2$ points $(y_i)_{i \in I} \subseteq \mathbb{R}^{d+1}$.

Because we have $d+2$ points in $\mathbb{R}^{d+1}$, the y_i are not linearly independent, so there are scalars α_i not all zero such that

$$\sum_{i \in I} \alpha_i y_i = \mathbf{0}$$

Define the sets

$$A \coloneqq \{i : \alpha_i > 0\}, \qquad B \coloneqq \{i : \alpha_i < 0\}$$

and define the scalars $\beta_i \coloneqq -\alpha_i$, positive for $i \in B$. Sorting positive and negative coefficients, we have

$$\begin{aligned} \sum_{\substack{i \in I \\ \alpha_i > 0}} \alpha_i y_i + \sum_{\substack{i \in I \\ \alpha_i < 0}} \alpha_i y_i &= \mathbf{0} \\ \sum_{\substack{i \in I \\ \alpha_i > 0}} \alpha_i y_i &= \sum_{\substack{i \in I \\ \alpha_i < 0}} -\alpha_i y_i \\ \sum_{i \in A} \alpha_i y_i &= \sum_{i \in B} \beta_i y_i \end{aligned}$$

Considering only the first d coordinates, we have

$$\sum_{i \in A} \alpha_i x_i = \sum_{i \in B} \beta_i x_i$$

and considering only the final coordinate, we have

$$\sum_{i \in A} \alpha_i = \sum_{i \in B} \beta_i \eqqcolon C$$

Because not all of the scalars are zero, A and B are non-empty, so these sums are non-empty, and we have $C > 0$.

Then, the vectors

$$x \coloneqq \sum_{i \in A} \frac{\alpha_i}{C} x_i, \qquad y \coloneqq \sum_{i \in B} \frac{\beta_i}{C} x_i$$

are convex combinations of two disjoint subsets of the x_i, and these two vectors are equal. ■

Instead of partitioning X into two parts whose convex hulls have non-empty intersection, one could ask if we can partition X into three or more subsets whose convex hulls have non-empty intersection. For three parts, we clearly need to start with more than $d+2$ points.

The right number turns out to be $2d+3$, and in general, if we are partitioning into k subsets, we need $(k-1)(d+1)+1$ points. This more general theorem holds, but is much harder to prove than Radon's lemma.

Theorem 8.3.18 (Tverberg). *Each set of $(k-1)(d+1)+1$ points in $\mathbb{R}^d$ can be partitioned into k subsets whose convex hulls have non-empty intersection.*

The simplest proof of Tverberg's theorem relies on the following result:

Theorem 8.3.19 (Colourful Caratheodory Theorem). *Let $C_1, \dots, C_{d+1}$ be arbitrary subsets of $\mathbb{R}^d$, each coloured with a different colour. Suppose that $\mathbf{0} \in \operatorname{conv}(C_i)$ for all $1 \leq i \leq d+1$. Then, there is a rainbow set $R \subseteq \bigcup_i C_i$ with precisely one point of each colour whose convex hull contains $\mathbf{0}$.*

The next theorem we will prove is a striking dual of Caratheodory's theorem.

Recall that a space X is compact if and only if for every open cover $\mathcal{U}$ of X, there exists a finite subcover $\mathcal{U}_0 \subseteq R$:

$$\operatorname{compact}(X) \quad \equiv \quad \bigcup \mathcal{U} = X \to \exists \text{ finite } \mathcal{U}_0 \subseteq \mathcal{U} : \bigcup \mathcal{U}_0 = X$$

We can De Morgan-dualise this statement by replacing open sets by closed sets, unions with intersections, and covers with empty-intersections:

$$\begin{aligned} &\equiv \quad \bigcap \mathcal{F} = \varnothing \to \exists \text{ finite } \mathcal{F}_0 \subseteq \mathcal{F} : \bigcap \mathcal{F}_0 = \varnothing \\ &\equiv \quad \bigcap \mathcal{F} \neq \varnothing \leftarrow \forall \text{ finite } \mathcal{F}_0 \subseteq \mathcal{F} : \bigcap \mathcal{F}_0 \neq \varnothing \end{aligned}$$

Taking the contrapositive in the second line, we have that a set is compact if and only if for every (non-empty) family $\mathcal{F}$ of closed subsets of X, every finite subfamily $\mathcal{F}_0 \subseteq \mathcal{F}$ having non-empty intersection implies that $\mathcal{F}$ has non-empty intersection. A set satisfying the hypotheses of this implication is said to have the *finite intersection property.*

The next theorem shows that in the convex case, we do not need to check all finite subfamilies, but only those of cardinality at most $d+1$.

Theorem 8.3.20 (Helly's Theorem). *Let $\mathcal{F} = (C_i)_{i=1}^m$ be a family of convex sets in $\mathbb{R}^d$, and suppose that every subfamily $\mathcal{F}_0 \subseteq \mathcal{F}$ of cardinality at most $d+1$ has non-empty intersection. Then, the whole family has a non-empty intersection:*

$$\bigcap \mathcal{F} \neq \varnothing$$

Proof. We induct on the number of sets m.

First, note that if $m \leq d+1$, there is nothing to prove, since the desired result is included in the hypotheses of the theorem.

For the base case, suppose $m = d+2$.

Then, for each C_i, the other $d+1$ sets $C_{k \neq i}$ have non-empty intersection by assumption, so we may select a point x_i in each of these intersections:

$$x_i \in \bigcap_{k \neq i} C_k$$

By Radon's lemma, $X := \{x_i\}_{i=1}^{d+2}$ has a partition into two subsets $X_1, X_2 \subseteq X$ whose convex hulls have non-empty intersection. Let u be a point in this intersection.

$$u \in \operatorname{conv}(X_1) \cap \operatorname{conv}(X_2)$$

We claim that u is contained in each C_i, and is hence in $\bigcap \mathcal{F}$.

Fix some $1 \leq j \leq m$, and without loss of generality, suppose that x_j is in the Radon subset X_1, so $x_j \notin X_2$. By construction of the x_i, we have $x_i \in C_j$ for all $i \neq j$, so $X_2 \subseteq X \setminus \{x_j\} \subseteq C_j$. Since C_j is convex, it also contains $\operatorname{conv}(X_2) \ni u$, so $u \in C_j$.

In the above, we have assumed that the x_i are all distinct. But if this were not the case, say $x_i = x_j$ for some $i \neq j$, then by construction, $x_i \in C_k$ for all $k \neq i$, but also $x_i = x_j \in C_i$ (since $i \neq j$), so $x_i \in C_k$ for all k, and the intersection $\bigcap \mathcal{F}$ is again non-empty.

For the inductive step, suppose $m > d + 2$ and that the result holds for $m - 1$. Consider a new family of $m - 1$ sets given by

$$\mathcal{F}' := \{C_1 \cap C_2, C_3, C_4, \ldots, C_m\}$$

and let $\mathcal{F}_0 \subseteq \mathcal{F}'$ be a subfamily of cardinality $d + 1$. Then, $\bigcap \mathcal{F}_0$ is the intersection of at most $d + 2$ of the original C_i, which is non-empty by the base case.

So, $\mathcal{F}$ satisfies the hypotheses of the result, so by the induction hypothesis, $\bigcap \mathcal{F}' \neq \varnothing$. Then,

$$\varnothing \neq \bigcap \mathcal{F}' = (C_1 \cap C_2) \cap C_3 \cap \cdots \cap C_m = \bigcap \mathcal{F}$$

which completes the inductive step. ■

If we also require that the C_i are compact, then Helly's theorem also holds for arbitrary collections $\mathcal{F}$, and not just finite collections:

Theorem 8.3.21 (Helly's Compactness Theorem). *Let $\mathcal{F} = (C_i)_{i \in I}$ be a family of compact convex sets in $\mathbb{R}^d$, and suppose that every subfamily $\mathcal{F}_0 \subseteq \mathcal{F}$ of cardinality at most $d+1$ has non-empty intersection. Then, the whole family has a non-empty intersection:*

$$\bigcap \mathcal{F} \neq \varnothing$$

Proof. For any finite $J \subseteq I$, consider the set

$$K_J = \bigcap_{j \in J} C_j$$

By assumption, the subfamily $\mathcal{F}_J = \{C_j\}_{j \in J}$ satisfies the hypotheses of the finite Helly's theorem, so each $\bigcap \mathcal{F}_J = K_J$ is non-empty.

Now, the family $\mathcal{K} = \{K_J\}_{J \subseteq I, J \text{ finite}}$ has the finite intersection property since the family is closed under intersection and each K_J is non-empty:

$$K_{J_1} \cap K_{J_2} = K_{J_1 \cup J_2} \neq \varnothing$$

So $\mathcal{K}$ has the finite intersection property. At this point, we would like to apply compactness to conclude that the intersection of $\mathcal{K}$ is non-empty, but $\mathbb{R}^d$ is not compact.

Instead, fix any finite indexing set $J_0 \subseteq I$, and consider the family

$$\mathcal{K}_0 = \{K_J : J \subseteq I, J \text{ finite}, J \supseteq J_0\} \subseteq \mathcal{K}$$

This family $\mathcal{K}_0$ also has the finite intersection property as a subset of $\mathcal{K}$. Furthermore, the set K_{J_0} is compact as the finite intersection of compact sets, and each $K_J \in \mathcal{K}_0$ is a closed subset of K_{J_0}, so the intersection of $\mathcal{K}_0$ is non-empty:

$$\bigcap \mathcal{K}_0 \neq \varnothing$$

We claim that

$$\bigcap \mathcal{K}_0 \subseteq \bigcap \mathcal{F}$$

(In fact, the two sets are equal, but we only need the forward containment here.)

Let $x \in \bigcap \mathcal{K}_0$ and for each $i \in I$, consider the set $J_i \coloneqq J_0 \cup \{i\}$. Note that J_i is a finite subset of I containing J_0, so $x \in K_{J_i} = \bigcap_{j \in J_i} C_j = C_i \bigcap_{j \in J_0} C_j$. So $x \in C_i$. Since i was arbitrary, $x \in C_i$ for all $i \in I$, and hence $x \in \bigcap_{i \in I} C_i = \bigcap \mathcal{F}$, which concludes the proof.

For completeness, the other containment is straightforward: if $x \in \bigcap \mathcal{F} = \bigcap_{i \in I} C_i$, then in particular $x \in C_j$ for all $j \in J$ for every finite $J \subseteq I$. So, $x \in \bigcap_{j \in J} C_j = K_J$ for each finite J, including those satisfying $J \supseteq J_0$. So,

$$x \in \bigcap \mathcal{K}_0$$

■

ldcldc

8.4 Partially Ordered Sets and Set Systems

A relation $\leq$ on a set X is a (*weak* or *non-strict*) *partial order* on X if it satisfies, for all $x,y,z \in X$:

(*i*) reflexivity: $x \leq x$;

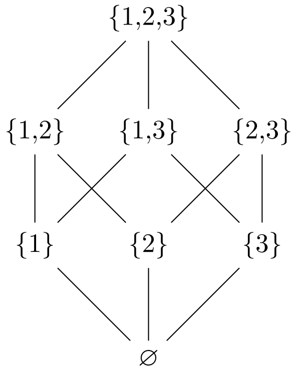
(*ii*) transitivity: $(x \leq y \wedge y \leq z) \to x \leq z$;

(*iii*) antisymmetry: $(x \leq y \wedge y \leq x) \to x = y$.

The pair $(X, \leq)$ is then called a *partially ordered set* or *poset.*

Note that not all elements in a poset may be *comparable* under the ordering:

Example. Consider $\big(\mathcal{P}\big([3]\big), \subseteq\big)$, illustrated as the *Hasse diagram* below, where an edge from a vertex x travelling upwards to a vertex y indicates that $x \leq y$.



In this example, $\{1\}$ and $\{2,3\}$ are *incomparable* in this poset, because neither $\{1\} \subseteq \{2,3\}$ nor $\{2,3\} \subseteq \{1\}$ hold. △

Vertices on the same horizontal level in a Hasse diagram are always incomparable.

If every pair of elements *are* comparable, then the ordering is *total.*

Example. $(\mathbb{R}, \leq)$ is a total ordering. △

Let $\leq$ be a partial order on a set X.

A *chain* is a subset $C \subseteq S$ such that $\leq$ is total on C. That is, every pair of elements in C are comparable under $\leq$:

$$\forall c_1,c_2 \in C : c_1 \leq c_2 \vee c_2 \leq c_1$$

Example. A chain in $\big(\mathcal{P}([3]), \subseteq\big)$ is given by the sequence of elements

$$\emptyset \subseteq \{1\} \subseteq \{1{,}2\} \subseteq \{1{,}2{,}3\}$$

or

$$\{3\} \subseteq \{1{,}3\}$$

△

An *antichain* is a subset $A \subseteq S$ such that every pair of elements in C are incomparable under $\leq$:

$$\forall a_1,a_2 \in A : a_1 \not\leq c_2 \wedge c_2 \not\leq c_1$$

Example. An antichain in $\big(\mathcal{P}([3]), \subseteq\big)$ is given by the set

$$\big\{\{2\},\{1{,}3\}\big\}$$

or

$$\big\{\{1\},\{2\},\{3\}\big\}$$

△

More generally, in $\big(\mathcal{P}([n]), \subseteq\big)$, any collection of subsets of a fixed cardinality form an antichain, since if two different sets have the same number of elements, then neither can be a subset of the other.

There are $\binom{n}{k}$ many subsets of $[n]$ of cardinality k, and this number is maximised when $k \approx n/2$. If n is even, then this has size precisely

$$\binom{n}{n/2}$$

and if n is odd, the two largest antichains are at level $k = (n-1)/2$ and $k = (n+1)/2$, and in either case, this has size

$$\binom{n}{\lfloor n/2 \rfloor}$$

Theorem 8.4.1 (Sperner). *The largest antichain in $\big(\mathcal{P}([n]), \subseteq\big)$ has cardinality*

$$\binom{n}{\lfloor n/2 \rfloor}$$

We will deduce Sperner's theorem from a stronger statement due to Lubell, Yamamoto, and Meshalkin:

Theorem 8.4.2 (LYM Inequality). *Let $\mathcal{F}$ be an antichain in $\big(\mathcal{P}([n]), \subseteq\big)$. Then,*

$$\sum_{A \in \mathcal{F}} \frac{1}{\binom{n}{|A|}} \leq 1$$

Proof of Sperner's Theorem. It is clear that such an antichain exists – just pick all subsets with $\lfloor n/2 \rfloor$ elements.

To show that such an antichain is maximal, let F be an antichain in $\big(\mathcal{P}([n]), \subseteq\big)$. Because $\binom{n}{\lfloor n/2 \rfloor} \geq \binom{n}{|A|}$ for any A, we have from the LWM inequality,

$$\sum_{A \in \mathcal{F}} \frac{1}{\binom{n}{|A|}} \leq 1$$
$$\sum_{A \in \mathcal{F}} \frac{1}{\binom{n}{\lfloor n/2 \rfloor}} \leq 1$$

$$|\mathcal{F}|\frac{1}{\binom{n}{\lfloor n/2 \rfloor}} \leq 1$$
$$|\mathcal{F}| \leq \binom{n}{\lfloor n/2 \rfloor}$$

■

Proof of the LYM Inequality. There is a bijection from the set of permutations on $[n]$ to the set of maximal chains in $\big(\mathcal{P}([n]),\, \subseteq\big)$, where each permutation gives the order in which to add elements to subsets in the chain.

For instance, for $n = 5$, the permutation [5,2,3,1,4] corresponds to the chain

$$\emptyset \subseteq \{5\} \subseteq \{5{,}2\} \subseteq \{5{,}2{,}3\} \subseteq \{5{,}2{,}3{,}1\} \subseteq \{5{,}2{,}3{,}1{,}4\}$$

We pick a maximal chain/permutation R uniformly at random. The probability that R is any given maximal chain C is

$$\mathbb{P}(R = C) = \frac{1}{n!}$$

since there are $n!$ permutations on $[n]$.

For each subset $A \subseteq [n]$, let E_A be the set of maximal chains that contain A. If A and B are in both in $\mathcal{F}$, then E_A and E_B must be disjoint, since A and B must be incomparable and cannot both belong to the same chain. So,

$$\mathbb{P}(R \cap \mathcal{F} \neq \varnothing) = \mathbb{P}\left(R \in \bigsqcup_{A \in \mathcal{F}} E_A\right)$$
$$= \sum_{A \in \mathcal{F}} \mathbb{P}(R \in E_A)$$

and since this is a probability, it is bounded above by 1:

$$\sum_{A \in \mathcal{F}} \mathbb{P}(R \in E_A) \leq 1$$

To finish the proof, it remains to show that

$$\mathbb{P}(R \in E_A) = \frac{1}{\binom{n}{|A|}}$$

First, we have

$$\mathbb{P}(R \in E_A) = \frac{\#\text{ of maximal chains containing } A}{\#\text{ of all maximal chains}}$$

A maximal chain containing A corresponds to a permutation which has the elements of A in any order as a prefix. For instance, any permutation starting with the numbers 1, 2, and 3 in any order will generate the subset {1,2,3} in the corresponding chain.

So, there are $|A|!$ many ways to arrange this prefix, then $(n - |A|)!$ ways to arrange the remaining numbers. So,

$$\mathbb{P}(R \in E_A) = \frac{\#\text{ of maximal chains containing } A}{\#\text{ of all maximal chains}} = \frac{|A|!(n - |A|)!}{n!} = \frac{1}{\binom{n}{|A|}}$$

as required. ■

8.4.1 Dilworth's Theorem

A chain and an antichain can have at most one common element, for if x and y were two common elements, then the chain would require x and y to be comparable, and the antichain would require x and y to be incomparable.

So, if a poset can be covered with m chains, then there cannot be any antichains with more than m elements by the pigeonhole principle. Hence, another method of proving Sperner's theorem would be to find a cover of $\big(\mathcal{P}([n]), \subseteq\big)$ using

$$\binom{n}{\lfloor n/2 \rfloor}$$

chains. In fact, the proof of the LYM inequality above uses a similar, but simpler, idea, since we only looked at the covering generated by all maximal chains, and then counted how many times each set was covered.

This process also works in reverse to deduce that there is a covering by a small number of chains if there are only small antichains:

Theorem 8.4.3 (Dilworth)**.** *Let $(\Omega, \leq)$ be a poset in which every antichain has at most m elements. Then, Ω can be covered by m chains (or fewer).*

Proof. We prove the case for finite Ω only.

We induct on $|\Omega|$. If $|\Omega| = 1$, then there is nothing to prove.

Suppose $|\Omega| > 1$, and that the result holds for all smaller posets. Let m be the size of the largest antichain in Ω, and choose a maximal chain $C = \{c_1 \leq c_2 \leq \cdots \leq c_n\}$ in Ω.

Suppose $\Omega \setminus C$ has no antichains of length m, so every antichain has at most $m - 1$ elements. Then, the smaller poset $\Omega \setminus C$ may be covered by $m - 1$ chains (or fewer) by the inductive hypothesis, so Ω may be covered by m chains by adding C to this cover, and we are done.

Otherwise, $\Omega \setminus C$ has (maximal) antichains of length m. Let $A = \{a_1, \ldots, a_m\} \subseteq \Omega \setminus C$ be such an antichain, and define the sets

$$A^- = \{x \in \Omega : \exists i, x \leq a_i\}, \qquad A^+ = \{x \in \Omega : \exists i, x \geq a_i\}$$

Note that these sets jointly cover Ω, since if there were an $x \in \Omega$ but not $A^- \cup A^+$, then this x would be incomparable to all the a_i, so we could extend A, contradicting the maximality of A.

However, we also have that neither of these sets can be all of Ω, since if $C \subseteq A^-$, then we would have $c_n \leq a_i$ for some i, so we could extend C by adding a_i, contradicting the maximality of C; and similarly, if $C \subseteq A^+$, we would have $c_1 \geq a_i$ for some i, again contradicting the maximality of C.

So, because A^- and A^+ are strict subsets of Ω, they are smaller posets, so the inductive hypothesis applies. So, A^- and A^+ can each be covered with m chains C_i^- and C_i^+, respectively:

$$A^- = \bigcup_{i=1}^{m} C_i^- \qquad A^+ = \bigcup_{i=1}^{m} C_i^+$$

Each of these decompositions partition the a_i, so, reindexing if necessary, we may assume that $a_i \in C_i^-$ and $a_i \in C_i^+$ for each i.

We claim that a_i is the maximal element of C_i^- and the minimal element of C_-^+.

Suppose otherwise, so there is an element x such that $a_i < x \in C_i^-$. Since $x \in C_i^- \subseteq A^-$, we have $x \leq a_j \neq a_i$. But then, $a_i < a_j$ are comparable, contradicting that A is an antichain. The proof that a_i is minimal in C_i^+ is entirely symmetric.

Then, the m chains $C_i^- \cup C_i^+$ cover Ω. ■

8.4.2 Covering by Chains

Along with our previous observation that there cannot be an antichain longer than the number of chains in a covering, Dilworth's theorem gives that:

Corollary 8.4.3.1. *For any poset $(\Omega, \leq)$, the length of the largest antichain is equal to the minimum number of chains required to cover Ω.*

We can also apply Dilworth's theorem to $\big(\mathcal{P}\big([n]\big), \subseteq\big)$ with $m = \binom{n}{\lfloor n/2 \rfloor}$ given by Sperner's theorem to obtain:

Corollary 8.4.3.2. *The poset $\big(\mathcal{P}\big([n]\big), \subseteq\big)$ can be covered using $\binom{n}{\lfloor n/2 \rfloor}$ chains.*

We can also show this directly, and conversely use this result with Dilworth's theorem to give another proof of Sperner's theorem. This proof will depend on Hall's theorem:

Theorem 8.4.4 (Hall)**.** *Let $G = (L \cup R, E)$ be a bipartite graph. For each subset $U \subseteq L$, let $N_G(U)$ denote the open neighbourhood of W in G:*

$$N_G(U) \coloneqq \{v \in R : \exists u \in U, (u,v) \in E\}$$

That is, the set of vertices in R that are adjacent to at least one element in U.

Then, there is an matching that covers L if and only if for every $U \subseteq L$,

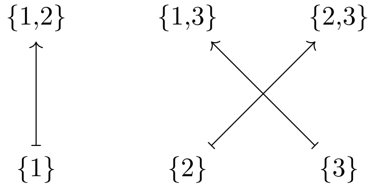
$$|U| \leq \big|N_G(U)\big|$$

That is, every subset $U \subseteq L$ must have sufficiently many neighbours in R for such a matching to exist.

Proof of Corollary 8.4.3.2. Let $r < n/2$, and consider the set R of subsets of cardinality r, and the set R^+ of subsets of cardinality $r + 1$:

$$R_r \coloneqq \big\{S \subseteq [n] : |S| = r\big\}, \qquad R_r^+ \coloneqq \big\{S \subseteq [n] : |S| = r + 1\big\}$$

We claim that there is an injection $f : R_r \rightarrowtail R_r^+$ such that $A \subseteq f(A)$ for all $A \in R_r$. For instance,



Consider the bipartite graph $G_r = (R_r \cup R_r^+, E)$, where $(A,B) \in E$ if and only if $A \subseteq B$ (i.e. the subgraph of the Hasse diagram induced by taking the r and $(r+1)$th rows). Note that an R_r-saturated matching in this bipartite graph precisely corresponds to the required injection.

Each set $S \in R_r$ has $n - r$ neighbours, since r of the n numbers we could add are already in the set. Conversely, each set $S^+ \in R_r^+$ has $r + 1$ neighbours, since we may remove any of its $r + 1$ elements.

Now, let $U \subseteq R_r$. Then, there are

$$\sum_{S \in U} \deg(S) = |U|(n - r)$$

edges incident to U, and each vertex in R_r^+ is incident to at most $r + 1$ of these edges, so

$$\big|N_G(U)\big| \geq \frac{|U|(n - r)}{r + 1}$$

Since $r < n/2$, we have

$$\begin{aligned} r+1 &\leq \frac{n}{2} \\ r+1 &\leq n - \frac{n}{2} \\ r+1 &\leq n-r \end{aligned}$$

so

$$\left|N_G(U)\right| \geq |U|$$

so Hall's condition is satisfied, and there exists an R_r-saturated matching M_r.

Repeating this construction on every layer, we obtain matchings from each layer to the next, up to layer $k \coloneqq \lfloor n/2 \rfloor$.

Note that $\bigcup_{r=0}^{k} M_r$ is a subgraph of $\bigcup_{r=0}^{k} G_r$ consisting of disjoint paths: there can be no vertices of degree 3 or higher since every layer consists of perfect matchings. By the construction of the edge set of the G_r, each such path defines a chain.

We can cover the rest of $\mathcal{P}\big([n]\big)$ by mirroring this construction as follows. Given $A \subseteq [n]$ with $|A| < k$, let $g(A) = [n] \setminus A$. Then, each above chain defines a chain $g(C)$.

If n is odd, then the middle two layers R_k and R_{k+1} have the same cardinality, and Hall's theorem gives a perfect matching, so the chains join up as desired. If n is even, then we may have some sets at $k = n/2$ uncovered, in which case, we may cover them with additional singleton chains.

Each chain contains a set in the middle layer k, so there are $\binom{n}{k} = \binom{n}{\lfloor n/2 \rfloor}$ total chains in this cover. ■

With the earlier observation that that a covering with m chains implies that every antichain has at most m elements, this provides another proof of Sperner's theorem.

There is a nice variation on this construction in which each chain is "symmetric": each chain consists of sets of sizes $0,1,\dots,n-1,n$ or of sizes $1,2,\dots,n-2,n-1$, etc. This improvement cannot be deduced from just Hall's theorem.

Theorem 8.4.5 (de Bruijn, Tengbergen, Kruyswijk)**.** *The poset $\big(\mathcal{P}\big([n]\big), \subseteq\big)$ can be covered using $\binom{n}{\lfloor n/2 \rfloor}$ symmetric chains.*

Proof. We induct on n. For $n = 1$, there is the unique chain $\big(\varnothing,\{1\}\big)$, which is symmetric.

Suppose $n > 1$, and that the result holds for $n-1$, so there is a decomposition of $[n-1]$ into symmetric chains. For each chain $C_i = (A_1,A_2,\dots,A_k)$ in this decomposition, we form two new chains

$$C_i^+ \coloneqq \big(A_1,A_2,\dots,A_k,A_k \cup \{n\}\big)$$

and

$$C_i^- \coloneqq \big(A_1 \cup \{n\},A_2 \cup \{n\},\dots,A_{k-1} \cup \{n\}\big)$$

The collection of these new chains covers $\mathcal{P}\big([n]\big)$ since for each old subset $A \in \mathcal{P}\big([n-1]\big)$, the new chains C_i^+ and C_i^- cover A and $A \cup \{n\}$, respectively.

If C_i consists of sets of size j to $n-j$, then C_i^+ has sizes j to $n-j+1 = (n+1)-j$, and C^- has sizes $j+1$ to $n-j = (n+1)-(j+1)$, so the new chains are also symmetric.

This construction also ensures that in each chain, each set has precisely one more element than the set before it, since the first transformation adds a new set one element larger to the top of an existing chain, and the second transformation adds one element to every set in an existing chain. In either case, this relation is preserved. So, each chain contains an element of cardinality $\lfloor n/2 \rfloor$, so there are $\binom{n}{\lfloor n/2 \rfloor}$ chains in the cover. ■

8.4.3 VC Dimension and the Sauer-Shelah Lemma

Given a finite set $U = \{u_1, u_2, \ldots, u_m\}$, a family of sets $\mathcal{F}$ *shatters* U if for every subset $V \subseteq U$, there is an element $A \in \mathcal{F}$ such that $A \cap U = V$.

Example. If U consists of the three vertices of a triangle in $\mathbb{R}^2$, and $\mathcal{F}$ is the family of half-spaces in $\mathbb{R}^2$, then each of the 8 subsets of U can be obtained by intersecting U with an appropriate half-space, so $\mathcal{F}$ shatters U.

However, if U consists of four points in the plane, then $\mathcal{F}$ cannot shatter U by Radon's lemma. △

Given a set Ω and a family of sets $\mathcal{F} \subseteq \mathcal{P}(\Omega)$, we define the *Vapnik–Cervonenkis* (*VC*) *dimension* $\mathrm{VC}(\mathcal{F})$ of $\mathcal{F}$ to be maximum cardinality of a subset of Ω that $\mathcal{F}$ can shatter.

Example.

- $\mathrm{VC}\big(\mathcal{P}([n])\big) = n$;
- $\mathrm{VC}\big(\{\text{half-spaces in } \mathbb{R}^n\}\big) = n + 1$;
- $\mathrm{VC}\big(\text{any FPP}\big) = 2$.

△

Fix integers n,k with $n \geq k$, and let $\Omega = [n]$. How large can $|\mathcal{F}|$ be before $\mathrm{VC}(\mathcal{F}) = k$? If $\mathcal{F}$ consists of all sets of size at most $k - 1$, then it cannot shatter a set of size k. In this case,

$$|\mathcal{F}| = \sum_{i=0}^{k-1} \binom{n}{i}$$

It turns out that this is the largest cardinality possible.

Theorem 8.4.6 (Sauer-Shelah Lemma)**.** *Suppose $\mathcal{F} \subseteq \mathcal{P}\big([n]\big)$ has cardinality*

$$|F| > \sum_{i=0}^{k-1} \binom{n}{i}$$

for some $k \leq n$. Then, $\mathcal{F}$ shatters a subset of $[n]$ of size k.

Proof. We induct on n. If $n = k = 1$, then $\mathrm{VC}(\mathcal{F}) = 1$ if and only if $|F| = 2$. So suppose $n > 1$ and that the result holds for any smaller sets.

Given a family $\mathcal{F}$, we create two families $\mathcal{F}_1$ and $\mathcal{F}_2$ of sets in $[n-1]$ as follows:

$$\mathcal{F}_1 \coloneqq \big\{A \subseteq [n-1] : A \in \mathcal{F} \text{ or } A \cup \{n\} \in \mathcal{F}\big\}$$
$$\mathcal{F}_2 \coloneqq \big\{A \subseteq [n-1] : A \in \mathcal{F} \text{ and } A \cup \{n\} \in \mathcal{F}\big\}$$

Note that the the condition in $\mathcal{F}_1$ is not exclusive, so $\mathcal{F}_2 \subseteq \mathcal{F}_1$.

We claim that

$$|\mathcal{F}| = |\mathcal{F}_1| + |\mathcal{F}_2|$$

Clearly,

$$|\mathcal{F}_1| = \sum_{A \subseteq [n-1]} \mathbf{1}_{A \in \mathcal{F}_1}, \qquad |\mathcal{F}_2| = \sum_{A \subseteq [n-1]} \mathbf{1}_{A \in \mathcal{F}_2}$$

and

$$|\mathcal{F}| = \sum_{A \subseteq [n-1]} \mathbf{1}_{A \in \mathcal{F}} + \sum_{A \subseteq [n-1]} \mathbf{1}_{A \in \mathcal{F}}$$

so it suffices to check that for every $A \subseteq [n-1]$,

$$\sum_{A \subseteq [n-1]} \mathbf{1}_{A \in \mathcal{F}} + \sum_{A \subseteq [n-1]} \mathbf{1}_{A \in \mathcal{F}} = \sum_{A \subseteq [n-1]} \mathbf{1}_{A \in \mathcal{F}_1} + \sum_{A \subseteq [n-1]} \mathbf{1}_{A \in \mathcal{F}_2}$$

If only one of A and $A \cup \{n\}$ are in $\mathcal{F}$, then A is in $\mathcal{F}_1$, but not $\mathcal{F}_2$. If both A and $A \cup \{n\}$ are in $\mathcal{F}$, then A is in both $\mathcal{F}_1$ and $\mathcal{F}_2$. In either case, A is counted the same number of times on each side, so the equality holds.

Now, we have

$$\begin{aligned} \sum_{i=0}^{k-1} \binom{n}{i} &= \sum_{i=0}^{k-1} \binom{n-1}{i} + \sum_{i=i}^{k-1} \binom{n-1}{i-1} \\ &= \sum_{i=0}^{k-1} \binom{n-1}{i} + \sum_{i=0}^{k-2} \binom{n-1}{i} \end{aligned}$$

by Pascal's formula, $\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$.

So, if $|\mathcal{F}| > \sum_{i=0}^{k-1} \binom{n}{i}$, then either $|\mathcal{F}_1| > \sum_{i=0}^{k-1} \binom{n-1}{i}$ or $|\mathcal{F})2| > \sum_{i=0}^{k-2} \binom{n-1}{i}$.

In the first case, the inductive hypothesis gives that $\mathcal{F}_1$ shatters a subset of $[n-1]$ of size k, in which case, $\mathcal{F}$ shatters the same subset.

In the second case, the family $\mathcal{F}_2$ shatters a subset S of $[n-1]$ of size $k-1$. For each set $B \in \mathcal{F}_2$, both B and $B \cup \{n\}$ are in $\mathcal{F}$, so $\mathcal{F}$ shatters $S \cup \{n\}$, which has size k. ■

We have a strengthening of the Sauer-Shelah theorem that implies that $\mathcal{F}$ in fact shatters at least $|\mathcal{F}|$ sets.

Theorem 8.4.7 (Pajor)**.** *Suppose $\mathcal{F} \subseteq \mathcal{P}\big([n]\big)$ has cardinality*

$$|F| > \sum_{i=0}^{k-1} \binom{n}{i}$$

for some $k \leq n$. Then, $\mathcal{F}$ shatters at least $|\mathcal{F}|$ sets.

This theorem immediately implies the Sauer-Shelah lemma, since only $\sum_{i=0}^{k-1} \binom{n}{i} < |\mathcal{F}|$ subsets have cardinality less than k.

Proof. We induct on n. If $n = 0$, the sum is empty. But, every family of only one set already shatters the empty set. So, suppose $n > 1$ and that the result holds for any smaller sets.

Given a family $\mathcal{F}$ satisfying the hypotheses of the result for n, we split $\mathcal{F}$ into disjoint subfamilies, $\mathcal{F}_1$ and $\mathcal{F}_2$, where $\mathcal{F}_1$ contains all the subsets containing n, and $\mathcal{F}$ is its complement, containing all the subsets that do not contain n.

By the inductive hypothesis, $\mathcal{F}_1$ and $\mathcal{F}_2$ each shatter two collections of sets whose sizes add to at least $|\mathcal{F}|$.

None of the sets S shattered by either family can contain n, since such sets cannot be shattered by $\mathcal{F}_1$, since any subset of S not containing n cannot be created by intersection; nor by $\mathcal{F}_2$, since any subset of S containing n cannot be created by intersection.

However, some of the shattered sets S may be shattered by both $\mathcal{F}_1$ and $\mathcal{F}_2$. If S is shattered by only one of $\mathcal{F}_1$ and $\mathcal{F}_2$, then it contributes one to the number of sets shattered by the subfamily and also to the number of sets shattered by $\mathcal{F}$. Otherwise, if S is shattered by both $\mathcal{F}_1$ and $\mathcal{F}_2$, then both S and $S \cup \{x\}$ are shattered by $\mathcal{F}$, so S contributes two to the number of shattered sets of the subfamilies and of $\mathcal{F}$.

Thus, $\mathcal{F}$ shatters at least as many sets as the number of set shattered by $\mathcal{F}_1$ and $\mathcal{F}_2$, which is at least $|\mathcal{F}|$. ■

8.5 Graph Colouring

A *proper* (*vertex*) *colouring* of a graph $G = (V,E)$ is a labelling of the vertex set $c : V \to [n]$ such that $c(u) \neq c(v)$ whenever $(u,v) \in E$, and the elements of $[n]$ are traditionally called *colours*.

The *chromatic number* $\chi(G)$ of a graph G is the minimum n for which such a labelling of G exists.

Example.

- For any n, $\chi(K_n) = n$, since all n vertices are adjacent to every other vertex.
- For any n, $\chi(C_{2n}) = 2$.
- For any n, $\chi(C_{2n+1}) = 3$.
- A graph G is bipartite if and only if $\chi(G) = 2$.

△

We write

$$\Delta(G) \coloneqq \max_{v \in V(G)} \deg(v)$$

for the maximum degree of G.

Lemma 8.5.1. *For any graph G,*

$$\chi(G) \leq \Delta(G) + 1$$

Proof. Pick any vertex of G, and greedily assign it any colour not present amongst its previously picked neighbours, then repeat. It will always be possible to assign a vertex a valid colour, since each vertex has at most $\Delta(G)$ neighbours that can already be coloured, and there are $\Delta(G) + 1$ colours available. ■

As we will see, this bound for the number of colours needed is rarely sharp.

Lemma 8.5.2. *Let G be a connected graph that has a vertex x of degree* $\deg(x) < \Delta(G)$. *Then,* $\chi(G) \leq \Delta(G)$.

Proof. For each vertex in G, determine the length of the shortest path to x. Because G is connected, this distance is well-defined.

Let k be the maximum distance, and for each $0 \leq i \leq k$, define the set

$$V_i \coloneqq \{v \in V : d(v,x) = i\}$$

Each vertex in V_i is adjacent to a vertex in V_{i-1} via an edge along a shortest path to x.

Now, consider the induced subgraph $G[V_k]$. Because every vertex in V_k has at least one edge to a vertex in V_{k-1}, each vertex has degree at most $\Delta(G) - 1$, so by the previous lemma, $G[V_k]$ is $\Delta(G)$-colourable.

Now, consider the induced subgraph $G[V_{k-1}]$. Again, every vertex in V_{k-1} has at least one edge to a vertex in V_{k-2}, so each vertex has degree at most $\Delta(G) - 1$, so we can greedily colour $G[V_{k-1}]$ with $\Delta(G)$ colours, taking the colours used for the previous layer into account.

The same argument continues for each V_i with $0 < i \leq k$ until only $V_0 = \{x\}$ remains. By assumption, $\deg(x) < \Delta(G)$, so we have a colour left for x. ■

A graph is *k-connected* if it requires the deletion of at least k vertices to disconnect it.

Example. Any connected graph is at least 1-connected. △

Example. The path graph P_3 is 1-connected but not 2-connected, as deleting the middle vertex disconnects the graph. △

Example. The cycle graph C_4 is 2-connected. △

Theorem 8.5.3 (Brooks)**.** *If G is a connected graph which is neither complete nor an odd cycle, then $\chi(G) \leq \Delta(G)$. Otherwise, $\chi(G) = \Delta(G) + 1$.*

Proof. If $G = K_n$, then $\chi(G) = n = \Delta(G) + 1$, and if $G = C_{2n+1}$, then $\chi(G) = 3 = \Delta(G) + 1$.

Also, if $\Delta(G) = 1$, then $G = K_2$, and if $\Delta(G) = 2$, then G is either a cycle or a path, and paths have chromatic number 2 via greedy colouring.

Otherwise, assume that G is neither complete nor an odd cycle, and that $\Delta(G) \geq 3$. We split into three cases:

(*i*) G is 1-connected but not 2-connected;

(*ii*) G is 2-connected but not 3-connected;

(*iii*) G is 3-connected.

(*i*) Let v be a vertex whose removal disconnects G into the connected components $G \setminus \{v\} = \bigcup_i G_i$.

Consider the induced subgraphs $G\big[G_i \cup \{v\}\big]$. In this induced subgraph, v has degree less than $\Delta(G)$, since it has edges to other connected components, so this induced subgraph can be coloured with $\Delta(G)$ many colours via the previous lemma.

By permuting the colourings in each induced subgraph, we can ensure that v has the same colour in each case, so the union of the colourings gives a proper colouring for G.

(*ii*) Let u,v be a pair of vertices whose removal disconnects G into the connected components $G\setminus\{u,v\} = \bigcup_i G_i$.

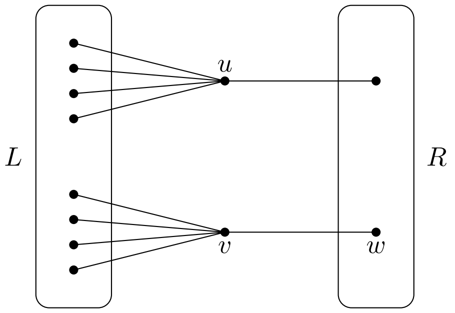
Note that both u and v have at least one edge incident to each component, since if, say, only u has such an edge, then deleting u alone disconnects G, contradicting that G is 2-connected.

Through identical arguments as in case (*i*), we may colour the induced subgraphs $G\big[G_i \cup \{u,v\}\big]$.

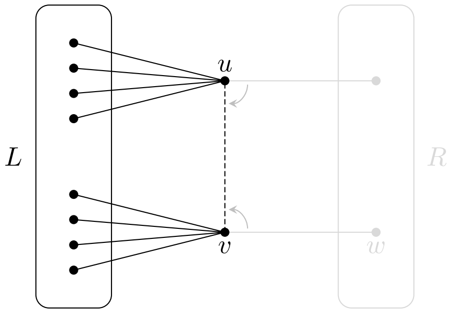
If u and v are adjacent, then in each of the colourings, they are assigned different colours, so by permuting the colourings in each induced subgraph, we can again take the union of these colourings to obtain a proper colouring of G.

Otherwise, u and v are not adjacent. Continue as before, but colour each induced subgraph as though there were an edge connecting u and v. Note that this cannot increase the maximum degree beyond $\Delta(G)$, since u and v previously had at least one other edge to a different connected component.

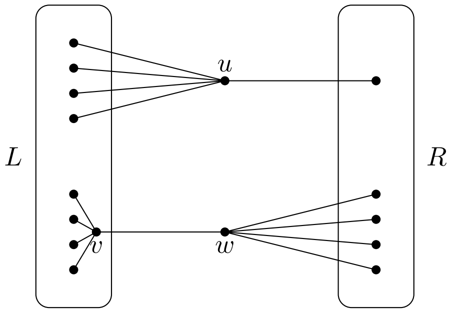
However, this might increase the degree of both u and v to exactly $\Delta(G)$, in which case we may not have a vertex of degree less than $\Delta(G)$ with which to apply the lemma. This happens if and only if $G \setminus \{u,v\} = L \cup R$ has two connected components, and u and v each only have one edge incident to one of them, say R,



since in this case, adding the edge between u and v in $G\big[L \cup \{u,v\}\big]$ leaves their degrees unchanged:



Instead, replace v by its neighbour w in R:



The vertices u and w disconnect G, and now in both components of $G \setminus \{u,w\}$, at least one of u and w has at most $\Delta(G) - 2$ neighbours, so adding in the edge (u,w) leaves each piece with maximum degree $\Delta(G)$, and at least one vertex of smaller degree.

(*iii*) We claim that there is an induced path in G of length 2, passing through vertices, say, u,x,v, such that u and v are not adjacent.

Let S be a maximal complete subgraph of G. By assumption, G is not complete, so there is a vertex $u \in G \setminus S$ adjacent to a vertex $x \in S$. There must also be a vertex $v \in S$ not adjacent to u, since if every vertex in S were adjacent in u, then $S \cup \{u\}$ would be a larger complete subgraph, contradicting that S is maximal. This proves the claim.

Colour the vertices u and v with the same colour. Since G is 3-connected, the graph $G \setminus \{u,v\}$ is connected, so each vertex in it has a well-defined distance from x. We proceed as in the proof of the previous lemma, greedily colouring in the subgraphs induced on the sets V_i of vertices at distance i from x. Once we reach x, it has two neighbours u and v with the same colour, so there is a spare colour for x.

■

8.5.1 The Chromatic Polynomial

Given a graph G, we define the function $P_G : \mathbb{N}_{>0} \to \mathbb{N}$ as:

$$P_G(k) := \# \text{ of proper } k\text{-colourings of } G$$

where two colourings are considered distinct if there is a vertex labelled with different colours in the two colourings.

Example. For the complete graph K_n on n vertices, we have:

$$P_{K_n}(k) = \prod_{i=0}^{n-1} (k-i)$$

If $k < n$, then K_n is not k-colourable, so $P_{K_n}(k) = 0$. If $k \geq n$, then k-colourings exist: we may select any of the k colours for the first vertex, any of the remaining $k-1$ for the second, etc.

So, there are $k(k-1)(k-2)\cdots\big(k-(n-1)\big)$ such colourings, and this formula also agrees with the $k < n$ case since one of the factors would vanish. △

Example. For the path graph P_n on n edges and $n+1$ vertices, we have:

$$P_{P_n}(k) = k(k-1)^n$$

If $k = 1$ and $n \geq 1$, then there are no colourings. If $k > 1$, we may choose any of the k colours for the first vertex. Then, traversing the path, for each of the n remaining vertices v_{i+1}, we may choose any of the $k-1$ colours distinct from the colour of the previous vertex v_i.

So, there are $k(k-1)(k-1)\cdots(k-1)$ such colourings, and this formula agrees with the $k = 1$ case, since every factor past the first would vanish. △

Example. For the empty graph E_n on n vertices, we have:

$$P_{E_n}(k) = k^n$$

Since there are no adjacencies, every vertex can be independently coloured with any of the k colours, and there are n vertices, so there are k^n total colourings. △

So far, we have seen that

$$\begin{aligned} P_{K_n}(k) &= \prod_{i=0}^{n-1} (k-i) \\ P_{P_n}(k) &= k(k-1)^n \\ P_{E_n}(k) &= k^n \end{aligned}$$

In each case, P_G is a polynomial in k. This turns out to be true for all finite graphs, and we call P_G the *chromatic polynomial.* This fact is not obvious:

Example. For the cycle graph C_n on n vertices, we can start similarly to the path graph: we choose a vertex v_0, and colour it with any of the k colours. Then, traversing the cycle in one direction, we can colour each vertex v_{i+1} with any of the $k-1$ colours distinct from the colour of the previous vertex v_i.

However, how can we colour the vertex v_{n-1} that is adjacent to v_0? The number of valid colours depends on whether v_{n-2} is the same colour as v_0 or not. At this point, it is unclear as to how we should proceed. △

Often, when proving a result on all finite graphs, we induct on the size of the vertex set. However, in the inductive step, the number of choices for the new vertex depends not only on the number of neighbours, but also on the colouring on the neighbours.

Instead, we might try to induct on the number of edges. Take a finite graph G and let $e = (x,y)$ be an edge in G. From G, we construct the graph $G \setminus e$ by deleting e, and the graph G/e by contracting e.

We can partition the possible k-colourings c of $G \setminus e$ into two cases:

- $c(x) \neq c(y)$;
- $c(x) = c(y)$.

In the former case, each such colouring is also an admissible colouring of G, since x and y are adjacent in G, but are assigned different colours. In the latter case, these colourings are not admissible. However, such colourings correspond precisely to the proper colourings of G/e, since x and y are the same vertex in the contraction.

This observation provides us with our inductive step.

Theorem 8.5.4. *For every finite graph $G = (V,E)$ containing an edge $e = (x,y) \in E$, and for every $k \geq 1$,*

$$P_G(k) = P_{G\setminus e}(k) - P_{G/e}(k)$$

Consequently, P_G coincides with a polynomial in $\mathbb{R}[k]$.

Proof. Every k-colouring of $G \setminus e$ either assigns different colours to x and y, in which case, it corresponds to a proper colouring of G, or it assigns them the same colour. So, it suffices to check that the number of colourings of $G \setminus e$ in which x and y are assigned the same colour is equal to the number of colourings of G/e.

Given a colouring of $G \setminus e$ with $c(x) = c(y)$ are the same colour, then we can construct a corresponding colouring of G/e by colouring the contracted vertex as $c(x) = c(y)$, and leaving all other vertices unchanged; conversely, given a colouring of G/e, we can construct a colouring of $G \setminus e$ by colouring x and y the same as the contracted vertex.

We deduce that P_G is a polynomial by induction on $|E|$.

If $|E| = 0$, then $G = E_{|V|}$ is an empty graph, and we have that $P_{E_n}(k) = k^n$ is a polynomial. Otherwise, G has an edge $e = (x,y)$, and $G \setminus e$ and G/e are graphs with fewer edges than G, so $P_G(k) = P_{G\setminus e}(k) - P_{G/e}(k)$ is the difference of two polynomials and is hence a polynomial. ■

We can use this recurrence relation to compute the chromatic polynomial:

Theorem 8.5.5. *The chromatic polynomial of the cycle graph C_n is*

$$P_{C_n}(k) = (k-1)^n + (-1)^n(k-1)$$

Proof. If $n = 3$, then $C_3 = K_3$, and

$$\begin{aligned}(k-1)^3 + (-1)^3(k-1) &= (k-1)^3 - (k-1)\\ &= k^3 - 3k^2 + 3k - 1 - k + 1\\ &= k^3 - 3k^2 + 2k\\ &= k(k-1)(k-2)\\ &= P_{K_3}(k)\\ &= P_{C_3}(k)\end{aligned}$$

as required.

Now suppose $n > 3$, and that the result holds for all smaller cycles. Let $e \in E(C_n)$. Then, $C_n \setminus e = P_{n-1}$ and $C_n/e = C_{n-1}$, so

$$\begin{aligned} P_{C_n}(k) &= P_{C_n \setminus e}(k) - P_{C_n/e}(k) \\ &= P_{P_{n-1}}(k) - P_{C_{n-1}}(k) \\ &= \left[k(k-1)^{n-1}\right] - \left[(k-1)^{n-1} + (-1)^{n-1}(k-1)\right] \\ &= (k-1)(k-1)^{n-1} + (-1)(-1)^{n-1}(k-1) \\ &= (k-1)^n + (-1)^n(k-1) \end{aligned}$$

completing the inductive step. ■

Note that if $k = 2$, then,

$$\begin{aligned} P_{C_n}(2) &= (2-1)^n + (-1)^n(2-1) \\ &= 1 + (-1)^n \\ &= \begin{cases} 0 & n \text{ odd} \\ 2 & n \text{ even} \end{cases} \end{aligned}$$

There are some other properties of the chromatic polynomial that we can immediately deduce from the recurrence relation.

Theorem 8.5.6. *For any finite graph $G = (V,E)$,*

- (i) *The degree of P_G is $|V|$;*
- (ii) *P_G is monic;*
- (iii) *The coefficients of P_G have alternating signs;*
- (iv) *$P_G(0) = 0$.*

Proof. In all cases, we induct on $|E|$ to use the chromatic polynomial recurrence relation.

If $|E| = 0$, then G is empty and $P_{E_{|V|}}(k) = k^{|V|}$ is a monic polynomial of degree $|V|$. Also, the coefficients are all zero apart from this first term, so they trivially alternate signs. We also have $P_{E_{|V|}}(0) = 0^{|V|} = 0$. This establishes the base case for all four claims.

Now, suppose $|E| > 0$, and that the result holds for all graphs with fewer edges, and let $e \in E$, so

$$P_G(k) = P_{G\setminus e}(k) - P_{G/e}(k)$$

- (i),(ii) The graph $G \setminus e$ has fewer edges than G, but the same number of vertices, so $P_{G\setminus e}(k)$ is a monic polynomial of degree $|V(G \setminus e)| = |V|$ by the inductive hypothesis. The graph G/e has fewer edges and fewer vertices than G, so its chromatic polynomial does not contribute to the $|V|$th order term in P_G. So, P_G is a monic polynomial of degree $|V|$.
- (iii) The graphs $G \setminus e$ and G/e have fewer edges than G, so their chromatic polynomial coefficients have alternating signs by the induction hypothesis. $P_{G/e}$ also has degree $|V| - 1$ by property (i), so its signs are opposite to that of $P_{G\setminus e}$, and this alternation is preserved in their difference.
- (iv) By the inductive hypothesis, $P_{G\setminus e}(0) = P_{G/e}(0) = 0$. So $P_G(0) = 0 - 0 = 0$.

■

Thus, if $G = (V,E)$ with $|V| = n$, then its chromatic polynomial is of the form:

$$P_G(k) = k^n - c_{n-1}k^{n-1} + c_{n-2}k^{n-2} - c_{n-3}k^{n-3} + \cdots$$

Theorem 8.5.7. *Let $G = (V,E)$ be a finite graph with chromatic polynomial*

$$P_G(k) = k^n - c_{n-1}k^{n-1} + c_{n-2}k^{n-2} - c_{n-3}k^{n-3} + \cdots$$

Then,

(*i*)
$$c_{n-1} = |E|$$

(*ii*) *and*
$$c_{n-2} = \binom{|E|}{2} - T(G)$$

where $T(G)$ is the number of triangles in G.

Proof. We induct on $|E|$. For $|E| = 0$, $P_{E_n}(k) = k^n$, and the two coefficients are 0, as required. Now, suppose $|E| > 0$, and that the result holds for all graphs with fewer edges.

Let $e = (x,y) \in E$ and let $c_n(G)$ denote the coefficient the nth degree term of P_G.

(*i*) Comparing the $(n-1)$th degree terms of the three polynomials, we have

$$\big[-c_{n-1}(G)\big] = \big[-c_{n-1}(G \setminus e)\big] - c_{n-1}(G/e)$$

Note that the first two terms are negative because c_{n-1} is the *second* coefficient of P_G and $P_{G\setminus e}$, but the *leading* coefficient of $P_{G/e}$, since $P_{G/e}$ degree one less than P_G.

By the previous theorem, the leading coefficient $c_{n-1}(G/e)$ is 1, and by the inductive hypothesis, $c_{n-1}(G \setminus e) = \big|E(G \setminus e)\big| = |E| - 1$. So,

$$\begin{aligned} c_{n-1}(G) &= \big(|E| - 1\big) + 1 \\ &= |E| \end{aligned}$$

This completes the inductive step.

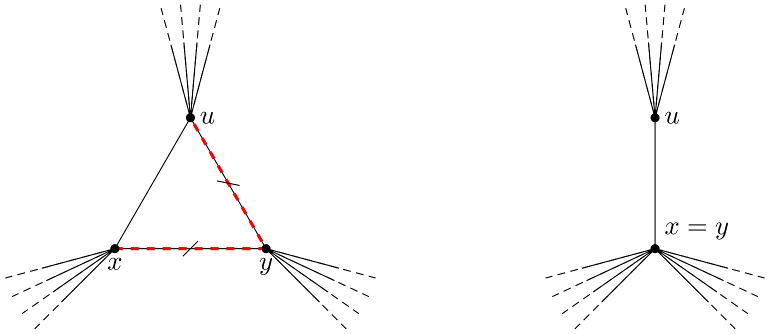
(*ii*) Comparing the $(n-2)$th degree terms of the three polynomials, we have

$$c_{n-2}(G) = c_{n-2}(G \setminus e) - \big[-c_{n-2}(G/e)\big]$$

This time, c_{n-2} is the third coefficient of P_G and $P_{G\setminus e}$, which is positive, but the second coefficient of $P_{G/e}$, which is negative. By part (*i*),

$$c_{n-2}(G/e) = \big|E(G/e)\big|$$

When we contract $e = (x,y)$, we lose one edge from E, but we also lose an edge for every vertex u adjacent to both x and y, since those two edges are combined into a single edge in the contraction:



Let T_0 be the number of such vertices, so

$$c_{n-2}(G/e) = |E| - 1 - T_0$$

Then, by the inductive hypothesis,

$$\begin{aligned} c_{n-2}(G \setminus e) &= \binom{|E(G \setminus e)|}{2} - T(G \setminus e) \\ &= \binom{|E| - 1}{2} - T(G \setminus e) \end{aligned}$$

$T(G \setminus e)$ is equal to $T(G)$ minus the number of triangles that were removed when deleting e. That is, the number of triangles that contain e. But this is exactly T_0, so

$$c_{n-2}(G \setminus e) = \binom{|E| - 1}{2} - \big(T(G) - T_0\big)$$

So the c_{n_2} coefficient of G is:

$$\begin{aligned} c_{n-2}(G) &= c_{n-2}(G \setminus e) + c_{n-2}(G/e) \\ &= \binom{|E| - 1}{2} - \big(T(G) - T_0\big) + |E| - 1 - T_0 \\ &= \binom{|E| - 1}{2} + |E| - 1 - T(G) \\ &= \binom{|E|}{2} - T(G) \end{aligned}$$

as required.

■

8.6 Matroids

A *matroid* $(E,\mathcal{I})$ consists of a *ground set* E, and a family $\mathcal{I} \subseteq \mathcal{P}(E)$ of its subsets satisfying

(i) $\varnothing \in \mathcal{I}$;

(ii) If $A \subseteq B$ and $B \in \mathcal{I}$, then $A \in \mathcal{I}$ (*hereditary property* or *downward-closedness*);

(iii) If $A,B \in \mathcal{I}$ and $|A| > |B|$, then there is an element $a \in A$ such that $\{a\} \cup B \in \mathcal{I}$ (*exchange condition*).

The sets in $\mathcal{I}$ are called the *independent sets* of the matroid.

Example. Let V be a vector space and $E \subseteq V$ be a set of vectors. If $\mathcal{I}$ is the collection of linearly independent subsets of E, then $(E,\mathcal{I})$ is a matroid called the *vector matroid.* △

Example. Fix integers n,r with $n < r$. Let $E = [n]$ and take $\mathcal{I} = \{S \subseteq E : |S| < r\}$ to be the set of subsets with cardinality at most r. Then, $(E,\mathcal{I})$ is a matroid called the *uniform matroid* $U_{n,r}$. △

Lemma 8.6.1 (Characterisation of Trees)**.** *For any graph $G = (V,E)$, any two of the following imply the third (and hence that G is a tree):*

(i) $|E| = |V| - 1$*;*

(ii) *G is connected;*

(iii) *G is acyclic.*

Theorem 8.6.2 (Graphic Matroids)**.** *Let $G = (V,E)$ be a graph, and let $\mathcal{I} \subseteq \mathcal{P}(E)$ be the set of acyclic subsets of E (i.e. the set of forests). Then, $(E,\mathcal{I})$ is a matroid.*

Proof. The empty set is acyclic, and any subset of an acyclic subset is still acyclic. Now, let $A,B \in \mathcal{I}$ with $|A| > |B|$, and consider the graph $G_B = (V,B)$. B is acyclic, so G_B is a forest.

If there is an edge $a \in A$ that connects distinct components of B, then we are done, as $\{a\} \cup B$ is acyclic. Otherwise, suppose there are no such edges. But then, every edge in A lies within the connected components of G_B. Since A is acyclic, it cannot have more edges in each component than a tree does, so it cannot have more elements than B. ■

A matroid $(E,\mathcal{I})$ is *representable* over a field K if there is a vector space V over K and a map $\phi : E \to V$ such that for each $A \subseteq E$, $A \in \mathcal{I}$ if and only if $\phi(A)$ is a linearly independent set in V.

Example. The uniform matroid $U_{4,2}$ cannot be represented over $\mathbb{Z}_2$.

Suppose otherwise, so there are 4 vectors $x_1,\ldots,x_4$ such that any pair are linearly independent, but any three are linearly dependent. The only possible linear dependency of three vectors is of the form

$$x_1 + x_2 + x_3 = 0$$

since the only coefficients available are 0 and 1..

But then, adding the linear dependencies $x_1+x_2+x_3 = 0$ and $x_1+x_2+x_4$ yields $x_3+x_4 = 0$, contradicting the linear independence of any pair of vectors. △

Example. Graphic matroids can be represented over any field. △

Example. $U_{n,r}$ is representable over $\mathbb{R}$ for any n and r. △

8.6.1 Rado's Theorem

We recall the set-theoretic statement of Hall's theorem.

A *transversal* or *system of distinct representatives* (*SDR*) of a family of subsets $\mathcal{F} \subseteq \mathcal{P}(X)$ is a subset of X obtained by selecting a distinct representative from each subset $S \in \mathcal{F}$.

Theorem 8.6.3 (Hall)**.** *Let $\{S_i\}_{i=1}^n$ be a collection of subsets of a set X. Then, there is a transversal of $\{S_i\}$ if and only for every subset of indices $\sigma \subseteq [n]$, we have*

$$\left|\bigcup_{i\in\sigma} S_i\right| \geq |\sigma|$$

Proof. Apply the graph-theoretic variant of Hall's theorem on the bipartite graph $G = (L \cup R,E)$, where $L = \{S_i\}$ and $R = X$, and $(S_i,e) \in E$ if and only if $e \in S_i$. ■

Suppose we have sets $S_i \subseteq E$ in a matroid $(E,\mathcal{I})$. Under what conditions can we find a transversal of the S_i that is an independent set?

For a set $A \subseteq E$, we define the *rank* $r(A)$ of A to be the cardinality of the largest independent set in A:

$$r(A) := \max\bigl\{|B| : B \subseteq A,B \in \mathcal{I}\bigr\}$$

Example. If $(E,\mathcal{I})$ is a vector matroid, then $r(E) = \dim(E)$, and $r(A) = \dim\bigl(\operatorname{span}(A)\bigr)$. △

Lemma 8.6.4 (Rank Submodularity)**.** *Given a matroid $(E,\mathcal{I})$, the rank function satisfies, for any $A,B \subseteq E$:*

$$r(A \cup B) + R(A \cap B) \leq r(A) + r(B)$$

Proof. Choose a maximal independent set $I_\cap \in \mathcal{I}$ in the intersection $A \cap B$. By definition of matroid rank, $|I_\cap| = r(A \cap B) =: n$. Through repeated applications of the exchange condition, extend this set to a maximal independent set $I_\cup$ in $A \cup B$ of size $|I_\cup| = r(A \cup B) =: m$.

Let $a := |I_\cup \setminus B|$ and $b := |I_\cup \setminus A|$, so

$$\begin{aligned} |I_\cup| &= |I_\cup \setminus B| + |I_\cup \setminus A| + |I_\cup \cap (A \cap B)| \\ |I_\cup| &= |I_\cup \setminus B| + |I_\cup \setminus A| + |I_\cap| \\ m &= a + b + n \end{aligned}$$

Then,

$$\begin{aligned} r(A \cup B) + r(A \cap B) &= m + n \\ &= (a + b + n) + n \\ &= (a + n) + (b + n) \end{aligned}$$

Note that $I_\cap \sqcup (I \cap \setminus B) = (I_\cup \cap A) \subseteq A$ is an independent subset of A of size $a + n$. The rank $r(A)$ is defined to be the size of a maximal independent subset of A, so we have $a + n \leq r(A)$. Similarly, $b + n \leq r(B)$, giving:

$$\leq r(A) + r(B)$$

as required. ■

Theorem 8.6.5 (Rado)**.** *Let $(E,\mathcal{I})$ be a matroid, and let $S_1, \dots, S_n \in \mathcal{P}(E)$ be arbitary subsets of E. If for every set $\sigma \subseteq [n]$ of indices,*

$$r\left(\bigcup_{i \in \sigma} S_i\right) \geq |\sigma|$$

then there is an independent traversal. That is, a set $\{e_1, \dots, e_n\} \in \mathcal{I}$ of n distinct elements of E with $e_i \in S_i$ for each i.

Theorem 8.6.6 (Horn)**.** *Let $X = \{x_1, \dots, x_n\}$ be vectors in a vector space V, and suppose that for each set $\sigma \subseteq [n]$ of indices,*

$$\dim\big(\operatorname{span}\big(\{x_i : i \in \sigma\}\big)\big) \geq \frac{|\sigma|}{2}$$

Then, the set of vectors can be partitioned into two linearly independent sets.

Proof. Let $E = X \sqcup X$, and denote the second copies of the x_i by x_i'. We declare a subset of E as independent in $\mathcal{I}$ if its x_i elements are linearly independent and its x_i' elements are linearly independent.

The empty set is independent in both cases, and a subset of a linearly independent set is still linearly independent, so $(E,\mathcal{I})$ satisfies downward-closure. Then, if $A,B \in \mathcal{I}$ and $|A| > |B|$, then A has more x_i than B, or A has more x_i' than B (or both). By considering only the larger set, this is effectively the ordinary vector matroid, so the exchange condition holds similarly in any case.

For each $i \in [n]$, define the set $S_i = \{x_i, x_i'\}$. Let $\sigma \subseteq [n]$. By the hypotheses of the theorem, there is a subset $\tau \subseteq \sigma$ of at least half the size for which the vectors $\{x_i : i \in \tau\}$ are linearly independent. But then,

$$\bigcup_{i \in \sigma} S_i \supseteq \bigcup_{i \in \tau}$$

which has at least $|\sigma|$ elements, and is independent in the matroid. So, the sets S_i satisfy the hypotheses of Rado's theorem, so there is an independent transversal of the S_i. That is, an independent selection of exactly one x_i or x_i' from each S_i. Then, the set of selected x_i and the set of selected x_i' give the required partition. ■

8.7 Random Graphs

Given $r \in \mathbb{N}$, we define $R(r,r)$ to be the smallest positive integer such that for every edge 2-colouring of K_n, there is a monochromatic K_r as a subgraph.

Theorem 8.7.1. *$R(3,3) = 6$.*

Proof. Suppose the edges of K_6 are coloured red and blue. Select a vertex u. There are five edges incident to u, so by the pigeonhole principle, at least three of these edges (u,v_1), (u,v_2), and (u,v_3) are the same colour, say, red. If any of the edges connecting the v_i are red, then this forms a red triangle including u. Otherwise, none of the edges are red, in which case, the v_i form a blue monochromatic triangle. ■

Theorem 8.7.2 (Erdős Lower Bound for $R(r,r)$)**.** *Let $r \geq 3$. Then,*

$$R(r,r) \geq 2^{\frac{r-1}{2}}$$

Proof. Colour the edges of K_n red or blue independently with probability 1/2 each. For any fixed set S_i of r vertices, define the random variable $X(S_i)$ to be 1 if the K_r induced on S_i is monochromatic, and 0 otherwise. For any S_i, the expectation of $X(S_i)$ is the probability that all $\binom{r}{2}$ edges are the same colour:

$$\mathbb{E}\big[X(S_i)\big] = 2 \cdot \left(\frac{1}{2}\right)^{\binom{r}{2}} = 2^{1-\binom{r}{2}}$$

There are $\binom{n}{r}$ many possible subsets S_i, so the number of monochromatic K_n is the sum

$$\sum_{i=1}^{\binom{n}{r}} X(S_i)$$

which has expected value

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{\binom{n}{r}} X(S_i)\right] &= \sum_{i=1}^{\binom{n}{r}} \mathbb{E}\big[X(S_i)\big] \\ &= \binom{n}{r} \cdot 2^{1-\binom{r}{2}} \\ &= \frac{2 \cdot n!}{r!(n-r)!2^{\frac{r(r-1)}{2}}} \\ &< \frac{2 \cdot n^r}{r!2^{\frac{r(r-1)}{2}}}\end{aligned}$$

If this is less than 1, then there are colourings without any monochromatic K_r. But, if $n \leq 2^{\frac{r-1}{2}}$, then this expectation is at least 1. ■

Given $n \in \mathbb{N}$ and $p \in [0,1]$, we write $G \sim G(n,p)$, or just $G_{n,p}$, if G is a random graph with vertex set $E(G) = [n]$ and each possible edge is included independently at random in $E(G)$ with probability p.

Lemma 8.7.3. *For any $p \in (0,1)$ and any m,*

$$\exp\left(\frac{-mp}{1-p}\right) \leq (1-p)^m \leq \exp(-mp)$$

Lemma 8.7.4. *For $1 \leq k \leq n$,*

$$\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$$

Theorem 8.7.5 (Linearity of Expectation)**.** *For any random variables $X_1, \ldots, X_n$, which may be dependent, and constants $c_1, \ldots, c_n$,*

$$\mathbb{E}\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i \cdot \mathbb{E}[X_i]$$

Example. We compute the expected number of triangles in $G_{n,p}$ using the linearity of expectation. There are $\binom{n}{3}$ possible triangles, and each triangle has probability p^3 of being included. For each triangle T, define the random variable X_T to be 1 if T is in $G_{n,p}$, and 0 otherwise. Then, the number of triangles in $G_{n,p}$ is

$$\sum_{T \in G} X_T$$

The X_T are not independent, since if T is in G, then any triangle sharing an edge with T is more likely to also be in G. But by the linearity of expectation, the expected number of triangles in G is

$$\begin{aligned}\mathbb{E}\left[\sum_{T \in G} X_T\right] &= \sum_{T \in G} \mathbb{E}[X_T] \\ &= \binom{n}{3} p^3\end{aligned}$$

△

Theorem 8.7.6 (Markov's Inequality)**.** *If X is a non-negative random variable, and $t > 0$, then*

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}[X]}{t}$$

8.7.1 Chromatic Numbers

Theorem 8.7.7 (Chromatic Number of a Random Graph)**.** *Let $k \in \mathbb{Z}$ and suppose p satisfies*

$$p \geq \frac{2k\log(k) + 4k}{n}$$

Then,

$$\mathbb{P}\big(\chi(G_{n,p}) \leq k\big) \leq \exp\left(-\frac{n}{2k}\right)$$

If a graph is k-coloured, then one of the colour classes has at least $\frac{n}{k}$ vertices in it by the pigeonhole principle, so it suffices to prove that if r is an integer satisfying

$$\frac{n}{k} \leq r \leq \frac{n}{k} + 1$$

then then probability that $G_{n,p}$ contains an independent set of r vertices is less than $\exp(-r/2)$.

Theorem 8.7.8 (Independence Number of a Random Graph)**.** *Let $k \in \mathbb{Z}$ and suppose p satisfies*

$$p \geq \frac{2k\log(k) + 4k}{n}$$

Then,

$$\mathbb{P}(G_{n,p} \textit{ has an independent set of size at least } r) \leq \exp\left(-\frac{r}{2}\right)$$

Lemma 8.7.9. *Let $g,n \in \mathbb{Z}$ and $p \in [0,1]$ satisfy*

$$\frac{5}{n} \leq p \leq \frac{n^{\frac{1}{g}}}{n}$$

and let X be the number of cycles of length at most $g-1$ in $G_{n,p}$. Then, $\mathbb{E}(X) \leq \frac{n}{4}$.

Corollary 8.7.9.1. *For large n, there is a graph on at most n vertices with chromatic number at least*

$$\frac{\log(n)}{4\log\big(\log(n)\big)} - 1$$

and no cycles shorter than

$$\frac{\log(n)}{\log\big(\log(n)\big)} - 1$$

8.7.2 Connectedness

Theorem 8.7.10 (Connectedness of Random Graphs)**.** *Let (c_n) be a sequence, and let*

$$p(n) = \frac{\log(n)}{n} + \frac{c_n}{n}$$

Then,

$$\mathbb{P}(G_{n,p} \textit{ is connected}) \to \begin{cases} 0 & c_n \to -\infty \\ 1 & c_n \to \infty \end{cases}$$

8.8 Regularity Method

WIP

Chapter 9

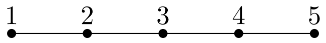
Graph Theory

9.1 Introduction

A *graph* $G = (V,E)$ consists of a finite set V of *vertices* or *nodes*, and a finite set $E \subseteq \binom{V}{2}$ of *unordered* pairs of distinct vertices, called *edges* or *arcs*.

Graphs have a natural visual representation in which each vertex is represented by a point, and each edge by a line connecting two points.

Example. We can draw the graph $G = \big(\{1,2,3,4,5\},\{\{1,2\},\{2,3\},\{3,4\},\{4,5\}\}\big)$ as



△

By various modifications, we obtain different types of graphs.

- If we instead have $E \subseteq V \times V$ such that the edges are *ordered* pairs, then the graph is *directed* or *oriented*, and can also be referred to as a *digraph*.
- If we allow both ordered and unordered edges $E \subseteq \binom{V}{2} \cup (V \times V)$, then we obtain *mixed graphs*.
- If we allow repeated or *parallel* edges by replacing E with a multiset, then we obtain *multigraphs*.
- If we allow edges to connect a vertex to itself (a *loop*), then we obtain *pseudographs*.
- If we allow the edges to be arbitrary subsets of vertices and not necessarily pairs, then we obtain *hypergraphs*.
- If we allow V and E to be infinite sets, we obtain *infinite graphs*.

A *simple graph* is a finite undirected graph without loops and multiple edges. In this chapter, every graph will be simple unless stated otherwise.

9.1.1 Terminology and Notation

9.1.1.1 Vertices and Edges

The vertex set of a graph G is also denoted by $V(G)$, and similarly, the edge set of G is denoted by $E(G)$. For notational convenience, an unordered edge $\{a,b\}$ will be shortened to just ab.

- Let $u,v \in V(G)$ be two vertices. If $uv \in E(G)$, then u and v are said to be *adjacent*, or that u is a *neighbour* of v.

- The (*open*) *neighbourhood* $N_G(v)$ of a vertex $v \in V(G)$ is the set of vertices adjacent to v:
$$N_G(v) \coloneqq \big\{u \in V(G) : uv \in E(G)\big\}$$
When the graph G is clear, we often suppress the subscript.
- The *closed neighbourhood* $N_G[v]$ of v is the neighbourhood of v, plus v itself:
$$N_G[v] \coloneqq N(v) \cup \{v\}$$
- If $e = uv$ is an edge of G, then G is *incident* to u and v. We also say that u and v are the *endpoints* of e.
- The *degree*, *valency*, or *order* of a vertex $v \in V(G)$ is the number of edges incident to v:
$$\deg(v) \coloneqq \big|N_G(v)\big|$$
In a digraph, we instead define the *indegree* and *outdegree* of a vertex to be the number of edges pointing into and out from the vertex, respectively.

  If

  - $\deg(v) = 0$, then v is *isolated.*
  - $\deg(v) = 1$, then v is a *leaf*, and v together with the only edge incident to v are called *pendant*;
  - $\deg(v) = |V(G)| - 1$, then v is *dominating.*
- The maximum vertex degree and minimum vertex degree in a graph G are denoted by $\Delta(G)$ and $\delta(G)$, respectively:
$$\Delta(G) \coloneqq \max_{v\in V(G)} \deg(v), \qquad \delta(G) \coloneqq \min_{v\in V(G)} \deg(v)$$
- The *degree sequence* of a graph is the sorted list of its vertex degrees. If every vertex has the same degree k, i.e. $\Delta(G) = \delta(G) = k$, G is said to be *k-regular.* In particular, 3-regular graphs are called *cubic.*

Lemma 9.1.1 (Euler's Handshaking Lemma)**.** *Let $G = (V,E)$ be a graph. Then,*
$$\sum_{u\in V} \deg(v) = 2|E|$$

Proof. Each edge is incident to two vertices, so each edge is counted twice in the sum. ■

Corollary 9.1.1.1. *The number of vertices of odd degree is even in any graph.*

9.1.1.2 Paths and Connectedness

- A *path* in a graph is a sequence of distinct vertices $v_1,v_2,\ldots,v_k$ such that v_iv_{i+1} is an edge for each $i = 1,\ldots,k-1$.
- The *length* of a path P is the number of edges connecting consecutive vertices of P.
- A *chord* in a path is an edge connecting two non-consecutive vertices. A *chordless path* is a path without chords.
- A graph is *connected* if every pair of distinct vertices is joined by a path, and is *disconnected* otherwise.

- The *distance* $d(x,y)$ between two vertices x and y is the length of a shortest path connecting them. This notion of distance defines a metric on any connected graph.
- The *diameter* of a connected graph is the maximum distance achieved between some pair of vertices:

$$\operatorname{diam}(G) \coloneqq \max_{x,y \in V(G)} d(x,y)$$

9.1.1.3 Special Graphs

- K_n is the *complete graph* on n vertices – the graph on n vertices with all possible edges.
- E_n is the *empty* (*edgeless*) *graph* on n vertices – the graph on n vertices with no edges.
- P_n is a *chordless path* on n vertices – $V(P_n) = \{v_1, \ldots, v_n\}$ and $E(P_n) = \{v_1v_2, \ldots, v_{n-1v_n}\}$.
- C_n is a *chordless cycle* on n vertices – $V(C_n) = \{v_1, \ldots, v_n\}$ and $E(C_n) = \{v_1v_2, \ldots, v_{n-1}v_n, v_nn_1\}$.
- Q_n is a *hypercube* – the graph whose vertex set is the set of all binary strings of length n where two vertices are adjacent if and only if they differ in precisely one coordinate.
- $G + H$ or $G \amalg H$ is the disjoint union of two graphs G and H – $V(G + H) = V(G) \sqcup H(G)$ and $E(G + H) = E(G) \sqcup E(H)$. In particular, nG denotes the disjoint union of n copies of G. For instance, $E_n = nK_1$.
- $G \times H$ is the *join* of G and H, obtained by adding all possible edges between G and H in the disjoint union $G + H$.
- $W_n \coloneqq K_1 \times C_n$ is the *wheel* on n vertices.

Two graphs $G_1 = (V_1,E_1)$ and $G_2 = (V_2,E_2)$ are *isomorphic* if there exists a bijection $f : V_1 \to V_2$ such that $uv \in E_1$ if and only if $f(u)f(v) \in E_2$, and we write $G_1 \cong G_2$ to denote this relation.

- The *complement* of a graph $G = (V,E)$ is a graph $\overline{G}$ with vertex set V and edge set E', where $e \in E'$ if and only if $e \notin E$.
- A graph is *self-complementary* if G is isomorphic to its complement.

9.1.1.4 Subgraphs

Given two graphs $G_1 = (V_1,E_1)$ and $G_2 = (V_2,E_2)$, G_1 is said to be:

- a *subgraph* of G_2 if $V_1 \subseteq V_2$ and $E_1 \subseteq E_2$, i.e. G_1 can be obtained from G_2 by deleting vertices and edges;
- a *spanning subgraph* of G_2 if $V_1 = V_2$ and $E_1 \subseteq E_2$, i.e. G_1 can be obtained from G_2 by deleting edges but not vertices;
- an *induced subgraph* of G_2 if G_1 is a subgraph of G_2 such that $uv \in E_1$ whenever $u,v \in V_1$, i.e. G_1 can be obtained from G_2 by deleting vertices.

Given a graph G and a subset $U \subseteq V(G)$, we write

- $G[U]$ for the subgraph of G induced by U, i.e. the graph with vertex set U whose vertices are adjacent if and only if they are adjacent in G;
- $G - U$ for the subgraph of G induced by $V(G) \setminus U$, i.e. the graph obtained from G by deleting the vertices in U.

We say that G contains a graph H as an induced subgraph if H is isomorphic to an induced subgraph of G, and we write $H < G$ to denote this relation. If $H \not< G$, then G is said to be *H-free.*

A maximal (with respect to inclusion) connected subgraph of G is called a *connected component* of G. A *co-component* in a graph is a connected component of its complement.

9.1.1.5 Cliques and Independent Sets

In a graph, a set of pairwise adjacent vertices is called a *clique*. The size of a maximum clique in G is called the *clique number* of G, and is denoted by $\omega(G)$.

A set of pairwise non-adjacent vertices is called an *independent set* or *stable set*. The size of a maximum independent set in G is called the *independence number* or *stability number* of G, and is denoted by $\alpha(G)$.

9.1.2 Exercises

1. Show that every graph has two vertices of the same degree.
2. Prove or disprove that there exist graphs in which all vertices are pendant.
3. Find the diameter of:
   (a) K_n;
   (b) P_n;
   (c) C_n;
   (d) Q_n;
   (e) W_n;
   (f) $P_n \times C_n$.
4. Find the length of the shortest path between $\mathbf{0} = [0,0,0,\ldots,0]$ and $[1,1,1,\ldots,1]$ in Q_n.
5. Determine $\delta(Q_n)$, $\Delta(Q_n)$, $|V(Q_n)|$, and $|E(Q_n)|$.
6.
7. Find all pairwise non-isomorphic graphs on 2, 3, 4, and 5 vertices.
8. Find all pairwise non-isomorphic $(n-2)$-regular graphs on n vertices.
9. Find all pairwise non-isomorphic with the degree sequence:
   (a) (0,1,2,3,4);
   (b) (1,1,2,3,4);
   (c) (2,2,3,3,4,4).
10. Let G be a self-complementary graph. Show that if the degree sequence of G is $(d_1,\ldots,d_n)$ with the d_i listed in non-increasing order, then

$$d_i = n - 1 - d_{n+1-i}$$

    for all $i = 1,2,\ldots,\lfloor n/2 \rfloor$
11. Prove that any self-complementary k-regular graph on n vertices satisfies $k = (n-1)/2$.
12. Prove that graph isomorphism is an equivalence relation the class of all graphs.
13. Prove that:
    (a) $C_4 \cong E_2 \times E_2$;
    (b) $K_4 \cong W_4$.
14. Find the complements of
    (a) C_4;

(b) C_5;

(c) P_4;

(d) P_5.

15. Show that

    (a) If $\operatorname{diam}(G) \geq 3$, then $\operatorname{diam}(\overline{G}) \leq 3$;

    (b) If $\operatorname{diam}(G) \geq 4$, then $\operatorname{diam}(\overline{G}) \leq 2$;

16. Determine $|E(\overline{P_n})|$.

17. Find examples of self-complementary graphs on 4, 5, and 6 vertices.

18. Is it possible for a self-complementary graph with 100 vertices to have exactly one vertex of degree 50?

19. Show that for every $n \in \mathbb{N}$, there exist self-complementary graphs with at least n vertices.

20. Prove that:

    (a) The complement of a connected graph is necessarily disconnected;

    (b) The complement of a disconnected graph is necessarily connected and hence deduce that a graph is connected if and its complement is disconnected.

21. Prove that a graph is connected if and only if for every partition of its vertex set into two non-empty sets A and B, there exist vertices $a \in A$ and $b \in B$ such that $ab \in E(G)$.

22. Prove that if a graph on n vertices has more than $\binom{n-1}{2}$ edges, then it is connected.

23. Prove that if a graph has exactly two vertices of odd degrees, then they are connected by a path.

24. Let G be a graph with $|V(G)|$ even, and for each vertex $v \in V(G)$, $\deg(v)$ is also even. Show that for each vertex $v \in V(G)$, there is a different vertex $u \in V(G)$ such that $|N_G(v) \cap N_G(u)|$ is even.

25. Show that if G is a graph with n vertices and m edges, then $\delta(G) \leq \frac{2m}{n} \leq \Delta(G)$.

26. Find the clique and independence number of:

    (a) K_n;

    (b) P_n;

    (c) C_n;

    (d) Q_n;

    (e) $P_n + C_n$;

    (f) $P_n \times C_n$.

27. Show that the vertices of Q_n can be partitioned into two independent sets.

28. Show that:

    (a) $\alpha(G + H) = \alpha(G) + \alpha(H)$;

    (b) $\alpha(G \times H) = \max\big(\alpha(G), \alpha(H)\big)$;

    (c) $\alpha(G) \geq \frac{|V(G)|}{1+\Delta(G)}$;

    (d) If G has no isolated vertices, then $\alpha(G) \leq \frac{|E(G)|}{\delta(G)}$.

9.2 Classes of Graphs

A *class of graphs* or a *graph property* is a set of graphs closed under isomorphism.

Example. We have already seen various examples of graph properties:

- complete graphs;
- cycles;
- paths;
- connected graphs.

△

Many more classes can be defined by via modifying various graph parameters. For instance, given $k \in \mathbb{N}$, we can define

- the class of graphs of maximum vertex degree at most k;
- the class of graphs of diameter k;
- the class of clique number at least k;
- the class of graphs whose vertices can be partitioned into k independent sets.

9.2.1 Hereditary Classes

A class of graphs X is *hereditary* if it is closed under taking induced subgraphs. That is, if $G \in X$, then $G \setminus v \in X$ for all $v \in V(G)$.

Given any class of graphs X, not necessarily hereditary, the unique minimal (with respect to inclusion) hereditary class containing X is called the *hereditary closure* of X. This class can be obtained by adding to X all induced subgraphs of graphs in X.

An important property of hereditary classes is that they admit *forbidden induced subgraph* characterisations. More precisely, given a set of graphs M, we define $\operatorname{Free}(M)$ to be the set of graphs containing no graphs from M as an induced subgraph, and we say that the graphs in M are *forbidden induced subgraphs* for the class $\operatorname{Free}(M)$, or that the graphs in $\operatorname{Free}(M)$ are *M-free.*

Theorem 9.2.1. *A class of graphs X is hereditary if and only if there is a set of graphs M such that $X = \operatorname{Free}(M)$.*

Proof. Suppose $X = \operatorname{Free}(M)$ for some set of graphs M. Let $G \in X$ and let H be an induced subgraph of G. Then, H is M-free, since otherwise G contains a forbidden graph from M. So, $H \in X$ and hence X is hereditary.

Conversely, if X is hereditary, then $X = \operatorname{Free}(M)$, where M is the set of all graphs not in X. ■

Example. Consider the set X of all complete graphs. Clearly, X is hereditary, and $X = \operatorname{Free}(M)$ with M being the set of all non-complete graphs.

However, we can also see that $X = \operatorname{Free}(\overline{K_2})$, since a graph G is complete if and only if $\overline{K_2} \not< G$. That is, if and only if G has no pair of non-adjacent vertices. △

Given a hereditary class X, a graph G is a *minimal* forbidden induced subgraph if $G \notin X$ and every proper induced subgraph of G belongs to X. We denote the set of all minimal forbidden induced subgraphs for a hereditary class X as $\operatorname{MFIS}(X)$.

Theorem 9.2.2. *For any hereditary class X,*

$$X = \text{Free}\big(\text{MFIS}(X)\big)$$

Moreover, MFIS(X) *is the unique minimal set with this property.*

Proof. Let $G \in X$. Then, by definition, all induced subgraphs of G belong to X, and hence no graph from MFIS(X) is an induced subgrpah of G, since none of them belong to X. So, $G \in \text{Free}\big(\text{MFIS}(X)\big)$, and hence $X \subseteq \text{Free}\big(\text{MFIS}(X)\big)$.

For the reverse containment, let $G \in \text{Free}\big(\text{MFIS}(X)\big)$, and suppose for a contradiction that $G \notin X$. Let H be a minimal induced subgraph of G not in X. Then, $H \in \text{MFIS}(X)$, contradicting that $G \in \text{Free}\big(\text{MFIS}(X)\big)$. So, $G \in X$, and hence $\text{Free}\big(\text{MFIS}(X)\big) \subseteq X$.

This establishes the required equality.

Now, suppose MFIS(X) is not minimal, so $X = \text{Free}(N)$ and $\text{MFIS}(X) \nsubseteq N$ for some set of graphs N. Let $H \in \text{MFIS}(X) \setminus N$. Because $H \in \text{MFIS}(X)$, it is minimal, so any proper induced subgraph of H belongs to $X = \text{Free}(N)$. But, we have $H \notin N$, so $H \in \text{Free}(N) = X = \text{Free}\big(\text{MFIS}(X)\big)$, contradicting that $H \in \text{MFIS}(X)$. ■

Theorem 9.2.3. $\text{Free}(M_1) \subseteq \text{Free}(M_2)$ *if and only if for every graph $G \in M_2$ there is a graph $H \in M_1$ such that H is an induced subgraph of G.*

Proof. Let $\text{Free}(M_1) \subseteq \text{Free}(M_2)$ and suppose for a contradiction that there exists $G \in M_2$ such that all induced subgraphs of G are not in M_1. Then, by definition, we have $G \in \text{Free}(M_1)$, so $G \in \text{Free}(M_2)$, contradicting that $G \in M_2$.

Conversely, suppose that every graph in M_2 contains an induced subgraph from M_1. Suppose for a contradiction that $\text{Free}(M_1) \nsubseteq \text{Free}(M_2)$ and let $G \in \text{Free}(M_1) \setminus \text{Free}(M_2)$. Since $G \notin \text{Free}(M_2)$, G contains an induced subgraph $H \in M_2$. Then, H contains an induced subgraph H' from M_1, so H' is an induced subgraph of G in M_1, contradicting that $G \in \text{Free}(M_1)$. ■

9.2.1.1 Exercises

- Determine which of the following classes of graphs are hereditary:
  - complete graphs;
  - connected graphs;
  - k-regular graphs;
  - chordless paths;
  - chordless cycles;
  - graphs of diameter k;
  - graphs of independence number at most k;
  - graphs whose vertices can be partitioned into two cliques.
- Show that the union of two hereditary classes is itself a hereditary class.
- Show that the intersection of two hereditary classes is itself a hereditary class.
- Show that if $X = \text{Free}(M)$ and $Y = \text{Free}(N)$, then $X \cap Y = \text{Free}(M \cup N)$.
- Show that if $X = \text{Free}(M)$, then $\overline{X} = \text{Free}(\overline{M})$.
- Show that the class $\text{Free}(K_3,\overline{K_3})$ does not contain any graphs with more than 5 vertices.

- Let X be the class of graphs with maximum vertex degree at most 2.
  - Is X hereditary?
  - What can we say about the structure of graphs in X? In particular, what do connected graphs in this class look like?
  - Find MFIS(X).
- Given a class of graphs X, we define $\overline{X}$ to be the class of complements of graphs in X. Determine for which of the following classes X the intersection $X \cap \overline{X}$ is hereditary and for which the intersection is finite:
  - the class of connected graphs;
  - the class of disconnected graphs;
  - the class of graphs with at most $3|V(G)|$ edges.

9.2.2 Hereditary Classes with Small Forbidden Induced Subgraphs

As we have already seen, Free($\overline{K_2}$) is the class of complete graphs. Similarly, Free(K_2) is the class of empty graphs. We describe some other simple classes of graphs defined by small forbidden induced subgraphs.

Lemma 9.2.4. *A graph is P_3-free if and only if it is a disjoint union of cliques. That is, every connected component of G is a clique.*

Proof. If G is a disjoint union of cliques, then any three vertices in a clique must form a $K_3 \neq P_3$ as an induced subgraph.

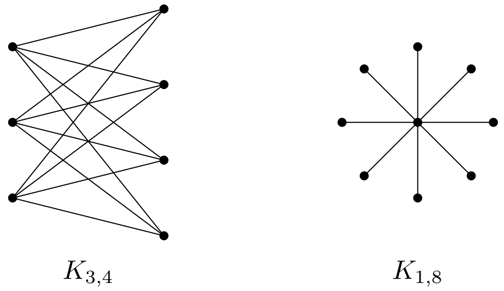
Conversely, let G be P_3-free and suppose that G contains a connected component C that is not a clique. Since C is not a clique, there are two vertices $u,v \in C$ that are not adjacent, but since C is connected, there are paths connecting them. Select a shortest (i.e. chordless) path connecting u and v. This path contains at least 2 edges, so G contains a P_3, contradicting that G is P_3-free. ■

A graph G is *complete multipartite* if the vertices of G can be partitioned into independent sets such that any two vertices belonging to different independent sets are adjacent. Equivalently, G is complete multipartite if and only if $\overline{G}$ is a disjoint union of cliques.

Corollary 9.2.4.1. *A graph is complete multipartite if and only if it is $\overline{P_3}$-free.*

If the number of parts in a complete multipartite graph is two, then the graph is called *complete bipartite.* A complete bipartite graph with two parts of size n and m is denoted by $K_{n,m}$. A complete bipartite graph of the form $K_{1,n}$ is called a *star.*

Example.



$K_{3,4}$ $K_{1,8}$

△

Corollary 9.2.4.2. *A graph is complete bipartite if and only if it is* ($\overline{P_3}$,K_3)*-free.*

Proof. Neither $\overline{P_3}$ nor K_3 are complete bipartite graphs, so a complete bipartite graph must be ($\overline{P_3}$,K_3)-free.

Conversely, if a graph is $\overline{P_3}$-free, then it is complete multipartite by the previous corollary, and if the graph is also K_3-free, then the number of parts cannot be larger than 2, since otherwise a K_3 arises, so the graph is complete bipartite. ■

Corollary 9.2.4.3. *A graph G is* (P_3,K_3)*-free if and only if* $\Delta(G) \leq 1$.

Proof. P_3 and K_3 both contain vertices of degree 2, so if $\Delta(G) \leq 1$, then G is necessarily (P_3,K_3)-free.

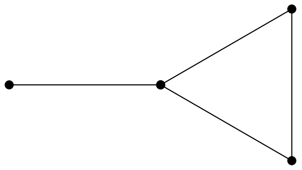
Conversely, if G is P_3-free, then every connected component of G is a clique, and if G is K_3-free, then every component has size at most 2, so every vertex has degree at most 1. ■

Corollary 9.2.4.4. *A graph G is* (P_3,$2K_2$)*-free if and only if its vertex set can be partitioned into two subsets C and I such that C is a clique and I is a set of isolated vertices.*

Proof. Neither P_3 nor $2K_2$ admit such a decomposition, so if G does, it is necessarily (P_3,$2K_2$)-free.

If G is P_3-free, then every connected component of G is a clique, and if G is $2K_2$-free, then at most one of the connected components of G has more than one vertex. This connected component forms C, and the rest of the vertices form I. ■

A *paw* is the unique (up to isomorphism) graph with the degree sequence (1,2,2,3):



Lemma 9.2.5. *Every connected paw-free graph is either K_3-free or $\overline{P_3}$-free.*

Proof. Suppose that a connected paw-free graph G contains a K_3 induced on vertices a,b,c, and let $x \in V(G) \setminus \{a,b,c\}$. Then, x cannot be adjacent to exactly one of a,b,c, since $G[a,b,c,x]$ would be an induced paw.

Suppose that x is adjacent to none of a,b,c. Since G is connected, there exists a path connecting x to the K_3. Without loss of generality, suppose that x is a vertex closest to the K_3 with no neighbours in $\{a,b,c\}$. Then, x is adjacent to a vertex y with neighbours in $\{a,b,c\}$. By the above, y cannot be adjacent to exactly one of the vertices, and x is the closest vertex adjacent to none of the vertices, so y is adjacent to at least 2 of $\{a,b,c\}$. But then, the two neighbours of y along with y and x induce a paw in G.

Denote by V_{ab}, V_{ac}, and V_{bc} the subsets of $V(G)$ consisting of vertices with two neighbours in $\{a,b,c\}$, and denote by V_3 the remaining vertices of G, i.e., those adjacent to all three vertices in $\{a,b,c\}$. Then,

- each of the V_{ij} are independent sets, since if, for example, V_{ab} contains two adjacent vertices x,y, then $G[a,c,x,y]$ is an induced paw;
- any two vertices in different sets are adjacent, since if, for example, $x \in V_{ab} \cup V_3$ is not adjacent to $y \in V_{bc}$, then $G[a,b,x,y]$ is an induced paw;
- $G[V_3]$ is $\overline{P_3}$-free, since if $G[V_3]$ contains a $\overline{P_3}$ induced by x,y,z, then $G[a,x,y,z]$ is an induced paw.

■

9.2.2.1 Exercises

- Characterise the structure of graphs in the class:
  - Free(P_3,K_4);
  - Free(P_3) $\cap$ Free($\overline{P_3}$);
  - Free(P_3) $\cup$ Free($\overline{P_3}$).
- Show that Free(P_3,$K_2 + 2K_1$) = Free(K_2) $\cup$ Free($\overline{K_2}$) $\cup$ Free(P_3,$\overline{K_3}$).
- Show that Free(P_3,$\overline{K_{1,3}}$) = Free($\overline{K_2}$) $\cup$ Free(P_3,K_3).
- Let X be the class of graphs in which the neighbourhood of each vertex is an independent set.
  - Prove that X is hereditary.
  - Determine MFIS(X).
- Show that Free(K_3) $\cup$ Free(P_3) $\cup$ Free($\overline{P_3}$) $\subseteq$ Free(paw). Does the reverse containment hold?

9.2.3 Speed of Graph Properties

There are two ways to count the number of graphs satisfying some given property. We can count *unlabelled* graphs, i.e. up to isomorphism, or *labelled* graphs. In a labelled n-vertex graph, the vertex set is given by $[n]$, and two labelled graphs are distinct if they have different sets of edges. That is, if there is at least one pair of vertices u,$v \in [n]$ that are adjacent in one graph but not in the other.

We are interested in counting the number of n-vertex graphs satisfying a graph property P.

Generally, counting unlabelled graphs is much more difficult, but even for labelled graphs, this question is often highly non-trivial.

We denote by $P(n)$ the set of n-vertex labelled graphs in P. The cardinality $|P(n)|$ considered as a function of n, is called the *speed* of P. Very few exact speeds are known, with only asymptotic values or bounds are available for most graph properties.

9.2.3.1 Exercises

- Find the number of labelled and unlabelled n-vertex graphs:
  - with $n = 20$ and 188 edges;
  - which are complete bipartite (an empty graph counts as complete bipartite);
  - which are complements of chordless paths;
  - which are chordless cycles, $n \geq 3$;
  - in the class Free(P_3,$\overline{P_3}$);
  - in the class Free(K_3,$\overline{K_3}$), $n > 5$.
- Determine the speed of:
  - all graphs;
  - complete graphs;
  - paths;
  - stars;
  - graphs with one edge;

- graphs with k edges;
- Free($\overline{P_3}$,K_3);
- Free(P_3,$2K_2$);
- Free(P_3,K_3).

9.2.4 Acyclic Graphs

A connected acyclic graph is called a *tree.*

Theorem 9.2.6. *The following statements are equivalent for a graph T:*

(*i*) *T is a tree;*

(*ii*) *Any two vertices in T are connected by a unique path;*

(*iii*) *T is minimally connected. That is, T is connected but $T \setminus e$ is disconnected for any $e \in E(T)$;*

(*iv*) *T is maximally acyclic. That is, T is acyclic but $T + uv$ contains a cycle for any non-adjacent vertices u,$v \in V(T)$;*

(*v*) *T is connected and $|E(T)| = |V(T)| - 1$;*

(*vi*) *T is acyclic and $|E(T)| = |V(T)| - 1$.*

Proof. (*i*) → (*ii*): Since trees are connected, any two vertices must be connected by at least one path. If there were two or more, then concatenating one path with another in reverse yields a cycle.

(*ii*) → (*i*): If any two vertices in T are connected by a path, T is connected, and since the path is unique, T is acyclic.

(*ii*) → (*iii*): Suppose T is not minimally connected, so there is an edge $e = ab$ in T such that $T \setminus e$ is connected. That is, a and b are connected in $T \setminus e$ by a path P. But then, P and e are two distinct paths connecting a and b in T.

(*iii*) → (*ii*): Since T is connected, any two vertices in T are connected by a path. This path is unique, since otherwise there would exist an edge e in T belonging to one of the paths but not the other(s) such that $T \setminus e$ is connected.

(*i*) → (*iv*): Let u,v be non-adjacent vertices in T. Since T is connected, there is a path P connecting u and v. Then, this path with the edge uv is a cycle in $T + uv$.

(*iv*) → (*i*): It suffices to show that the T is connected. Suppose otherwise, and let u,v be vertices of T in different connected components. Then, $T + uv$ has no cycles.

(*iii*) → (*v*): We induct on $n := |V(T)|$. For $n = 1$,2 this is obvious. Suppose that $n > 2$ and that the result holds for all smaller graphs.

Let T be a tree on n vertices, and let $e = ab \in E(T)$. By (*iii*), $T \setminus e$ is disconnected. Let T_1 and T_2 be the connected components of $T \setminus e$. T_1 and T_2 are trees, and since each of them have fewer vertices than T, we have by the inductive hypothesis that $|E(T_1)| = |V(T_1)| - 1$ and $|E(T_2)| = |V(T_2)| - 1$. Then,

$$\begin{aligned} |E(T)| &= |E(T_1)| + |E(T_2)| + 1 && (9.1)\\ &= \big(|V(T_1)| - 1\big) + \big(|V(T_2)| - 1\big) + 1 && (9.2)\\ &= \big(|V(T_1)| + |V(T_2)|\big) - 1 && (9.3)\\ &= |V(T)| - 1 && (9.4) \end{aligned}$$

completing the induction.

$(v) \to (i)$: Given T, we construct the graph T' by deleting any edge from any cycles in T, and repeating until T is acyclic. Since deleting edges in a cycle does not disconnect the graph, T' is connected, so $|E(T')| = |V(T')| - 1$.

However, we did not delete any vertices from T when constructing T', so $|V(T)| = |V(T')|$. Then,

$$\begin{aligned} |E(T)| &= |V(T)| - 1 \\ &= |V(T')| - 1 \\ &= |E(T')| \end{aligned}$$

so $T = T'$, and T is acyclic, and is hence a tree.

$(i) \to (vi)$: It suffices to show that $|E(T)| = |V(T)| - 1$. We induct on $n := |V(T)|$. For $n = 1$, the trivial graph on 1 vertex has 0 edges, so the result holds. Suppose $n > 1$ and that the result holds for all smaller graphs.

Let T be a tree on $n+1$ vertices. Since T is acyclic, it contains at least one leaf vertex v. Let T' be the graph obtained from T by deleting v. By the inductive hypothesis, T' has $n-1$ edges. Adding v and its edge back in, we have that T has n edges.

$(vi) \to (i)$: It suffices to show that T is connected. Denote by $T_1, \ldots, T_k$ the connected components of T. Each component is a tree, and hence for each component T_i, we have $|E(T_i)| = |V(T_i)| - 1$. Then,

$$\begin{aligned} |E(T)| &= \sum_{i=1}^{k} |E(T_i)| \\ &= \sum_{i=1}^{k} \big(|V(T_i)| - 1\big) \\ &= \sum_{i=1}^{k} \big(|V(T_i)|\big) - k \\ &= |V(T)| - k \end{aligned}$$

But, $|E(T)| = |V(T)| - 1$, so $k = 1$ and T is connected. ■

Corollary 9.2.6.1. *Every tree T on at least 2 vertices has at least 2 vertices of degree 1.*

Proof. By Euler's handshaking lemma and the previous theorem,

$$\begin{aligned} \sum_{v \in V(T)} \deg(v) &= 2|E(T)| \\ &= 2\big(|V(T)| - 1\big) \\ &= 2|V(T)| - 2 \end{aligned}$$

so at least 2 vertices are of degree 1. ■

A graph with all connected components trees is called a *forest*. That is, a forest is an acyclic graph without any connectedness requirements.

9.2.5 Exercises

- Let $T_1,T_2,\ldots,T_k$, $k > 1$, be pairwise intersecting subtrees of a tree T. Show that there is a vertex of T contained in all the subtrees.
- Let $T_1 = (V,E_1)$ and $T = (V,E_2)$ be two trees on the same vertex set. Show that the graph $G = (V,E_1 \cup E_2)$ has a vertex of degree at most 3.
- We define the *mean degree* of a graph G as

  $$d_{\text{mean}}(G) := \frac{1}{|V(G)|} \sum_{v\in V(G)} \deg(v)$$

  Express the number of vertices of a tree in terms of the mean degree.
- Let T be a tree with n vertices, and suppose that each vertex of T has degree at least 1. Find a formula for the number of vertices of degree 1 in T.
- Show that a tree has exactly two vertices of degree 1 if and only if it is a chordless path.
- Find the hereditary closure of the class of trees.
- Determine which of the following graphs are forests:
  - P_n;
  - C_n;
  - K_n;
  - $K_{1,n}$;
  - paw;
  - $\overline{P_n}$;
  - $\overline{C_n}$;
  - $\overline{K_n}$;
  - $\overline{K_{1,n}}$;
  - co-paw.
- Prove that the class of forests is hereditary.
- Characterise the structure of P_4-free forests and trees.
- Characterise, in terms of minimal forbidden induced subgraphs, the class of:
  - forests whose connected components are stars;
  - forests whose connected components are chordless paths;
- Let $\mathcal{Q}$ be the hereditary closure of the set of all hypercubes. Show that $\mathcal{Q}$ is a superclass of the class of forests.

9.2.6 The Prüfer Code

Every labelled graph can be described by listing its edges, or pairs of vertices. So, every tree on n vertices can be described by listing $2(n-1)$ labels. In fact, we can do better.

Let T be a tree with vertices $\{1,2,\ldots,n\}$. Let a_1 be the smallest leaf in T, and let b_1 be its unique neighbour. By deleting a_1 from T, we obtain a new tree, T_1. Now, let a_2 be the smallest leaf in T_1, and let b_2 be its unique neighbour. By deleting a_2 from T_1, we obtain a new tree T_2.

Proceeding in this way, in $n-2$ steps, we obtain the tree T_{n-2} with a single edge $a_{n-1}b_{n-1}$. Also, note that the sequence $a_1b_1,a_2b_2,\dots,a_{n-1}b_{n-1}$ contains all edges of T. We define the *Prüfer code* of T to be the sequence of vertices

$$P(T) := b_1,b_2,\dots,b_{n-2}$$

We claim that the Prüfer code $P(T)$ is sufficient to compute the numbers $a_1,a_2,\dots,a_{n-1},b_{n-1}$, and therefore, to recover the tree T.

Lemma 9.2.7. *A tree is uniquely recoverable from its Prüfer code.*

Proof. By the construction of a_1, no vertex smaller than a_1 can be a leaf in T.

Also, each vertex v that is not a leaf in T has at least two neighbours in T (by definition). Since at the last step we are left with a single edge and two vertices, at least one of these neighbours must be deleted in say, the ith step. Then, this deleted neighbour is a_i and hence $v = b_i$ is in $P(T)$. So, every vertex that is not a leaf in T appears in $P(T)$.

It follows that a_1 is precisely the smallest label not in $P(T)$, since any smaller labels cannot be leaves, and any non-leaf vertices are in $P(T)$.

Suppose now that we have recovered $a_1,\dots,a_{i-1}$, and consider the tree T_{i-1}. This tree corresponds to the suffix $P(T_{i-1}) = b_i,b_{i+1},\dots,b_{n-2}$ of the code.

Arguing similarly to before, no vertex smaller than a_i can be a leaf in T_{i-1}, and any vertex that is not a leaf in T_{i-1} appears in $P(T_{i-1}) = b_i,b_{i+1},\dots,b_{n-2}$. It follows that a_i is precisely the smallest unused label (i.e. not in $\{a_1,a_2,\dots,a_{i-1}\}$) not in $P(T_{i-1}) = \{b_i,b_{i+1},\dots,b_{n-2}\}$. That is, a_i is the smallest label in $\{1,2,\dots,n\} \setminus \big(\{a_1,\dots,a_{i-1}\} \cup \{b_i,b_{i+1},\dots,b_{n-2}\}\big)$.

This allows us to recursively recover the $a_1,\dots,a_{n-2}$, and the remaining two labels a_{n-1} and b_{n-1} are then the two remaining numbers in $\{1,2,\dots,n\} \setminus \{a_1,\dots,a_{n-2}\}$. ■

Theorem 9.2.8 (Cayley)**.** *The number of labelled trees with n vertices is n^{n-2}.*

Proof. The Prüfer code of a tree with n vertices is a word of length $n-2$ over an alphabet of n letters, so the total number of such words is n^{n-2}. We have already seen above that a tree can be recovered from its code uniquely, so P is injective, but we claim further than the Prüfer code is furthermore a bijection between trees and these words.

The decoding procedure above is applicable to any sequence $b_1,b_2,\dots,b_{n-2}$ of labels from $\{1,2,\dots,n\}$, producing numbers $a_1,a_2,\dots,a_{n-1},b_{n-1}$ from the same set from which we may we construct the graph with edges $a_1b_1,a_2b_2,\dots,a_{n-1}b_{n-1}$. It remains to show that this graph is a tree.

Denote by G_i the graph with vertices $\{1,2,\dots,n\}$ and edges $\{a_ib_i,\dots,a_{n-1}b_{n-1}\}$. The graph G_{n-1} has a single edge and no cycles, so it is a tree. The number a_i is distinct from the following numbers $a_{i+1},\dots,a_{n-1}$ by construction, and also from the numbers $b_{i+1},\dots,b_{n-1}$ by the decoding algorithm, so a_i has degree 1 in G_i, so adding a_i to G_{i+1} to form G_i does not add any cycles. So, G_1 has no cycles. Also, since G_1 has n vertices and $n-1$ edges, it is connected and hence a tree. ■

9.2.7 Exercises

- Determine the number of labelled and unlabelled trees with 5 vertices.
- Determine the number of labelled forests with n vertices and at most 2 connected components.

9.3 Rooted Trees

A *rooted tree* is a tree with a designated vertex v_0 called the *root.* Each vertex v_i is connected to the root by a unique path $v_i,v_{i-1},\ldots,v_1,v_0$, and we say that v_{i-1} the *parent* of v_i, and v_i is the *child* of v_{i-1}.

The *height* of a rooted tree is the distance of from the root to the furthest leaf.

A *binary tree* is a rooted tree in which every vertex has at most two children. A binary tree is *full* if every non-leaf has exactly two children, and a full binary tree is *complete* if all leaves are at the same distance from the root.

9.3.1 Exercises

- Determine the number of leaves in a complete binary tree of heigh h.
- Determine the maximum number of vertices in a binary tree of height h.
- Determine the minimum height of a binary tree with n vertices.
- Determine the number of vertices of degree 3 in a binary tree with t leaves.
- What is the number of labelled rooted trees with n vertices?
- Let T be a binary tree with k leaves. For $i = 1,2,\ldots,k$, let ℓ_i denote the length of the path connecting the ith leaf to the root. Show that

$$\sum_{i=1}^{k} 2^{-\ell_i} \leq 1$$

9.4 Cographs and Modular Decomposition

9.4.1 P_4-free Graphs – Cographs

A graph G is called *complement reducible* or *cograph*, if every induced subgraph of G with at least 2 vertices is either disconnected or the complement to a disconnected graph.

Theorem 9.4.1. *A graph is a cograph if and only if it is P_4-free.*

Proof. Since neither P_4 nor its complement is disconnected, every cograph is P_4-free.

Conversely, let G be P_4-free. We show that G is a cograph by induction on $n := |V(G)|$. ■

Chapter 10

Number Theory

> *"Mathematics is the queen of sciences, and number theory is the queen of mathematics."*
>
> — Carl Friedrich Gauss

Number theory is the branch of mathematics that studies the natural numbers, and by extension, the integers and integer-valued functions.

Unlike in set theory, the set of natural numbers excludes 0 by convention in number theory. This is because many theorems in number theory would otherise require an additional clause stipulating "except 0". To denote the positive elements of a set, we superscript a plus symbol to the set. For instance, we write $\mathbb{N}^+ = \{n \in \mathbb{N} : n > 0\}$ to represent this set of natural numbers that excludes zero. This is the same set as $\mathbb{Z}^+ = \{x \in \mathbb{Z} : x > 0\}$, the positive integers. We also write superscript a star to represent the non-zero elements of a set. For example, $\mathbb{Z}^*$ represents the non-zero integers. The non-zero naturals and the positive naturals coincide, so $\mathbb{N}^+ = \mathbb{N}^*$. Note that this star notation is not completely standardised, and sometimes represents the non-negative elements of a set. It can also denote the unit group of a field. Here, it will always represent the non-zero elements of a set, to match with its usage in abstract algebra and ring theory.

The operations of addition and multiplication are both associative and commutative over $\mathbb{N}$, making it a commutative semiring. Adding in inverses to get $\mathbb{Z}$, we have a commutative ring.

10.1 Divisibility

Apart from the identity elements 0 and 1, natural numbers do not generally have additive or multiplicative inverses. The lack of multiplicative inverses in particular means that we cannot generally divide a natural number b by another natural number a to get another natural number, k. That is, given b and $a \neq 0$, there is no guarantee that we can find some natural k such that $b = ka$.

If such a k does exist, we say that b is *divisible* by a, though, in number theory, we prefer to reverse this and say that *a divides b*, and we write $a|b$. If k does not exist, then a does not divide b, and we write $a \nmid b$.

If $a|b$, then a is a *factor* or *divisor* of b, while b is called the *dividend*. A number greater than 1 whose only factors are 1 and itself is called *prime*. Non-primes greater than 1 are called *composite*. The remaining naturals, 0 and 1, are neither prime nor composite by convention, again allowing us to avoid "except 0 and 1" clauses in theorems that concern prime numbers. In some situations – particularly algebraic ones

– we prefer to classify 1 separately as a *unit.* If a number has exactly two prime factors, then the number is *semiprime.*

The same definition of divisiblity extends readily to the integers, by letting a divide b if there exists an integer k such that $b = ka$. This gives $a|b$ if and only if $|a|$ divides $|b|$.[*]

This definition does have some odd consequences. For instance, -7 is considered a prime with this definition of divisibility. -1 is another number which gets a special classification as a unit, along with 1.

For any a,b,$c \in \mathbb{Z}$ with $a \neq 0$,

1. $a|a$;

   Proof. $a = 1 \cdot a$. ■

2. $a|0$;

   Proof. $0 = 0 \cdot a$. ■

3. (Transitivity) If $a|b$ and $b|c$, then $a|c$;

   Proof. $a|b$, so $b = ka$ for some integer k. $b|c$, so $c = jb$ for some integer j. Then, $c = jb = j(ka) = jka = (jk)a$, and jk is an integer, so $a|c$. ■

4. If $a|b$ and $a|c$, then $a|bn + cm$ for all m,$n \in \mathbb{Z}$;

   Proof. $b = ka$ and $c = ja$. Then, $bn + cm = kan + jam = a(kn + jm)$, and $(kn + jm)$ is integer for all integers n,m, so $a|bn + cm$. ■

5. $a|b$ if and only if $an|bn$ for all $n \in \mathbb{Z}^*$;

   Proof. Suppose $a|b$, so $b = ka$ for some integer k. Then, for all $n \neq 0$, $bn = k(an)$, so $an|bn$, completing the forward direction.

   Now suppose $an|bn$, so $bn = kan$. Then, if $n \neq 0$, we have $b = ka$, so $a|b$, completing the backward direction. ■

6. If a,$b \in \mathbb{Z}$, $b \neq 0$ and $a|b$, then $|a| \leq |b|$;

   Proof. $a|b$, so $b = ka$ and $|b| = |ka| = |k||a|$. Because $b \neq 0$, $k \neq 0$ so $|k| \geq 1$. Suppose $|a| > |b|$. Since $|k| \geq 1$, $|b| = |k||a| > |k||b| > |b|$ implying $|b| > |b|$, which is a contradiction. It follows that $|a| \not> |b|$ so $|a| \leq |b|$. ■

7. (*Euclid's Lemma*) If p is prime, then $p|ab$ if and only if $p|a$ or $p|b$.

   Proof. Difficult. The proof is left for later (§10.1.5), as a corollary of Bézout's identity (§10.1.3). ■

10.1.1 Division Algorithm

If m does not divide n, then trying to divide n things into m equal piles will always leave some things left over. In this case, we use an extended version of division that allows for these remainders. We write $n = qm + r$, where q is the *quotient* of m and n, and r is the *remainder* satisfying $0 \leq r < m$. If $r = 0$, then we just have the same thing as before. This form of integer division with quotients and remainders is known as *Euclidean division.*

The equation $n = qm + r$ is guaranteed to have solutions in q and r by the division algorithm, which gives unique values for both q and r that satisfy $0 \leq r < |m|$ given any integers n and $m \neq 0$.

[*] We can already see a problem with the divisibility symbol: $|a|||b|$ is not particularly readable, even if we make the middle line bigger, $|a|\big||b|$. For this reason, when dealing with absolute values, we generally just say $|a|$ divides $|b|$, rather than writing it symbolically.

For positive m, we can separately write the quotient as $\left\lfloor \frac{n}{m} \right\rfloor$, the floor functions making it clear that the quotient should be an integer. For negative m, we use the ceiling instead, $\left\lceil \frac{n}{m} \right\rceil$.[*] For the remainder, we use *modular notation*, which we will explore in depth later, and we write $n \pmod{m}$.

For non-negative n and m, we can find q and r by setting q to the greatest integer less than $\left|\frac{n}{m}\right|$, and r to $n - qm$.

We can do this more methodically, however, with the *division algorithm*.

Algorithm 1 Division Algorithm

```
procedure DIVISION($n$,$m$)
    if $n < m$ then
        $q \leftarrow 0$
        $r \leftarrow n$
        return $q$,$r$
    else
        $q$,$r \leftarrow$ DIVISION($n - m$,$m$)
        return $q + 1$,$r$
    end if
end procedure
```

If $n < m$, we set $q = 0$ and $r = n$. Otherwise, we compute q_1 and r_1 for $n - m$ and m. Every time we return a value from the call stack, we increment the quotient by 1.

In other words, we repeatedly subtract m from n until we get $n < m$, in which case, we know $q = 0$ and $r = n$. Then, we count up how many times we subtracted m away from n, which is exactly the value of q.

This algorithm is very much unoptimised, as we are dividing by recursively subtracting, but the main thing is that it works, and that we can prove that the algorithm works correctly.

We need to prove that q and r exist, and are unique.

Theorem (Division Algorithm)**.** *Let n,m be integers with $m \neq 0$. Then, there exist unique integers q and r such that $0 \leq r < |m|$ and $n = qm + r$.*

Proof. We prove existence through strong induction on n, considering three cases.

Suppose $n \geq 0$ and $m > 0$. If $n < m$, then $q = 0$ and $r = n$ satisfies $n = qm + r$ and $0 \leq r < m$. Now suppose values for q and r exist for all numbers up to but not including some fixed arbitrary $n \geq m$.

So, if $n \geq m$, then $n - m \geq 0$ and $n - m < n$, so from the induction hypothesis, there exists q' and r such that $n - m = q'm + r$ and $0 \leq r < m$. Then, letting $q = q' + 1$, we have,

$$\begin{aligned} n &= (n - m) + m \\ &= q'm + r + m \\ &= (q' + 1)m + r \\ &= qm + r \end{aligned}$$

so q and r exist when $n \geq 0$ and $m > 0$.

[*] This function is sometimes written as $[x]$ (§34). This division definition is one of the few cases where this "round towards zero" function is used.

Next, suppose $n < 0$ and $m > 0$. Then, there exists q' and r' with $0 \leq r' < m$ such that $-n = q'm + r'$ by the induction hypothesis. If $r' = 0$, we let $q = -q'$ and $r = 0$, so,

$$\begin{aligned} n &= -(-n) \\ &= -(q'm + r') \\ &= qm + r \end{aligned}$$

If $r' \neq 0$, then let $q = -(q' + 1)$ and $r = m - r'$, so,

$$\begin{aligned} n &= -(-n) \\ &= -(q'm + r') \\ &= -(q'm + m - m + r') \\ &= -((q' + 1)m - (m + r')) \\ &= -(-qm - r) \\ &= qm + r \end{aligned}$$

so in both cases, q and r that satisfy the requirements exist.

Finally, suppose $m < 0$. By the inductive hypothesis, there exist q' and r such that $n = q'(-m) + r$, where $0 \leq r < -m$. Let $q = -q'$. Then,

$$\begin{aligned} n &= q'(-m) + r \\ &= (-q')m + r \\ &= qm + r \end{aligned}$$

So, in all cases, q and r exist.

However, we have yet to show that q and r are unique. For uniqueness, suppose $n = qm + r = q'm + r'$, where $0 \leq r < |m|$ and $0 \leq r' < |m|$. Without loss of generality, suppose $r \leq r'$. Then,

$$\begin{aligned} q'm + r' &= qm + r \\ r' - r &= qm - q'm \\ r' - r &= (q - q')m \end{aligned}$$

so $m|r' - r$. Then, there exists some integer k such that $r' - r = k|m|$. If $k = 0$, then $r' = r$, so $q' = q$.

Otherwise, $k \neq 0$. $r' \geq r$, so the left side is non-negative. As $|m|$ is non-negative, if $k \neq 0$, then $k \geq 1$. So, $r' \geq r' - r = k|m| \geq |m|$ so $r' \geq |m|$, contradicting that $r' < |m|$. It follows that $k = 0$, $r' = r$, and $q' = q$, so r and q are unique. ■

10.1.2 Euclidean Algorithm

If a,b,$c \in \mathbb{Z}$, and $c \neq 0$, then c is a *common divisor* of a and b if $c|a$ and $c|b$. Additionally, if at least one of a and b is non-zero, then we can define the *greatest common divisor* or *gcd* of a and b as the largest possible integer which divides both a and b. That is, d is the greatest common divisor of a and b if,

- $d|a$;
- $d|b$;
- If $c|a$ and $c|b$, then $c|d$.

We can extend the division algorithm to find the greatest common divisor of two numbers by applying it repeatedly. This algorithm is the *Euclidean algorithm.*

Algorithm 2 Euclidean Algorithm

```
procedure GCD($n$,$m$)
    $q$,$r$ ←DIVISION($n$,$m$)
    if $r = 0$ then
        return $m$
    else
        $q$,$r$ ←DIVISION($m$,$r$)
    end if
end procedure
```

We call the division algorithm on n and m to find q_0 and r_0 such that $n = q_0 m + r_0$ and $0 \leq r_0 < m$. If $r_0 = 0$, then we know $m|n$ so $\gcd(n,m) = m$. Otherwise, we call the division algorithm on m and r_0 to find q_1 and r_1 such that $m = q_1 r_0 + r_1$ and $0 \leq r_1 < r_0$. If $r_1 = 0$, then $\gcd(n,m) = r_1$. If again, $r_1 \neq 0$, we continue the process, giving the system of equations,

$$\begin{aligned}
n &= q_0 m + r_0 \\
m &= q_1 r_0 + r_1 \\
r_0 &= q_2 r_1 + r_2 \\
r_1 &= q_3 r_2 + r_3 \\
&\vdots \\
r_{n-2} &= q_n r_{n-1} + r_n \\
r_{n-1} &= q_{n+1} r_n + 0
\end{aligned}$$

The last non-zero remainder, r_n, is the greatest common divisor of a and b.

A related notion to common divisors are *common multiples*. For two integers a and b, a common multiple is an integer k such that both $a|k$ and $b|k$. The *least common multiple* or *lcm* of a and b is the smallest possible integer that both a and b divide. That is, m is the least common multiple of a and b if,

- $a|m$;
- $b|m$;
- If $a|c$ and $b|c$, then $m|c$.

If the greatest common divisor of two numbers is 1, then the two numbers are *relatively prime* or *coprime.*

10.1.3 Bézout's Identity

Bézout's identity says that the greatest common divisor of two numbers can be written as a linear combination of those two numbers. That is, there always exists integers x and y such that $\gcd(a,b) = ax + by$. x and y are called the *Bézout coefficients* of a and b.

Theorem (Bézout's Identity)**.** *If a and b are non-zero integers, then there exists integers x and y such that* $\gcd(a,b) = ax + by$.

Proof. Let a and b be non-zero integers. Let $S = \{ax + by : x,y \in \mathbb{Z} \land ax + by > 0\}$. S is non-empty, because it contains at least one of a and $-a$ with $x = \pm 1$ and $y = 0$. Since S is a non-empty set of positive integers, it has a least element by the well-ordering principle (§6.11.4), $d = ax' + by'$.

Using the division algorithm, we are guaranteed that there exists integer q and r such that $a = dq + r$ with $0 \leq r < d$. Then,

$$r = a - qd$$

$$\begin{aligned}&= a - q(ax' + by')\\&= a(1 - qx') - b(qy')\end{aligned}$$

so $r \in S \cup \{0\}$. However, d is the smallest element of S, but $0 \leq r < d$, so $r \notin S$, $r \in \{0\}$, and $r = 0$.

It follows that d is a divisor of a. Similarly, d is also a divisor of b, so d is a common divisor of a and b.

Now, suppose c is also a common divisor of a and b. That is, there exists u and v such that $a = cu$ and $b = cv$. Then,

$$\begin{aligned}d &= ax' + by'\\&= cux' + cvy'\\&= c(ux' + vy')\end{aligned}$$

so $c|d$. Since $d > 0$, $c \leq d$, so d is the greatest common divisor of a and b. ■

With Bézout's identity in mind, we can also define two numbers n and m to be coprime if and only if there exists integers x and y such that $ax + by = 1$.

Note that Bézout's identity doesn't actually give us values for x and y, only guaranteeing that such values exist. We can extend the Euclidean algorithm further into the aptly named *extended Euclidean algorithm* to find these values for x and y.

10.1.4 Extended Euclidean Algorithm

We can find x and y through a series of backsubstitutions through the quotients and remainders, which is generally how it is done by hand, but we also have an algorithm to do this more efficiently.

In the standard Euclidean algorithm, only the remainders are kept after each iteration, while the quotients are discarded. In the extended Euclidean algorithn, we use the quotients to generate two other sequences which give us the Bézout coefficients.

With a and b as inputs, the standard Euclidean algorithm computes a sequence, $q_1, \ldots, q_k$ of quotients, and a sequence, $r_0, \ldots, r_{k+1}$ of remainders such that,

$$\begin{aligned}r_0 &= a\\r_1 &= b\\&\vdots\\r_{n+1} &= r_{n-1} - q_n r_n\\&\vdots\end{aligned}$$

with the constraint $0 \leq r_{n+1} < |r_n|$ uniquely defining q_n and r_{n+1} from r_{n-1} and r_n.

In the extended Euclidean algorithm, we add two additional sequences,

$$\begin{aligned}r_0 &= a & s_0 &= 0 & t_0 &= 1\\r_1 &= b & s_1 &= 1 & t_1 &= 0\\&\vdots & &\vdots & &\vdots\\r_{n+1} &= r_{n-1} - q_n r_n & s_{n+1} &= s_{n-1} - q_n s_n & t_{n+1} &= t_{n-1} - q_n t_n\end{aligned}$$

again, with the constraint $0 \leq r_{n+1} < |r_n|$.

These sequences similarly stop when $r_{n+1} = 0$ and gives r_k as the greatest common divisor, as before. However, the values of s_n and t_n can also be returned to give the Bézout coefficients. That is, $\gcd(a,b) = as_n + bt_n$.

Additionally, we can find the quotients of a and b with these coefficients with $s_{n+1} = \pm\frac{b}{\gcd(a,b)}$ and $t_{n+1} = \pm\frac{a}{\gcd(a,b)}$.

Furthermore, if $a,b > 0$ and $\gcd(a,b) \neq \min(a,b)$, then,

$$|s_i| \leq \left\lfloor \frac{b}{2\gcd(a,b)} \right\rfloor \qquad |t_i| \leq \left\lfloor \frac{a}{2\gcd(a,b)} \right\rfloor$$

for all $0 \leq i < n$, implying that the coefficients the algorithm returns are the minimal pair of coefficients.

For a more recursive implementation: if $a = 0$, $\gcd(a,b) = b$ with $x = 0$ and $y = 1$. Otherwise, if $a > 0$, let $b = qa + r$, with $0 \leq r < m$, and recursively call the algorithm on r and a toget x' and y' such that $rx' + by' = \gcd(r,b) = \gcd(a,b)$. Then, substituting $r = b - qa$ gives $\gcd(a,b) = x'(b - qm) + y'a = (y' - x'q)m + x'a$, so $x = y' - x'q$ and $y = x'$.

10.1.5 Euclid's Lemma

Theorem (Euclid's Lemma)**.** *Let p be prime. Suppose $p|ab$. Then, $p|a$ or $p|b$.*

Some equivalent formulations of the lemma are as follows:

- If $p \nmid a$ and $p \nmid b$, then $p \nmid ab$.
- If $p \nmid a$ and $p|ab$, then $p|b$.

We instead prove a generalisation of Euclid's lemma, from which the original immediately follows.

Theorem (Euclid's Lemma)**.** *If $n|ab$ and n is coprime with a, then $n|b$.*

Proof. Suppose $n|ab$ and that n and a are coprime. Then, by Bézout's identity, there exists integers x and y such that $nx + ay = 1$, so $nxb + ayb = b$. The first term is divisible by n, while the second is divisible by ab, which is assumed to be divisible by n. It follows that their sum, b, is also divisible by n.

If n is prime, then either $n|a$ or n and a are coprime, so $n|b$, giving the original lemma. ■

Bézout's identity was not known at Euclid's time. The original proof is rather difficult to read, partially due to the lack of modern algebraic notation, the lemma being proved by comparing ratios of lengths. However, the lemma can also be proven just by using the Euclidean algorithm and strong induction. We give a shorter proof using ideals in a later section (§10.3.3).

Euclid's lemma also shows that $\mathbb{Z}_p$ has no zero divisors: non-zero numbers a and b such that $ab = 0$. In $\mathbb{Z}_m$ with non-prime m, then m is composite,[*] so two factors a and b exist such that $ab = m \equiv 0$.

10.2 Modular Arithmetic

Modular arithmetic is a system of a arithmetic for integers that is restricted to remainders under division by some fixed integer called the *modulus*. One familiar example is in timekeeping, where the numbering of hours wraps back around once you go past 12. In mathematics, we would describe this as arithmetic *modulo 12*.

From the division algorithm, for every pair of integers n and $m \neq 0$, there is a unique remainder r with $0 \leq r < |m|$, and $n = qm + r$ for some q. We can define an equivalence relation (§4.4.9) called *congruence* using these remainders.

For some fixed modulus, two numbers, n and n', are *congruent* they have the same remainder when divided by the modulus, and we write $n \equiv_m n'$ or $n \equiv n' \pmod{m}$, where m is the modulus. Or

[*] m can technically be a unit, i.e. 1, if the ring is the trivial ring, but then the multiplicative and additive identities coincide, so no non-zero elements exist, so zero divisors also do not exist for the trivial ring.

equivalently, $n \equiv n' \pmod{m}$ if and only if there exists some integer k such that $n = n' + km$ – they differ only by exact integer multiples of the modulus.

Congruence is a very important equivalence relation, and thus has some additional terminology to distinguish it from others. The equivalence class of an integer n under congruence modulo m is called the *congruence class*, *residue class*, or *residue* of that integer, and is written $[n]_m$, or $\bar{n}_m$. From the definition of congruence, we see $[n]_m = \{\ldots,n-2m,n-m,n,n+m,n+2m,\ldots\}$. Furthermore, the sets $[0]_m,[1]_m,\ldots,[m-1]_m$ partition the integers, and the set of residue classes, $\{[0]_m,[1]_m,\ldots,[m-1]_m\}$, defines the *integers mod m*, and is denoted* $\mathbb{Z}_m$ or $\mathbb{Z}/m\mathbb{Z}$.

$\mathbb{Z}_m$ behaves very similarly to $\mathbb{Z}$, with addition, subtraction, and multiplication all being well-defined, as we will soon show. This makes $\mathbb{Z}_m$ a commutative ring, but unlike $\mathbb{Z}$, $\mathbb{Z}_m$ is a finite ring. As we will see, when m is prime, division is also well-defined, so $\mathbb{Z}_p$ is a *finite* or *Galois field* for any prime p. This also means that the set $\mathbb{Z}_m$ with addition always has an (abelian) group structure, while $\mathbb{Z}_m$ with multiplication is only a group for prime m.

We define arithmetic operations on residue classes in $\mathbb{Z}_m$ just as we defined arithmetic operations on integers – as equivalence classes on ordered pairs of naturals. Given residue classes $[x]_m$ and $[y]_m$, we define $[x]_m + [y]_m = [x+y]_m$, where the addition on the right side is normal integer addition in $\mathbb{Z}$. Because every element of a residue class is equivalent in every way we care about, we tend to just use a single element in the class to *represent* the entire class. Above, x and y are *representatives* of their respective residue classes. Generally, we select the representative to be in the range $0 \leq x \leq m$ (only one element per class lies within that range), making it clear that we are using remainders modulo m. This allows us to write things like $10 + 4 = 2 \pmod{12}$. Another notation that is common in group theory is $10 +_{12} 4 = 2$, as this moves the "modulo" into the operation itself, so we're operating on ordinary numbers. That is, instead of "regular" addition on residue classes, $[a] + [b] \pmod{m}$, it's "modular" addition on regular numbers, $a +_m b$. These structures are, however, isomorphic, so for number theory, the distinction is immaterial.

But first, we should verify that this definition of addition is indeed well-defined. In particular, the definition should work regardless of which representative is picked.

To prove this, we start with an alternative characterisation of congruence.

Lemma 10.2.1. *Let $x,y \in \mathbb{Z}$ and $m \in \mathbb{N}^+$. Then, $x \equiv y \pmod{m}$ if and only if $m|x-y$.*

Proof. Suppose $x \equiv y \pmod{m}$, so x and y have the same remainder under divison by m. That is, $x = qm+r$ and $y = sm+r$ for some integers q and s, and $0 \leq r < m$. Then, $x-y = (q-s)m+(r-r) = (q-s)m$, so $m|x-y$, completing the forward direction.

Now, suppose $m|x-y$, so $x-y = km$ for some integer k. From the division algorithm, we can write $x = qm+r$ and $y = sm+t$, where $0 \leq r < m$ and $0 \leq t < m$. Then, $x-y = (q-s)m+(r-t) = km$, so $r-t = 0$ and $r = t$, so $x \equiv y \pmod{m}$, completing the backward direction. ∎

Theorem 10.2.2. *If $x \equiv x' \pmod{m}$ and $y \equiv y' \pmod{m}$, then $x+y \equiv x'+y' \pmod{m}$.*

Proof. From the previous lemma, $m|x-x'$ and $m|y-y'$, so $m|(x-x')+(y-y')$, which rearranges to $m|(x+y)-(x'+y')$. Applying the previous lemma in reverse, we have $x+y \equiv x'+y' \pmod{m}$, as required. ∎

* The former notation can be confusing, because another number system in number theory, the p-adic numbers, are also denoted $\mathbb{Z}_p$, where p is an integer.

The latter notation also nice because it suggests the structure of the set itself: it is the set of integers, divided by a multiple of the integers. This notation also has connections to quotient groups and quotient rings in abstract algebra (§12.9.1).

However, the former notation is also much shorter to write, and is equally, if not even more, popular than the latter – at least in number theory.

We similarly define $-[x]_m = [-x]_m$ and $[x]_m \cdot [y]_m = [x \cdot y]_m$. Similar arguments to the ones above show that these definitions also give well-defined operations on residue classes.

All the usual properties of addition, subtraction, and multiplication are inherited from $\mathbb{Z}$; commutativity and associativity of addition and multiplication, distributivity, etc., all apply, making $\mathbb{Z}_m$ a commutative ring.

Because $[x]_m + [y]_m = [x + y]_m$ and $[x]_m \cdot [y]_m = [x \cdot y]_m$ for all x and y, the remainder operation, $x \mapsto x \pmod{m}$ is a homomorphism (§12.4) from $\mathbb{Z}$ to $\mathbb{Z}_m$. This means that it doesn't matter when we perform modulo operations when converting an expression in $\mathbb{Z}$ to the corresponding equation in $\mathbb{Z}_m$, so, instead of doing $(185 + 512) \cdot (23 + 16) + 256 = 27\,439 \equiv 1 \pmod{2}$, we can apply the modulus before addition, instead giving $(1 + 0) \cdot (1 + 0) + 0 \pmod{2} \equiv 1 \pmod{2}$, which you may find easier.

This property is what defines a general congruence relation: a congruence relation is any equivalence relation over an algebraic structure that is compatible with the structure, in the sense that the operations on that structure when applied to equivalent elements yield equivalent elements. So, the modulo congruence relation satisfies:

- $a \equiv a \pmod{m}$ (reflexivity);
- $a \equiv b \pmod{m}$ if and only if $b \equiv a \pmod{m}$ (symmetry);
- If $a \equiv b \pmod{m}$ and $b \equiv c \pmod{m}$, then $a \equiv c \pmod{m}$ (transitivity);
- If $a \equiv b \pmod{m}$ and $c \equiv d \pmod{m}$, then $a + c \equiv b + d \pmod{m}$ (compatibility with ring addition);
- If $a \equiv b \pmod{m}$, then $ac \equiv bc \pmod{m}$ for $c \in \mathbb{Z}$ (compatibility with ring multiplication);
- If $a \equiv b \pmod{m}$, then $a^n \equiv b^n \pmod{m}$ for $n \in \mathbb{Z}$;
- $a \equiv 0 \pmod{m}$ if and only if $m|a$.

Proving these properties is left as an exercise for the reader.

With these operations, we can solve congruence equations. The set $\{n : 0 \leq n < m - 1\}$ is called the set of *least residues modulo n*. Solving a congruence equation means finding the least residues that satisfy that equation.

For example, the equation $x \equiv 23 \pmod{3}$ is solved by $x \equiv 2 \pmod{3}$, because 23 and 2 are equivalent modulo 3. We can also add or subtract any multiple of m from either side of an equivalence, because that doesn't change the remainder modulo m.

Addition and subtraction are also easy to deal with because the operations are inverse and are both are well-defined over $\mathbb{Z}_m$. For instance, $x + 6 \equiv 12 \pmod{4}$ can be reduced to $x \equiv 6 \pmod{4}$ by subtracting 6 from both sides, so we see $x \equiv 2 \pmod{4}$ solves the equation.

Now, recall that division is not always well defined over $\mathbb{Z}_m$. This makes multiplication trickier to deal with, because solutions may not exist, or multiple solutions exist. For example, $10x \equiv 5 \pmod{12}$ does not have any solutions, while the equation $2x \equiv 10 \pmod{14}$ has solutions $x \equiv 5$ and $x \equiv 12$ ($2 \cdot 12 = 24 \equiv 10 \pmod{14}$). Note that simply dividing both sides of the original equation by 2 misses the $x \equiv 12$ solution.

We can actually tell which will be the case using the greatest common divisor of the modulus and the coefficient of the desired variable.

Let a,$b \in \mathbb{Z}$, $m \in \mathbb{N}^+$ and let $d = \gcd(a,m)$. Then,

- If $d \nmid b$, then $ax \equiv b \pmod{m}$ has no solutions.
- If $d|b$, then $ax \equiv b \pmod{m}$ has exactly d solutions in the set of least residues modulo m.

and the rule for cancelling is as follows:

- If $ka \equiv kb \pmod{m}$ and $\gcd(k,m) = d$, then $a \equiv b \left(\operatorname{mod} \frac{m}{d}\right)$.

So, for the equation, $2x \equiv 10 \pmod{14}$, we find $\gcd(2,14) = 2$, so we divide everything, including the modulus, by 2 to get $x \equiv 5 \pmod{7}$, which is equivalent to $x \equiv 5,12 \pmod{14}$.

If we know $ax \equiv b \pmod{m}$ has solutions, and we have already reduced it down by cancelling, we can also find the *multiplicative inverse* of a. The multiplicative inverse of a modulo m is an integer p that satisfies $ap \equiv 1 \pmod{m}$. As in abstract algebra, we also write a^{-1} for the multiplicative inverse of a.

The multiplicative inverse exists if and only if a and m are coprime, or equivalently, $\gcd(a,m) = 1$. This is because, if there exists some $d > 1$ that divides both a and m, then it will continue to divide aa' and m for any $a' \neq 0$. So in particular, xx' cannot be congruent to 1 modulo m since $qm + 1$ and m don't share any common factors for any value of q.

The set of residue classes $[x]_m$ where $\gcd(x,m) = 1$ is writen as $\mathbb{Z}_m^*$. For prime p, $\mathbb{Z}_p^*$ includes all non-zero elements of $\mathbb{Z}_p$ since $\gcd(x,p) = 1$ for all x not equal to 0 or a multiple of p. This means that $\mathbb{Z}_p$ is a field. Specifically, because it is finite in cardinality, it is a finite or Galois field. $\mathbb{Z}_p$ is not, however, an ordered field. Because numbers wrap around the modulus, there is no way to define an ordering relation $\prec$ such that $\prec$ has translational and scaling invariance.

Unlike addition, subtraction, and multiplication, division in $\mathbb{Z}_p$ doesn't map directly from a correpsonding operation in $\mathbb{Z}$ or $\mathbb{Q}$ the way addition, subtraction, and multiplication do. For example, $4 \cdot 4 = 16 \equiv 1 \pmod{5}$, so $4^{-1} = 4$, so we could write $\frac{3}{4} = 3 \cdot 4^{-1} \equiv 3 \cdot 4 \equiv 2 \pmod{5}$. However, if we compute $\frac{3}{4}$ first, in $\mathbb{Q}$, for example, there is no natural mapping from $\mathbb{Q}$ to $\mathbb{Z}_5$ that sends $\frac{3}{4}$ to 2. We could try define a function,

$$f\left(\frac{p}{q}\right) = pq^{-1} \pmod{5}$$

which would work for many rationals, including $\frac{3}{4}$, but we run into problems with fractions like $\frac{3}{5}$, where the denominator does not have an inverse in $\mathbb{Z}_5$.

When solving congruence equations, we can generally reduce equations down to situations where the coefficient on the desired variable and the modulus are coprime, as $\gcd(a,m)|b$ whenever the solution has equations.

We can then find these inverses using Bézout's identity.

If a and m are coprime, then there exists integers p and q such that

$$\begin{aligned} ap + mq &= 1 \\ ap &= 1 - mq \\ ap &\equiv 1 \pmod{m} \end{aligned}$$

so a and p are multiplicative inverses.

We can then multiply both sides of the equation, $ax \equiv b \pmod{m}$ by p. Because a and p are multiplicative inverses, they are equivalent to 1, giving $x \equiv bp \pmod{m}$.

Example. Solve $75x \equiv 12 \pmod{237}$.

$\gcd(75,237) = 3$, so we can cancel the equation down to $25x \equiv 4 \pmod{79}$. Now, we use the Euclidean algorithm to find $\gcd(25,79)$,

$$\begin{aligned} 79 &= 3 \cdot 25 + 4 \\ 25 &= 6 \cdot 4 + 1 \\ 4 &= 4 \cdot 1 + 0 \end{aligned}$$

so $\gcd(25,79) = 1$ and there is a unique solution modulo 79.

Next, we use backsubstitution to find the Bézout coefficients.

$$\begin{aligned}
25 - 6(4) &= 1 \\
25 - 6(79 - 3(25)) &= 1 \\
25 - 6(79) + 18(25) &= 1 \\
19(25) - 6(79) &= 1 \\
19(25) &= 1 + 6(79) \\
19(25) &\equiv 1 \pmod{79}
\end{aligned}$$

so 19 and 25 are multiplicative inverses modulo 79. So,

$$\begin{aligned}
25x &\equiv 4 \pmod{79} \\
19 \cdot 25x &\equiv 19 \cdot 4 \pmod{79} \\
x &\equiv 76 \pmod{79}
\end{aligned}$$

△

10.2.1 Chinese Remainder Theorem

The *Chinese remainder theorem* states that, if you know the remainders from the Euclidean division of an integer n by several other pairwise coprime integers, then you can determine uniquely the remainder of the division of n by the product of those integers.

For example, if we know that the remainder of n divided by 3 is 2, the remainder of n divided by 5 is 3, and the remainder of n divided by 7 is 2, since 3, 5 and 7 are pairwise coprime, then even without knowing the value of n, we can determine that the remainder of n divided by $3 \cdot 5 \cdot 7 = 105$ is 23. More importantly, this tells us that if n is a natural less than 105, then n must be exactly 23.

Theorem (Chinese Remainder Theorem)**.** *Let $\{m_i\}_{i=1}^k$ be integers greater than* 1 *such that* $\gcd(m_i,m_k) = 1$ *for all i,j, $i \neq j$. That is, the m_i are pairwise coprime.*

If $\{n_i\}_{i=1}^k$ are integers such that $0 \leq n_i < m_i$ for all $1 \leq i \leq k$, then there is a unique integer x such that the remainder of the Euclidean division of x by m_i is n_i for all $1 \leq i \leq k$ and $0 \leq x < \prod_{i=1}^k m_i$.

This statement is helpful because it tells us what a solution has to be whenever it is smaller than the product of the moduli.

Example. Solve $x^3 \equiv 53 \pmod{120}$.

We factor 120 into $3 \cdot 5 \cdot 8$, so,

$$\begin{array}{ccc}
x^3 \equiv 53 \pmod{3} & x^3 \equiv 53 \pmod{5} & x^3 \equiv 53 \pmod{8} \\
x^3 \equiv 2 \pmod{3} & x^3 \equiv 3 \pmod{5} & x^3 \equiv 5 \pmod{8} \\
x \equiv 2 \pmod{3} & x \equiv 2 \pmod{5} & x \equiv 5 \pmod{8}
\end{array}$$

The first two congruences together give $x \equiv 2 \pmod{15}$, so $x - 2 \equiv 0 \pmod{15}$ and $x - 2 \equiv 3 \pmod{8}$, so we're looking for a number of the form $8n + 3$ that is divisible by 15, vastly reducing the search space. Quickly plugging in values for n, we have, 3,11,19,27,35,43,51,59,67,75, so 75 is our number. It follows that $x \equiv 77 \pmod{120}$. △

The Chinese remainder theorem can also be stated in terms of congruence relations, which is the form that we will prove.

Theorem (Chinese Remainder Theorem)**.** *Let $\{m_i\}_{i=1}^k$ be pairwise coprime integers greater than* 1*, and let $N = \prod_{i=1}^k m_i$. If $\{n_i\}_{i=1}^k$ are any integers, then the system,*

$$\begin{aligned} x &\equiv n_1 \pmod{m_1} \\ x &\equiv n_2 \pmod{m_2} \\ &\vdots \\ x &\equiv n_k \pmod{m_k} \end{aligned}$$

has solutions, and solutions are congruent modulo N.

Proof. Suppose x and y are solutions to the congruences. As x and y give the same remainder when divided by the m_i, their difference, $x - y$ is a multiple of each m_i. Because the m_i are pairwise coprime, their product, N, also divides $x - y$, so x and y are congruent modulo N. If x and y are non-negative and less than N, then they can only be congruent modulo N if $x = y$, so this solution is unique.

The map $x \pmod{N} \mapsto (x \pmod{m_1}, x \pmod{m_2}, \ldots, x \pmod{m_k})$ maps the congruence classes of division modulo N to sequences of congruence classes modulo m_i. Because solutions are unique up to congruence modulo N, this map is injective. Furthermore, because the domain and codomain of the map have the same cardinality, the map is also surjective, proving existence of solutions. ■

The proof above shows the existence of solutions non-constructively, and is included here because it is short. Constructive proofs that provide algorithms to calculate x do exist, but the algorithms are generally rather complicated to do by hand. One such algorithm is provided at the end of this section.

In abstract algebra, the theorem is commonly stated in terms of rings and morphisms.

Theorem (Chinese Remainder Theorem)**.** *The map,*

$$x \pmod{N} \mapsto (x \pmod{n_1}, x \pmod{n_2}, \ldots x \pmod{n_k})$$

is a ring isomorphism, between the ring of integers modulo N, and the direct product of the rings of integer modulo n_i. That is,

$$\mathbb{Z}/N\mathbb{Z} \cong \mathbb{Z}/n_1\mathbb{Z} \times \mathbb{Z}/n_2\mathbb{Z} \times \cdots \times \mathbb{Z}/n_k\mathbb{Z}$$

For example, 3 and 4 are coprime, so every integer $n \in [0,11]$ can be represented uniquely as pairs of numbers (n_1, n_2), where $n_1 = n \pmod 3$ and $n_2 = n \pmod 4$. We can give these in a table as follows:

n	n_1	n_2
0	0	0
1	1	1
2	2	2
3	0	3
4	1	0
5	2	1
6	0	2
7	1	3
8	2	0
9	0	1
10	1	2
11	2	3

This gives a factorisation of $\mathbb{Z}_{12}$ as $\mathbb{Z}_3 \times \mathbb{Z}_4$. This doesn't just mean we can represent elements of $\mathbb{Z}_{12}$ as elements in $\mathbb{Z}_3 \times \mathbb{Z}_4$ – the Chinese remainder theorem states that this factorisation is an isomorphism, so it is compatible with the ring structure. So, we can do arithmetic on these pairs in $\mathbb{Z}_3 \times \mathbb{Z}_4$ and get the same answers as if we did the arithmetic in $\mathbb{Z}_12$. For example, $7 \in \mathbb{Z}_{12}$ is represented as $(1,3) \in \mathbb{Z}_3 \times \mathbb{Z}_4$, and 5 is represented by (2,1). So, to multiply 7 by 5 in $\mathbb{Z}_{12}$, we could instead multiply (1,3) by (2,1) componentwise in $\mathbb{Z}_3 \times \mathbb{Z}_4$, giving $(1 \cdot 2{,}3 \cdot 1) = (2{,}3)$, which we see represents 11 in $\mathbb{Z}_12$, matching the expected result of $7 \cdot 5 = 35 \equiv 11 \pmod{12}$.

This formulation of the theorem is very powerful, particularly in computer science, as it allows us to transform equations with a very large modulus into exceedingly simple systems of equations with much smaller moduli.

10.2.1.1 Constructive Proof

For a constructive proof, we first ease in with a proof of the two equation variant of the congruence equation form.

Theorem (Chinese Remainder Theorem)**.** *Let m_1 and m_2 be integers greater than* 1 *such that* $\gcd(m_1,m_2) = 1$*. That is, the m_1 and m_2 are coprime. Then, any pair of equations,*

$$\begin{aligned} x &= n_1 \pmod{m_1} \\ x &= n_2 \pmod{m_2} \end{aligned}$$

has a unique solution x with $0 \leq x \leq m_1 m_2$.

Proof. We observe that, if $a|b$, then $(x \pmod{b}) \pmod{a} \equiv x \pmod{a}$, because $x \equiv x - qb \pmod{b}$ for some integer q, so $(x \pmod{b}) \pmod{a} \equiv (x \pmod{a}) - (qb \pmod{a}) \equiv x \pmod{a}$, since any multiple of b is also a multiple of a, so $qb \equiv 0 \pmod{a}$ for all q.

Now, since m_1 and m_2 are coprime, Bézout's identity allows us to find multiplicative inverses for $m_1 \pmod{m_2}$, and $m_2 \pmod{m_1}$. So, we have m_1' and m_2' such that $m_1' m_1 \equiv 1 \pmod{m_2}$ and $m_2' m_2 \equiv 1 \pmod{m_1}$.

We claim that the solution is given by $n = (n_1 m_2' m_2 + n_2 m_1' m_1) \pmod{m_1 m_2}$.

We verify that this n satisfies the first equation as follows:

$$\begin{aligned} n \pmod{m_1} &= ((n_1 m_2' m_2 + n_2 m_1' m_1) \pmod{m_1 m_2}) \pmod{m_1} \\ &= (n_1 m_2' m_2 + n_2 m_1' m_1) \pmod{m_1} \\ &= (n_1 \cdot 1 + n_2 m_1' \cdot 0) \pmod{m_1} \\ &= n_1 \pmod{m_1} \\ &= n_1 \end{aligned}$$

and, through an almost identical calculation, we verify that $n \pmod{m_2} = n_2$.

This shows existence of solutions.

We have just given an algorithm for generating a solution for any pair, so we know that our function is surjective. There are also exactly $m_1 m_2$ possible choices for (n_1,n_2) and for solutions n, so, if some pair has more than one solution, then another must have none, so it follows that our function must be injective, and is therefore bijective, so solutions are unique. ■

The main idea is that $m_2' m_2$ acts like $1 \pmod{m_1}$, but like $0 \pmod{m_2}$, and vice versa for $m_1' m_1$, so we have the "basis vectors" (1,0) and (0,1), and we can then get arbitrary solutions for (n_1,n_2) just by adding up sufficient copies of each basis vector.

We can now constructively prove the general congruence equation form of the theorem.

Theorem (Chinese Remainder Theorem)**.** *Let $\{m_i\}_{i=1}^k$ be integers greater than* 1 *such that* $\gcd(m_i,m_k) = 1$ *for all i,j, $i \neq j$. That is, the m_i are pairwise coprime. Then, any system of equations,*

$$x = n_i \pmod{m_i}$$

has a unique solution x with $0 \leq x \leq \prod_{i=1}^k m_i$.

Proof. The solution can be computed using the formula,

$$x = \left(\sum_{i=1}^k n_i \prod_{i \neq j} \left(m_j^{-1} \pmod{m_i} \right) m_j \right) \left(\text{mod} \prod_{i=1}^k m_j \right)$$

For any fixed l,

$$\begin{aligned} x \pmod{m_l} &= \left(\sum_{i=1}^k n_i \prod_{i \neq j} \left(m_j^{-1} \pmod{m_i} \right) m_j \right) \left(\text{mod} \prod_{i=1}^k m_j \right) \\ &= \left(\sum_{i=1}^k n_i \prod_{i \neq j} \left(m_j^{-1} \pmod{m_i} \right) m_j \right) (\text{mod}\, m_l) \\ &= \left(n_k \cdot 1 + \sum_{i \neq l} (n_i \cdot 0) \right) \\ &= n_k \end{aligned}$$

For uniqueness, the same argument from before still applies. ■

10.2.2 Fermat's Little Theorem

To deal with powers, we have another theorem to help us.

Theorem (Fermat's Little Theorem)**.** *If p is prime and $p \nmid a$, then $a^{p-1} \equiv 1 \pmod{p}$, or equivalently, $a^p \equiv a \pmod{p}$.*

Proof. Consider the set of least residues $G = \{1,2,...,p-1\}$ under multiplication modulo p ($\times_p$). Multiplication is closed and associative over $\mathbb{Z}/p\mathbb{Z}$, 1 is the identity element, and Bézout's identity guarantees that every element has an inverse as p is prime, so $(G,\times_p)$ is a group. Let $a \in G$, $k = |a|$ and $H = \langle a \rangle = \{1,a,a^2...,a^{k-1}\}$. $(H,\times_p)$ forms a subgroup of G of order k. By Lagrange's theorem (§12.4.10), k divides $|G| = p-1$, so $p-1 = nk$ for some $n \in \mathbb{Z}^+$. Thus, $a^{p-1} = a^{nk} = (a^k)^n \equiv 1^n = 1 \equiv 1 \,(\text{mod}\, p)$. ■

Fermat's little theorem can help us find multiplicative inverses with powers. If a is a positive natural, p is prime, and $p \nmid a$, then, a^{p-2} is a multiplicative inverse of a modulo p;

Example. What is the remainder when 2^{1000} is divided by 13?

$2^{12} = 1 \pmod{13}$, and $12 \cdot 83 = 996$, so,

$$\begin{aligned} 2^{1000} &= 2^{996} \cdot 2^4 \\ &= (2^{12})^{83} \cdot 2^4 \\ &\equiv 1 \cdot 2^4 \pmod{13} \\ &= 16 \\ &\equiv 3 \pmod{13} \end{aligned}$$

△

Example. Solve $x^{103} = 4 \pmod{11}$.

$x^{11} \equiv x \pmod{11}$ and $11 \cdot 9 = 99$, so,

$$\begin{aligned} x^{103} &= (x^{11})^9 \cdot x^4 \\ &\equiv x^9 \cdot x^4 \pmod{11} \\ &= x^{13} \\ &= x^{11} \cdot x^2 \\ &\equiv x \cdot x^2 \pmod{11} \\ &= x^3 \end{aligned}$$

so now, we need to solve $x^3 \equiv 4 \pmod{11}$, which we can do by inspection, giving $x \equiv 5 \pmod{11}$ as the solution. △

10.2.3 Euler's Theorem

We can extend this further with the use of *Euler's totient function*, $\phi(n)$. $\phi(n)$ counts the number of naturals up to n that are coprime to n. That is, $\phi(n) = |\mathbb{Z}_n^*|$. These coprime numbers are called the *totatives* of n.

There are several formulae for computing $\phi(n)$, but we first prove several lemmata.

Lemma (Multiplicativity of $\phi(n)$)**.** *If* $\gcd(m,n) = 1$*, then* $\phi(m)\phi(n) = \phi(mn)$

Proof. Let A, B and C be the sets of positive integers coprime to and less than m, n, and mn, respectively, so $|A| = \phi(m)$, $|B| = \phi(n)$, and $|c| = \phi(mn)$. By the Chinese remainder theorem, there is a bijection between $A \times B$ and C. ■

Lemma (Prime arguments of $\phi(n)$)**.** *If p is prime and $k \geq 1$, then,*

$$\begin{aligned} \phi(p^k) &= p^k - p^{k-1} \\ &= p^{k-1}(p-1) \\ &= p\left(1 - \frac{1}{p}\right) \end{aligned}$$

Proof. Because p is prime, $\gcd(p^k,m)$ can only be powers of p, $1,p,p^2,\ldots,^k$, with $\gcd(p^k,m) \neq 1$ if and only if $p|m$. That is, $m \in \{p,2p,3p,\ldots,p^{k-1}p = p^k\}$, and there are p^{k-1} such multiples not greater than p^k. It follows that the other $p^k - p^{k-1}$ numbers are all relatively prime to p^k. ■

Theorem (Euler's Product Formula)**.**

$$\phi(n) = n \prod_{p|n} \left(1 - \frac{1}{p}\right)$$

Proof. If $n > 1$, then, by the fundamental theorem of arithmetic, there is a unique factorisation of $n = p_1^{k_1} p_2^{k_2} \cdots p_r^{k_r}$, where $p_1 < p_2 < \cdots < p_r$ are prime numbers, and every $k_i \geq 1$.

Repeatedly applying the multiplicative property and the formula for prime arguments yields,

$$\begin{aligned} \phi(n) &= \phi(p_1^{k_1})\phi(p_2^{k_2}) \cdots \phi(p_r^{k_r}) \\ &= p_1^{k_1}\left(1 - \frac{1}{p_1}\right) p_2^{k_2}\left(1 - \frac{1}{p_2}\right) \cdots p_r^{k_r}\left(1 - \frac{1}{p_r}\right) \end{aligned}$$

$$
\begin{aligned}
&= p_1^{k_1} p_2^{k_2} \cdots p_r^{k_r} \left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right) \cdots \left(1 - \frac{1}{p_r}\right) \\
&= n \left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right) \cdots \left(1 - \frac{1}{p_r}\right)
\end{aligned}
$$

■

Exercise. Prove Euler's product formula without using the multiplicative property of $\phi(n)$, instead using the inclusion-exclusion principle on the set of least residues modulo n.

The totient function gives a generalisation of Fermat's little theorem:

Theorem (Euler's Theorem)**.** *If n and a are coprime, then $a^{\phi(n)} \equiv 1 \pmod{n}$.*

Proof. Consider the set of residue classes modulo n coprime to n under multiplication modulo n ($\times_n$). Multiplication is closed and associative, 1 is the identity element, and Bézout's identity guarantees that every element has an inverse as every number is coprime to n. It follows that $(G,\times_n)$ is a group with $|G| = \phi(n)$. Let $a \in G$, $k = |a|$ and $H = \langle a \rangle = \{1,a,a^2...,a^{k-1}\}$. $(H,\times_p)$ forms a subgroup of G of order k. By Lagrange's theorem (§12.4.10), k divides $|G| = \phi(n)$, so $\phi(n) = nk$ for some $n \in \mathbb{Z}^+$. Thus, $a^{\phi(n)} = a^{nk} = (a^k)^n \equiv 1^n = 1 \equiv 1 \pmod{p}$. ■

Corollary 10.2.2.1. *If $x \equiv y \pmod{\phi(n)}$, then $a^x \equiv a^y \pmod{n}$.*

Proof. $x \equiv y \pmod{\phi(n)}$, so $x = y + k\phi(n)$ for some integer k. Then,

$$
\begin{aligned}
a^x &= a^{y+\phi(n)k} \\
&= a^y (a^{\phi(n)})^k \\
&\equiv a^y 1^k \pmod{n} \\
&\equiv a^y \pmod{n}
\end{aligned}
$$

If n is prime, this is also a corollary of Fermat's little theorem. ■

10.2.4 The Fundamental Theorem of Arithmetic

Theorem 10.2.3 (Fundamental Theorem of Arithmetic)**.** *Let $n > 0$ be an integer. Then, there is a unique sequence of primes $p_1 \leq p_2 \leq \cdots \leq p_k$ such that $n = p_1 p_2 \cdots p_k$.*

In this case, we call the sequence $p_1, p_2, \ldots, p_k$ the *prime factorisation* of n.

Proof. Showing that at least one such sequence exists is easily done with strong induction. If $n = 1$, the empty sequence suffices, as the empty product returns the multiplicative identity, 1. If n is prime, then we take $p_1 = n$, and we are done. Otherwise, n is composite, so $n = ab$ for some naturals a and b. Then, $n = p_1 p_2 \ldots p_k q_1 q_2 \ldots q_l$ where the p_i are the prime factorisation of a, and q_i are the prime factorisation of b, which are given by the inductive hypothesis.

This argument shows existence, but not uniqueness, of the sequence. We can show uniqueness with another strong induction combined with Euclid's lemma. If $n = 1$, then any non-empty sequence of primes gives a product greater than 1, so the empty sequence is the unique factorisation of 1. Now, if n is prime, then any sequence that differs from $p_1 = n$ would imply that n has multiple factors, so n would be composite, which is a contradiction. It follows that the prime factorisation of any prime is just the prime itself. This provides base cases for $n = 1$, $n = 2$, and $n = 3$, as well as any other larger prime values of n.

Now, suppose n is the least integer such that there exists two sequences $\{p_i\}_{i=1}^k$ and $\{q_i\}_{i=1}^l$ of primes such that $n = \prod_{i=1}^k p_i = \prod_{i=1}^l q_i$. p_1 divides $\prod_{i=1}^l q_i$, so p_1 divides some q_i by Euclid's lemma. Without loss of generality, suppose $p_1 \mid q_1$. As p_1 and q_1 are both prime, we have $p_1 = q_1$, so we may cancel them in the original factorisations, obtaining, $n = \prod_{i=2}^k p_i = \prod_{i=2}^l q_i$. But these are distinct factorisations of an integer smaller than n, contradicting the construction of n. It follows that no such n exists, and every integer has just one unique factorisation. ■

10.2.4.1 FTA and gcd

Because factorisations are unique, we can compute gcd(a,b) by factorising a and b into their prime sequences, then calculate the product of the primes that lie in the intersection of the two sequences. This is the algorithm taught at school. Without uniqueness, this algorithm doesn't work: we could factorise a and b in the wrong way such that the intersection doesn't give us the *greatest* common divisor. For very large integers, computing prime factorisations quickly becomes impractical, so the Euclidean algorithm is a better option.

We similarly compute the least common multiple lcm(a,b) by taking the maximum of the exponents on each prime that appears in the factorisation of a and b, which is again, the algorithm generally taught at school. For larger integers, it is more efficient to calculate $\text{lcm}(a,b) = \frac{ab}{\gcd(a,b)}$, since that avoids having to factorise a and b.

One way to think about this, is that we can represent any $n \in \mathbb{N}^+$ as a sequence of exponents, where the ith element is the exponent of the ith prime number. For example, $21 = 3 \cdot 7 = 2^0 \cdot 3^1 \cdot 5^0 \cdot 7^1$ (followed by an infinite tail of primes with 0 exponent), so we can represent 21 with the sequence $(0,1,0,1,0,\ldots)$, and $120 = 2^3 \cdot 3^1 \cdot 5^1$, so we can represent 120 with the sequence $(3,1,1,0,0,\ldots)$.

Taking the gcd of two numbers is then the same as taking the componentwise min of the corresponding sequences, while taking the lcm corresponds to the componentwise max. So, $\gcd(21,120) = (0,1,0,0,0,\ldots) = 3^1 = 3$, while $\text{lcm}(120,126) = (3,1,1,1,0,\ldots) = 2^3 \cdot 3^1 \cdot 5^1 \cdot 7^1 = 840$.

10.2.4.2 Prime Factorisations & RSA Encryption

For very, *very*, large integers, even the Euclidean algorithm quickly becomes much too slow, and computing prime factorisations, particularly of semiprimes, turns out to be a very difficult problem. So difficult, in fact, that it lies at the core of many encryption algorithms. The idea is that, if you have two very large prime numbers, then you can check their product very easily, but if you are just given the product, there is no algorithm to find the two factors better than just guessing and checking: it is a *one-way function* – something (relatively) quick to calculate given some inputs, but is extremely difficult to reverse back the other way to find the inputs, given just the output.

For instance, a 250 digit (829 bits) number[*] took 2 700 CPU core-years to factorise. The most commonly used encryption algorithm, RSA, uses 1 024 bits as a minimum, with 2 048 or 4 096 bit primes being common.

RSA relies on the fact that $(x^e)^d \equiv x \pmod{m}$ when m is semiprime, $de \equiv 1 \pmod{\phi(m)}$ and $0 \leq x < m$.[†]

[*] This number is RSA-250. The RSA numbers are a set of very large semiprimes that were released to encourage research into computational number theory and the practical difficulty of factoring large integers. More than half of the RSA numbers remain unfactorised, despite the list being over 30 years old.

[†] This fact doesn't quite follow immediately from Euler's theorem, because Euler's theorem only says that $x^{\phi(m)} \equiv 1 \pmod{m}$ when $\gcd(x,m) = 1$. However, we can use the Chinese remainder theorem to prove that $x^{de} = x^{k(p-1)(q-1)+1} \equiv x \pmod{m}$ holds even if $\gcd(x,pq \neq 1)$, as long as p and q are distinct primes.

The idea is that $\mathbb{Z}_{pq}$ factorises into $\mathbb{Z}_p \times \mathbb{Z}_q$, so we can represent $x \in \mathbb{Z}_{pq}$ as an ordered pair (x_p,x_q), where $x_p = x \pmod{p}$ and $x_q = x \pmod{q}$. Then, $x_p^{de} = \left(x_p^{p-1}\right)^{k(q-1)} \cdot x_p \equiv x_p \pmod{p}$, because either $x_p \equiv 0 \pmod{p}$ and the product is also 0, or $x_p \not\equiv 0 \pmod{p}$, so Euler's theorem gives $x^{p-1} \equiv 1 \pmod{p}$. The same reasoning applies to q, so we have, $\left(x_p^{de},x_q^{de}\right) = (x_p,x_q)$, so $x^{de} \equiv x \pmod{m}$ by the Chinese remainder theorem.

All of the above assumes that $\gcd(x,m) \neq 1$, as is overwhelmingly likely to be the case. But what happens if our message

So, we can encrypt some information, encoded as an integer $0 \leq x < m$, by raising it to the power of e modulo m, then decrypt it by raising the result to the power of d modulo m. As far as we understand, releasing e and m reveals no useful information about d, provided that e and m are chosen carefully.

The protocol for RSA specifically is as follows: the receiver selects extremely large primes p and q (remember that x has to be less than their product, so smaller primes necessitate more data chunking), then calculates d and e such that $de = 1 \pmod{(p-1)(p-2)}$, and $m = pq$. They then publish m, without revealing the factors p and q, and e.

A sender encrypts a message x by calculating x^e modulo m. Because x and e are generally extremely large numbers, this is computationally expensive, but there exist algorithms that speed up exponentiation and Euclidean division. For instance, we can compute x^e in stages by repeatedly squaring x and taking the product of the appropriate powers to reach e in binary. To decrypt x^e, the recipient similarly computes $(x^e)^d \pmod{m}$, which is often done with the Chinese remainder theorem in most implementations.

If you don't have the private key however, to get x back from x^e, you would need to calculate $\phi(m)$, which requires knowing the factors of m. Interestingly enough, the difficulty of this problem is only presumed, and hasn't been mathematically proven – it is completely possible that there exists an algorithm that efficiently computes the factors of a number, or an algorithm that computes $\phi(m)$ without factoring m. Discovering one would immediately break almost all modern asymmetric encryption, but as of yet, no such algorithms have been discovered.

This cryptographic system means that we can always keep d, the *private key*, to ourselves, and we never have to send it to whoever we want to talk to privately, so there is no opportunity for it to be intercepted or damaged. Meanwhile, the person who wants to send us an encrypted message just needs our values of e and m, the *public keys*. Because everyone knows them, we don't need to care about them being hidden away or encrypting the public keys themselves. This system is a type of encryption system called *public-key cryptography* or *asymmetric cryptography*.

As an analogy, we have a padlock which can be locked by any public key, which we give to everyone, but only unlocked by our private key. In particular, the public key *cannot* unlock the padlock. For someone to send us a message, they just get ahold of our public keys, lock their message with a padlock, and send it off to us. If we want to send a reply, we just get ahold of the sender's public keys, and send it in the same way. One of the strengths of this system is that the private keys never leave our hands, making them more secure, and the public keys don't have to be communicated secretly, or through an expensive secure channel.*

In contrast, *symmetric-key cryptography* only has a single key, that has to be transmitted between the sender and recipient before any messages are sent. In this system, there is a padlock that can both locked and unlocked by a single key. For a someone to send a message to us, we need to give them the key, across a hopefully secure channel, which they can use to lock their message before sending it to us. To reply, we just use the same key to lock our messages.

As a practical example of using RSA, let us pick $p = 7$ and $q = 13$, so $m = 91$. $\phi(m) = (p-1)(q-1) = 72$. Next, we pick e coprime to $\phi(m)$, say, $e = 5$. $5 \cdot 29 = 72 \cdot 2 + 1 \equiv 1 \pmod{\phi(m)}$, so $d = 29$ works. Note

just happens to give a gcd not equal to 1? Because m is semiprime, the only way that can happen is if $m|x$, and by Euclid's lemma, we find that $p|x$ or $q|x$. This actually breaks the encryption, as we, the sender, can now recover the factors of m by taking gcd(x,m), computing the other factor, then using them to compute d.

* If you are wondering if someone could replace the public keys of the sender with their own public keys, then decrypt, read, and re-encrypt the message before sending it to us, the recipient, the answer is yes.

This is called a *man-in-the-middle attack*, and is generally defended against through the use of *signed certificates* or the *Transport Layer Security* protocol, where we allow a trusted third party to authenticate our keys, so an interceptor cannot re-encrypt messages with their own unauthenticated keys.

The point is, even if the public key is known by other parties, any attacker still wouldn't be able to decrypt an already-encrypted message, whereas a compromised shared key in symmetric-key encryption is disastrous.

So, while some security is still needed, it's nowhere near as much as would be required to send a key in symmetric-key cryptography.

that computing d required us knowing $\phi(m)$, which also required us knowing p and q. There is no known way to compute d given just m and e.*

Now, say our message, x, is the number 11. Using $e = 5$ and $m = 91$, we want to compute $11^5 \pmod{91}$. We could do this by calculating 11^5, then dividing by 91 and checking the remainder, but this is computationally expensive, both in terms of memory and resources, since we're calculating and storing massive numbers, only to reduce the result to something less than 91. This is more of a problem for actual implementations using primes bigger than 5 or 11, but it's still preferable to reduce our workload when computing by hand.

To calculate the remainder, we instead use the *square and multiply algorithm.* This process basically works by repeatedly squaring our number to quickly reach large exponents, and performing the modulo reduction at every stage to keep the numbers (relatively) small in memory. Then, once we have enough powers of powers of 2, we multiply the right ones together. In practice, this is done extremely using bit-shifts and binary expansions, but we'll just do it by inspection here.

$$\begin{aligned}
11^1 &= 11 \\
11^2 &= 121 \\
&\equiv 30 \\
11^4 &\equiv 30^2 \\
&= 900 \\
&\equiv 81 \\
11^5 &= 11^4 \cdot 11 \\
&\equiv 81 \cdot 11 \\
&= 891 \\
&\equiv 72 \pmod{91}
\end{aligned}$$

When the recipient, who knows d, receives the encrypted message 72, they can decrypt the message into the original by computing 72^{29}:

$$\begin{aligned}
72^1 &= 72 \\
72^2 &= 5184 \\
&\equiv 88 \\
72^4 &\equiv 88^2 \\
&\equiv (-3)^2 \\
&= 9 \\
72^8 &\equiv 9^2 \\
&= 81 \\
72^{16} &\equiv 81^2 \\
&\equiv (-10)^2 \\
&= 100 \\
&\equiv 9 \\
72^{29} &= 72^{16} \cdot 72^8 \cdot 72^4 \cdot 72^1
\end{aligned}$$

* No known *classical* way. We have quantum algorithms (see *Shor's algorithm*) for factoring numbers in polynomial time, but sufficiently powerful quantum computers have not yet been constructed to implement this algorithm in practice.

We can simulate quantum computations on classical computers, but qubits tend to grow in size exponentially when stored classically, and we begin to run into other problems, making this simulation just as bad as other brute-force factorisation methods.

$$\begin{aligned}
&\equiv 9 \cdot 81 \cdot 9 \cdot 72 \\
&= 81^2 \cdot 72 \\
&\equiv (-10)^2 \cdot 72 \\
&= 100 \cdot 72 \\
&\equiv 9 \cdot 72 \\
&= 648 \\
&\equiv 11 \pmod{91}
\end{aligned}$$

Note that all of our computation is done in $\mathbb{Z}_{91}$, saving us from having to actually compute 72^{29} in $\mathbb{Z}$ and only taking the remainder at the end.

For actual usage, m needs to be large enough that it is computationally infeasible to recover p and q from m. As mentioned earlier, m is usually 2 048 or 4 096 bits long, and p and q are generally between 10^{308} and 10^{617} in size.

10.3 Ideals of the Integers

In the previous section, we proved the fundamental theorem of arithmetic, which states that natural numbers have prime factorisations, and that those factorisations are unique up to reordering and multiplication by a unit.

A similar property holds in the integers. However, the fundamental theorem of arithmetic does not hold in all number systems. For instance, in the system $\mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} : a,b \in \mathbb{Z}\}$, we have $6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5})$, so the number 6 has two distinct factorisations.

To solve this, we use *ideal numbers*, or *ideals*.

Ideals are covered more generally and abstractly in §12.9.1, but for now, we will deal with ideals in the ring of integers.

A subset $S \subseteq \mathbb{Z}$ is an ideal in the integers if,

- $S \neq \varnothing$: S is non-empty,
- $\forall m,n \in S : m + n \in S$: S is closed under addition,
- $\forall k \in \mathbb{Z}\, \forall m \in S : k \cdot m \in S$: S is closed under multiplication by an integer.

Lemma 10.3.1. *Ideals are closed under negation.*

Proof. Since S is closed under multiplication, if $m \in S$, then $(-1) \cdot m = -m \in S$. ■

This lemma gives a much simpler characterisation of ideals in $\mathbb{Z}$:

Lemma 10.3.2. *A subset $S \subseteq \mathbb{Z}$ is an ideal in the integers if and only if*

- *$S \neq \varnothing$: S is non-empty,*
- *$\forall m,n \in S : m - n \in S$: S is closed under subtraction.*

Proof. Suppose S is an ideal, so S is non-empty. Let $m,n \in S$. By Lemma 10.3.1, $-n \in S$, and since S is closed under addition by definition, we have $m + (-n) = m - n \in S$, completing the forward direction.

Now, suppose $S \subseteq \mathbb{Z}$ is non-empty and closed under subtraction. Fix $m \in S$. Then, $m - m = 0 \in S$, so $0 - m = -m \in S$ and S is closed under negation.

Let $m,n \in S$, By closure under negation, $-n \in S$, and by closure under subtraction, $m-(-n) = m+n \in S$, so S is also closed under addition.

It remains to show that S is closed under multiplication any integer k. Fix $m \in S$. We induct on k. As shown above, $0 \in S$, so the proposition holds for $k = 0$. Now, suppose $km \in S$. Since S is closed under addition, $km + m = k(m + 1) \in S$, completing the inductive step, so $km \in S$ for any $k \in \mathbb{N}$. But since S is closed under negation, $-km \in S$ for all $k \in \mathbb{N}$. ■

Here are a few subsets of the integers which are ideals:

- The trivial ring, $\{0\}$.
- The set of integers, $\mathbb{Z}$.
- The set of even integers, $2\mathbb{Z}$.
- The set of integers divisible by 3, $3\mathbb{Z}$.

In fact, all ideals in the integers are of the form $k\mathbb{Z}$,$k \in \mathbb{N}$. We also write this ideal as (k), and we call k the *generator* of the ideal, because the ideal can be generated by multiplying every integer by k.

Theorem 10.3.3. *If $S \subseteq \mathbb{Z}$ is an ideal, then there exists a unique $k \in \mathbb{N}$ such that $S = k\mathbb{Z} = \{kn : n \in \mathbb{Z}\} = (k)$.*

Proof. If $S = \{0\}$, then take $k = 0$ and we are done.

Otherwise, there is a non-zero element $m \in S$, so $-m \in S$ by closure under negation. Let $S^+ \subset S$ be the set of positive elements of S. Exactly one of m and $-m$ lies in S^+, so S^+ is non-empty. By the well-ordering principle (§6.11.4) there exists a unique least element, k.

Since S is an ideal, S is closed under multiplication by integers, so for all $n \in \mathbb{Z}$, $nk \in S$, so $k\mathbb{Z} \subseteq S$.

Now, fix $m \in S$, so $m = qk + r$ for some $q \in \mathbb{Z}$ and $0 \leq r < k$. Since $k \in S$, $qk \in S$ by closure of multiplication by integers and $m - qk = r \in S$ by closure of subtraction. If $r \geq 0$, then $r \in S^+$, but $r < k$, contradicting that k is the least element of S. It follows that $r = 0$, so $m = qk \in k\mathbb{Z}$ and $S \subseteq k\mathbb{Z}$.

Because both $k\mathbb{Z} \subseteq S$ and $S \subseteq k\mathbb{Z}$, $S = k\mathbb{Z}$. It remains to show that this k is unique. Suppose $S = k\mathbb{Z} = l\mathbb{Z}$ with $k \neq l$. Without loss of generality, suppose $0 < k < l$. Then, $k \in k\mathbb{Z}$ but $k \notin l\mathbb{Z}$ as l is the least positive element of $l\mathbb{Z}$. It follows that $k\mathbb{Z} \neq l\mathbb{Z}$, contradicting that $S = k\mathbb{Z} = l\mathbb{Z}$, so $k = l$. ■

This theorem shows that there is a bijection between the ideals of $\mathbb{Z}$ and the set of natural numbers:

Theorem (Ideals of Integers Equivalent to Natural Numbers)**.** *Let S be the set of ideals of $\mathbb{Z}$. Then, the mapping $\psi : \mathbb{N} \to S$ defined by,*

$$\forall k \in \mathbb{N} : \psi(k) = (k)$$

is a bijection.

That is, every ideal is generated by a unique natural, and every natural generates a unique ideal.

There is an interesting relationship between divisibility of integers and set inclusion of ideals.

Theorem 10.3.4. *Let $m,n \in \mathbb{Z}$, Then, $m|n$ if and only if $(m) \supseteq (n)$.*

Proof. If $m|n$, then $m|kn$ for any $k \in \mathbb{Z}$, so $(m) \supseteq (n)$, completing the forward direction.

Conversely, if $(m) \supseteq (n)$, then $n \in (m)$, so $m|n$, completing the backward direction. ■

10.3.1 Operations on Ideals

We now give two operations on ideals.

Let I and J be ideals in $\mathbb{Z}$.

- $I \cap J = \{n \in \mathbb{Z} : n \in I \wedge n \in J\}$ is the *intersection* of I and J.
- $I + J = \{i + j \in \mathbb{Z} : i \in I, j \in J\}$ is the *sum* of I and J.

The intersection of two ideals is a subset of the two ideals, but the sum of two ideals is a superset of the two ideals.

Lemma 10.3.5.

1. *$I \cap J$ is a subset of both I and J.*
2. *I and J are subsets of $I + J$.*

Proof.

1. Let x be in $I \cap J$. Then, $x \in I$ and $x \in J$ by the definition of $I \cap J$, so we have $x \in I$. As the choice of x was arbitrary, $\forall x \in I \cap J$, $x \in I$ so $I \cap J \subseteq I$. By symmetry, $I \cap J \subseteq J$.
2. Because J is an ideal, it is closed under negation, so $0 \in J$. Then, $I = \{i+0 \in \mathbb{Z} : i \in I\} \subseteq \{i+j \in \mathbb{Z} : i \in I, j \in J\} = I + J$, so $I \subseteq I + J$. By symmetry, $J \subseteq I + J$.

■

Theorem 10.3.6. *Suppose I, J, and K are ideals of $\mathbb{Z}$. Then,*

1. *$I \cap J$ and $I + J$ are ideals of $\mathbb{Z}$.*
2. *$K \subseteq I$ and $K \subseteq J$ if and only if $K \subseteq I \cap J$.*
3. *$I \subseteq K$ and $J \subseteq K$ if and only if $I + J \subseteq K$.*

These latter two properties can also be stated informally as,

2. *$I \cap J$ is the "largest" ideal contained in both I and J.*
3. *$I + J$ is the "smallest" ideal containing both I and J.*

Proof.

1. Both I and J are closed under subtraction, so 0 is in both I and J. It follows that $I \cap J$ and $I + J$ both also contain 0 and are therefore non-empty.

   Suppose $m,n \in I \cap J$, so $m,n \in I$ and $m - n \in I$. Similarly, $m - n \in J$. It follows that $m - n \in I \cap J$, so $I \cap J$ is closed under subtraction. Since $I \cap J$ is non-empty and is closed under subtraction, it is an ideal.

   Suppose $m,n \in I + J$, so $m = a + c$ and $n = b + d$ with $a,b \in I$ and $c,d \in J$. By closure of subtraction, $a - b \in I$ and $c - d \in J$, so $(a - b) - (c - d) \in I + J$.

   $(a - b) - (c - d) = (a + c) - (b + d) = m - n$, so $m - n \in I + J$. Since $I + J$ is non-empty and is closed under subtraction, it is an ideal.
2. Suppose K is a subset of both I and J, so for any $k \in K$, $k \in I$ and $k \in J$. Then, $k \in I \cap J$ so $K \subseteq I \cap J$.

   Suppose $K \subseteq I \cap J$. By Lemma 10.3.5, $I \cap J$ is a subset of both I and J, so K is also a subset of both I and J.

3. Suppose that both I and J are subsets of K. For any $i \in I \subseteq K$ and $j \in J \subseteq K$, we have $i \in K$ and $j \in K$ so $i + j \in K$ by closure of addition. It follows that $I + J \subseteq K$.

   Suppose $I + J \subseteq K$. By Lemma 10.3.5, I and J are both subsets of $I + J$, so I and J are both subsets of K.

■

10.3.2 GCDs and LCMs with Ideal Operations

We can link the notion of an ideal with greatest common divisors and least common multiples with the following theorem:

Theorem 10.3.7. *Let* $m,n \in \mathbb{Z}$*,* $\gcd(m,n) = g$*, and* $\operatorname{lcm}(m,n) = l$

1. $m\mathbb{Z} + n\mathbb{Z} = g\mathbb{Z}$

2. $m\mathbb{Z} \cap n\mathbb{Z} = l\mathbb{Z}$

Proof.

1. $m\mathbb{Z}$ and $n\mathbb{Z}$ are both ideals, so their sum is also an ideal, and by Theorem 10.3.3, there exists a unique natural k that generates this ideal, so $m\mathbb{Z} + n\mathbb{Z} = k\mathbb{Z}$. By Lemma 10.3.5, $m\mathbb{Z} \subseteq k\mathbb{Z}$ and $n\mathbb{Z} \subseteq k\mathbb{Z}$, so $k|m$ and $k|n$, and k is a common divisor of both m and n. By Theorem 10.3.6, $m\mathbb{Z} + n\mathbb{Z}$ is contained in every ideal containing $m\mathbb{Z}$ and $n\mathbb{Z}$, so k is divisible by any other common divisor of m and n. It follows that k is the greatest common divisor of m and n.
2. By Theorem 10.3.3, $m\mathbb{Z} \cap n\mathbb{Z} = k\mathbb{Z}$ for some unique natural k. By Lemma 10.3.5, we deduce that k is a common multiple of m and n, and by Theorem 10.3.6, k must be the least common multiple.

■

10.3.3 Bézout's Identity with Ideals

Using ideals, we can give a much shorter proof of Bézout's identity.

Theorem (Bézout's Identity)**.** *If a and b are non-zero integers, then there exists integers x and y such that* $\gcd(a,b) = ax + by$*.*

Proof. By Theorem 10.3.7, we have $a\mathbb{Z} + b\mathbb{Z} = g\mathbb{Z}$, where $g = \gcd(a,b)$, so $g \in a\mathbb{Z} + b\mathbb{Z}$ and there exists $p \in m\mathbb{Z}$ and $q \in \mathbb{Z}$ such that $g = p + q$. Since $p \in m\mathbb{Z}$, there exists some $x \in \mathbb{Z}$ such that $p = ax$, and similarly, there exists some $y \in \mathbb{Z}$ such that $q = by$, so $g = ax + by$, as required. ■

10.4 The Integers

10.4.1 Prime Numbers

10.4.2 Prime Number Theorem

10.4.3 Integer Partitions

10.4.4 Four Square Theorem

10.4.5 Diophantine Equations

10.4.6 Diophantine Approximation

10.5 Modular Arithmetic

10.5.1 Fermat's Little Theorem

10.5.2 Fundamental Theorem of Algebra

10.6 Analytic Number Theory

10.7 Algebraic Number Theory

10.8 Arithmetic Combinatorics

10.9 p-adic Numbers

Chapter 11

The Real Numbers

> *"In mathematics the art of proposing a question must be held of higher value than solving it."*
>
> — Georg Cantor, *Doctoral thesis*

Abstract algebra is the study of sets equipped with operations, called algebraic structures. In this chapter, we will be introduced to the notion of a *field* through the axiomatisation of the real numbers. In the next chapter, we will explore some other structures known as *groups* and *rings*, before combining these in with fields. These two chapters should be read in succession, as a pair.

11.1 Axiomatisation of the Real Numbers

Due to their ubiquity, we should take an aside to further discuss the characterisation of the real numbers.

The set of real numbers, $\mathbb{R}$, is probably the most commonly used number system in all of mathematics, their importance overshadowed only possibly by the complex numbers. Some important subsets of the real numbers are the naturals, $\mathbb{N} = \{0,1,2,\ldots\}$, the integers, $\mathbb{Z} = \{\ldots,-2,-1,0,1,2,\ldots\}$, and the rationals, $\mathbb{Q}$, which include all the real numbers which can be written as the ratio of two integers, $\frac{p}{q}$. Any real number with a terminating or recurring decimal is representable as such a ratio. However, some numbers, such as $\sqrt{2} = 1.414\ldots$, $e = 2.718\ldots$, and $\pi = 3.141\ldots$ don't ever terminate or repeat.

We also have the irrationals, which are the real numbers which are not rational. There isn't a standard symbol for the set of irrationals, and we often just write $\mathbb{R} \setminus \mathbb{Q}$ for this set, but $\mathbb{Q}'$ would also generally be understood to represent this set, given sufficient context. We also have algebraic numbers,[†] which are numbers that can be written as the root of a polynomial with rational coefficients, and the transcendental numbers, which are numbers that are non-algebraic. Some transcendental numbers we commonly use are π, e and $\ln 2$.

Beyond the reals, we have the complex numbers, $\mathbb{C}$, which are intimately linked to many functions we normally use. Next, we have the quaternions, $\mathbb{H}$,[‡] which adds three imaginary units to form a non-

[†] The algebraic numbers are sometimes denoted $\mathbb{A}$, but this symbol is usually denoted for a different structure called the *adele ring*.

[‡] $\mathbb{Q}$ is already reserved for the rationals, so $\mathbb{H}$ is used to honour their discoverer, *William Hamilton*.

Famously, Hamilton had been struck by a flash of inspiration whilst out on a walk, and carved the quaternion formula

$$i^2 = j^2 = k^2 = ijk = -1$$

into the Brougham bridge as he paused on it. It is now an annual event called the Hamilton Walk to walk from Dunsink Observatory to the bridge in remembrance of this discovery.

At the time of their discovery, quaternions were the main language used to describe topics such as kinematics and commutative 4 dimensional number system highly useful in modelling rotations and computer graphics. Next are the octonions, $\mathbb{O}$, which are a non-associative 8 dimensional number system with 7 imaginary components. There are further general extensions to the complex numbers, called the *hypercomplex numbers*, each one with twice the dimension of the previous extension.

As with any mathematical structure, the real numbers may be characterised by a list of axioms. There are many equivalent axiomatisations of the reals. We have already constructed the reals with Dedekind cuts in §4.5.4, and we will see another construction in §34.3.4. Here, we will give an axiomatic construction of the reals as a type of *field*.

11.2 Field Axioms

We will quickly list the axioms symbolically, then focus on each one individually.

Given a set S, a *binary operation* on S is a function that takes two elements of S, called the *operands* or *arguments* of the operation, and returns another element of S: it is *closed* over S. That is, it is a binary function $S \times S \to S$.

A *field* is a set, K, together with two elements, $0_K \neq 1_K \in K$, and two binary operations, $\cdot : K \times K \to K$ and $+ : K \times K \to K$, called *multiplication* and *addition*, respectively, that satisfies the following axioms:

(A1) $\forall a, b \in K : a + b = b + a$ (commutativity of addition);

(A2) $\forall a, b, c \in K : a + (b + c) = (a + b) + c$ (associativity of addition);

(A3) $\exists 0_K \in K$ such that $\forall a \in K, a + 0_K = 0_K + a = a$ (existence of additive identity);

(A4) $\forall a \in K : \exists(-a) \in K$ such that $a + (-a) = (-a) + a = 0_K$ (existence of additive inverses);

(M1) $\forall a, b \in K : a \cdot b = b \cdot a$ (commutativity of multiplication);

(M2) $\forall a, b, c \in K : a \cdot (b \cdot c) = (a \cdot b) \cdot c$ (associativity of multiplication);

(M3) $\exists 1_K \in G$ such that $\forall a \in K, a \times 1_K = 1_K \times a = a$ (existence of additive identity);

(M4) $\forall a \in K : \exists(a^{-1}) \in K \setminus \{0\}$ such that $a * (a^{-1}) = (a^{-1}) * a = 1_K$ (existence of multiplicative inverses);

(D) $\forall a, b, c \in K : (a + b)c = ac + bc$ (distributivity of multiplication over addition);

(ND) $0_K \neq 1_K$ (non-degeneracy).

Where there is no room for confusion, we write ab for $a \cdot b$, and 0 and 1 for 0_K and 1_K, respectively. We often denote general fields with K or F. The symbol $\mathbb{F}$ is reserved for a certain type of finite field.

Additionally, we may call elements of a field, *numbers*, to distinguish them from elements of other structures, say, groups.

11.2.1 Axioms for Addition

The first four field axioms are about the binary operation called addition, denoted with the symbol $+$. Again, it should be emphasised that this is just a symbol, and that any operation that satifies these axioms is a valid addition over a field.

These four axioms can more concisely expressed as "Addition over a field satisfies the axioms of an abelian group" (§12.3).

Maxwell's equations – which we now describe with vectors. In fact, the real component of a quaternion is called the *scalar* part, and the imaginary components, the *vector* part.

A1 (Commutativity of addition)**.** *For all numbers in a field,*

$$a + b = b + a$$

Any operation that satisfies this axiom is *commutative.* Commutativity allows us to ignore the order of arguments of an operation.

We've seen a few examples of commutative operations before – $\wedge$, $\vee$, $\cap$, $\cup$, etc. – and a few of non-commutative as well – $\rightarrow$, $\times$ (Cartesian product), $-$, etc.

A2 (Associativity of addition)**.** *For all numbers in a field,*

$$a + (b + c) = (a + b) + c$$

Any operation that satisifes this axiom is *associative.* Associativity means that repeated applications of an operation chained together can be evaluated in any order we want. This allows us to ignore brackets and write $a + b + c$ for $a + (b + c) = (a + b) + c$.

Again, we've seen a few examples of associative and non-associative operations, for example, $\wedge$, and $\times$ (Cartesian product), respectively.

A3 (Existence of additive identity)**.** *There exists a number, denoted* 0_K*, such that, for all numbers* a*,*

$$a + 0_K = 0_K + a = a$$

When the field and operation are clear, we may just write 0 for 0_K.

Any object which satisfies this axiom is called an *identity element* or a *neutral element* for its operation. The term identity element is often shortened to just *identity* when there is no room for confusion, as in the case of *additive identity* or *multiplicative identity*, but the identity implicitly depends onthe binary operation it is associated with.

For example, the identity element for function composition is the aptly-named identity map, which is $x \mapsto x$. Composing this function with any other function from either side leaves the function unchanged, so it is the identity element for function composition.

This definition of identity can actually be further split into two types. Let S be a set, and $+$ be a binary operation over S. An element, $e \in S$ is a *left identity* if $e + s = s$ for all $s \in S$, and a *right identity* if $s + e = s$ for all $s \in S$. If e is both a left and right identity, we call it a two-sided identity, or just an identity. Because addition is commutative, additive identities must be two-sided.

Identities for an operation are unique:

Lemma 11.2.1. $\forall a : e + a = a + e = a \wedge f + a = a + f = a \rightarrow e = f$

Proof. Suppose e and f are distinct identities for the operation denoted by $+$. Then, $e + f = e$, because f, being an identity, is a right identity. But $e + f = f$ because e, being an identity, is a left identity, so both $e + f = e = f$, so $e = f$, and the identity is unique. ■

A4 (Existence of additive inverses)**.** *For each number,* a*, there exists a number* $(-a)$ *such that,*

$$a + (-a) = (-a) + a = 0_K$$

where 0_K is the additive identity.

For convenience, we write $a + (-b)$ as $a - b$ (a *minus* b), encoding our idea of subtraction.

The number $(-a)$ is called the (*additive*) *inverse* of a. However, since we often like to use the word "inverse" for multiplicative inverses, we call the additive inverse, the *negative* or *minus* of a.

Like identities, inverses can be categorised into left and right inverses. Let $+$ be a binary operation over a field K. If $a + b = 0_K$, then a is a *left inverse* of b, and b is a *right inverse* of a.

Inverses are also unique:

Lemma 11.2.2. *If $a + b = b + a = 0$, then $a = -b$.*

Proof.

$$\begin{aligned} a + b &= 0 \\ a + b + (-b) &= 0 + (-b) \\ a + (b + (-b)) &= 0 + (-b) \\ a + 0 &= -b \\ a &= -b \end{aligned}$$

■

Every element is also equal to it the inverse of its inverse: that is,

Lemma 11.2.3. *$\forall a : a = -(-a)$.*

Proof.

$$\begin{aligned} a + (-a) &= 0 \\ a + (-a) + (-(-a)) &= 0 + (-(-a)) \\ a + ((-a) + (-(-a))) &= (-(-a)) \\ a + 0 &= -(-a) \\ a &= -(-a) \end{aligned}$$

This proof is perhaps clearer in multiplicative notation, as shown below. ■

11.2.2 Axioms for Multiplication

The next four axioms concern a binary operation called multiplication, denoted with the symbol $\cdot$. We often omit the symbol $\cdot$, and just write ab for $a \cdot b$. This convention will be used wherever there is no room for confusion. Also note that we always denote multiplication on fields with $\cdot$, and never $\times$.

These four axioms can more concisely expressed as "Multiplication over a field, minus the additive identity element, satisfies the axioms of an abelian group" (§12.3).

M1 (Commutativity of multiplication)**.** *For all numbers in a field,*

$$a \cdot b = b \cdot a$$

M2 (Associativity of multiplication)**.** *For all numbers in a field,*

$$a \cdot (b \cdot c) = (a \cdot b) \cdot c$$

M3 (Existence of multiplicative identity)**.** *There exists a number, denoted 1_K, such that, for all numbers a,*

$$a \cdot 1_K = 1_K \cdot a = a$$

The proof for the uniqueness of multiplicative identities works exactly the same as for additive identities:

Lemma (Uniqueness of additive identities)**.** $\forall a : ea = ae = a \land fa = af = a \to e = f$.

Proof. Suppose e and f are both multiplicative identities. Then, $ef = e$ as f is the identity. But $ef = f$, as e is also the identity, so $ef = e = f$, so $e = f$ and the identity is unique. ■

M4 (Existence of multiplicative inverses)**.** *For each number, $a \neq 0_K$, there exists a number a^{-1} such that,*

$$a \cdot a^{-1} = a^{-1} \cdot a = 1_K$$

where 1_K is the multiplicative identity.

For convenience, we often write $a \cdot b^{-1}$ as $\frac{a}{b}$, encoding our idea of division. 0 is not guaranteed to have a multiplicative inverse, so expressions such as $\frac{a}{0}$ are undefined.

The number a^{-1} is called the *multiplicative inverse* of a, often shortened to just *inverse*.

Multiplicative inverses are similarly unique:

Lemma (Uniqueness of multiplicative inverses)**.** *For all a and b, if $ab = ba = 1$, then $a = b^{-1}$.*

Proof.

$$\begin{aligned} ab &= 1 \\ abb^{-1} &= 1b^{-1} \\ a(bb^{-1}) &= b^{-1} \\ a(1) &= b^{-1} \\ a &= b^{-1} \end{aligned}$$

■

Every element is also equal to it the inverse of its inverse:

Lemma 11.2.4. $\forall a : a = (a^{-1})^{-1}$

Proof.

$$\begin{aligned} a(a^{-1}) &= 1 \\ a(a^{-1})(a^{-1})^{-1} &= 1(a^{-1})^{-1} \\ a1 &= (a^{-1})^{-1} \\ a &= (a^{-1})^{-1} \end{aligned}$$

■

11.2.3 Axiom of Distributivity

D (Distributivity of multiplication over addition)**.** *For all numbers a, b and c,*

$$(a+b)c = ac + bc$$

The additive identity 0_K also has a special role in multiplication as a consequence of this distribution axiom: it is an *annihilator*, an element that gives 0_K when multiplied by anything.

Proof.

$$\begin{aligned} 0+0 &= 0 \\ a\cdot(0+0) &= a\cdot 0 \\ a\cdot 0 + a\cdot 0 &= a\cdot 0 \\ a\cdot 0 + a\cdot 0 + (-(a\cdot 0)) &= a\cdot 0 + (-(a\cdot 0)) \\ a\cdot 0 + (a\cdot 0 - a\cdot 0) &= a\cdot 0 - a\cdot 0 \\ a\cdot 0 &= 0 \end{aligned}$$

■

A similar argument shows that if $a\cdot b = 0$, then $a = 0$ or $b = 0$.

Proof. Suppose $a\cdot b = 0$, but $a \neq 0$, so a^{-1} exists by axiom A4. Then,

$$\begin{aligned} a\cdot b &= 0 \\ a^{-1}\cdot a\cdot b &= a^{-1}\cdot 0 \\ b &= 0 \end{aligned}$$

■

We can also show

$$a\cdot(-b) = -(a\cdot b) \tag{1}$$

$$(-a)\cdot b = -(a\cdot b) \tag{2}$$

$$(-a)\cdot(-b) = a\cdot b \tag{3}$$

Proof. For (1),

$$\begin{aligned} a\cdot 0 &= 0 \\ a\cdot(b+(-b)) &= 0 \\ a\cdot b + a\cdot(-b) &= 0 \\ -(a\cdot b) + (a\cdot b + a\cdot(-b)) &= -(a\cdot b) + 0 \\ (-(a\cdot b) + a\cdot b) + a\cdot(-b) &= -(a\cdot b) \\ 0 + a\cdot(-b) &= -(a\cdot b) \\ a\cdot(-b) &= -(a\cdot b) \end{aligned}$$

(2) is identical, with a and $-b$ replaced with their negations.

For (3),

$$(-a)\cdot 0 = 0$$

$$
\begin{aligned}
(-a) \cdot (b + (-b)) &= 0 \\
(-a) \cdot b + (-a) \cdot (-b) &= 0 \\
-(a \cdot b) + (-a) \cdot (-b) &= 0 \\
(a \cdot b) + (-(a \cdot b)) + (-a) \cdot (-b) &= (a \cdot b) + 0 \\
(-(a \cdot b) + a \cdot b) + (-a) \cdot (-b) &= a \cdot b \\
0 + (-a) \cdot (-b) &= a \cdot b \\
(-a) \cdot (-b) &= a \cdot b
\end{aligned}
$$

Noting that we use (2) to move from the third line to the fourth. ■

A special case of (2) is that multiplying by -1 is equivalent to negation:

Corollary 11.2.4.1. $\forall a : (-1) \cdot a = -a$

Proof. Using (3), $(-1) \cdot a = -(1 \cdot a) = -a$. ■

0 being the annihilator element is one of the reasons why we can't define 0^{-1}, thus disallowing division by 0. If 0^{-1} was an element of the field, then, for any a and b in the field, we would have,

$$
\begin{aligned}
a \cdot 0 &= b \cdot 0 \\
(a \cdot 0) \cdot 0^{-1} &= (b \cdot 0) \cdot 0^{-1} \\
a \cdot (0 \cdot 0^{-1}) &= b \cdot (0 \cdot 0^{-1}) \\
a \cdot 1 &= b \cdot 1 \\
a &= b
\end{aligned}
$$

11.2.4 Axiom of Non-Degeneracy

ND (Non-Degeneracy).

$$0_K \neq 1_K$$

This condition prevents the set $\{e\}$ from being a field, where e is both the additive and multiplicative identity. This use useful because it also prevents $e = 0$ from having an inverse (itself). This is similar to 1 not being a prime number: if we let the single element set be a field, many field theorems end up having to stipulate "unless $F = \{e\}$".

11.2.5 Examples of Fields

These axioms we have stated characterise a field. The real numbers are a field, but not all fields are the real numbers.

For instance, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are all fields. If you are familiar with classical constructions, the set of numbers you can construct with a straightedge and compass also forms a field.

On the other hand, $\mathbb{N}$ is not a field as additive inverses do not exist. (The additive identity doesn't exist either, if you exclude 0 from the naturals.) $\mathbb{Z}$ is not a field either, as not every non-zero element has a multiplicative inverse.

However, the integers modulo p is a (finite, or *Galois*) field for any prime p, commonly denoted $\mathbb{F}_p$ or $\mathbb{Z}_p$. In particular, $\mathbb{F}_2$ is frequently used in computer science, as many logic operations that apply to bits can be converted to operations applying to elements of $\mathbb{F}_2$ (for example, adding two elements is the same as taking the XOR of those elements, and multiplying two elements is the same as taking the AND of those elements).

11.3 Order Axioms

To obtain the real numbers, and the real numbers only, we add a few more axioms.

This means that, as well as the binary operations of addition and multiplication, the field is equipped with a total ordering relation (§4.4.8).

Notice that $\mathbb{C}$ and $\mathbb{F}_p$ do not have any relations defined on them that are total orders. For the former, in particular, suppose $0 \prec i$. Then, $0i \prec ii$, so $0 \prec -1$, which already looks somewhat problematic, but, suppose we continue with this anyway. Then, $0 \prec -1$ implies $1 = 0 + 1 \prec -1 + 1 = 0$, so $1 \prec 0 \prec 1$, and our ordering falls apart. Assuming $i \prec 0$ runs into similar problems, so there can't be an ordering over the complex numbers.

The axioms we use to describe this ordering is as follows:

(C) $\forall a,b \in K : a \leq b \vee b \leq a$ (Comparability*);

(A) $\forall a,b \in K : a + (b + c) = (a + b) + c$ (Antisymmetry);

(T) $\forall a,b,c \in K : a \leq b \wedge b \leq c \to a \leq c$ (Transitivity);

(TI) $\forall a,b,c \in K : a \leq b \to a + c \leq b + c$ (Translational invariance);

(SI) $\forall a,b,c \in K : a \leq b \wedge 0 \leq c \to a \cdot c \leq b \cdot c$ (Scaling invariance).

For further discussion of the first three axioms, see §4.4.5.

The latter two axioms describe how $\leq$ interacts with addition and multiplication – addition and multiplication both preserve order.

We only define the single relation $\leq$ like this. $a \geq b$ is defined to be $b \leq a$, $a < b$ is $a \leq b \wedge a \neq b$, and $a > b$ is $b \leq a \wedge a \neq b$.

Additionally, if $a > 0$, we say that a is *positive*. If $a < 0$, a is *negative*. If $a \geq 0$, a is *non-negative*. If $a \leq 0$, a is *non-positive*, though this term does seem to be rarer.

Other properties of $\leq$ can be derived from these axioms.

Lemma 11.3.1 (Reflexivity). *$\forall x : x \leq x$.*

Proof. Apply the comparability axiom with $a = b = x$ ■

Lemma 11.3.2 (Trichotomy). *$\forall x,y : x < y \veebar x = y \veebar x > y$. That is, for any numbers a and b, exactly one of $x < y$, $x = y$ and $x > y$ holds.*

Proof. We first show that at least one holds. From the comparability axiom, at least one of $x \leq y$ and $x \leq y$ holds. If $x = y$, we are done. Otherwise, $x \neq y$, so exactly one of $x \leq y$ and $y \leq x$ holds as well, which is the definition of $<$, so exactly one of $x < y$ and $y < x$ holds, given that $x \neq y$.

If $x = y$, then $x \not< y$ and $y \not< x$, because they are both defined to hold only when $x \neq y$, so if $x = y$, then it is the only one which holds. Next, suppose $x < y$ and $y < x$ both hold, but $x \neq y$. Then, $x \leq y$ and $y \leq x$, so by antisymmetry, $x = y$, contradicting our assumption. So, at most one holds. ■

Trichotomy allows us to treat $x \not< y$ and $x \geq y$ as equivalent.

Lemma 11.3.3. *$\forall a : a \geq 0 \to -a \leq 0$: if $a \geq 0$, then $-a \leq 0$.*

* As mentioned previously (§4.4.2), we can alternatively write this as $x \lesseqqgtr y$ for all x and y under $\leq$.

Proof.

$$\begin{aligned} a &\geq 0 \\ a + (-a) &\geq 0 + (-a) \\ 0 &\geq -a \\ -a &\leq 0 \end{aligned}$$

The second line holds by the axiom of translational invariance, while last line is justified by the definition of $\geq$. ■

Lemma 11.3.4. $\forall a,b : a \geq b \leftrightarrow a - b \geq 0$: *for all a and b, $a \geq b$ if and only if $a - b \geq 0$.*

Proof.

$$\begin{aligned} a &\geq b \\ a + (-b) &\geq b + (-b) \\ a - b &\geq 0 \end{aligned}$$

and the reverse direction,

$$\begin{aligned} a - b &\geq 0 \\ a - b + b &\geq 0 + b \\ a &\geq b \end{aligned}$$

■

Lemma 11.3.5. $\forall a,b : a \geq 0 \wedge b \geq 0 \rightarrow a + b \geq 0$: *if $a \geq 0$ and $b \geq 0$, then $a + b \geq 0$.*

Proof. $a \geq 0$, so $0 \geq -a$ by Lemma 11.3.3. So, $b \geq 0 \geq -a$, and $b \geq -a$ by transitivity. By TI, $b + a \geq -a + a$, and with A1, $a + b \geq 0$. ■

Theorem 11.3.6. $\forall a,b,c,d : a \geq b \wedge c \geq d \rightarrow a + c \geq b + d$: *for all a, b, c and d, if $a \geq b$ and $c \geq d$, then $a + c \geq b + d$.*

Proof. By Lemma 11.3.4, $a \geq b \rightarrow a - b \geq 0$ and $c \geq d \rightarrow c - d \geq 0$. Now, apply Lemma 11.3.5 to obtain $(a - b) + (c - d) \geq 0$, and add $b + d$ to both sides (TI) to get $a + c \geq b + d$. ■

This theorem allows us to add together different inequalities.

Lemma 11.3.7. $\forall a,b : a \leq b \rightarrow -b \leq -a$: *if $a \leq b$, then $-b \leq -a$.*

Proof. Subtract $a + b$ from both sides. ■

Theorem 11.3.8. $\forall a,b,c \leq 0 : a \leq b \rightarrow a \cdot c \geq b \cdot c$: *if $a \leq b$ and $c \leq 0$, then $a \cdot c \geq b \cdot c$.*

Proof. From Lemma 11.3.3, $-c \geq 0$, so by SI, $-c{\cdot}a \leq -c{\cdot}b$. By Lemma 11.3.7, $c{\cdot}a \geq c{\cdot}b$, as required. ■

This theorem tells us that multiplying an inequality by a negative number reverses the direction of the inequality.

11.4 Completeness Axiom

Total ordering alone still doesn't fully determine set of reals: the set of rationals, as well as being a field, also has a well-defined total ordering.

The property that characterises the reals, is *completeness*: the idea that there aren't any "gaps" or "missing points" in the real numbers, as opposed to the rationals, which has gaps at every irrational number.

There are many ways to add one last axiom to separate the reals from the rationals. We have already seen Dedekind completeness when we defined real numbers to be generated by some Dedekind cut. We will also see *Cauchy completeness* in §34, but here, we will use the least upper bound property.

11.4.1 Least Upper Bound Property

Let S be a non-empty set of real numbers. A real number x is an *upper bound* of S if $x \geq s$ for all $s \in S$. A real number y is the *least upper bound* or *supremum* of S if y is an upper bound, and $y \leq x$ for all upper bounds x of S. The supremum of a set is unique (see §34.3.2) so we are justified in saying *the* least upper bound, rather than *a* least upper bound.

We define the *lower bound* and *greatest lower bound* or *infimum* similarly.

LU (Least upper bound property)**.** *A set, R, has the least upper bound property if for every subset $S \subseteq R$, if S is non-empty and has an upper bound, then S has a supremum in R.*[*]

The least upper bound of a set S, if it exists, is written $\sup S$. Similarly, the greatest lower bound is written $\inf S$, and is equal to $-\sup\{-s : s \in S\}$ (proof §34.3.2). A consequence of the least upper bound property is that any non-empty subset of the reals that has a lower bound also has a real infimum. Within the real numbers, neither the supremum nor infimum are defined for empty or unbounded sets.

The infimum and supremum of a set S do not have to lie within the set. For example, if $S = \{x \in \mathbb{R} : x < 0\}$, then $\sup S = 0$, but $0 \notin S$.

The rationals do not satisfy the least upper bound property. For example, the set $\{x \in \mathbb{Q} : x^2 < 2\}$ is a subset of the rationals. However, its least upper bound is $\sqrt{2}$, which is not within the set of rationals.

A consequence of having least upper bounds is that real numbers do not get too big or too small.

Theorem 11.4.1 (Archimedean property)**.** $\forall x{,}y \in \mathbb{R} : 0 < x < y \to \exists n \in \mathbb{N} : n \cdot x > y$*: For any two real numbers $0 < x < y$, there exists a natural n such that $n \cdot x > y$.*

Proof. Suppose there exists two real numbers $0 < x < y$ such that $nx \leq y$ for all $n \in \mathbb{N}$. Then, $n \leq \frac{x}{y}$ for all $n \in \mathbb{N}$, so $\frac{x}{y}$ is an upper bound of the naturals. From the least upper bound property, it follows that there exists $z = \sup \mathbb{N}$. The number $z - 1$ is less than z, so it is not an upper bound of $\mathbb{N}$, so there exists natural numbers such that $n > z - 1$. However, this implies that $n + 1 > z$, contradicting that z is the supremum of $\mathbb{N}$. It follows that our original assumption is false. ■

This property excludes *infinitesimals* – non-zero numbers that are smaller than every positive rational number – from existing. It also prevents the existence of a real number that is larger than every positive rational number from existing.

Completeness is discussed further in the chapter for real analysis, in §34.3

[*] This one is a mess symbolically: a set, R, has the least upper bound property if,

$$\forall S \subseteq R : [S \neq \varnothing \wedge (\exists x \in R : \forall s \in S : x \geq s)] \to \exists y \in R : \forall s \in S : y \geq s \wedge (\forall x \forall s \in S\ : x \geq s \leftrightarrow y \leq x)$$

11.4.1.1 Existence of Real Roots

Using the least upper bound property, we can prove many theorems about real numbers. For instance, we can prove that every positive number, x, has a unique positive nth root. That is, for every positive x, there exists $r > 0$ such that $r^n = x$, and we write $x^{\frac{1}{n}} = \sqrt[n]{x} = r$ to represent this number.

In fact, without the least upper bound property (or another axiom equivalent to completeness), we can't prove this at all, since our number system would then only consist of the rational numbers, where positive roots do not always exist.

Theorem 11.4.2. *Let $x \in \mathbb{R}^+$ be a positive real number. For each natural number, $n \geq 1$, there exists $r \in \mathbb{R}^+$ such that $r^n = x$. That is, every positive real number has a unique positive real nth root.*

Proof. Given x, let $S = \{s \in \mathbb{R}^+ : s^n \leq x\}$.

Consider the number $t = \frac{x}{x+1}$. Note that $t < 1$ and $t < x$. From the former, it follows that $t^{n-1} < 1$, so $t^n < t$, and from the latter, we have $t^n < t < x$, so $t \in S$ and S is non-empty.

Now, consider $u = x + 1$. Now, $u > 1$ and $u > x$, so $u^{n-1} > 1$ and $u^n > u$, similarly giving $u^n > u > x$, so u is an upper bound of S.

As S is non-empty, and has an upper bound, it has a supremum by completeness. Let $r = \sup S$. We claim that $r^n = x$. We will prove this with trichotomy, and showing that the other two options, $r^n < x$ and $r^n > x$ lead to contradictions.

First, a lemma.

Lemma 11.4.3. $\forall a,b \in \mathbb{R} : 0 < a < b \to \forall n : b^n - a^n < (b - a)nb^{n-1}$.

Proof. If $0 < a < b$, then $a^k < b^k$ for all k. Then,

$$\begin{aligned} b^k - a^k &= (b-a)(b^{k-1} + b^{k-2}a + \cdots + ba^{k-2} + a^{k-1}) \\ &< (b-a)(\underbrace{b^{k-1} + b^{k-1} + \cdots + b^{k-1} + b^{k-1}}_{k}) \\ &< (b-a)kb^{k-1} \end{aligned}$$

■

Now, suppose $r^n < x$. Let $h < 1$ be a real number such that

$$0 < h < \frac{x - r^n}{n(r+1)^{n-1}}$$

h always exists, because $r^n < x$ and $x - r^n > 0$ is positive, and the denominator is the product of positive terms, so the whole expression is positive.

Now, we apply the lemma with $a = r$ and $b = (r + h)$, and choose $k = n$, giving,

$$\begin{aligned} (r+h)^n - r^n &< hn(r+h)^{n-1} \\ &< \left(\frac{x - r^n}{n(r+1)^{n-1}}\right) n(r+h)^{n-1} \\ &< \left(\frac{x - r^n}{n(r+1)^{n-1}}\right) n(r+1)^{n-1} \\ &< x - r^n \\ (r+h)^n &< x \end{aligned}$$

so $(r+h) \in S$. However h is positive, so $r < r+h$, contradicting that r is an upper bound of S. It follows that the assumption that $r^n < x$ is false.

Instead, suppose $r^n > x$, and let

$$h = \frac{r^n - x}{nr^{n-1}}$$

We note that,

$$r > \frac{r}{n} = \frac{r^n}{nr^{n-1}} > \frac{r^n - x}{nr^{n-1}} = h > 0$$

Again, we apply the lemma, this time with $a = r - h$ and $n = r$, and choose $k = n$, giving,

$$\begin{aligned} r^n - (r-h)^n &< hnr^{n-1} \\ &< \left(\frac{r^n - x}{nr^{n-1}}\right) nr^{n-1} \\ &< r^n - x \\ -(r-h)^n &< -x \\ x &< (r-h)^n \end{aligned}$$

so $(r-h)$ is an upper bound of S. However, since $h > 0$, $r - h < r$, contradicting that r is the supremum. It follows that the assumption that $r^n > x$ is false.

Since $r^n \not< x$ and $r^n \not> x$, it follows that $r^n = x$ by trichotomy. ■

11.4.2 Arithmetic

It is technically possible to show that all the arithmetic operations and algorithms taught at school all work in $\mathbb{R}$, purely working from the axioms.

This is, however, extremely tedious, and even rather difficult. For example, even proving that 1 is a positive number takes a bit of work.*

Just as we mostly skipped over the construction of the rationals in the introductory chapter for set theory, we'll similarly skip over this process here and assume that all the algorithms and methods previously learnt all work. For instance, we don't need to use the definition of multiplicative inverses and distributivity to conclude that $\frac{1}{2} + \frac{1}{3} = \frac{5}{6}$.

11.4.3 Algebraic Closure

In each of the number systems we have constructed so far, $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$, there exist polynomials with coefficients written in those sets which have roots that lie outside of those sets.

As some specific examples,

$$\begin{aligned} x + 1 &= 0 && \text{Coefficients in } \mathbb{N}\text{, but roots in } \mathbb{Z} \\ 2x - 1 &= 0 && \text{Coefficients in } \mathbb{Z}\text{, but roots in } \mathbb{Q} \\ x^2 - 2 &= 0 && \text{Coefficients in } \mathbb{Q}\text{, but roots in } \mathbb{R} \\ x^2 + 1 &= 0 && \text{Coefficients in } \mathbb{R}\text{, but roots in } \mathbb{C} \end{aligned}$$

But we stop here. This is because the field of complex numbers, $\mathbb{C}$ is *algebraically closed*: if you write a polynomial with complex coefficients, all the roots will be at most, complex.

* Suppose $1 \leq 0$. Then, by Theorem 11.3.8, $1 \cdot 1 \geq 0 \cdot 1$, so $1 \geq 0$. Since $1 \neq 0$ by the axiom of non-degeneracy, $1 \leq 0$ and $1 \geq 0$ cannot both hold, contradicting the axiom of antisymmetry. It follows that $0 < 1$, so 1 is a positive number.

11.5 Algebraic Structures

A slight distinction should be made here. The “real numbers” can both refer to the set of numbers themselves, and the ordered field, as we have constructed them in the previous chapter.

The real numbers, as in the ordered field, is an example of an *algebraic structure*, or just an *algebra.*[*] An algebraic structure consists of a non-empty set called the *underlying set*, *carrier set* or *domain*, a collection of operations on the set of finite arity, typically binary, and a finite set of identities, or axioms, that these operations must satisfy. In the case of the real numbers, the set is the set of real numbers, the operations are 0,[†], 1, +, and $\cdot$ and the axioms are as defined previously.

A *substructure* is a structure whose domain is a subset of a bigger structure, and whose functions and relations are also restricted to this domain. We can also call the main structure an *extension* or *superstructure* of its substructure. A *subalgebra* is a substructure that is closed.

Substructures inherit all the properties of their superstructures that don't depend on the existence of specific elements in the domain of the superstructure that aren't otherwise specified to exist by the operations or axioms. For example, any general substructure of the ordered field of real numbers will still have; addition and multiplication being commutative and associative; multiplication distributing over addition; and 0 and 1 are still identities, but other axioms may start to fail.

Some notable substructures of the reals are:

- {0}. If we don't specify that $0 \neq 1$, but include both addition and multiplication, we get an algebraic structure called the *trivial ring.* If we also only include one of the operations we get a different algebraic structure called the *trivial group.*
- {0,1}. If we do require that $0 \neq 1$, and include only multiplication, we get an algebraic structure called a *semigroup*. A semigroup is a set equipped with an associative operation, and no other requirements. It turns out that there are only 5 structurally distinct semigroups on two element sets. This one is called semigroup ({0,1},$\wedge$).
- The natural numbers $\mathbb{N}$. If you specify your structure to have 0, 1, and addition, closure then forces you to add the rest of the naturals numbers into your set, as addition gives 0,1,$1+1$,$1+1+1,\ldots$, forming the naturals.

  Additive and multiplicative inverses do not exist, but the naturals are still totally ordered.

  The natural numbers along with addition and multiplication is a type of algebraic structure called a *semiring.*

  The naturals with just addition or just multiplication is a structure called a *commutative monoid*. A monoid is a set equipped with an associative operation that has identity elements – in other words, a semigroup, but with identities; or, a group, but without inverses.
- The integers $\mathbb{Z}$. We get this structure by additionally requiring additive inverses, so closure forces us to get the negative integers by repeatedly adding the inverse of 1 to itself, as $-1, -1+(-1), -1+(-1)+(-1),\ldots$

  We still don't have multiplicative inverses, but the order axioms are still satisfied.

  The integers along with addition and multiplication is a *commutative ring.* A ring is a generalisation of a field which doesn't require multiplication to be commutative (in this case, our multiplication does happen to be commutative, hence the *commutative* ring), and doesn't require multiplicative inverses. The integers with just addition is a *group.* Groups and rings are discussed in more detail in §12.

[*] We will not use this second name, because there is an algebraic structure similar to a vector space called an “algebra”. We will continue using “algebraic structure”.

[†] Again, this is similar to how we defined constants in logic as zero-arity function symbols (§2.3.2).

- The dyadic rationals $\mathbb{D}$. These are numbers of the form $m2^{-n}$ with integer m and n – rational numbers where the denominator are a power of 2.

  These numbers aren't used as much in maths outside of very specialised and advanced topics, but they are important for computer science, because almost all numbers are represented like this in computers, given their binary nature.

  Like the integers, multiplicative inverses still don't exist, so the dyadic rationals are also a (commutative) ring. This also makes it an overring of the integers, or equivalently, the integers are a subring of the dyadic rationals.

- The rationals, $\mathbb{Q}$. If we insist on multiplicative inverse, we get the numbers that can be written as $\frac{p}{q}$ with integer p and q. Unless we put a few restrictions on p and q, we actually get a few duplicates, since these representations are not unique.

  The rationals form an ordered field, like the reals, which makes it a *subfield* of $\mathbb{R}$ – and is in fact the smallest such subfield. It can also be called the *field of fractions of* $\mathbb{Z}$, which is a notion formalised in §12.13.2.

One issue that should be addressed is that the natural numbers as we defined in §4.5 are not elements of $\mathbb{R}$ as defined in terms of Dedekind cuts. The former definition of natural numbers are finite ordinals while the latter are downward-closed sets of rationals, themselves expressed as ordered pairs of $\mathbb{N} \times \mathbb{N}$. Similarly, the integer elements of $\mathbb{Q}$ are ordered pairs of the form $(n,1)$ with $n \in \mathbb{N}$, rather than elements of $\mathbb{N}$ itself. Additionally, the Peano construction of the naturals builds them up from 0 and repeated applications of a successor operation S, without any sets in sight. So how can we say things like $\mathbb{N} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}$?

We resolve this with *isomorphisms* (§12.4.1). The idea is that there are bijections between each different method of construction that preserve the behaviour of 0, 1, $+$, and $\cdot$, and that's all that really matters in terms of the *structure* of these sets – if they all behave in exactly the same way in every way that matters for our purposes, it doesn't really make sense to distinguish them by their *composition*. Otherwise, it would be like saying two sentences are only the same in meaning if they are written in the exact same font – the differences with respect to semantics are only superficial in nature.

$$\begin{array}{l}
\{\varnothing,\{\varnothing\},\{\varnothing,\{\varnothing\}\},\{\varnothing,\{\varnothing\},\{\varnothing,\{\varnothing\}\}\},\ldots\} \\
\{(0,1),(1,1),(2,1),(3,1),\ldots\} \\
\{\{(p,q) : p < 0\},\{(p,q) : p < 1\},\{(p,q) : p < 2\},\{(p,q) : p < 3\},\ldots\} \\
\{0,S0,SS0,SSS0,\ldots\} \\
\\
\qquad\quad \Updownarrow \\
\\
\{0, \quad 1, \quad 2, \quad 3, \quad \ldots\}
\end{array}$$

These sets are all isomorphic.

Isomorphisms are an integral part of abstract algebra, and mathematics as a whole. In fact, they're why all of these axioms exist: if we can prove some random object follows a set of axioms, then every theory built from those axioms will apply to that object. And conversely, if we want to prove a new result, we do it in terms of axioms, or objects that are defined entirely in terms of those axioms. Doing it this way, anyone else can use your theorems, provided they can prove that their object follows your axioms, or the axioms of the objects you used to build your theorem.

So, “which of the sets above is *really* the natural numbers?” isn’t really the right question. None of them are. The natural numbers are a set of axioms, which can be exhibited by various distinct – but equivalent – structures.

Chapter 12

Introduction to Abstract Algebra

"There is no branch of mathematics, however abstract, which may not some day be applied to phenomena of the real world."

— Nikolai Lobachevsky, *The Foundations of Geometry*

We begin this chapter with an exploration into the algebraic structures known as *groups*. It is recommended that you read the immediately preceding chapter (or at least, the final section of the last chapter) before reading this chapter, as the final section provides some context for the study of abstract algebra.

12.1 Introduction

12.1.1 Groups as Symmetries

When we say that a face is symmetric, we mean that you can reflect it across a vertical line, and the resulting face looks the same as the original. A symmetry describes any type of transformation, an *action*, that can be performed on an object such that the object is *invariant* in some way.

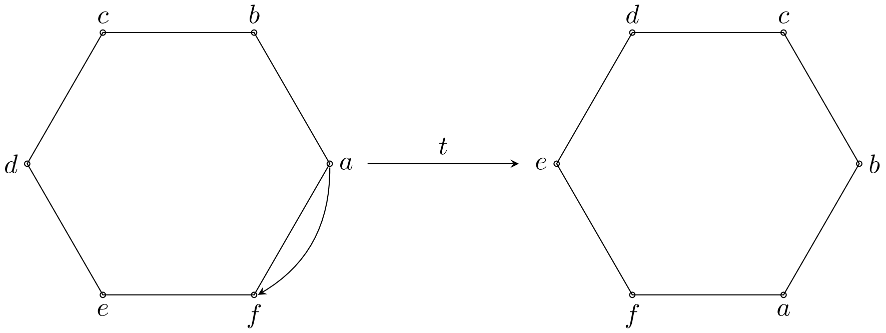
But moving on, something like a hexagon is also symmetric, but in more ways. We can rotate it in 5 distinct ways, and reflect it in 6. Even something like an line has translational symmetry. Infinitely many, in fact.

The set of these actions on an object, is a *group* (kind of). The fact that such a generic name is reserved for this rather seemingly specific type of collection hints at just how significant and fundamental they are.

For a face, we take the reflection action, and the *identity* action of doing nothing, and we have a group called C_2. A hexagon, we take the 5 rotations, 6 reflections, and again, the identity, and we have a group called D_6.

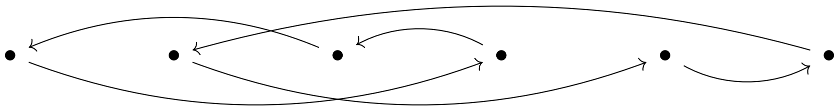
Now, when we said that the object has to be invariant under the action, we didn't really define precisely what structure has to be invariant. That's because this definition can vary, resulting in different groups.

For D_6, we only allow rigid transformations of the hexagon. We could be more restrictive, and say we only allow rotations; we care about the orientation of the hexagon. This smaller collection of only 6 actions also forms a group, called C_6.



An example of an action of C_6 and D_6 acting on 6 points,
preserving the hexagonal structure of the points.

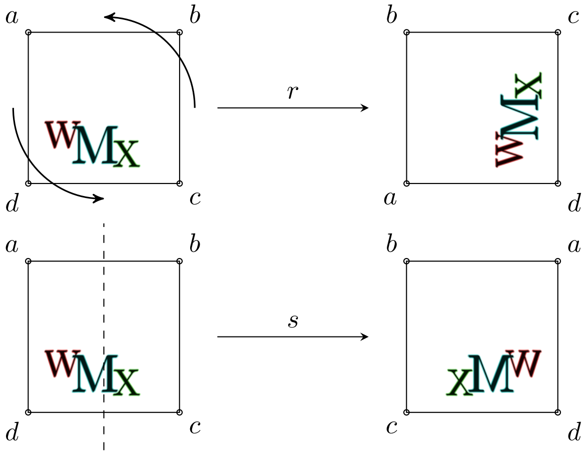
But we could be looser with our definition of invariance. The most general (or rather, lack of) structure we could have, is to simply consider the ways we can rearrange the six vertices of the hexagon without actually caring about the hexagon itself; that is, we consider the ways to permute six points amongst themselves. This is a larger group with 720 actions, called S_6.



An example of an action of S_6 acting on 6 points.

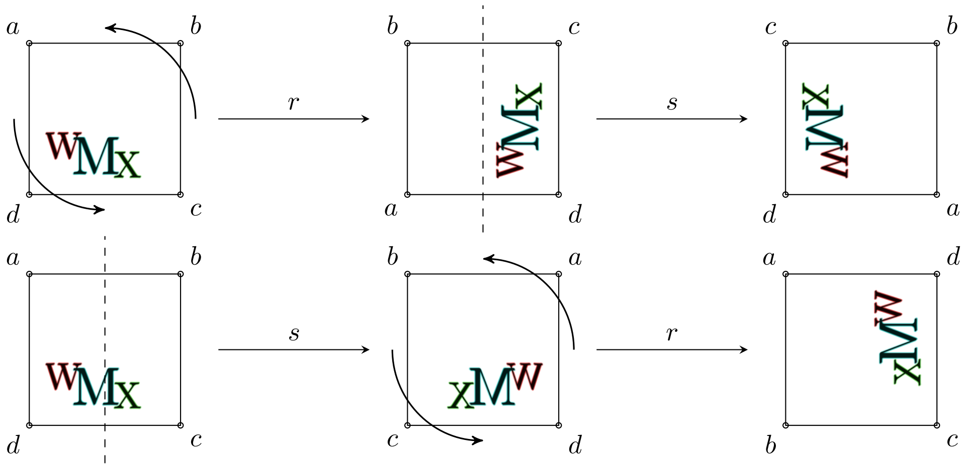
While this is nice and all, what is far more interesting is how we can combine actions together.

Let's simplify our object of consideration down to a square, our group of interest now being D_4. To make things clearer, an asymmetric chiral image has also been placed into the square to help keep track of transformations. Here are two transformations, a 90° anticlockwise rotation and a reflection in the vertical axis, applied to a square.

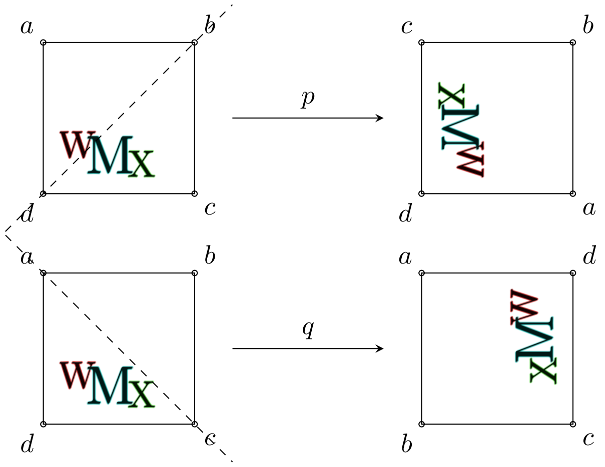


The two actions have been labelled as r and s for convenience.

This little bit of abstraction allows us to do some more interesting things. What happens if we apply them one after the other?



Notice how we get different results depending on the order in which we apply the transformations. That is, these particular transformations are not *commutative*. We also note that final squares can both also be reached in a single transformation of the original square:



So, because these diagonal reflections give the same overall effects as the rotation and vertical reflection, we could say "vertical reflection, plus 90° anticlockwise rotation is the same as a reflection in the upwards diagonal".

We could do this for every possible transformation on a square. For compactness, we do this on a multiplication grid, filling in each square with little diagrams. For the sake of me not having to draw 80 little squares with arrows in LaTeX, this table is omitted.

Instead, we can use the labels, and write,

$$r \circ s = p$$
$$s \circ r = q$$

(We read right to left for composition. This notation stems from function notation; $(r \circ s)(\mathrm{S}) = r(s(\mathrm{S}))$, so we apply s first.)

More generally, we...

12.1.2 Abstraction

Denote the identity transformation as e, rotations of 90°, 180° and 270° as ρ_0, ρ_1 and ρ_2, respectively, and the reflections in the vertical axis, upwards diagonal, horizontal axis and downwards diagonals as, σ_0, σ_1, σ_2 and σ_3, respectively.

Then, we have,

	e	ρ_0	ρ_1	ρ_2	σ_0	σ_1	σ_2	σ_3
e	e	ρ_0	ρ_1	ρ_2	σ_0	σ_1	σ_2	σ_3
ρ_0	ρ_0	ρ_1	ρ_2	e	σ_1	σ_2	σ_3	σ_0
ρ_1	ρ_1	ρ_2	e	ρ_0	σ_2	σ_3	σ_0	σ_1
ρ_2	ρ_2	e	ρ_0	ρ_1	σ_3	σ_0	σ_1	σ_2
σ_0	σ_0	σ_3	σ_2	σ_1	e	ρ_2	ρ_1	ρ_0
σ_1	σ_1	σ_0	σ_3	σ_2	ρ_0	e	ρ_2	ρ_1
σ_2	σ_2	σ_1	σ_0	σ_3	ρ_1	ρ_0	e	ρ_2
σ_3	σ_3	σ_2	σ_1	σ_0	ρ_2	ρ_1	ρ_0	e

with the elements that label the rows, the *nearer factor*, being applied first, and the elements that label the columns, the *further factor*, applied second.

Notice how the identity transformation is in every row and every column. That corresponds to the fact that every transformation can be undone by another action, which seems like a fairly obvious result: we can always undo an action by playing it in reverse, which is just another action. We also note that the identity transformation, combined with anything else just gives that other transformation back.

This *Cayley table* presents all the information we could need about the possible actions on the square, all together in a compact form.

Now, we forget about the square. Forget that we defined ρ_0 as the label for the 90° rotation of a square, and treat each element purely symbolically, as an abstract object in and of itself.

This is analogous to how we write regular multiplication tables. We don't draw n dots in the rows and m dots in the columns, then rectangles of n by m dots to represent their products; we write them purely symbolically using numbers.

You probably even find these symbols easier to deal with than the dots they came from. This abstraction for multiplication, or more generally, for numbers and counts, lets us think about numbers in new and different ways. For instance, if 4×5 comes from adding up 4 sets of 5 objects, what does 1.5×2 mean? Or 3×-12? Or, if exponentiation comes from repeated multiplication, what does $3^{\frac{1}{2}}$ or even $e^{i\pi}$ mean?

The relationship numbers have with counts, and all the associated operations, is very much analogous to groups and the symmetry actions we considered in the previous section. In fact, all of the sets and transformations in the previous sections are not technically groups (though we will continue to refer to them as such just for now), instead being *group actions*. When we talk about groups, we really mean this purely abstract table of relationships of elements, without the underlying object and actions.

When we write "3", we often don't refer to a literal collection of 3 specific objects. The symbol, "3" is just that – an abstract symbol. The symbol isn't really helpful by itself unless we define it in relation to other numbers, like the way it adds or multiplies with other numbers. Again, you could do this all with counts and triplets of things, but most of us are comfortable with just manipulating the symbols. In much the same way, the elements in the table above, to a group theorist, doesn't represent a specific transformation on a square that preserve some given structure. They're symbols, useful only when defined in relation to other symbols, like $\sigma_2 \circ \rho_0 = \sigma_1$. What makes a group, a group, is the way these elements combine with each other.

This point is crucial for a good intuitive understanding of groups. In the next section, we formalise the definition of a group using the group axioms.* Out of context, they seem extremely arbitrary and specific, but, with the knowledge of group actions, they are trivial consequences of this underlying idea of symmetric actions.

12.1.3 Isomorphisms

We said earlier that using numbers instead of counts lets us do more interesting things. But what can we do with abstract groups over group actions?

Consider the group of rotations on a cube, and the group of permutations on 4 points.

These groups, at first, might seem very different. The former, you could think of as a set of rotations acting upon 8 vertex points in three dimensions in such a way that preserves the distance and orientation structure between all of them. The latter, we have no structure at all being preserved on just 4 points.

It turns out, however, that these two groups are the same, in the sense that their Cayley tables are identical. Anything you can say about one of these groups, will also apply to the other. For example, there are 8 distinct permutations that cycle 3 elements, so you get back to the identity after three applications of that permutation. There are also 8 rotations of the cube which have this property of returning to the identity after 3 applications. If you want to explore this connection a bit more, try considering the 4 inner diagonals of the cube (§12.6.2).

We say that the group of rotations of a cube and the permutation group of 4 points are *isomorphic*.

More formally, two groups are isomorphic if there is a bijective map between the elements of the first group and the second that *preserves the group operation*, somewhat similarly to order-types of sets being equivalent if there is a bijective map that preserves order.

In this case, preserving the group operation just means that there exists a map such that, if we compose two rotations of a cube, a and b, to get c, then composing the matching permutations a' and b' gives c', for all possible choices of a and b.

Now, the group of permutations seems a lot easier to deal with than the group of rotations of a cube. We can store each permutation as a list of 4 numbers, and drawing each permutation is a lot easier than a cube rotating, especially when composing them together.

Because abstract groups and isomorphisms don't represent the symmetries of a specific object, instead representing an abstract way that things can even be symmetric, groups come up in lots of places that don't immediately bring symmetry to mind. Similar to how vector spaces can be useful anywhere you have some notion of adding and scaling some objects, groups are often useful anywhere you have some notion of multiplying two things together to get a third. Abstraction is discussed in more detail in §33.6.

Group isomorphisms let us prove powerful and very general results about a wide variety of groups, by proving they are isomorphic to others.

For example, one proof for the insolubility of the quintic, the *Abel-Ruffini theorem*, relies on group theory.

Recalling the factor theorem, we can rewrite a polynomial in terms of its roots:

$$\begin{aligned} ax^5 + bx^4 + cx^3 + dx^2 + ex + f = 0 \\ (x - r_0)(x - r_1)(x - r_2)(x - r_3)(x - r_4) = 0 \end{aligned}$$

We can permute the order of these brackets without changing the equation itself, so these permutations on the roots of a quintic form a group isomorphic to S_5.

* This usage of the word "axiom" is different from the axioms in symbolic logic and set theory. There, an axiom meant a statement that is assumed to be true. Here, it just means a list of rules that define a type of object. If we can prove that something follows those rules, then all the theorems of group theory will apply.

Similar to how integers break down into products of primes, we have ways of breaking down groups into products of smaller, indivisible *simple* groups, though this is far beyond the scope of this document.

If a permutation group decomposes into the product of certain groups, C_n (we met one of these earlier!), then a formula for the roots of the polynomial can exist. Quadratics, cubics and quartics all do this, and general formulae for the roots of those polynomials do indeed exist.

However, S_5 has a different type of group in its decomposition – one which can never be made from polynomial solutions built from elementary functions, proving that a quintic formula *cannot exist.* In fact, the roots of almost all quintics cannot be written in closed form using elementary functions.

Obviously a massive amount of detail is being glossed over, but the point is, we can prove an extremely obscure fact about polynomials by determining the structure of the prime decomposition of a group.

And even more on abstraction, note that we've really abstracted the word "symmetry" as well. In common parlance, it usually just means a line you can reflect an object over or a point to rotate around to make it "look the same". But here, it's just any type of transformation that preserves some property of some object. This might seem like mathematicians have taken a perfectly descriptive word, and generalised it until it is meaningless outside of this application, but this turns out to be a very helpful idea in general. Bijections, isomorphisms, homeomorphisms and diffeomorphisms all fit the definition of symmetry, allowing us to apply theorems about symmetries to this wide range of transformations.

12.1.4 Symmetries & Conservations

You might not be surprised that groups, being fundamentally about symmetries, apply widely in physics. *Noether's theorem*, says that every conservation law corresponds to some kind of symmetry – to some kind of group.

Remember when we said that a symmetry is just any transformation that preserves some kind of invariant? Well, we can take energy to be our invariant, and consider a system to have a symmetry under a transformation if the total energy of the objects in the system remains the same.

Noether's theorem tells us that spatial translational symmetry corresponds to conservation of momentum, rotational symmetry with angular momentum, and temporal translational symmetry to conservation of energy. Using this, we can easily determine whether a given system will conserve some given quantity.

For example, consider a bunch of particles all travelling through space. A shifted version of the same system of particles with the same velocities has the same energy, so this system is symmetric with respect to translation. Noether's theorem then tells us that this system of particles as a whole will conserve momentum, regardless of whether they collide with each other or not. Shifting a particle orbiting the Earth in its orbit doesn't change its energy state either, so we know that angular momentum is preserved. Shifting an object closer to the Earth, however, changes its gravitational potential energy, so we know that momentum is *not* conserved when dealing with a gravity field. The temporal translational symmetry is harder to demonstrate, but has many applications in quantum mechanics.

And it isn't just these conservations: the conserved quantities are "generators" of the transformation, and we can calculate what generator gives any given transformation. If you find some new exotic system, and discover that it is symmetric with respect to some transformation, Noether's theorem allows you to calculate some quantity that is being conserved in that system.

12.2 Terminology for Groups

Before we move onto groups proper, there is some preamble and background necessary to get out of the way first. We also recall some terminology from buried within previous sections.

12.2.1 Sets

- A *set* is a collection of *elements.*
- The *empty set*, denoted $\varnothing$, is the set with no elements.
- You should be familiar with $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ and $\mathbb{R}^n$ and related operations over those sets.
- $\mathbb{R}[x]$ is the set of polynomials with real coefficients. $\mathbb{Z}[x]$, etc., are defined similarly.

See §4 or §6 for a more introductory discussion of sets.

12.2.1.1 Binary Operations

Given a set S, a *binary operation* on S is a function that takes two elements of S, called the *operands* or *arguments* of the operation, and returns another element of S: it is *closed* over S. That is, it is a binary function $S \times S \to S$.

If $*$ is some binary operation on S we say that $*$ is,

- *commutative* on S if $a * b = b * a$ for all $a,b \in S$;
- *associative* on S if $(a * b) * c = a * (b * c)$ for all $a,b,c \in S$.

If a binary operation is commutative, its Cayley table over a set is symmetric across the diagonal.

12.2.1.2 Functions

- Given two sets, X and Y, a *function*, f, maps unique elements from X to Y. This is written as $f : X \to Y$. X is the *domain* of f, and Y is the *codomain* of f.
- Two functions, $f : X \to Y$ and $g : A \to B$ are equal if $X = A$, $Y = B$ and $f(x) = g(x)$ for all $x \in X$.
- Let A, B and C be sets, and $f : A \to B$, $g : B \to C$ be functions. The *composition* of f and g, written $g \circ f$ is defined as $g(f(x))$. Note that the function on the right of the composition is applied first, as per function notation.
- Composition is an associative operation.
- A function, $f : X \to Y$, is *injective* if, for all $a,b \in X$, $a \neq b \to f(a) \neq f(b)$, or equivalently, $f(a) = f(b) \to a = b$.
- A function, $f : X \to Y$, is *surjective* if for all $y \in Y$, $\exists x \in X$ such that $f(x) = y$.
- A function is *bijective* if it is both injective and surjective.

Theorem 12.2.1. *A function is invertible if and only if it is bijective.*

Proof. Let $f : X \to Y$ be a function, and let g be the inverse of f. Let $a,b \in X$ such that $f(a) = f(b)$. Then, $g(f(a)) = g(f(b)) \to a = b$, so f is injective. Now, let $y \in Y$. As g is the inverse of f, $g(y)$ is a member of X. But $f(g(y)) = y$, so y has an origin element in X, so f is surjective. It follows that f is bijective, completing the forward direction.

Now, let $f : X \to Y$ be a bijective function. As f is bijective, $\forall y \in Y, \exists x \in X$ such that $f(x) = y$. Define $g(y) = x$.

Let $x \in X$ so $f(x) = y \in Y$. Then, $g(f(x)) = g(y) = x$. Let $y \in Y$. Then, $g(y) = x \in X$, so $f(g(y)) = f(x) = y$. It follows that g is the inverse of f, completing the backwards direction. ■

12.2.2 Matrices

We assume some background knowledge of matrix algebra. See §33 for a more introductory approach to matrices.

Matrix arithmetic is as usual:

- Addition is associative and commutative;
- Multiplication is associative and non-commutative;
- Multiplication distributes over addition.

Theorem 12.2.2. *For any square matrices* $\mathbf{A}$ *and* $\mathbf{B}$*,*

$$\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$$

Proof. §33.2.3 ■

Theorem 12.2.3. *If* $\mathbf{A}$ *is invertible, then it has non-zero determinant.*

Proof. $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$, and $\mathbf{I}$ has determinant 1, which is non-zero, so $\mathbf{A}$ and $\mathbf{A}^{-1}$ must both have non-zero determinant. ■

Theorem 12.2.4. *If* $\mathbf{A}$ *has non-zero determinant, then it is invertible.*

Proof. Exercise.

(Method) Multiply a generic matrix by its generic inverse and show that you get the identity. ■

12.3 Group Axioms

A *group*, $(G,*)$ is a set, G, equipped with a binary operation, $* : G \times G \to G$, that obeys the following axioms:

- $\forall a,b \in G, a * b \in G$ (closure);
- $\forall a,b,c \in G, a * (b * c) = (a * b) * c$ (associativity);
- $\exists e \in G$ such that $\forall a \in G, a * e = e * a = a$ (existence of identity);
- $\forall a \in G, \exists(a^{-1}) \in G$ such that $a * (a^{-1}) = (a^{-1}) * a = e$ (existence of inverses).

We can also write id_G for the identity for clarity (and also to mark which group the identity is from if multiple groups are being considered). If the operation is additionally commutative, that is, $\forall a,b \in G, a * b = b * a$, then the group is *abelian*.

Recalling our group action perspective from the first section, you can see where all of these properties come from. Performing one action followed by another is just another action, giving closure. Associativity is a trivial property as well; doing action a, then (b and c), is clearly the same as doing (a and b), then c. Similarly, doing nothing always preserves your structure, since your object already has to have the structure in the first place, and you can always undo an action just by playing it in reverse, giving identities and inverses.

Example.

- $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, any ring R, and any field K form an (abelian) group under addition.
- The set of non-zero elements of a field K, $K^* = K \setminus \{0_K\}$, forms a group under multiplication.

△

We define multiplicative notation for groups as follows:

- The group operation is omitted, so $a * b$ is written as ab;
- The identity is often written as 1 or 1_G, instead of e or id;
- If $n \in \mathbb{N}$, then $a^n = \underbrace{a * a * \cdots * a}_{n}$;
- If $n = 0$, $a^n = 1_G$;
- If n is a negative integer, $a^n = (a^{-}n)^{-1}$;
- $(a^{-1})^n = a^{-n}$;
- $(a^m)^n = a^{mn}$;
- $(a^m)(a^n) = a^{m+n}$;
- If the group is abelian, $(ab)^n = (a^n)(b^n)$.

Abelian groups are more commonly written in additive notation:

- The group operation is written as $+$;
- The identity is often written as 0 or 0_G, instead of e or id;
- If $n \in \mathbb{N}$, then $na = \underbrace{a + a + \cdots + a}_{n}$;
- If $n = 0$, $na = 0_G$;
- The inverse of g is written as $-g$ instead of g^{-1}.
- If n is a negative integer, $na = -n(-a)$;
- $n(-a) = -na$;
- $n(ma) = (m \times n)a$;
- $(ma) + (na) = (m + n)a$;
- If the group is abelian, $n(a + b) = na + nb$.

12.3.1 Basic Properties

Theorem (Cancellative Property). *Let G be a group and let a,b,$g \in G$. Then,*

(i) $ga = gb \to a = b$;

(ii) $ag = bg \to a = b$.

Proof. For (i),

$$\begin{aligned} ga &= gb \\ g^{-1}(ga) &= g^{-1}(gb) && \text{[Existence of inverses]} \\ (g^{-1}g)a &= (g^{-1}g)b && \text{[Associativity]} \\ \mathrm{id}_G\, a &= \mathrm{id}_G\, b \\ a &= b && \text{[Identity]} \end{aligned}$$

(ii) is proved similarly by right multiplying by g^{-1}. ■

In future proofs, we will omit brackets and not explicitly refer to associativity in the interest of space.

Lemma (Uniqueness of Identity)**.** *The identity of a group is unique.*

Proof. Suppose e and f are identities of a group, G. $ef = e$, as f is the identity. But $ef = f$, as e is also the identity, so $ef = e = f$, so $e = f$ and the identity is unique. ■

Lemma (Uniqueness of Inverse)**.** *Every element of a group has a unique inverse.*

Proof. Suppose a and b are both inverses of g, so $ga = \mathrm{id}_G = gb$. By the cancellative property, $a = b$. ■

Lemma (Two-Sided Identity)**.** *If e_ℓ is a left identity for a group G – that is, $e_\ell g = g$ for all $g \in G$ – and e_r is a right identity for G, then $e_\ell = e_r = \mathrm{id}_G$.*

Proof. $e_\ell e_r = e_r$ as e_ℓ is a left identity, and $e_\ell e_r = e_\ell$ as e_r is a right identity, so $e_\ell = e_\ell e_r = e_r = \mathrm{id}_G$. ■

Lemma (Two-Sided Inverse)**.** *If ℓ is a left inverse for an element g – that is, $\ell g = \mathrm{id}_G$ – then ℓ is the (two-sided) inverse of g. Similarly, if r is a right inverse for g, then it is a (two-sided) inverse of g.*

Proof. $\ell g = \mathrm{id}_G$ as ℓ is a left inverse of g, so,

$$\begin{aligned}
\ell g &= \mathrm{id}_G \\
\ell g &= g^{-1}g \\
\ell g g^{-1} &= g^{-1}gg^{-1} \\
\ell &= g^{-1}
\end{aligned}$$

As the choice of ℓ was arbitrary, all left inverses of g are equal. The proof for right inverses is similar. ■

Theorem (Distribution of Inversion)**.** *For all $a,b \in G$, $(ab)^{-1} = b^{-1}a^{-1}$.*

Proof.

$$\begin{aligned}
(ab)^{-1}ab &= \mathrm{id}_G \\
(ab)^{-1}abb^{-1} &= \mathrm{id}_G\, b^{-1} \\
(ab)^{-1}a &= b^{-1} \\
(ab)^{-1}aa^{-1} &= b^{-1}a^{-1} \\
(ab)^{-1} &= b^{-1}a^{-1}
\end{aligned}$$

■

Theorem (Involutivity of Inversion)**.** *For all $a,b \in G$,*

$$(a^{-1})^{-1} = a$$

Proof.

$$\begin{aligned}
\mathrm{id}_G &= aa^{-1} \\
\mathrm{id}_G(a^{-1})^{-1} &= aa^{-1}(a^{-1})^{-1} \\
(a^{-1})^{-1} &= a(a^{-1}(a^{-1})^{-1}) \\
(a^{-1})^{-1} &= a\,\mathrm{id}_G \\
(a^{-1})^{-1} &= a
\end{aligned}$$

■

Theorem 12.3.1. *For all $a \in G$ and all $n \in \mathbb{N}$, $a^n \in G$.*

Proof sketch. $a^0 = 1$ by the definition of a^0, which is in G by the existence of identity axiom. For positive n, induct on n using closure, and for negative n, use the previous result combined with the existence of inverses axiom. ■

12.3.2 Order

Let $(G, *\,)$ be a group. The cardinality of the underlying set G is called the *order* of the group, denoted $|G|$.

Let $g \in G$. The *order* of g, denoted $|g|$ or $o(g)$ is the least integer $n > 0$ such that $g^n = \mathrm{id}_G$. If no such n exists, then g has *infnite order* and we write $|g| = \infty$.

Note that if g has infinite order, then $g^i \neq g^j$ for all $i \neq j$, or else if $g^i = g^j$ for some $i < j$, then $g^{j-i} = \mathrm{id}_G$ by the cancellative property, so the order of g divides $j - i$ and is hence finite. Similarly, if g has finite order n, then $g^i \neq g^j$ for all $i \neq j \in [0,n]$.

Lemma 12.3.2. *$|g| = 1$ if and only if $g = \mathrm{id}_G$.*

Proof. $\mathrm{id}_G^1 = \mathrm{id}_G$. Conversely, for all $\mathrm{id}_G \neq g \in G$, $g^1 = g \neq \mathrm{id}_G$. ■

Lemma 12.3.3. *If $|g| = n$, then $g^k = 1$ if and only if $n|k$.*

Proof. Suppose $n \nmid k$, so $k = qn + r$ for some $q,r \in \mathbb{N}$ with $0 < r < n$ (by the division algorithm). Then,

$$\begin{aligned} g^k &= g^{qn+r} \\ &= (g^n)^q g^r \\ &= \mathrm{id}_G^q\, g^r \\ &= g^r \end{aligned}$$

and since $0 < r < n$, $g^r \neq g^n = \mathrm{id}_G$.

Conversely, suppose $n|k$ so $k = qn$. Then,

$$\begin{aligned} g^k &= g^{qn} \\ &= (g^n)^q \\ &= \mathrm{id}_G^q \\ &= \mathrm{id}_G \end{aligned}$$

■

Theorem 12.3.4. *For every $g \in G$, $|g|$ divides $|G|$.*

Proof. Follows from Lagrange's theorem (§12.4.10). ■

Theorem 12.3.5. *If $|G| = n$, then $g^n = 1$ for all $g \in G$.*

Proof. Let a be the order of g, so $g^a = 1$. By Lagrange's theorem, a divides n, so $n = ab$ for some integer b. So, $g^n = g^{ab} = (g^a)^b = 1^b = 1$. ■

12.3.3 Subgroups

Let $(G,*)$ be a group, and let H be a subset of G. Furthermore, suppose that $(H,*)$ is also a group. $(H,*)$ is then a *subgroup* of $(G,*)$, and we write $H \leq G$ to denote this relation.

To show that a subset $H \subseteq G$ is a subgroup of G, it suffices to show that H is non-empty, is closed under $*$, and that every element has an inverse in H.

Theorem (Two-Step Subgroup Test)**.** *If $(G,*)$ is a group and $H \subseteq G$, then $(H,*)$ is a subgroup of G if and only if,*

(*i*) $H \neq \varnothing$*;*

(*ii*) $a,b \in H \to a * b \in H$*;*

(*iii*) $a \in H \to a^{-1} \in H$*.*

Proof. Every subgroup H clearly fulfils these three conditions for the forward implication.

For the reverse implication, we verify the four axioms. Closure is given by the condition (*ii*), while associativity is inherited from the main group, as the operation in H is just the restriction of the operation in G. The existence of an inverse element follows from condition (*iii*). The existence of the identity element follows from taking a,b to both be the identity in condition (*ii*), or by taking b to be a^{-1}. ■

The test is named the two-step test because H is often assumed to be non-empty, so the first condition need not be checked.

This suggests a shorter test still:

Theorem (One-Step Subgroup Test)**.** *If $(G, *)$ is a group and $H \subseteq G$, then $(H, *)$ is a subgroup of G if and only if,*

1. $H \neq \varnothing$*;*
2. $a,b \in H \to ab^{-1} \in H$*;*

Proof. Every subgroup H clearly fulfils these three conditions for the forward implication.

For the reverse implication, we verify the four axioms. Associativity is again inherited from the main group.

Since H is non-empty, there exists an element $x \in H$. Taking $a = x$ and $b = x$ gives $x * x^{-1} = \mathrm{id}_G \in H$, so the identity element is in H.

Inverses follow from taking $a = \mathrm{id}_G$ and $b = x$, giving $\mathrm{id}_G * x^{-1} = x^{-1} \in H$.

Let $x,y \in H$. Then, as inverses exist, $y^{-1} \in H$, and so we may take $a = x$ and $b = y$, giving $x * (y^{-1})^{-1} = x * y \in H$, and hence H is closed. ■

Theorem 12.3.6. *The following results hold for all groups:*

(*i*) *The intersection of two subgroups is also a subgroup.*

(*ii*) *The union of two subgroups is generally not a subgroup.*

(*iii*) *The group itself, G, and the trivial group, $\{\mathrm{id}_G\}$, are always subgroups of G.*

Proof. (*i*) Let $H \leq G$ and $K \leq G$. $\mathrm{id}_G \in H$ and $\mathrm{id}_G \in K$, so $H \cap K$ is non-empty as it also contains id_G. Since $H \leq G$, $xy^{-1} \in H$ for all $x,y \in H$, and similarly for K. Suppose $a,b \in H \cap K$ so $a,b \in H$

and $a,b \in K$. Then, $ab^{-1} \in H$ and $ab^{-1} \in K$, and hence $ab^{-1} \in H \cap K$, so $H \cap K$ is a subgroup by the one-step test. ■

Any subgroup not equal to G is a *proper* subgroup, while any subgroup not equal to $\{\mathrm{id}_G\}$ is a *non-trivial* subgroup.

Theorem 12.3.7. *If H is a subgroup of G, and $|G|$ is finite, then $|H|$ divides $|G|$.*

Proof. Corollary of Lagrange's theorem (§12.4.10). ■

12.4 Homomorphisms

A homomorphism between two arbitrary structures is a map that *preserves* the stucture.

Homomorphism sometimes have additional names for specific structures. For example, a homomorphism between sets is called a *function*, a homomorphism between vector spaces is called a *linear transformation*, and a homomorphism between probability spaces is called a *measurable function.* For groups, we simply have *group homomorphisms.*

Suppose that A and B are sets, and that $*$ and $\circ$ are binary operations defined over A and B, respectively. If the map $f : A \to B$ obeys,

$$f(x * y) = f(x) \circ f(y)$$

for all $x,y \in A$, then we say that f *preserves* the operation or is *compatible* with the operation.

More generally, a map $f : A \to B$ preserves operations μ_A and μ_B of arity k defined on A and B if,

$$f(\mu_A(a_1,a_2,\ldots,a_k)) = \mu_B(f(a_1),f(a_2),\ldots,f(a_k))$$

for all $a_1,a_2,\ldots a_k \in A$.

If a map preserves *all* operations over an algebraic structure, then it is a *homomorphism.* Note that this includes nullary functions – that is, constants. For example, if a structure requires an identity element, then the identity element of the first structure must be mapped to the corresponding identity element of the second. For instance, all vector space homomorphisms must preserve the zero element, which is why all linear transformations fix the origin in place. Field homomorphisms must preserve both addition and multiplication operations, and map the additive and multiplicative identities to other additive and multiplicative identities.

Let $(G,\circ)$ and $(H,*)$ be groups. A function, $\phi : G \to H$ is a *group homomorphism* between G and H if $\phi(a \circ b) = \phi(a) * \phi(b)$ for all $a,b \in G$. Note that we do not specifically demand that the group homomorphism preserves the identity in this case, as this is already implied by the structure preserving requirement:

$$\begin{aligned}\mathrm{id}_H * \phi(g) &= \phi(g)\\ &= \phi(\mathrm{id}_G \circ g)\\ &= \phi(\mathrm{id}_G) * \phi(g)\end{aligned}$$

so $\mathrm{id}_H = \phi(\mathrm{id}_G)$ by the cancellative property.

Example. The real numbers have group structure under addition, $(\mathbb{R},+)$, and the positive real numbers have group structure under multiplication, $(\mathbb{R}^+,\cdot)$. The exponential function, $x \mapsto e^x$ satisfies $e^{x+y} = e^x \cdot e^y$, so the exponential function is a group homomorphism from $(\mathbb{R},+)$ to $(\mathbb{R}^+,\cdot)$. △

Group homomorphisms also map inverses to inverses:

Theorem 12.4.1. *If $\phi : G \to H$ is an isomorphism, then $\phi(g^{-1}) = \phi(g)^{-1}$ for all $g \in G$.*

Proof. For all $g \in G$,

$$\begin{aligned} \mathrm{id}_H &= \phi(\mathrm{id}_G) \\ &= \phi(gg^{-1}) \\ &= \phi(g) \cdot \phi(g^{-1}) \end{aligned}$$

so $\phi(g^{-1})$ is the inverse of $\phi(g)$ in H, giving $\phi(g^{-1}) = \phi(g)^{-1}$. ■

Lemma 12.4.2. *If $H \leq G$, then the inclusion map $\phi : H \hookrightarrow G$ defined by $h \mapsto h$ for all $h \in H$ is a homomorphism. If $H = G$, then it is furthermore an (identity) isomorphism.*

Some specific kinds of homomorphisms have a special names. For instance, an injective homomorphism is also called a *monomorphism*, and a surjective homomorphism is called an *epimorphism*. In further abstract algebra, when considering structures too large to be sets (and hence injectivity and surjectivity are not well-defined notions), monomorphisms and epimorphisms are instead defined in terms of left and right cancellative properties.

12.4.1 Isomorphisms

If the inverse of a homomorphism is a homomorphism, or equivalently, if the homomorphism is a bijection,[*] then it is called an *isomorphism*. If an isomorphism exists between G and H, we say that G and H are *isomorphic*, and we write $G \cong H$ to denote this relation. It is easy to check that isomorphism is an equivalence relation.

The Cayley tables for isomorphic finite groups look identical, up to relabelling of variables. In fact, isomorphic objects are completely indistinguishable from the viewpoint of the structure that is being preserved.

Example. In the previous example, the homomorphism between $(\mathbb{R},+)$ and $(\mathbb{R}^+,\cdot)$ given by the exponential function is actually an isomorphism, because the inverse function, the natural logarithm, satisfies $\ln(x \cdot y) = \ln(x) + \ln(y)$, and is also a homomorphism. We then say that $(\mathbb{R},+)$ and $(\mathbb{R}^+,\cdot)$ are isomorphic groups and we write $(\mathbb{R},+) \cong (\mathbb{R}^+,\cdot)$. △

Isomorphisms capture the idea that objects can be functionally equivalent, where this function is just whatever property we care about. For groups, this property is the group structure. Earlier, when we constructed the naturals in several ways, this property is that the numbers obey the Peano axioms or Robinson arithmetic, or whatever. If two sets behave in the same way in every way that matters for whatever we're trying to do, we don't really need to distinguish them, so we can just slap on the label of isomorphic, and prove results in terms of the properties that they share. Otherwise, it would be like saying two sentences are only the same in meaning if they are written in the exact same font – the differences with respect to semantics are only superficial in nature.

Lemma 12.4.3. *If $\phi : G \to H$ is an isomorphism, then $\phi(g^{-1}) = \phi(g)^{-1}$ for all $g \in G$.*

Proof. For all $g \in G$,

$$\begin{aligned} \mathrm{id}_H &= \phi(\mathrm{id}_G) \\ &= \phi(gg^{-1}) \\ &= \phi(g) \cdot \phi(g^{-1}) \end{aligned}$$

[*] This is only equivalent for algebraic structures. For non-algebraic structures, bijectivity and having an inverse homomorphism are not necessarily the same thing. For instance, a homomorphism between topological spaces is a continuous map, but the inverse of a bijective continuous map is not necessarily continuous.

so $\phi(g^{-1})$ is the inverse of $\phi(g)$ in H, giving $\phi(g^{-1}) = \phi(g)^{-1}$. ■

Theorem 12.4.4. *If $\phi : G \to H$ is an isomorphism, then $|g| = |\phi(g)|$ for all $g \in G$.*

Proof. If $|g|$ is infinite, then g^k is distinct for all $k \in \mathbb{Z}$. Then, $\phi(g^k) = \phi(g)^k$ must also be distinct for all $k \in \mathbb{Z}$, so $|\phi(g)|$ is infinite.

Conversely, suppose $n = |g|$ is finite.

$$\begin{aligned} \phi(g)^n &= \phi(g^n) \\ &= \phi(\mathrm{id}_G) \\ &= \mathrm{id}_H \end{aligned}$$

so $|\phi(g)| \leq n = |g|$. Now, let $m = |\phi(g)|$, so

$$\begin{aligned} \phi(g^m) &= \phi(g)^m \\ &= \mathrm{id}_H \\ &= \phi(\mathrm{id}_G) \end{aligned}$$

and since ϕ is an isomorphism, it is injective, so $g^m = \mathrm{id}_G$ and hence $|\phi(g)| = m \leq |g|$. Then,

$$|\phi(g)| \leq |g| \leq |\phi(g)|$$

so $|\phi(g)| = |g|$. ■

12.4.2 Endomorphisms

An *endomorphism* is a homomorphism whose domain and codomain coincide. That is, a homomorphism from an algebraic structure to itself. We don't use these as much in group theory , but they are especially important in linear algebra, where endomorphisms are changes of basis transformations (§33.5), and furthermore, the set of endomorphisms on any algebraic structure X has monoidal structure (or even group or ring structure in certain cases) under composition, denoted $\mathrm{End}(X)$.

Furthermore, the set of endomorphisms of a vector space itself has ring structure. For vector spaces of finite dimension, this ring of endomorphisms is isomorphic to the ring of square matrices of the same dimension – this is exactly what allows us to write every linear transformation as a square matrix (proving this fact is set as an exercise in §33.7).

12.4.3 Automorphisms

An *automorphism* is an endomorphism that is also an isomorphism.

The set of automorphisms of an algebraic structure X has group structure under composition, and is called the *automorphism group* of the structure, denoted $\mathrm{Aut}(X)$. The automorphism group is a subset of the endomorphism monoid.

For example, the general linear group, $GL_n(K)$ can be characterised as the automorphism group of a vector space of dimension n, over a field, K.

12.4.4 Morphisms

More generally, we can consider any kind of mappings between any kind of objects that compose associatively and admit an identity mapping. This is the topic of study of *category theory*, which is discussed in §51.

12.4.5 Cyclic Groups

Let $S \subseteq G$ be a set of elements of G. $H = \langle S \rangle$ is then defined to be the minimal group that contains all of S. That is, there are no subgroups of H that contain every element of S. S is then called the *generating set* of H, or equivalently, we say that H is *generated* by S.

If $S = \{g\}$ is a singleton set, then $H = \langle S \rangle = \langle g \rangle$ is given by $\{g^n : n \in \mathbb{N}\} = \{\cdots,(g^{-2}),g^{-1}, \mathrm{id}_H ,g,g^2,g^3, \cdots\}$. If $g \in G$, then $\langle g \rangle$ is a subgroup of G.

If a group G can be written in this form – that is, $G = \langle g \rangle$ for a single element g – then we say that G is *cyclic*, and that g is the *generator* of G. That is, a group is cyclic if and only if it is generated by a single element (so cyclic groups are a special case of generated groups).

Example. All subgroups of $\mathbb{Z}$ under addition are cyclic, and are of the form $k\mathbb{Z}$. △

Note that generators for a cyclic groups are not necessarily unique. That is, there may exist two distinct elements of G, g and h, such that $\langle g \rangle = \langle h \rangle$. For instance, $\mathbb{Z}$ is generated by both 1 and -1, and $\mathbb{Z}/p\mathbb{Z}$ with p prime is generated by every non-identity element.

Theorem 12.4.5. *Cyclic groups are abelian.*

Proof. Let $G = \langle g \rangle$ and let $a,b \in G$. Then,

$$\begin{aligned} a \cdot b &= g^n \cdot g^m \\ &= \underbrace{g \cdots g}_{n} \cdot \underbrace{g \cdots g}_{m} \\ &= \underbrace{g \cdots g}_{m} \cdot \underbrace{g \cdots g}_{n} \\ &= g^m \cdot g^n \\ &= b \cdot a \end{aligned}$$

by associativity. ■

Theorem 12.4.6. $|\langle g \rangle| = |g|$.

Proof. Obvious from definition. ■

Lemma 12.4.7. *In an infinite cyclic group, every generator has infinite order. In a finite cyclic group of order n, every generator has order n.*

We write C_n for the finite cyclic group of order n.

Theorem (Infinite Cyclic Groups)**.** *Every infinite cyclic group is isomorphic to the group of integers under addition.*

Proof. Suppose $(G,\times)$ is an infinite cyclic group with generator g. Define the map $\phi : (\mathbb{Z},+) \to (G, \cdot)$ by $n \mapsto g^n$.

$$\begin{aligned} \phi(a+b) &= g^{a+b} \\ &= g^a \cdot g^b \\ &= \phi(a) \cdot \phi(b) \end{aligned}$$

so ϕ is a homomorphism. Then, as G has infinite order, so does g and hence $g^a \neq g^b$ for all $a \neq b$, so ϕ is injective. As G is cyclic, every element can be written in the form g^n for some $n \in \mathbb{Z}$, which is exactly the statement of surjectivity for ϕ. It follows that ϕ is an isomorphism. ■

Corollary 12.4.7.1. *Any two infinite cyclic groups are isomorphic.*

Proof. As G was arbitrary, all infinite cyclic groups are isomorphic. ■

Theorem 12.4.8. *Any two cyclic groups of equal order are isomorphic.*

Proof. Let G and H be cyclic groups of finite order k with generators g and h, respectively. Define the map $\phi : G \to H$ by $g^n \mapsto h^n$. This map is clearly bijective by construction.

Let $a,b \in G$. As G is cyclic, $a = g^s$ and $b = g^t$ for some integers s,t.

$$\begin{aligned}\phi(ab) &= \phi(g^s g^t)\\ &= \phi(g^{s+t})\\ &= h^{s+t}\\ &= h^s h^t\\ &= \phi(g^s)\phi(g^t)\\ &= \phi(a)\phi(b)\end{aligned}$$

so ϕ is a homomorphism, and is hence an isomorphism. ■

Theorem 12.4.9. *Cyclic groups are abelian.*

Proof. Let $G = \langle g\rangle$ and let $a,b \in G$. Then,

$$\begin{aligned}ab &= g^n g^m\\ &= g^{n+m}\\ &= g^{m+n}\\ &= g^m g^n\\ &= ba\end{aligned}$$

by associativity. ■

Theorem 12.4.10. *If a group G has prime order p, then it is cyclic. That is, $G \cong C_p$.*

Proof. $|G| \geq 2$ as $p \geq 2$ is prime. Let $g \in G \setminus \{\mathrm{id}_G\}$. As $g \neq \mathrm{id}_G$, $|\langle g\rangle| > 1$. By Lagrange's theorem, $|\langle g\rangle|$ divides $|G| = p$, but p is prime, so $|\langle g\rangle| = |G|$, and hence $\langle g\rangle = G$. ■

12.4.6 Dihedral Groups

Let P be a regular n-sided polygon in the plane with $n \geq 3$. The collection of isometries on P has group structure under composition. This group is called the *dihedral group* of order $2n$, and is denoted D_n.

These isometries consist of:

(*i*) n rotations through the angles $2\pi k/n$ for $0 \leq k < n$;

(*ii*) n reflections.

We label the vertices of P in order and consider these isometries as permutations on these vertices. Then, the rotations are the elements a^k, $0 \leq k < n$, where $a = (1,2,\ldots,n)$ is the cyclic permutation corresponding to the rotation by $2\pi/n$, and the reflections are the elements $a^k b$, $0 \leq k < n$, where $b = (2,n)(3,n-1)(4,n-2)\ldots$ is the reflection that passes through the vertex 1.

In all cases, we have $ba = a^{n-1}b = a^{-1}b$, so $ba^k = a^{n-k}b = a^{-k}b$ for $0 \leq k < n$. This allows us to find the full Cayley table of this group expressed in this form as we can then perform any of the four basic types of products:

(i) $(a^k)(a^l) = a^{k+l}$

(ii) $(a^k)(a^lb) = a^{k+l}b$

(iii) $(a^kb)(a^l) = a^k(ba^l) = a^ka^{-l}b = a^{k-l}b$

(iv) $(a^kb)(a^lb) = a^k(ba^l)b = a^ka^{-l}bb = a^{k-l}$

with all exponents taken modulo n.

12.4.7 Symmetric Groups

If X is any set, then the collection of permutations on X has group structure under composition. This group is called the *symmetric group* on X, and is denoted $\mathrm{Sym}(X)$.

It doesn't really matter what the elements of X actually are, since they just label the inputs and outputs of the functions we're interested in, so the structure really only depends on the cardinality of X:

Theorem 12.4.11. *Suppose $|X| = |Y|$ for two sets X and Y. Then,* $\mathrm{Sym}(X) \cong \mathrm{Sym}(Y)$.

We then write $\mathrm{Sym}(n)$ or S_n for the symmetric group on n elements.

The symmetric group is of extreme importance in theory, as they are some of the most general groups possible and as such, it may not be surprising that every group is in some sense contained within a symmetric group. More precisely, we will see later on that every group is isomorphic to a subgroup of a symmetric group.

Because of its importance, we have specialised notation for writing elements of S_n.

12.4.7.1 Permutation Notation

We can write a permutation in S_n in *Cauchy's two-line notation* as,

$$\begin{pmatrix} 1 & 2 & 3 & \cdots & n \\ a & b & c & \cdots & d \end{pmatrix}$$

where the first line lists the elements of S, and the second lists their image. For example, a permutation in S_5 could be,

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 5 & 4 & 3 & 1 \end{pmatrix}$$

which represents the map defined by $\sigma(1) = 2$, $\sigma(2) = 5$, $\sigma(3) = 4$, $\sigma(4) = 3$, and $\sigma(5) = 1$.

To compose permutations in this notation, we write them next to each other, applying them right to left, as per function notation. Simply follow where each element goes. For example,

$$\rho = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}, \quad \mu = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}$$

$$\rho\mu = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ a & b & c \end{pmatrix}$$

Looking at the rightmost permutation, 1 maps to 1, then 1 maps to 3, so $a = 3$. 2 maps to 3, then 3 maps to 2, so $b = 2$. 3 maps to 2, then 2 maps to 1, so $c = 1$.

To work out the inverse of a permutation written in Cauchy two-line notation, we just swap the first and second rows.

We can also write permutations in S_n in *cycle notation.*

Let A_1,A_2,A_3,…,A_m be distinct elements of $\{1,2,\ldots,n\}$. The *cycle* $(A_1,A_2,A_3,\ldots,A_m)$ represents the permutation that maps A_1 to A_2, A_2 to A_3, $\cdots$, A_{m-1} to A_m, A_m to A_1; and any elements not in the cycle are fixed in place.

The number of elements in the cycle is the *length* of the cycle. A cycle of length 2 is additionally called a *transposition.*

So, in S_5, the cycle of length 3, (1,4,5) would map the ordering [1,2,3,4,5] to [5,2,3,1,4]. Note that these are **not** the same as the rows from Cauchy's two-line notation. The same permutation in two-line notation is,

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 2 & 3 & 5 & 1 \end{pmatrix}$$

Cycles are equivalent up to circular shifts, so, for example, $(1,2,3) = (3,1,2) = (2,3,1)$ as in all 3 cases, the cycle represents the mappings $1 \mapsto 2$, $2 \mapsto 3$, and $3 \mapsto 1$.

Two cycles are *disjoint* if they do not contain any numbers in common. Disjoint cycles additionally commute.

Every permutation can be written as a product of disjoint cycles.

Example. Write the following permutation as a product of disjoint cycles:

$$\rho = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 5 & 7 & 1 & 4 & 8 & 2 & 6 & 3 \end{pmatrix}$$

We follow where 1 is mapped, then see where its image is mapped, etc., obtaining the chain of mappings, $1 \mapsto 5 \mapsto 8 \mapsto 3 \mapsto 1$, so the first cycle is (1,5,8,3). Now, check the next element which doesn't appear in the cycle - in this case, 2. We then have the chain $2 \mapsto 7 \mapsto 6 \mapsto 2$, so (2,7,6) is the next cycle. Continuing, we have $4 \mapsto 4$, and now every element is in some cycle, so we write $\rho = (1,5,8,3)(2,7,6)(4)$.

Cycles of length 1 may be omitted as they do not affect the permutation, so $\rho = (1,5,8,3)(2,7,6)$ is another valid answer. △

Example. (Composing disjoint cycles)

$$\sigma = (1,3,10,9)(2,5,6)$$
$$\tau = (4,3,10)(1,5,8)$$

What is $\sigma\tau$?

Follow where 1 goes. Remember we read right to left as per function notation, so $1 \mapsto 5$ in τ. Now, apply σ to 5, so $5 \mapsto 6$. Overall, we have $\sigma\tau(1) = 6$.

Our cycle is $(1,6,\cdots)$ so far. Now, we want to see where 6 maps to under $\sigma\tau$, so we find $\tau(6) = 6$, and $\sigma(6) = 2$, so the cycle is now $(1,6,2,\cdots)$, and we then follow 2. Repeat until every element is in a cycle.

Continuing in this way, we obtain $\sigma\tau = (1,6,2,5,8,3,9)(4,10)$. △

To invert a permutation given as a product of not necessarily disjoint-cycles, just write it backwards. That is, reverse each cycle, then reverse the order of cycles.

Example. Let $\rho = (1,4,3,7,6)(5,9,4,1)(9,2,4,8)$. What is ρ^{-1}?

Writing out the cycles in reverse, we have $\rho^{-1} = (8,4,2,9)(1,4,9,5)(6,7,3,4,1)$. △

12.4.8 The Alternating Group & Transpositions

Theorem 12.4.12. *Every permutation can be written as a product of transpositions.*

Proof. Every permutation can be written as a product of disjoint cycles, so it suffice to show that cycles can be written as products of transpositions. Then, $(A_1,A_2,A_3,\dots,A_m) = (A_1,A_m)\cdots(A_1,A_3)(A_1A_2)$ ■

Example.

$$(1,2,3,4,5) = (1,5)(1,4)(1,3)(1,2)$$

△

Note that these transpositions are not disjoint, and do not commute. Furthermore, the transposition decomposition of a permutation is not unique.

Let $n \geq 2$ be an integer. Let $x_1,x_2,\dots,x_n$ be variables, and let P_n be the polynomial $P_n = \prod_{1\leq i\leq j\leq n} x_i - x_j$.

For example,

$$\begin{aligned}
P_2 &= x_1 - x_2 \\
P_3 &= (x_1-x_2)(x_1-x_3)(x_2-x_3) \\
P_4 &= (x_1-x_2)(x_1-x_3)(x_1-x_4)(x_2-x_3)(x_2-x_4)(x_3-x_4) \\
&\vdots
\end{aligned}$$

P_n is called the nth *Vandermonde polynomial.*

Theorem 12.4.13. *Let $\sigma \in S_n$, so $\sigma(P_n) = \prod_{1\leq i\leq j\leq n} x_{\sigma(i)} - x_{\sigma(j)}$. Let $\tau \in S_n$ be a transposition. Then $\tau(P_n) = -P_n$.*

Proof. Let $\tau = (a,b)$. Any factor $(x_i - x_j)$ without a or b inside is unchanged by τ. If both $i = a$ and $j = b$, then $(x_i - x_j) \mapsto (x_j - x_i) = -(x_i - x_j)$. Now, consider all other factors where only one of i and j are equal to a or b. While ($i < a$ and $i < b$) or ($i > a$ and $i > b$), τ just swaps the position of the two factors. Otherwise, the sign is switched. But the situation is symmetric, so each factor has a mirrored pair, so no total sign change is effected from these factors. Thus, the only sign change is from when $i = a$ and $j = b$. ■

Every permutation can be written as a product of an even number of transpositions, or an odd number of transpositions, but crucially, not both.

A permutation is *even* if it can be written as a product of an even number of transpositions, and similar for *odd.*

The alternating group $\mathrm{Alt}(X)$ on a set X is the set of even permutations on X under composition. As with $\mathrm{Sym}(X)$, the isomorphism classes of the alternating groups depend only on the cardinality of X, so we write $\mathrm{Alt}(n)$ or A_n for the alternating group on n elements.

A_n is a clearly a subgroup of S_n as we can write $A_n = \{\sigma \in S_n : \sigma \text{ is even}\}$, and it has order $\frac{n!}{2}$.

12.4.9 Common Groups & Sets

We give a table of commonly occuring groups:

- D_n (the *dihedral group*) – the group of symmetries of a regular n-gon. $|D_n| = 2n$.
- S_n (the *symmetric group*) – the group of permutations of n points. $|S_n| = n!$.

- A_n (the *alternating group*) – te group of even permutations of n points. $|A_n| = \frac{n!}{2}$.
- $\mathbb{Z}/n\mathbb{Z}$ – set of integers mod n under addition, or possibly multiplication if n is prime.[*]
- Nth roots of unity – solutions of $z^n = 1$ over the complex numbers under multiplication, sometimes denoted U_n, though this is non-standard notation.
- $\mathbb{S}^1$ or $\mathbb{T}$ (the *circle group*) – the set of complex numbers with magnitude 1 under multiplication.
- $\text{Map}(A)$ – the set of functions from a set, A, to itself.
- $\text{Sym}(A)$ – the set of bijections from a set, A, to itself. $S_n = (\text{Sym}(\{1,2,\ldots,n\}),\circ)$.
- $M_{m\times n}(\mathbb{R})$ is the set of matrices with real entries. $M_{m\times n}(\mathbb{Z})$, etc., are defined similarly.
- $GL_n(\mathbb{R})$ (the *general linear group*) is the set of $n \times n$ matrices with non-zero determinants and real entries, under matrix multiplication.
- $SL_n(\mathbb{R})$ (the *special linear group*) is the set of $n\times n$ matrices with unit determinant and real entries, under matrix multiplication.
- $SL_2(\mathbb{Z})$ (the *modular group*) is the set of 2×2 matrices with unit determinant and integer entries, under matrix multiplication.
- $SO_n(\mathbb{R})$ (the *special orthogonal group*) is the set of $n \times n$ rotation matrices under matrix multiplication.

12.4.10 Cosets

Let G be a group, H be a subgroup of G, and g be an element of G. The set $gH = \{gh : h \in H\}$ is a *left coset* of H, and $Hg = \{hg : h \in H\}$ is a *right coset* of H. In the case of abelian groups written in additive notation, we denote cosets by $g + H$ rather than gH.

A coset of a subgroup has the same order as the subgroup, as inverses are unique.

Theorem 12.4.14. *The following statements are equivalent for all $g,k \in G$:*

(i) $k \in gH$

(ii) $gH = kH$

(iii) $gk^{-1} \in H$

Corollary 12.4.14.1. *Two left cosets g_1H and g_2H in G are either equal or disjoint.*

Proof. If g_1H and g_2H are not disjoint, then there exists some element $k \in g_1H \cap g_2H$. But then $g_1H = kH = g_2H$ by the above theorem. ■

Example. $2\mathbb{Z}$ is a subgroup of $\mathbb{Z}$. What are the left cosets of $2\mathbb{Z}$ in $\mathbb{Z}$?

First, pick an element of $\mathbb{Z}$. Say, 0. Add it to every element of $2\mathbb{Z}$:

$$0 + 2\mathbb{Z} = \{\cdots,0 + (-2),0 + (0),0 + (2),\cdots\} = 2\mathbb{Z}$$

so $2\mathbb{Z}$ is a left coset of $2\mathbb{Z}$ in $\mathbb{Z}$.

[*] The groups $(\mathbb{Z}n/\mathbb{Z},+)$ and $(\mathbb{Z},+_n)$ are technically slighly different, though they are isomorphic §12.4.1 and are functionally identical. The set underlying the first group contains congruence classes with modularity built into the elements themselves, while the set underlying the second group is just the integers, and the modularity is built into the operation instead.

Now, take 1 and add it to every element of $2\mathbb{Z}$:

$$1+2\mathbb{Z}=\{\cdots,1+(-2),1+(0),1+(2),\cdots\}$$

This is distinct from the previous set, so this is a new coset.

Now, if we try 2 or anything else, we'll find that we just land in one of our two previous cosets. In fact, these two cosets partition $\mathbb{Z}$, so we know we have them all. Thus, the left cosets of $2\mathbb{Z}$ in $\mathbb{Z}$ are $2\mathbb{Z}$ and $1+2\mathbb{Z}$. △

Lemma 12.4.15. *If H is finite, then all left cosets have exactly $|H|$ elements. That is, $|gH|=|H|$ for all $g\in G$.*

Proof. The map $\phi:H\to gH$ defined by $\phi(h)=gh$ is a bijection by the cancellative property. ■

Let G be a group and H be a subgroup of G. The *index* $[G:H]$ is defined to be the number of left cosets (or right cosets, but not counting both) of H in G.

Example. What is the index $[\mathbb{Z}:2\mathbb{Z}]$?

In the previous part, we found two cosets, so $[\mathbb{Z}:2\mathbb{Z}]=2$. △

Theorem (Lagrange)**.** *If H is a subgroup of a group G, then $|G|=[G:H]|H|$.*

Proof. Let H be a subgroup of a group G, and define an equivalence relation R on all pairs of elements $x,y\in G$ such that xRy holds if and only if there exists $h\in H$ such that $x=yh$. Under this equivalence relation, the left cosets of H in G are equivalence classes, and therefore partition G into disjoint sets. The mapping $x\mapsto ax$ is inverted by $y\mapsto a^{-1}y$, and therefore defines a bijection $H\to aH$, so each left coset aH has the same cardinality as H. The number of left cosets is the index, $[G:H]$, so $|G|=[G:H]|H|$, as required. ■

If the index and sizes of each set are interpreted as cardinal numbers, Lagrange's theorem holds even if some of the sets are infinite in size.

Corollary (Lagrange's Theorem)**.** *The order of any element a of a finite group divides the order of the group. Or equivalently, the order of any subgroup of a group divides the order of the group.*

12.5 Normal Subgroups

A subgroup N of a group G is *normal* in G if $gN=Ng$ for all $g\in G$, and we write $N\triangleleft G$ to denote this relation.

For any group, G, the trivial subgroup, $\{\mathrm{id}_G\}$, is always a normal subgroup of G. G itself is also always a normal subgroup of G. If these are the only normal subgroups, then G is a *simple* group.

Theorem 12.5.1. *If H is a subgroup of a group G such that $[G:H]=2$, then H is normal in G.*

Proof. Since H has index 2, it has exactly two left cosets; H itself, and $G\setminus H$. H also has exactly two right cosets; H, and $G\setminus H$. Thus, the left and right cosets of H coincide and H is normal. ■

We give an alternative characterisation of normal subgroups:

Theorem 12.5.2. *If H is a subgroup of a group G such that $ghg^{-1}\in H$ for all $g\in G$ and $h\in H$, then H is normal in G.*

That is, a subgroup N of a group G is normal if and only if it is invariant under conjugation (§12.6.2). That is, the conjugation of any element of N by any element of G is always in N; $ghg^{-1} \in H$ for all $g \in G$ and $h \in H$. For this reason, normal subgroups are also sometimes called *invariant* or *self-conjugate* in G.

This then gives various equivalent conditions for a subgroup to be normal:

- For all $g \in G$, the left and right cosets gN and Ng are equal;
- The set of left and right cosets of N in G are equal;
- N is a union of conjugacy classes of G;
- The image of conjugation of N by any element of G is a subset of N;
- The image of conjugation of N by any element of G is equal to N.

(Some of these will be proved later.)

Theorem 12.5.3. *Every subgroup of an abelian group is normal.*

Proof. Let H be a subgroup of an abelian group G, and let $g \in G$. Let $x \in gHg^{-1}$ so $x = ghg^{-1}$ for some $h \in H$. Then,

$$\begin{aligned} x &= ghg^{-1} \\ &= hgg^{-1} \\ &= h \\ &\in H \end{aligned}$$

so H is invariant under conjugation by any g and is hence normal. ■

12.5.1 Direct Products

Let G and H be groups. The *direct product* (*group*) $G \times H$ of G and H is the group on the Cartesian product of G and H,

$$\{(g,h) : g \in G, h \in H\}$$

of ordered pairs of elements from G and H, under the operations of G and H applied componentwise. That is, we define the group operation $\star$ on $G \times H$ to be,

$$(g_1,h_1) \star (g_2,h_2) = (g_1 * g_2, h_1 \cdot h_2)$$

where $*$ is the group operation on G, and $\cdot$ is the group operation on H. The identity element $\mathrm{id}_{G\times H}$ is then given by $(\mathrm{id}_G, \mathrm{id}_H)$, and the inverse of (g,h) is (g^{-1},h^{-1}).

Theorem 12.5.4. *Any group of order* 4 *is isomorphic to either C_4 or $C_2 \times C_2$.*

Theorem 12.5.5. *Any group of order* 6 *is isomorphic to either C_6 or D_3.*

The *quaternion group* Q_8 is a non-abelian group of order 8, isomorphic to the set of quaternion units (and their inverses) under quaternion multiplication. That is, the set $\{1,i,j,k,-1,-i,-j,-k\}$ where $i^2 = j^2 = k^2 = ijk = -1$.

Theorem 12.5.6. *Any group of order* 8 *is isomorphic to either C_8 or $C_4 \times C_2$, $C_2 \times C_2 \times C_2$, D_4, or Q_8.*

12.5.2 Quotient Groups

A *quotient group* or *factor group* is a group obtained by identifying similar elements of a larger group together using an equivalence relation that preserves some of the group structure, with the rest of the structure being "factored" out. For instance, the group of integers under addition modulo n, $(\mathbb{Z}/n\mathbb{Z},+)$ or equivalently, $(\mathbb{Z},+_n)$, can be obtained from the group of integers under addition, $(\mathbb{Z},+)$, by identifying elements that differ by a multiple of n, and defining a group structure that operates on congruence classes (§10.2) rather than individual elements.

Subgroups and quotient groups are dual notions, the two being the primary ways of constructing smaller groups from a larger one. Any normal subgroup has a corresponding quotient group, formed by eliminating the distinction between elements of the subgroups. For any congruence relation on a group G, the equivalence classes of the identity element is always a normal subgroup, N, of the original group, while the other classes are precisely the cosets of that normal subgroup, and the corresponding quotient group is G/N.

The reason why G/N is called a "quotient" group comes from an analogy with division of integers. When dividing 12 by 3, we obtain the answer 4 because we can split a collection of 12 objects into 3 subcollections each containing 4 objects. Quotient groups follow a similar idea, but when "dividing" groups, we end up with another group as the answer rather than a number, because groups have more structure than arbitrary collections of objects.

When we have a quotient group, G/N, with N being a normal subgroup of a group G, the group structure is used to form our subcollections – the cosets of N in G. Because we started with a group and normal subgroup, the final quotient contains more structure than just the number of cosets (which is what regular division yields), but instead has group structure itself (given an appropriate binary operation on cosets, as we will now show).

Let N be a normal subgroup of a group G. We define $G/N = \{gN : g \in G\} = \{\{gn : n \in N\} : g \in G\}$ to be the set of all left cosets of N in G. Since the identity element $e \in N$, we have $a \in aN$. We define a binary operation, $\cdot$, on G/N as $aN \cdot bN = (ab)N$, where ab is the group operation applied to a and b.

Lemma 12.5.7. *Let N be normal in G, and let $g,h \in G$. Then, the product of any element in the coset gN with any element in the coset hN is an element in the coset $(gh)N$.*

Proof. Let $gn_1 \in gN$ and $hn_2 \in hN$. Then, by normality of N, $gN = Ng$, so $n_1h \in Nh$ is equal to some element $hn \in hN$, and hence $(gn_1)(hn_2) = g(n_1h)n_2 = g(hn)n_2 = (gh)(nn_2) \in (gh)N$. ■

We claim that this binary operation is well-defined if and only if N is normal.

Suppose that N is normal, and $xN = aN$ and $yN = bN$ for some $x,y,a,b \in N$. That is, x and a (and y and b, respectively), are possibly distinct representatives of the same coset. Because N is normal, every left coset is equal to its corresponding right coset, so we can "commute" these products:

$$\begin{aligned}(ab)N &= a(bN)\\ &= a(yN)\\ &= a(Ny)\\ &= (aN)y\\ &= (xN)y\\ &= x(Ny)\\ &= x(yN)\\ &= (xy)N\end{aligned}$$

so the operation does not depend on choice of representative of each left coset when N is normal.

For the backward direction, suppose that the operation is well-defined for some subgroup N of a group G. That is, for all $xN = aN$ and $yN = bN$ with $x,y,a,b \in N$, $(ab)N = (xy)N$.

Let $n \in N$ and $g \in G$. Since $eN = nN$,

$$\begin{aligned} gN &= (eg)N \\ &= (eN)(gN) \\ &= (nN)(gN) \\ &= (ng)N \end{aligned}$$

so $N = (g^{-1}ng)N$ and $(g^{-1}ng) \in N$. Since the choice of n and g was arbitrary, $(g^{-1}ng) \in N$ for all $n \in N$ and $g \in G$, so N is normal.

If A and B are subsets of a group G, we define their (*internal*) *product* AB to be the set $\{ab : a \in A, b \in B\}$. This allows us to more concisely state the previous discussion:

Lemma 12.5.8. *If N is normal in G and gN and hN are cosets of N in G, then $(gN)(hN) = (gh)N$.*

Proof. By the previous lemma, $(gN)(hN) \subseteq (gh)N$. Then, let $n \in N$, so $(gh)n = (g\,\mathrm{id}_G)(hn) \in (gN)(hN)$ and $(gh)N \subseteq (gN)(hN)$. ■

We can also verify that this operation on G/N is associative, that G/N has identity element N, and the inverse of an element aN under this operation can always be represented by $a^{-1}N$. It follows that the set G/N with the operation $aN \cdot bN = (ab)N$ has group structure, and we call this group the quotient group of G by N.

Theorem 12.5.9. *Let N be normal in G. Then, the set G/N of left cosets gN of N in G forms a group under internal multiplication called the quotient group of G by N.*

Proof. By the previous lemma, $(gN)(hN) = (gh)N$, giving closure, and associativity is inherited from associativity in G. Then, $(1N)(gN) = (1g)N = gN = (g1)N = (gN)(1N)$ for all $g \in G$, so $1N$ is the identity element, and $(g^{-1}N)(gN) = (g^{-1}g)N = 1N$, so $(g^{-1})N$ is the inverse element of gN. ■

Additionally, because N is normal, the above definition with left cosets replaced with right cosets is equivalent.

Also note that if G is finite, then $|G/N| = [G : N] = |G|/|N|$.

Example. Consider the group of integers under addition modulo 6, $G = \{0,1,2,3,4,5\}$, and the subgroup, $N = \{0,3\}$. Because G is abelian, N is normal. The set of left cosets has cardinality 3:

$$G/N = \{a + N : a \in G\} = \{0 + N, 1 + N, 2 + N\} = \{\{0,3\},\{1,4\},\{2,5\}\}$$

Along with the binary operation as defined above, this set has group structure. In this case, this quotient group is isomorphic to the cyclic group of order 3. △

Theorem 12.5.10. *If G is abelian or cyclic, then so is G/N*

Proof. Exercise. ■

12.5.3 Kernels and Images

Let $\phi : G \to H$ be a group homomorphism. Then, the *kernel* $\ker(\phi)$ of ϕ is the set of elements mapped to id_H. That is,

$$\ker(\phi) = \{g \in G : \phi(g) = \mathrm{id}_H\}$$

The *image* $\mathrm{im}(\phi)$ of ϕ is just its image as a function.

Theorem (Trivial Kernel (Groups))**.** *Let $\phi : G \to H$ be a group homomorphism. Then, ϕ is injective if and only if* $\ker(\phi) = \{\mathrm{id}_G\}$.

Proof. Since $\mathrm{id}_G \in \ker(\phi)$, $\phi(\mathrm{id}_G) = \mathrm{id}_H$. If ϕ is injective, then $\ker(\phi) = \{\mathrm{id}_G\}$. Conversely, suppose $\ker(\phi) = \{\mathrm{id}_G\}$. Let $g_1,g_2 \in G$ such that $\phi(g_1) = \phi(g_2)$. Then,

$$\begin{aligned}\mathrm{id}_H &= \phi(g_1)^{-1}\phi(g_1)\\ &= \phi(g_1)^{-1}\phi(g_2)\\ &= \phi(g_1^{-1}g_2)\end{aligned}$$

so $g_1^{-1}g_2 \in \ker(\phi)$, and hence $g_1^{-1}g_2 = \mathrm{id}_G$ and $g_1 = g_2$, so ϕ is injective. ■

Theorem 12.5.11. *Let $\phi : G \to H$ be a group homomorphism. Then,* $\ker(\phi)$ *is a normal subgroup of G.*

Theorem 12.5.12. *Let $N \triangleleft G$ be a normal subgroup. Then the map $\pi : G \to G/N$ defined by $g \mapsto gN$ is a surjective homomorphism with kernel* $\ker(\pi) = N$.

Proof. For any $a,b \in G$, $\pi(ab) = (ab)N = (aN)(bN) = \pi(a)\pi(b)$, so π is a homomorphism. Then, for any $gN \in G/N$, $gN = \pi(g)$, so π is surjective. Now, suppose $\pi(g) = \mathrm{id}_{G/N}$. Then,

$$\begin{aligned}\pi(g) &= \mathrm{id}_{G/N}\\ gN &= \mathrm{id}_G N\end{aligned}$$

Since $gN = \mathrm{id}_G N$, ${\mathrm{id}_G}^{-1}g = g \in N$, so $\ker(\pi) = N$. ■

This homomorphism is called the *quotient map*, or *natural* or *canonical* homomorphism from G to G/N.

Theorem 12.5.13. *Let $\phi : G \to H$ be a group homomorphism. Then,* $\mathrm{im}(\phi)$ *is a (not necessarily normal) subgroup of H.*

Proof. Let $h_1,h_2 \in \mathrm{im}(\phi)$, so there exist $g_1,g_2 \in G$ such that $\phi(g_1) = h_1$ and $\phi(g_2) = h_2$. Then,

$$h_1h_2^{-1} = \phi(g_1)\phi(g_2)^{-1} = \phi(g_1g_2) \in \mathrm{im}(\phi)$$

so $\mathrm{im}(\phi)$ is a subgroup by the one-step test. ■

12.5.4 The Isomorphism Theorems

Theorem (First Isomorphism Theorem)**.** *Let $\phi : G \to H$ be a homomorphism with kernel* $\ker(\phi) = K$*. Then $G/K \cong$* $\mathrm{im}(\phi)$*, and more precisely, there is a homomorphism $\bar{\phi} : G/K \to$* $\mathrm{im}(\phi)$ *defined by $\bar{\phi}(gK) = \phi(g)$ for all $g \in G$.*

Proof. Clearly, $\mathrm{im}(\bar{\phi}) = \mathrm{im}(\phi)$, so $\bar{\phi}$ is surjective. Now, suppose $gK = hK$, so $gh^{-1} \in K$. Let $k = gh^{-1}$, so $g = kh$. Then, because $k \in K = \ker(\phi)$, $\phi(g) = \phi(k)\phi(h) = \phi(h)$, so $\bar{\phi}$ is a well-defined map.

Let $aK,bK \in G/K$. Then,

$$\begin{aligned}\bar{\phi}\big((aK)(bK)\big) &= \bar{\phi}\big((ab)K\big)\\ &= \phi(ab)\\ &= \phi(a)\phi(b)\\ &= \bar{\phi}(aK)\bar{\phi}(bK)\end{aligned}$$

so $\bar{\phi}$ is a homomorphism.

Finally, suppose $gK \in \ker(\bar{\phi})$, so,

$$\begin{aligned} \bar{\phi}(gK) &= \mathrm{id}_H \\ \phi(g) &= \mathrm{id}_H \end{aligned}$$

so $g \in \ker(\phi) = K$ ■

We can restate this theorem more precisely with a commutative diagram:

Theorem (First Isomorphism Theorem). *Let $\phi : G \to H$ be a homomorphism with kernel* $\ker(\phi) = K$ *and let $\pi : G \to G/K$ be the quotient map. Then, there is an isomorphism $\bar{\phi} : G/K \to \mathrm{im}(\phi)$ such that the following diagram commutes:*

$$\begin{tikzcd} G \arrow[r, "\pi"] \arrow[dr, "\phi"'] & G/\ker(\phi) \arrow[d, "\bar{\phi}"] \\ & \mathrm{im}(\phi) \end{tikzcd}$$

Proof. Suppose $aK = bK$. Then, $\phi(aK) = \phi(a)\phi(K) = \phi(a)$, and similarly for bK, so $a = b$. The universal property of quotients then yields the unique well-defined map $\bar{\phi} : G/K \to \mathrm{im}(\phi)$ such that the diagram above commutes, and since ϕ and π are surjective, $\bar{\phi} = \phi \circ \pi$ is also surjective. Now, suppose $\pi(g) \in \ker(\bar{\phi})$. Then, from commutativity, $\mathrm{id}_K = \bar{\phi}(\pi(g)) = \phi(g)$, so $g \in \ker(\phi)$, and hence $\ker(\bar{\phi}) = \{\ker(\phi)\}$, so $\bar{\phi}$ is injective. ■

The next two isomorphism theorems are less important, and are used mainly in more advanced group theory.

Theorem (Second Isomorphism Theorem). *Let G be a group, $H \leq G$ be a subgroup, and $K \triangleleft G$ be a normal subgroup. Then,*

(i) $HK = KH$ is a subgroup of G;

(ii) $H \cap K$ is a normal subgroup of H;

(iii) $H/(H \cap K) \cong HK/K$.

Proof. Exercise. See problem 49 in §12.7 for steps. ■

Theorem (Third Isomorphism Theorem). *Let G be a group and let $K \subseteq H \subseteq G$. Suppose K and H are both normal in G. Then,*

(i) K is normal in H;

(ii) H/K is a normal subgroup of G/K;

(iii) $(G/K)/(H/K) \cong G/H$.

Proof. Exercise. See problem 50 in §12.7 for steps. ■

12.6 Group Actions

Many groups we have used so far arise naturally from sets of functions from some set to itself. For instance, $\mathrm{Sym}(X)$ is the set of permutations on a set X; $GL_n(\mathbb{R})$ is the set of endofunctions on $\mathbb{R}^n$; and D_n is the set of isometries on the set of vertices of a regular n-gon. Informally, we'd might say that Sym "acts on" the set X, $GL_n(\mathbb{R})$ "acts on" $\mathbb{R}^n$; and D_n "acts on" the vertices of a regular n-gon. We can formalise this notion with *group actions*.

Let G be a group, and X a set. A (left) *action* of G on X is a map $\cdot : G \times X \to X$ satisfying,

$(A1)$ $\mathrm{id}_G \cdot x = x$ for all $x \in X$;

$(A2)$ $(gh) \cdot x = g \cdot (h \cdot x)$ for all $g,h \in G$ and $x \in X$.

Right group actions are defined similarly as maps $X \times G \to X$ satisfying analogous properties, but we will only consider left actions here.

Example.

- $\mathrm{Sym}(X)$ (and any subgroups, such as $\mathrm{Alt}(X)$) acts on X by the map $\rho \cdot x = \rho(x)$.
- $GL_n(\mathbb{R})$ (and any subgroups, such as $SL_n(\mathbb{R})$) acts on $\mathbb{R}^n$ by the matrix multiplication $\mathbf{A} \cdot \mathbf{v} = \mathbf{A}\mathbf{v}$.

△

In these examples, every element of the group induces a permutation on X, which is an element of $\mathrm{Sym}(X)$. In fact, this is always the case:

Theorem 12.6.1. *Let $\cdot$ be an action of a group G on a set X. For $g \in G$, define the map $\phi(g) : X \to X$ by $\phi(g)(x) = g \cdot x$. Then, $\phi(g) \in \mathrm{Sym}(X)$, and furthermore, $\phi : G \to \mathrm{Sym}(X)$ is a group homomorphism.*

This suggests an alternative characterisation of group actions as a homomorphism from a group to the symmetric group on some target set.

The *kernel* of an action $\cdot$ of G on X is defined to be the kernel $K = \ker(\phi)$ of the homomorphism $\phi : G \to \mathrm{Sym}(X)$ as defined in the above theorem. That is,

$$K = \{g \in G : g \cdot x = x \text{ for all } x \in X\}$$

If $K = \{\mathrm{id}_G\}$, we say that the action $\cdot$ is *faithful*.

Let $(G, *)$ be a group. Then, taking X to be the set G underlying the group, the left *regular action* of G on itself is the faithful action defined by $g \cdot x = g * x$.

For a faithful action with kernel K, $G \cong G/K$, as the quotient is trivial. Then, the first isomorphism theorem gives $G/K \cong \operatorname{im}\phi \leq \mathrm{Sym}(X)$, so $G \leq \mathrm{Sym}(X)$.

Theorem (Cayley). *Every group is isomorphic to a subgroup of a symmetric group. Specifically, for each $g \in G$, the left-multiplication map $\ell_g : G \to G$ defined by $x \mapsto gx$ is a permutation on G, and the map $G \to \mathrm{Sym}(G)$ defined by $g \mapsto \ell_g$ is an injective homomorphism, thus embedding G into a subgroup of* $\mathrm{Sym}(G)$.

12.6.1 Orbits and Stabilisers

Let $\cdot$ be an action of G on X. Define the relation $\sim$ on $x,y \in X$ by $x \sim y$ if and only if there exists a $g \in G$ such that $y = g \cdot x$. Then, $\sim$ is an equivalence relation, and the equivalence classes are called the *orbits* of G on X. In particular, the orbits of a specific element $x \in X$, denoted by $G \cdot x$ or $\mathrm{Orb}_G(x)$ is,

$$\begin{aligned}\mathrm{Orb}_G(x) &= \{y \in X : (\exists g \in G : g \cdot x = y)\} \\ &= \{g \cdot x : g \in G\}\end{aligned}$$

An action of G on X is *transitive* if there is only a single orbit. Equivalently, an action is transitive if for every $x,y \in X$, there exists $g \in G$ such that $y = g \cdot x$.

Given $g \in G$ and $x \in X$ such that $g \cdot x = x$, we say that x is a *fixed point* of g, or that g *fixes* x. For each $x \in X$, the *stabiliser* (*subgroup*) of G with respect to x, denoted G_x or $\operatorname{Stab}_G(x)$, is the set of elements in G that fix x. That is,

$$\operatorname{Stab}_G(x) = \{g \in G : g \cdot x = x\}$$

This is a subgroup of G, but not necessarily a normal one.

Theorem 12.6.2. *Let G act on X and let $x \in X$. Then, $\bigcap_{x \in X} \operatorname{Stab}_G(x)$ is the kernel of the action of G on X.*

Proof. For any $g \in G$, $g \in \bigcap_{x \in X} \operatorname{Stab}_G(x)$ if and only if $g \cdot x = x$ for all $x \in X$, which is the definition of being in the kernel. ■

Theorem (Orbit-Stabiliser)**.** *Let a finite group G act on X, and let $x \in X$. Then,*

$$|G| = |\operatorname{Orb}_G(x)| \times |\operatorname{Stab}_G(x)|$$

Proof. Let $y \in \operatorname{Orb}_G(x)$, so there exists $g \in G$ such that $y = g \cdot x$, and let $H = \operatorname{Stab}_G(x)$. Now, suppose an element $g' \in G$ satisfies $y = g' \cdot x$. Then,

$$\begin{aligned} g' \cdot x &= y \\ g' \cdot x &= g \cdot x \\ g^{-1}g' \cdot x &= x \end{aligned}$$

so $g^{-1}g'$ fixes x, giving $g^{-1}g' \in \operatorname{Stab}_G(x) = H$. Then, $g' \in gH$, so the elements satisfying $g' \cdot x = y$ are exactly the elements of the coset gH, and as cosets of a set are equal in size, we have $|gH| = |H| = |\operatorname{Stab}_G(x)|$. It follows that for each $y \in \operatorname{Orb}_G(x)$, there are exactly $|\operatorname{Stab}_G(x)|$ elements g' of G such that $g' \cdot x = y$, so the total number of such y must be $|G|/|\operatorname{Stab}_G(x)|$. ■

12.6.2 Conjugation

Recall that the (left) regular action of a group $(G, *)$ is the action of the group on itself under the group operation, so $g \cdot x = g * x$. Another important action of G on itself is the *conjugation* action defined by,

$$g \cdot x = gxg^{-1}$$

for $g,x \in G$. The orbits of this action are called the *conjugacy classes* of G, and elements in the same conjugacy class are said to be *conjugate* in G. We write $\operatorname{Cl}_G(x)$ for the orbit of x, or equivalently, the conjugacy class containing x. That is,

$$\operatorname{Cl}_G(x) = \{gxg^{-1} : g \in G\}$$

The stabiliser for this action with respect to x is the set of elements $g \in G$ such that $g \cdot x = x$, so,

$$\begin{aligned} g \cdot x &= x \\ gxg^{-1} &= x \\ gx &= xg \end{aligned}$$

so the stabiliser is exactly the set of elements that commute with x. This subgroup is called the *centraliser* of x in G, and is denoted $C_G(x)$. That is,

$$C_G(x) = \{g \in G : gx = xg\}$$

Applying the orbit-stabiliser theorem then yields,

Theorem 12.6.3. *Let G be a finite group and let $x \in G$. Then,*

$$|G| = |\operatorname{Cl}_G(x)| \times |C_G(x)|$$

The kernel K of this action then consists of the elements that fix, and hence commute with, all elements $g \in G$. This is called the *centre* of G, and is denoted $Z(G)$. So,

$$Z(G) = \{f \in G : fg = gf \text{ for all } g \in G\}$$

Note that $g \in Z(g)$ if and only if $\operatorname{Cl}_G(g) = \{g\}$.

Example. For any abelian group G,

- $Z(G) = G$;
- $C_G(g) = G$;
- $\operatorname{Cl}_G(g) = \{g\}$.

for all $g \in G$. △

Example. The symmetric group S_3 has three conjugacy classes that partition its six permutations of three objects:

- Identity ($abc \mapsto abc$);
- Transposing two elements ($abc \mapsto acb$,$abc \mapsto bac$,$abc \mapsto cba$);
- Cyclic permutations of three elements ($abc \mapsto cab, abc \mapsto cab$).

These three classes also correspond to the three ways of transforming a equilateral triangle: identity, reflections and rotations, respectively. △

As mentioned previously (§12.1.3), the group of permutations on 4 points correspond to the group actions of proper rotations on a cube. We can phrase this more precisely by saying that the proper rotations of the cube, which can be characterised by the permutations of the inner diagonals, are described by conjugations in S_4. In this case, the 24 permutations are partitioned into 5 conjugacy classes.

In general, the number of conjugacy classes in the symmetric group S_n is equal to the number of integer partitions (§10.4.3) of n, because each conjugacy class corresponds to exactly one partition of $\{1,2,\dots,n\}$ into cycles, up to permutation of the elements of $\{1,2,\dots,n\}$.

12.6.3 Conjugacy Classes in Symmetric Groups

Consider two permutations f,$g \in \operatorname{Sym}(X)$. Suppose one of the cycles in g is $(x_1,x_2,\dots,x_r)$, so $g(x_1) = x_2$, $g(x_2) = x_3$, etc. Then, $fg(x_1) = f(x_2)$, so $fgf^{-1}\big(f(x_1)\big) = fg(x_1) = f(x_2)$, and more generally, $fgf^{-1}\big(f(x_i)\big) = f(x_{i+1})$ for i taken modulo r. So, fgf^{-1} has a cycle $(f(x_1),f(x_2),\dots,f(x_r))$. This applies to any cycle in g, so we obtain:

Theorem 12.6.4. *Given a permutation g as a product of cycles, the conjugate fgf^{-1} of g by f is the permutation given by the same product of cycles with each $x \in X$ replaced with $f(x)$.*

Example. Let $X = \{1,2,3,4,5,6,7\}$, $g = (1,5)(2,4,7,6)$, and $f = (1,3,5,7,2,4,6)$. Then,

$$\begin{aligned} fgf^{-1} &= \big(f(1),f(5)\big)\big(f(2),f(4),f(7),f(6)\big) \\ &= (3,7)(4,6,2,1) \end{aligned}$$

△

A permutation has *cycle type* $2^{r_2}3^{r_3}4^{r_4}\dots n^{r_n}\dots$ if it has exactly r_i cycles of length i, for $i \geq 2$.

Example. The permutation (1,2,3)(4,5)(6,7)(8,9,10)(11,12,13,14),(15,16) has cycle type $2^3 3^2 4^1$ because it has 3 cycles of length 2, 2 cycles of length 3, and 1 cycle of length 4. △

Theorem 12.6.5. *Two permutations in* $\operatorname{Sym}(X)$ *are conjugate in* $\operatorname{Sym}(X)$ *if and only if they have the same cycle type.*

12.6.4 Conjugacy Classes in Alternating Groups

Recall that the alternating group A_n is the subgroup of S_n that consists of even permutations. The odd and even permutations partition S_n, so the index of A_n in S_n is 2, so A_n is normal in S_n.

Theorem 12.6.6. *Let* $g \in A_n$*. Then, either,*

$$\operatorname{Cl}_{A_n}(g) = \operatorname{Cl}_{S_n}(g)$$

or

$$|\operatorname{Cl}_{A_n}(g)| = \frac{1}{2}|\operatorname{Cl}_{S_n}(g)|$$

hold.

12.6.5 Simple Groups

Recall that a non-trivial group G is *simple* if the only subgroups normal in G are G itself, and the trivial group $\{\mathrm{id}_G\}$.

Theorem 12.6.7. *Cyclic groups of prime order are simple.*

Proof. By Lagrange's theorem, the only possible order of their subgroups are 1 and p. Normality follows from cyclic groups being abelian. ■

In fact, these are the only abelian simple groups possible:

Theorem 12.6.8. *A simple abelian group is cyclic with prime order.*

Proof. Let G be simple and abelian, and let $g \in G \setminus \{\mathrm{id}_G\}$. If $|g|$ is infinite, then the subgroup generated by g^2 is non-trivial, as it contains $g^2 \neq \mathrm{id}_G$; and proper, as it does not contain g; so G is not simple. If $|g|$ is finite but composite, so $|g| = ab$, then the subgroup generated by g^a is similarly non-trivial and proper, so G is not simple. It follows that $|g|$ is finite and prime, and furthermore, we have $\langle g \rangle = G$, or else $\langle g \rangle$ would be a non-trivial proper subgroup. ■

There are also finite non-abelian groups that are simple. General simple groups have been classified into three main infinite families (with cyclic groups of prime order forming one of the families), and 26 separate groups that do not fit into any of the families, called the *sporadic groups.*

One of the other infinite families of simple groups consists of the alternating groups A_n for $n \geq 5$.

Lemma 12.6.9. *A subgroup* H *of a group* G *is normal in* G *if and only if* H *consists of a union of conjugacy classes of* G.

Proof. Recall that H is normal in G if and only if it is invariant under conjugation . That is, $ghg^{-1} \in H$ for all $g \in G$, $h \in H$. But this is just the statement that H is normal in G if and only if $\operatorname{Cl}_G(h) \subseteq H$ for all h. ■

12.6.6 Sylow's Theorems

One corollary of Lagrange's theorem is that the order of any subgroup H of a finite group G always divides the order of G. One obvious converse question to ask is if a group G has subgroups of all orders that divide $|G|$. This is true for some groups, like finite cyclic grops. However, it is not true in general:

Theorem 12.6.10. *A_4 has no subgroup of order 6.*

Proof. Suppose A_4 has a subgroup H of order 6. Groups of order 6 must be cyclic or dihedral, and A_4 has no elements of order 6, so $H \cong S_3$, so H must have 3 elements of order 3. Specifically, H must contain the identity element and 3 pairs of transpositions. But then these elements form a subgroup of A_4, so H contains a subgroup of order 4, contradicting Lagrange's theorem. ■

Let G be a finite group of order $p^n m$, where n is the largest power of the prime p that divides $|G|$, so m is not divisible by p. A subgroup of G of order p^n is a *Sylow p-subgroup* of G.

Theorem (Sylow's Theorems)**.** *Let G be a finite group, p a prime, and $|G| = p^n m$, where $p \nmid m$. Then,*

(i) *G has a Sylow p-subgroup, and any subgroup of G of order p^a for $1 \leq a \leq n$ is contained in a Sylow p-subgroup of G.*

(ii) *Any two Sylow p-subgroups of G are conjugate in G. That is, if H and K are Sylow p-subgroups of G, then there exists an element $g \in G$ such that $gHg^{-1} = K$.*

(iii) *The number r of Sylow p-subgroups of G satisfies $r \equiv 1 \pmod{p}$ and $r|m$.*

Let G be a group of order $p^n m$ with $n \geq 1$ and $p \nmid m$. We define $\mathrm{Syl}_p(G)$ to be the set of Sylow p-subgroups of G,

$$\mathrm{Syl}_p(G) = \{H \leq G : |H| = p^n\}$$

and by Sylow's first theorem, this set is always non-empty. It turns out that this set is closed under conjugation:

Lemma 12.6.11. *If $P \in \mathrm{Syl}_p(G)$ and $g \in G$, then $gPg^{-1} \in \mathrm{Syl}_p(G)$.*

Now, consider the map $\cdot : G \times \mathrm{Syl}_p(G) \to \mathrm{Syl}_p(G)$ defined by $g \cdot H = gHg^{-1}$ for $H \in \mathrm{Syl}_p(G)$. The above lemma verifies the correctness of the codomain, but this map can furthermore be shown to be a group action of G on $\mathrm{Syl}_p(G)$. Now, $\mathrm{Orb}_G(P) = \{gPg^{-1} : g \in G\}$, and by Sylow's second theorem, this action is transitive, so,

$$\mathrm{Orb}_G(P) = \mathrm{Syl}_p(G)$$

Then, by the orbit-stabiliser theorem and Lagrange's theorem, we have,

Lemma 12.6.12. *$|\mathrm{Syl}_p(G)|$ divides $|G|/|P|$.*

Theorem 12.6.13. *If there is only one Sylow p-subgroup of G, then it is normal in G.*

12.6.7 Sylow's Theorem and Simple Groups

Theorem 12.6.14. *There are no simple groups of order 2 552.*

Proof. Let G be a group of order $2\,552 = 8 \cdot 11 \cdot 29$.

Take $p = 11$, so $|G| = 11 \times (8 \times 29) = 11^1 \times 232$. The number of Sylow 11-subgroups, r, must divide 232 and satisfy $r \equiv 1 \pmod{11}$. Consider the factorisation $232 = 2^3 \times 29$; the factors of 232 are then: 1, 2, 4, 8, $29 \equiv 7$, $58 \equiv 3$, $116 \equiv 6$, and $232 \equiv 1$, so $r = 1{,}232$ are the possible solutions.

Now, if G has more than 1 Sylow 11-subgroup, then it must have 232 Sylow 11-subgroups. As 11 is prime, these subgroups must be cyclic, so every non-identity element generates the group. It follows that these subgroups intersect only at the identity element, so each subgroup contributes 10 elements of order 11, so there must be $232 \times 10 = 2\,320$ elements of order 11 in G.

Now, take $p = 29$, so $|G| = 29 \times (8 \times 11) = 29^1 \times 88$. By identical arguments as before, the number of Sylow 29-subgroups must be 1 or 88, and again, as 29 is prime, each subgroup must be cyclic, so if there is more than 1 Sylow 29-subgroup, then there are $88 \times 28 = 2\,464$ elements of order 28.

Now, by Sylow's first theorem, there exist Sylow 29 and 11-subgroups. If there are more than one of each, then we have 2 320 and 2 464 elements of order 11 and 29, respectively. But these values sum to more than $2\,552 = |G|$, so we cannot simultaneously have more than 1 Sylow 29 and 11-subgroups. But then, any unique Sylow p-subgroup is normal, so G cannot be simple. ■

12.7 Exercises

These questions are in no particular order of subject. Some of the questions are significantly more difficult than others, mostly those at the end, while some can be done in a single sentence – some are solvable just by recalling and stating definitions of algebraic structures. Questions on permutations or cycles have not been included, as you can easily come up with some random cycles of your own, and check them using a CAS.

Don't worry if you can't complete some of the questions without help; they are designed to encourage you to research and learn more by yourself.

1. Prove that the empty set cannot form a group.
2. Prove that $\mathbb{R}^*$ (the set of non-zero reals) forms a group under multiplication.
3. Prove that $\mathbb{Z}$ forms a cyclic group under addition.
4. Prove that the identity element of a group is unique.
5. Prove that the inverse of an element in a group is unique.
6. Suppose that G is a group such that $(ab)^2 = a^2b^2$ for all $a,b \in G$. Prove that G is abelian.
7. Suppose that G is a group such that $(ab)^3 = a^3b^3$ for all $a,b \in G$, and that there are no elements of order 3. Prove that G is abelian.
8. Suppose G is a group with prime order. Prove G is cyclic.
9. Suppose G is a cyclic group. Prove that G is abelian.
10. Suppose that G is a group such that g is self-inverse for all $g \in G$. Prove that G is abelian.
11. Let G be a group. Prove that $|g| = |g^{-1}|$ for all $g \in G$.
12. Let G be a group, and let $a \in G$. Prove that a commutes with a^2.
13. Suppose G has even order.
    (a) Prove there exists an element $a \in G \setminus \{\mathrm{id}_G\}$ such that $a^2 = \mathrm{id}_G$.
    (b) Prove that there are an odd number of such elements.
14. Prove that the identity is the only idempotent element in a group. That is, prove that if $g \cdot g = g$ for an element $g \in G$, then $g = \mathrm{id}_G$.
15. Prove that every element of a finite group has finite order.
16. Let G be a group with order n. Prove that $g^n = e$ for all $g \in G$.
17. Prove that, if G has no non-trivial subgroups, then G is finite with prime order.
18. Let G be a group of order p, where p is prime. Prove that G has $p-1$ elements of order p.
19. Let G be a group, and denote by $\mathrm{Aut}(G) = \{\phi : G \to G \text{ is an isomorphism}\}$. Prove that $\mathrm{Aut}(G)$ is a group under composition.
20. Let G be a group, and let $a,b \in G$ such that a and b are self-inverse. Prove that a and b commute if and only if ab is also self-inverse.
21. Prove that every finite group with more than two elements has a non-trivial automorphism.
22. Consider the set of 2×2 matrices with entries in $\mathbb{F}_2$ and non-zero determinant.
    (a) Prove that this set is a group under matrix multiplication.

(b) Prove that this group is isomorphic to S_3.

23. Prove that two cyclic groups of the same order are isomorphic to each other.
24. Find an example of a group G and an infinite subset H of G such that H is closed under the group operation, but not under inversion.
25. Prove that $(\mathbb{Q},+)$ is not cyclic.
26. Let H and K be subgroups of a group G. Prove that $H \cup K$ is a subgroup of G if and only if $H \subseteq K$ or $K \subseteq H$.
27. Let $G = \{x \in \mathbb{R} | x \neq -1\}$, and $x * y \coloneqq x + y + xy$.
    (a) Prove that $(G,*)$ is a group.
    (b) Prove that $(G,*)$ is abelian.
    (c) Prove that $(G,*)$ is isomorphic to $(\mathbb{R}^*,\times)$.
28. Prove that every infinite cyclic group is isomorphic to $(\mathbb{Z},+)$. (This proof is given earlier in this document, but do give it a try yourself. It is an extremely useful result.)
29. Prove that $(\mathbb{R}^*,\times)$ is not isomorphic to $(\mathbb{Z},+)$.
30. Prove that $(\mathbb{R}^+,\times)$ is isomorphic to $(\mathbb{R},+)$.
31. Prove that $\mathbb{R}/\mathbb{Z}$ is isomorphic to $\mathbb{S}^1$.
32. Prove that $(\mathbb{R},+)$ and $(\mathbb{R}^2,+)$ are isomorphic (but do not attempt to construct the isomorphism).
33. Prove that $(\mathbb{Z},+)$ and $(\mathbb{Z}^2,+)$ are not isomorphic.
34. Prove that $\mathbb{C}^*$ is isomorphic to $\mathbb{S}^1$ (but do not attempt to construct the isomorphism).
35. Does an infinite group exist such that every element of the group has finite order? If so, give an example. Otherwise, prove the non-existence of such a group.
36. Let G_1 and G_2 be groups.
    (a) Prove that $H_1 = G_1 \times \{\mathrm{id}_{G_2}\}$ and $H_2 = \{\mathrm{id}_{G_1}\} \times G_2$ are both subgroups of $G_1 \times G_2$.
    (b) Prove that H_1 and H_2 are both normal in $G_1 \times G_2$.
    (c) Prove that if $h_1 \in H_1$ and $h_2 \in H_2$, then $h_1h_2 = h_2h_1$.
37. Define $f,g : \mathbb{R} \to \mathbb{R}$ by $f(x) = \frac{1}{x}$ and $g(x) = \frac{x-1}{x}$. These functions generate a group G with the binary operation given by composition. Prove that $G \cong S_3$.
38. Let m and n be coprime. Prove that there is no non-trivial group homomorphism from $\mathbb{Z}_m$ to $\mathbb{Z}_n$.
39. Prove that $GL_2(\mathbb{R})/SL_2(\mathbb{R}) \cong (\mathbb{R}^*,\times)$.
40. Prove that any group with order 9 is abelian. More generally, prove that any group with order p^2, where p is prime, is abelian. Give an example of a group of order p^3 that is not abelian.
41. Prove that A_5 is a simple group.
42. Prove that there are no simple groups of order 24.
43. Prove that the centre of a group is always a normal subgroup.
44. Let G be a group of order $2n$. Suppose that exactly n elements of G have order 2, and that the other n elements form a subgroup, $H \subset G$ of order n.
    (a) Prove that n is odd.

(b) Prove that H is abelian.

45. Prove that if a group G has only one element of order 2, then this element is in the centre of G.

46. Let H and K be subgroups of a group G.

    (a) Prove that $H \cap K$ is a subgroup of G.

    (b) By means of counterexample, prove that $H \cup K$ is not necessarily a subgroup of G.

    (c) Prove that if G is finite and the orders of H and K are coprime, then $H \cap K$ is the trivial subgroup.

47. Let G be a finite group of order n such that every non-identity element has order 2.

    (a) Prove that G is abelian.

    (b) Let H be a subgroup of G, and let $g \in G \setminus H$. Prove that $H \cup gH$ is a subgroup of G.

    (c) Prove that $|H \cup gH| = 2|H|$.

    (d) Deduce that the order of G is a power of 2.

48. In this exercise, we prove Cayley's theorem.

    (a) For any element g of a group $(G,*)$, define the function $f_g : G \to G$ by $f_g(x) = g * x$. Prove that this function is a bijection by considering the function induced by the inverse element g^{-1}, and hence deduce that these functions are elements of $\mathrm{Sym}(G)$.

    (b) Prove that the set $K = \{f_g : g \in G\}$ is a subgroup of $\mathrm{Sym}(G)$.

    (c) Define the function $T : G \to \mathrm{Sym}(G)$ by $g \mapsto f_g$. Prove that T is a group homomorphism.

    (d) Prove that T is injective by considering the identity element of G and hence deduce Cayley's theorem.

49. In this exercise, we prove the second isomorphism theorem. Let G be a group, H a subgroup of G, and K a normal subgroup of G.

    (a) Prove that $HK = KH$ and $H \cap K$ are subgroups of G.

    (b) Prove that the mapping $\phi : H \to HK/K$ defined by $\phi(h) = hK$ is a group homomorphism.

    (c) Find the kernel of ϕ and deduce that ϕ is surjective.

    (d) Apply the first isomorphism theorem and obtain the second isomorphism theorem.

50. In this exercise, we prove the third isomorphism theorem. Let G be a group, and let $K \subseteq H \subseteq G$ be subgroups. Suppose that K and H are normal in G.

    (a) Prove that K is normal in H.

    (b) Define the mapping $\phi : G/N \to G/H$ by $\phi(gN) = g(H)$. Prove that ϕ is well-defined.

    (c) Prove that ϕ is a group homomorphism, and by considering the subset relation of K and H, further deduce that ϕ is surjective.

    (d) Find the kernel of ϕ, and apply the first isomorphism to obtain the third isomorphism theorem.

12.7.1 Solutions

1. Groups require an identity element, so the empty set is not a group.

2. Multiplication of non-zero reals yields reals, so $\mathbb{R}^*$ is closed. Associativity is a basic property of multiplication, the identity is given by $1 \in \mathbb{R}^*$, and the multiplicative inverse of a real number $r \in \mathbb{R}^*$ is given by $\frac{1}{r}$, which is also in $\mathbb{R}^*$.

3. Any element $n \in \mathbb{Z}$ can be written as $\underbrace{1+1+\cdots+1}_{n}$, so $\mathbb{Z} = \langle 1 \rangle$ is cyclic.
4. Suppose e and f are identity elements of a group G. $ef = e$ because f is an identity, and hence a right identity for e, and $ef = f$ because e is an identity, and hence a left identity for f. Then, $e = ef = f$, and the identity is unique.
5. Let b and c be inverses of a. Then, $ab = \mathrm{id}_G = ac$ and by the cancellative property, $b = c$, so the inverse of a is unique. Alternatively,

$$\begin{aligned} b &= b\,\mathrm{id}_G \\ &= b(ac) \\ &= (ba)c \\ &= \mathrm{id}_G c \\ &= c \end{aligned}$$

6. $abab = (ab)^2 = a^2b^2 = aabb$. Then,

$$\begin{aligned} ab &= (a^{-1}a)ab(bb^{-1}) \\ &= a^{-1}(aabb)b^{-1} \\ &= a^{-1}(abab)b^{-1} \\ &= (a^{-1}a)ba(bb^{-1}) \\ &= ba \end{aligned}$$

7. $ababab = (ab)^3 = a^3b^3$. Then,

$$\begin{aligned} ababab &= a^3b^3 \\ a^{-1}ababab^{-1} &= a^{-1}a^2b^2b^{-1} \\ baba &= a^2b^2 \\ (ba)^2 &= a^2b^2 \qquad (1) \end{aligned}$$

for any a,$b \in G$. Now, consider the expression $xyx^{-1}y^{-1}$ (this is the *commutator* of x and y). Applying (1) with $b = xy$ and $a = x^{-1}y^{-1}$, we have:

$$(xyx^{-1}y^{-1})^2 = (x^{-1}y^{-1})^2(xy)^2$$

Now apply (1) with $b = x^{-1}$ and $a = y^{-1}$ on the left, and $b = x$ and $a = y$ on the right:

$$\begin{aligned} &= \big((y^{-1})^2(x^{-1})^2\big)(y^2x^2) \\ &= y^{-2}x^{-2}y^2x^2 \\ &= y^{-2}(x^{-2}y^2)x^2 \end{aligned}$$

Apply (1) with $a = x^{-2}$ and $b = y^2$:

$$\begin{aligned} &= y^{-2}(yx^{-1})^2x^2 \\ &= y^{-2}(yx^{-1})(yx^{-1})x^2 \\ &= y^{-2}yx^{-1}yx^{-1}x^2 \\ &= y^{-1}x^{-1}yx \end{aligned}$$

so we have

$$(xyx^{-1}y^{-1})^2 = y^{-1}x^{-1}yx \qquad (2)$$

for any $x,y \in G$. Swapping variables again, we square (2) to obtain:

$$(ghg^{-1}h^{-1})^4 = (h^{-1}g^{-1}hg)^2$$

Applying (2) on the right side with $x = h^{-1}$ and $y = g^{-1}$, we have,

$$\begin{aligned}(ghg^{-1}h^{-1})^4 &= ghg^{-1}h^{-1}\\(ghg^{-1}h^{-1})^3 &= \mathrm{id}_G\end{aligned}$$

for any $g,h \in G$. Since G does not contain any elements of order 3, it follows that $ghg^{-1}h^{-1} = \mathrm{id}_G$ and hence $gh = hg$.

8. Let G be a group of prime order, and let $g \in G \setminus \{\mathrm{id}_G\}$. By Lagrange's theorem, the order of $\langle g\rangle$ divides the order of G. As G is of prime order, $\langle g\rangle$ is either equal to G or the trivial subgroup, but $\langle g\rangle$ contains $g \neq \mathrm{id}_G$, so the latter cannot hold.

9. Let $a,b \in G$. Because G is cyclic, $a = g^n$ and $b = g^m$ for some integers n and m and generator element g. Then,

$$\begin{aligned}ab &= g^n g^m\\&= \underbrace{g\cdot g\cdot g\cdots g}_{n}\cdot\underbrace{g\cdot g\cdot g\cdots g}_{m}\\&= g^{n+m}\\&= \underbrace{g\cdot g\cdot g\cdots g}_{m}\cdot\underbrace{g\cdot g\cdot g\cdots g}_{n}\\&= g^n g^m\\&= ba\end{aligned}$$

10. Consider the element $ab \in G$.

$$\begin{aligned}ab &= (ab)^{-1}\\&= b^{-1}a^{-1}\\&= ba\end{aligned}$$

11. Let $a \in G$ and $a^n = \mathrm{id}_G$ for some integer n. Then,

$$\mathrm{id}_G = (aa^{-1})^n = a^n(a^{-1})^n = \mathrm{id}_G(a^{-1})^n = (a^{-1})^n$$

so $|a^{-1}| \leq |a|$. The same argument with a and a^{-1} reversed shows $|a| \leq |a^{-1}|$.

12. $a \cdot a^2 = a\cdot(a\cdot a) = (a\cdot a)\cdot a = a^2\cdot a$.

13. We prove the two statements together. Note that the requirement that $g^2 = \mathrm{id}_G$ is equivalent to $g = g^{-1}$ for any element g, so the identity element and the elements of order 2 are the only elements in G equal to their own inverse elements.

$$G = \{\mathrm{id}_G\} \cup \{g : |g| = 2\} \cup \{g_1,g_1^{-1},g_2,g_2^{-1},\ldots,g_k,g_k^{-1}\}$$

where g_i are elements of order greater than 2. Because the g_i have orders greater than two, the elements g_i and g_i^{-1} are distinct, so the final set has even cardinality, and $\{\mathrm{id}_G\}$ has odd cardinality. Because G has odd order, it follows that the set $\{g : g^2 = \mathrm{id}_G\}$ has an odd number of members.

In particular, 0 is not an odd number, so $\{g : g^2 = \mathrm{id}_G\}$ cannot be the empty set, and hence contains at least one element, a.

14. Suppose g is idempotent. Then,

$$\begin{aligned} gg &= g \\ ggg^{-1} &= gg^{-1} \\ g &= \mathrm{id}_G \end{aligned}$$

15. Let $g \in G$, and consider the sequence of elements

$$g,g^2,g^3,g^4,g^5,\dots$$

Because G is finite, $g^n = g^m$ for at least one pair of distinct $n,m \in \mathbb{Z}$. Without loss of generality, suppose $n < m$. Then,

$$\begin{aligned} g^n &= g^m \\ g^n g^{-n} &= g^m g^{-n} \\ \mathrm{id}_G &= g^{m-n} \end{aligned}$$

16. Let $g \in G$ have order a, so $g^a = \mathrm{id}_G$. By Lagrange's theorem, a divides $|G|$, so $|G| = ab$ for some integer b and hence $g^{|G|} = g^{ab} = (g^a)^b = \mathrm{id}_G^b = \mathrm{id}_G$.

17. The trivial group only has a proper subgroup, so G cannot be trivial.

    Suppose G is infinite, so it contains non-identity elements. Let $g \in G \setminus \{\mathrm{id}_G\}$. Then, $\langle g \rangle$ is a non-trivial subgroup of G.

    Otherwise, suppose G is finite with composite order $n = ab$, and again let $g \in G \setminus \{\mathrm{id}_G\}$. If $\langle g \rangle \neq G$, we are done. Otherwise, the order of g is ab, so $\langle g^a \rangle$ is a non-trivial subgroup of G.

18. Let $g \in G \setminus \{\mathrm{id}_G\}$. The order of g must divide p, but p is prime so $|g| = 1$ or $|g| = p$. g is not the identity, so the former cannot hold, and hence all non-identity elements, of which there are $p - 1$, have order p.

19. The composition of two isomorphisms is again an isomorphism; isomorphisms are just special functions, so associativity is inherited from set functions; the identity mapping is an automorphism and acts as an identity under composition; and isomorphisms also have, by definition, an inverse isomorphism. Hence $\mathrm{Aut}(G)$ is a group.

20. Suppose that a and b commute. Then,

$$\begin{aligned} ab &= ba \\ &= b^{-1}a^{-1} \\ &= (ab)^{-1} \end{aligned}$$

    Conversely, suppose ab is self-inverse. Then,

$$\begin{aligned} ab &= (ab)^{-1} \\ &= b^{-1}a^{-1} \\ &= ba \end{aligned}$$

21. Suppose G is non-abelian, so there exists $g,h \in G$ such that $gh \neq hg$. Consider the conjugation map $\phi : G \to G$ defined by $x \mapsto gxg^{-1}$ for a fixed g. Proving this is a group homomorphism is straightforward, and its inverse is given by conjugation by g^{-1}. Now, suppose this conjugation

mapping is trivial. Then, $h = \phi(h) = ghg^{-1}$, and hence $gh = hg$, contradicting our choice of g and h. Hence ϕ is a non-trivial automorphism of G.

Now, suppose G is a finite abelian group of order $n > 2$. Since G is abelian, the map $\phi : G \to G$ given by $x \mapsto x^{-1}$ is an automorphism. If ϕ is trivial, then $x = \phi(x) = x^{-1}$, so $x^2 = \mathrm{id}_G$. If G has an element of order greater than 2, then we are done. Otherwise, all elements of G have order at most 2, so

$$G \cong (\mathbb{Z}/2\mathbb{Z})^n$$

and since $|G| > 2$, we have $n > 1$. Then, the map $\psi : (\mathbb{Z}/2\mathbb{Z})^n \to (\mathbb{Z}/2\mathbb{Z})^n$ defined by $(x_1,x_2,x_3 \dots ,x_n) \mapsto (x_2,x_1,x_3 \dots ,x_n)$ is a non-trivial automorphism of G.

22.

12.8 Rings

A *ring* is a triple, $(R, + , \cdot)$, where R is a set and $+$ and $\cdot$ are binary operations $R \times R \to R$ such that:

(R0) R is closed under $\times$;

(R1) R is an abelian group under $+$;

(R2) $\cdot$ is associative on R;

(R3) $\cdot$ left and right distributes over $+$;

(R4) R contains an identity under $\times$.

or in more detail,

(R0) $\forall a,b \in R, a \cdot b \in R$ (closure of $\cdot$);

(R1) $(R,+)$ is an abelian group (additive group);

(R2) $\forall a,b,c \in R, a \cdot (b \cdot c) = (a \cdot b) \cdot c$ (associativity of $\cdot$);

(R2) $\forall a,b,c \in R, (a + b) \cdot c = a \cdot c + b \cdot c$ and $a \cdot (b + c) = a \cdot b + a \cdot c$ (left and right distributivity);

(R3) $\exists 1_R \in R$ such that $\forall a \in R, a \cdot 1_R = 1_R \cdot a = a$ (existence of multiplicative identity).

We call the operation denoted by $+$ *addition*, and the operation denoted by $\times$ *multiplication* or *product* (regardless of what the operations actually are). We also call the additive identity 0_R the *ring zero*, as it is also the zero element for the multiplication operation.

Triples satisfying only axioms R0 to R3 are sometimes called *rngs* (as in, rings without identity), and in contrast, rings *with* identity are called *unital rings* to distinguish them from rngs. Whenever "ring" is used without qualification, we will assume that it is a unital ring.

A ring $(R, + ,\times)$ is furthermore a *commutative ring* if it satisfies:

(R5) $\times$ is commutative on R.

Note that the "commutative" part of the name "commutative ring" refers to commutativity of multiplication, as commutativity of addition is required in all rings regardless. However, rings notably do *not* require multiplicative inverses.

Example.

- The set $\{0\}$ under the trivial operations $0 + 0 = 0$ and $0 \cdot 0 = 0$ forms the *zero* or *trivial* ring.
- $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are commutative rings under their usual addition and multiplication operations.
- $\mathbb{Z}/n\mathbb{Z}$ or $\mathbb{Z}_n$ is a commutative ring under addition and multiplication modulo n for all naturals $n \in \mathbb{N}$.
- If R is a ring, the set $R[x]$ of polynomials in indeterminate x and coefficients in R is another ring under the usual addition and multiplication of polynomials, called a *polynomial ring.*
- If R is a ring, then the set $M_{n\times n}(R)$ of $n \times n$ matrices with entries in R is another ring. Matrix rings are generally non-commutative, and in fact, are commutative if and only if R is the trivial ring, or R is commutative and $n = 1$.

△

Let $(R, + , \cdot)$ be a ring, and let S be a subset of R. Furthermore, suppose that $(S, + , \cdot)$ is also a ring. $(S, + , \cdot)$ is then a *subring* of $(R, + , \cdot)$.

To show that S is a subring of R, it suffices to show that S contains the identity of $+$ and $\cdot$, is closed under $+$ and $\cdot$, and that every element has an inverse in S under $+$. More symbolically, if R is a ring, then $S \subseteq R$ is a subring if and only if,

- $0_R \in S$ (additive identity);
- $1_R \in S$ (multiplicative identity);
- If $a,b \in S$ then $a + b \in S$ (closure under $+$);
- If $a,b \in S$ then $a \cdot b \in S$ (closure under $\times$);
- If $a \in S$ then $(-a) \in S$ (additive inverses).

Associativity is inherited from the main ring, and you do not have to check for multiplicative inverses.

We can collapse some of these properties together:

Theorem (Subring Test). *If $(R, +, \cdot)$ is a ring and $S \subseteq R$, then $(S, +, \cdot)$ is a subring of R if and only if,*

1. *$(S,+)$ is a subgroup of $(R,+)$;*
2. *$a,b \in S \to ab \in S$;*
3. *$1_R \in S$.*

Proof. The reverse direction is trivial. Conversely, suppose the three conditions above hold for a subset $S \subseteq R$. We verify the ring axioms:

(R0) Closure follows directly from condition 2.

(R1) $(S,+)$ is an abelian group as it is a subgroup of an abelian group by condition 1.

(R2) Associativity is inherited from R as $S \subseteq R$.

(R3) Distributivity is inherited from R as $S \subseteq R$.

(R4) Multiplicative identity follows directly from condition 3.

■

Example.

- $\mathbb{Z}[i] = \{a + bi : a,b \in \mathbb{Z}\}$ is a subring of $\mathbb{C}$ called the ring of *Gaussian integers*.
- $\mathbb{Z}\left[\sqrt{2}\right] = \left\{a + b\sqrt{2} : a,b \in \mathbb{Z}\right\}$ is a subring of $\mathbb{R}$.
- The set

$$\left\{\frac{a}{2^n} : a \in \mathbb{Z}, n \in \mathbb{Z}_{\geq 0}\right\}$$

  is a subring of $\mathbb{Q}$ called the ring of *dyadic rationals*.

△

These examples show that it can be easier to describe a ring by expressing it as a subring of a different known ring, as we avoid having to define the multiplication and addition operations, and do not have to verify associativity and distributivity.

Theorem 12.8.1. *The intersection of subrings of a ring R is itself a subring of R.*

12.8.1 Morphisms

A *(ring) homomorphism* between two rings $(R, +, \cdot)$ and $(S, \oplus, \odot)$ is a function $\phi : R \to S$ that preserves the structure of R. That is,

- $\phi(a + b) = \phi(a) \oplus \phi(b)$;
- $\phi(a \cdot b) = \phi(a) \odot \phi(b)$;
- $\phi(1_R) = 1_S$.

Additive inverses and the additive identity are also part of the preserved structure, but they are not explicitly specified as they follow from these three conditions.

If the inverse of a ring homomorphism is a homomorphism, or equivalently, if the homomorphism is a bijection, then it is called a (ring) isomorphism. If an isomorphism exists between R and S, we say that R and S are isomorphic (rings), and we write $R \cong S$ to denote this relation. Again, isomorphism is an equivalence relation.

Like with groups, an injective ring homomorphism is also called a *monomorphism*, and a surjective homomorphism is called an *epimorphism*.

Example.

- For each $n \in \mathbb{N}$, the map $x \mapsto x \pmod{n}$ is a ring homomorphism $\mathbb{Z} \to \mathbb{Z}_n$.
- The map $z \mapsto \overline{z}$ is a ring isomorphism $\mathbb{C} \to \mathbb{C}$.
- If R is any ring and S is a subring of R, then for each element $\alpha \in R$, the map $\phi_\alpha : S[x] \to R$ defined by $f \mapsto f(\alpha)$ is a ring homomorphism known as the *evaluation map* (at α).
- If $\phi : R \to S$ is a ring homomorphism, then there is an *induced* homomorphism $\psi : R[x] \to S[x]$, defined by,
- $$\psi(a_n x^n + \cdots + a_1 x + a_0) = \phi(a_n)x^n + \cdots + \phi(a_n)x + \phi(a_0)$$

△

Let $\phi : R \to S$ be a ring homomorphism. Then, the *kernel* $\ker(\phi)$ of ϕ is its kernel when treated as a group homomorphism between the additive groups of R and S. That is, the set of elements that are mapped to the additive identity:

$$\ker(\phi) = \{r \in R : \phi(r) = 0_S\}$$

The *image* $\operatorname{im}(\phi)$ of ϕ is just its image as a function.

We have similar results for ring homomorphisms as we had for group homomorphisms:

Theorem (Trivial Kernel (Rings))**.** *Let $\phi : R \to S$ be a ring homomorphism. Then, ϕ is injective if and only if* $\ker(\phi) = \{0_r\}$.

Proof. See §12.5.3. ■

Theorem 12.8.2. *Let $\phi : R \to S$ be a ring homomorphism. Then,* $\operatorname{im}(\phi)$ *is a subring of S.*

Proof. Follows from the subring test. ■

Note that the kernel of a ring homomorphism is *not* necessarily a subring of the target ring. For example, the kernel of the homomorphism $\phi : \mathbb{Z} \to \mathbb{Z}_n$ is the set $n\mathbb{Z}$, which does not contain 1 for all $n \geq 2$.

Let R and S be rings. The *direct product* (*ring*) $R \times S$ of R and S is the ring on the Cartesian product of R and S,

$$\{(r,s) : r \in R, s \in S\}$$

of ordered pairs of elements from R and S, under the two operations of R and S both applied componentwise. That is,

$$\begin{aligned}(r_1,s_1) + (r_2,s_2) &= (r_1 + r_2, s_1 + s_2)\\(r_1,s_1) \cdot (r_2,s_2) &= (r_1 \cdot r_2, s_1 \cdot s_2)\end{aligned}$$

where $+$ and $\cdot$ on the left are the ring operations on $R \times S$, and the two $+$ and $\cdot$ operations on the right are the appropriate ring operations on R and S. The multiplicative identity element $1_{R\times S}$ is then given by $(1_R, 1_S)$; the additive identity $0_{R\times S}$ by $(1_R, 1_S)$; and the additive inverse of (r,s) by $(-r, -s)$.

Notice that R and S are not generally isomorphic to subrings of $R \times S$ in general, even under the obvious projection mapping. For instance, R can be thought of as the elements of $R \times S$ of the form $(r, 0_S)$, and these elements do indeed define a ring isomorphic to R, but its multiplicative identity element is $(1_R, 0_S)$, which is not the identity of $R \times S$, so this ring is not a subring of $R \times S$.

Theorem (Chinese Remainder Theorem)**.** *$Z_n \times Z_m \cong Z_{nm}$ if and only if n and m are coprime.*

By induction, we can extend this result to,

Corollary 12.8.2.1. *If $n = p_1^{a_1} \cdot p_2^{a_2} \cdots p_k^{a_k}$ is a factorisation of n into k distinct primes, then,*

$$\mathbb{Z}_n \cong \mathbb{Z}_{p_1^{a_1}} \times \mathbb{Z}_{p_2^{a_2}} \times \cdots \times \mathbb{Z}_{p_k^{a_k}}$$

Theorem 12.8.3. *Let R be a ring and $a,b \in R$. Then,*

(i) $a \cdot 0 = 0 \cdot a = 0$;

(ii) $a \cdot (-1) = (-1) \cdot a = -a$.

Proof. For (*i*),

$$\begin{aligned}a \cdot 0 &= a \cdot (0 + 0)\\&= a \cdot 0 + a \cdot 0\end{aligned}$$

so $a \cdot 0 = 0$ by the cancellative property in the group $(R,+)$, and similarly, $0 \cdot a = 0$.

For (*ii*),

$$\begin{aligned}(-1) \cdot a + 1 \cdot a &= (-1 + 1) \cdot a\\&= 0 \cdot a\\&= 0\end{aligned}$$

so $(-1) \cdot a = -a$ by uniqueness of inverses in the group $(R,+)$, and similarly, $a \cdot (-1) = -a$. ■

Theorem (Uniqueness of Multiplicative Identity)**.** *The multiplicative identity of a ring is unique.*

Proof. Suppose 1 and $1'$ are multiplicative identities of R. Then, $1 = 1 \cdot 1' = 1'$. ■

Theorem (Coinciding Identities)**.** *Let R be a ring, and suppose that the additive and multiplicative identities coincide, so $0 = 1$. Then, R is the trivial ring.*

Proof. For all $a \in R$, $a = a \cdot 1 = a \cdot 0 = 0$. ■

If a ring is not the trivial ring, we also say that it is a *non-zero* ring.

12.9 Quotient Rings

12.9.1 Ideals

We previously covered ideals as subsets of the integers (§10.3), but ideals are defined more generally as special subsets of any ring.

For an arbitrary ring, $(R, + ,\cdot)$, let $(R,+)$ be its additive group. A subset $I \subseteq R$ is a *left ideal* in R if,

- $(I,+)$ is a subgroup of $(R,+)$,
- For every $r \in R$ and every $x \in I$, $r \cdot x \in I$,

A *right ideal* is defined similarly, with $r \cdot x \in I$ being replaced with $x \cdot r \in I$ in the second requirement, and a *two-sided ideal*, or just *ideal*, is a left ideal that is also a right ideal. If the ring is commutative, then the definitions of left, right and two-sided ideals coincide.

So, an ideal is a subset of the ring that is a group under the ring addition restricted to the subset and absorbs multiplication from one or both sides.

Theorem 12.9.1. *An ideal I of a ring R contains 1_R only when $I = R$.*

Theorem 12.9.2. *If $\phi : R \to S$ is a ring homomorphism, then $\ker(\phi)$ is an ideal in R.*

Proof. $\ker(\phi)$ is an additive subgroup of R when ϕ is considered as a group homomorphism. Then, if $r \in \ker(\phi)$ and $x \in R$, then,

$$\begin{aligned}\phi(x \cdot r) &= \phi(x) \cdot \phi(r)\\ &= \phi(x) \cdot 0_S\\ &= 0_S\end{aligned}$$

so $x \cdot r \in \ker(\phi)$. Similarly, $r \cdot x \in \ker(\phi)$, so $\ker(\phi)$ absorbs multiplication as well, and is hence an ideal in R. ■

When R is a commutative ring, the subset,

$$\{ra : r \in R\}$$

consisting of all multiples of a in R is an ideal of R. This ideal is called the *principal ideal* generated by a, and is denoted (a), aR, or Ra.

For an arbitrary ring, the principal ideal (a) is equal to the set of finite sums,

$$\left\{\sum_{i=1}^{k} r_i a s_i : r_i,s_i \in R\right\}$$

Theorem 12.9.3. *If R is commutative, then $(a) = R$ if and only if a is a unit of R.*

We now consider how ideals arise in the specific case of construction of integer subrings under modulo arithmetic.

Consider the ring of integers modulo n, $\mathbb{Z}/n\mathbb{Z}$, given $n \in \mathbb{Z}$, also noting that the ring of integers is commutative. We obtain $\mathbb{Z}/n\mathbb{Z}$ by "wrapping" the line of integers around into a loop such that various integers become identified together, subject to two constraints:

- n must identify with 0, since n is congruent to 0 modulo n.
- The resulting structure must be a ring.

The second constraint forces additional identifications, and determines the precise way in which $\mathbb{Z}$ is wrapped around. The notion of an ideal arises when we ask what set of integers is forced to identify with 0.[*] Unsurprisingly, the set of integers congruent to 0 modulo n, $n\mathbb{Z} = \{nm : m \in \mathbb{Z}\}$, satisfies this. That is, $\mathbb{Z}$ must be wrapped around itself infinitely many times so that the integers $\ldots, n \cdot (-2), n \cdot (-1), n \cdot (+1), n \cdot (+2) \ldots$ all align with 0. If we consider what properties this set must satisfy to ensure that $\mathbb{Z}/n\mathbb{Z}$ is a ring, then we obtain the above definition of an ideal.

We can construct similar structures from any commutative ring, R: start with an arbitrary element $x \in R$, then identify with 0 all elements of the set $xR = \{xr : r \in R\}$. This set is always an ideal, and furthermore, it is the smallest ideal that contains x, called the ideal *generated* by x, denoted (x) or $\langle x \rangle$. More generally, we can take any subset $S \subseteq R$, and identify with 0 all elements in the ideal generated by S the smallest ideal (S) such that $S \subseteq (S)$. The ring we obtain after identification depends only on the ideal (S), and not on the set S. That is, several subsets may generate the same ideal, but each ideal generates a unique ring.

So, an ideal I of a commutative ring R captures the information required to obtain the ring of elements R modulo a given subset $S \subseteq R$. The elements of I are, by definition, the elements of R that are congruent to 0. That is, they are the elements that are identified with 0 in the resulting ring. This ring is called the *quotient* of R by I, and is denoted R/I.

We can see that ideals are to rings what normal subgroups are to groups in that we can quotient a ring by an ideal to generate another ring, just like how groups can be factored through by a normal subgroup.

Intuitively, the definition of an ideal gives two conditions necessary for I to contain all elements designated as "zeros" by R/I:

- I is an additive subgroup of R, so the zero 0_R of R is a zero 0_R in I as well. Furthermore, if $x_1, x_2 \in I$ are "zeros" (they obey the ring axioms in the same manner as 0), then $x_1 - x_2 \in I$ is a "zero" too.
- Any element $r \in R$ multiplied by a "zero" $x \in I$ returns another "zero", $rx \in I$.

It turns out that these two conditions are also sufficient for I to contain all the necessary "zeros". That is, no other elements have to be designated as "zero" in order for R/I to be a ring.

We can formalise this with relations. Given a ring, R, and a two-sided ideal I in R, we define a equivalence relation $\sim$ on R such that $a \sim b$ if and only if $a - b$ is in I. It turns out that $\sim$ is actually a congruence relation on R, and if $a \sim b$, we say that a and b are *congruent modulo I*. The congruence class of an element $a \in R$ is given by $[a] = a + I = \{a + r : r \in I\}$. This congruence class is also sometimes written as $a \bmod I$, and is called the *residue class of a modulo I*.

Since an ideal I of a ring R is a subgroup of $(R,+)$, we can consider its cosets $I + a$ for $a \in R$. We already know that these form a quotient group under the addition operation defined by,

$$(I + a_1) + (I + a_2) = I + (a_1 + a_2)$$

But to define a ring structure, we also require a multiplication operation.

Theorem (Quotient Ring). *Let I be an ideal of R. Then, the set R/I of cosets $I + a$ of I in R forms a ring under the addition operation in the quotient group, and the multiplication,*

$$(I + a) \cdot (I + b) = I + (a \cdot b)$$

[*] Elements other than 0 must also be identified – for example, the elements $1 + n\mathbb{Z}$ must be identified with 1, the elements $2 + n\mathbb{Z}$ with 2, and so on. However, these are determined uniquely by $n\mathbb{Z}$ since $\mathbb{Z}$ is an additive group.

where $+$ and $\cdot$ on the right hand side are the ring operations. The proof that these operations are well-defined is omitted, as it is almost identical to the earlier proof (§12.5.2) for the operation on quotient groups.

The zero element (additive identity) of R/I is then given by $[0] = (0_R + I) = I$, and the multiplicative identity is $[1_R] = (1_R + I)$.

Example. The quotient ring $\mathbb{Z}/(n) = \mathbb{Z}/n\mathbb{Z}$ is isomorphic to the ring $\mathbb{Z}_n$ of residues modulo n, with the isomorphism $\mathbb{Z}_n \to \mathbb{Z}/n\mathbb{Z}$ given by $x \mapsto x + n\mathbb{Z}$. △

Theorem 12.9.4. *Let I be an ideal of a ring R. Then, the map $\pi : R \to R/I$ defined by $\pi(a) = I + a$ is a surjective ring homomorphism with kernel I called the quotient map.*

The isomorphism theorems for groups also apply similarly to rings, but we only state the first one here:

Theorem (First Isomorphism Theorem)**.** *Let $\phi : R \to S$ be a homomorphism with kernel $\ker(\phi) = I$. Then $R/I \cong \operatorname{im}(\phi)$, and more precisely, there is a homomorphism $\bar{\phi} : R/I \to \operatorname{im}(\phi)$ defined by $\bar{\phi}(I+a) = \phi(a)$ for all $a \in R$.*

12.9.2 Integral Domains

Let R be a ring, and let $a,b \in R$. If a and b are both non-zero and satisfy $ab = 0$, then a and b are called (*left* and *right*, respectively) *zero divisors.*

A ring R is an (*integral*) *domain* if,

(*i*) R is commutative;

(*ii*) R is not the trivial ring;

(*iii*) R has no zero divisors; that is, if $a,b \in R$, then $a \cdot b = 0 \to (a = 0 \vee b = 0)$.

That is, an integral domain is a non-zero commutative ring in which the product of any two non-zero elements is non-zero.

Example.

- The rings $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are integral domains.
- Subrings of integral domains are also integral domains, so $\mathbb{Z}[i]$ and $\mathbb{Z}\left[\sqrt{2}\right]$ are also integral domains.

△

Again, it can be easier to describe integral domains as subrings of other known integral domains.

Theorem 12.9.5. *$\mathbb{Z}_n$ is an integral domain if and only if n is prime.*

Proof. If $n = 1$, then $\mathbb{Z}_n \cong \{0\}$. If $n = ab$ is composite, then $ab = 0$ with $a,b \neq 0$ in Z_n. If n is prime and $a,b \in \mathbb{Z}_n$, then a and b are coprime to n, and hence ab is coprime to n by multiplicativity of gcd, so n does not divide ab, and $ab \neq 0$ in $\mathbb{Z}_n$. ■

12.9.3 Units

An element, a, of a ring R is a *unit* if it has a two-sided inverse under multiplication. That is, there exists some $b \in R$ such that $a \cdot b = b \cdot a = 1$.

Note that in any non-trivial ring, the additive identity 0_R is not a unit.

The *unit group* of R is the group formed by the set $\{a \in R : a \text{ is a unit in } R\}$ under the ring multiplication operation, denoted R^*.

Example. In $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$, every non-zero element, k, has a multiplicative inverse, $\frac{1}{k} \in \mathbb{Q},\mathbb{R},\mathbb{C}$, so the units are the non-zero elements. $\mathbb{Q}^*$, $\mathbb{R}^*$ and $\mathbb{C}^*$ are therefore $\mathbb{Q} \setminus \{0\}$, $\mathbb{R} \setminus \{0\}$ and $\mathbb{C} \setminus \{0\}$, respectively.

However, in $\mathbb{Z}$, $\frac{1}{k}$ is an integer only for $k = \pm 1$, so the units in $\mathbb{Z}$ are ± 1. $\mathbb{Z}^*$ is therefore $\{-1,1\}$.

In $\mathbb{Z}_n$, an element $a \in \mathbb{Z}_n$ is a unit in $\mathbb{Z}/n\mathbb{Z}$ if and only if a and n are coprime (by the Euclidean algorithm and Bézout's identity), so $\mathbb{Z}_n^* = \{a : \gcd(a,n) = 1\}$. △

A non-trivial ring R is called a *division ring* if $R \setminus \{0_R\}$ is a group under multiplication. That is, if every non-zero element is a unit, or, if $R \setminus \{0_R\} = R^*$.

12.10 Fields

Using our new terminology from ring theory, we can shorten the list of axioms given in §11.2 by defining fields in terms of rings or groups.

A *field* is a commutative division ring. So, in total, $(F, +, \times)$ is a field if,

- $(F,+)$ is an abelian group with additive identity 0_F;
- $(F \setminus \{0_F\},\times)$ is an abelian group with multiplicative identity 1_F;
- $0_F \neq 1_F$ (the *non-degeneracy* condition);
- Multiplication distributes over addition.

The non-degeneracy condition is there just to exclude the trivial ring $\{0\}$ from being a field, as many field theorems do not apply well to the trivial ring.

Equivalently, $(F, +, \times)$ is a field if $(F,+)$ is an abelian group with additive identity 0_F, $(F \setminus \{0_F\},\times)$ is an abelian group with multiplicative identity 1_F, $0_F \neq 1_F$ and multiplication distributes over addition.

Theorem 12.10.1. *Let $a \in \mathbb{Z}/n\mathbb{Z}$. Then $\overline{a}$ is a unit in $\mathbb{Z}/n\mathbb{Z}$ if and only if a and n are coprime.*

This implies that for any prime p, the quotient division ring $\mathbb{Z}/p\mathbb{Z}$ is always a field.

Example.

- $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are fields.
- For any prime p, $\mathbb{Z}/p\mathbb{Z}$ is a finite or *Galois* field, sometimes denoted $\mathbb{F}_p$ or $\mathrm{GF}(p)$.
- $\mathbb{F}_2$ in particular, is known as the *binary field.* In the context of computer science and Boolean algebra, 0 and 1 in this field are often alternatively denoted by *true* and *false*, or $\top$ and $\bot$ instead, respectively. Addition in this field is simply the XOR operation §2.2.1.
- The smallest non-prime Galois field is the field with four elements, denoted $\mathbb{F}_4$ or $\mathrm{GF}(4)$, consisting of four elements, 0, 1, α, and $1+\alpha$, where 0 is the additive identity, 1 is the multiplicative identity, α and $1+\alpha$ satisfy $\alpha^2 = 1+\alpha$, and $x + x = 0$ for all $x \in \mathbb{F}_4$.

  0 and 1 correspond exactly to their counterparts in $\mathbb{F}_2$, so $\mathbb{F}_4$ is a field extension of $\mathbb{F}_2$, and furthermore, $\mathbb{F}_4$ can also be constructed as the quotient field, $\mathbb{F}_2[X]/(X^2 + X + 1)$.
- A number that is a root of a non-zero polynomial in one variable with integer (or equivalently, rational) coefficients is called an *algebraic number.* These are discussed in more detail later, but the countable set $\mathbb{A}$ of algebraic numbers forms a field and satisfies $\mathbb{Q} \subset \mathbb{A} \subset \mathbb{R}$.
- In classical construction geometry, we can construct geometric figures using only an infinite idealised unmarkable ruler called a *straightedge* and an idealised compass that has no minimum or maximum radius. Given a line segment of unit length, the set of lengths that can be constructed using a finite sequence of straightedge-and-compass constructions is called the set of *constructible numbers*, and indeed, this set is a field. Not all real numbers are constructible; for instance, the length of a side

of a cube of volume 2 – that is, $\sqrt[3]{2}$ – is famously* not a constructible number; and $\sqrt[3]{2}$ is even algebraic, so the constructible numbers are a subfield of the algebraic numbers as well as the real numbers.

It turns out that the constructible numbers can equivalently be characterised algebraically as the subset of the real numbers that can be described by formulae that combine integers using finitely many operations of addition, subtraction, multiplication, multiplicative inverse, and square roots of non-negative integers. In fact, we can restrict the integers in these formulae with to be only 0 or 1. With this characterisation of constructible numbers, we see that numbers like $\sqrt{2} = \sqrt{1+1}$ are constructible† (and geometrically, this is simply the diagonal of a unit square), so this field is also a field extension of the rationals.

△

Theorem 12.10.2. *Every field is an integral domain.*

Proof. Let F be a field. Suppose there exist $x,y \in F \setminus \{0\}$ such that $xy = 0$. As F is a field, $x \neq 0$ has a multiplicative inverse x^{-1}, so,

$$\begin{aligned} xy &= 0 \\ x^{-1}xy &= 0 \\ y &= 0 \end{aligned}$$

contradicting the definition of y. ■

Lemma (Cancellative Properties in Domains). *Let R be an integral domain, and let $x,y,c \in R$. If,*

- *$c \neq 0$;*
- *$cx = cy$ or $xc = yc$,*

then $x = y$.

Proof.

$$\begin{aligned} cx &= cy \\ cx - cy &= 0 \\ c(x-y) &= 0 \end{aligned}$$

Since R is a domain, and $k \neq 0$, we must have $x - y = 0$, so $x = y$. The proof for $xc = yc$ is similar. ■

Theorem 12.10.3. *Every finite integral domain is a field.*

Proof. Let $R = \{0_R = r_0, r_1, r_2, \ldots, r_n\}$ be a finite domain. By the previous lemma, for a fixed $i > 0$, the n products $r_i r_j$ for $1 \leq j \leq n$ are distinct and non-zero, and since there are only n possible values, they all occur exactly once. In particular, this means that $r_i r_j = 1_R$ for some j, so R is a field. ■

* This is known as *doubling the cube* or the *Delian problem*, and is an ancient geometric problem.

This problem often comes along with another two geometric problems: *squaring the circle*, and *angle trisection*, both of which have also been since proven impossible.

Squaring the circle, or the *quadrature of the circle*, is the task of constructing a square with the same area as a given circle, again, only using straightedge-and-compass constructions. Doing so requires constructing the square root of π – which is a non-algebraic, or *transendental*, number. As the constructible numbers are a subfield of the algebraic numbers, this is impossible.

Angle trisection is the task of constructing an angle that is one-third of a given arbitrary angle. In some special cases, this is possible, but in general, the required angle is not constructible. For instance, if we are given an angle of $\pi/3$ radians – 60° – then we would again have to construct a quantity dependent on π.

† The non-constructibility of $\sqrt[3]{2}$ in the doubling the cube problem follows from this algebraic characterisation. In particular, its minimal polynomial over $\mathbb{Q}$ is degree 3, so this cube root cannot be computed from integers by a finite sequence of these operations.

Let R be a ring. If there exists a positive integer n such that $nx = 0$ for all $x \in R$, then, then we call the minimal such positive integer the *characteristic* of R. If no such positive integer exists, then the characteristic is 0.

Example.

- $\mathbb{Q}$ and $\mathbb{Z}$ have characteristic 0.
- $\mathbb{Z}_n$ has characteristic n.
- The polynomial ring $R[x]$ has the same characteristic as R.

△

Let R be a non-trivial commutative ring. An element $r \in R$ is *nilpotent* if $r^n = 0$ for some positive integer n.

12.11 Polynomial Rings

Let $R[x]$ be a polynomial ring over a ring R. If an element $f \in R[x]$ has the form,

$$f = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

with $a_n \neq 0$, then we define the degree $\deg(f)$ of f to be n, and a_n is the *leading coefficient* of f. If $a_n = 1$, then f is a *monic* polynomial.

Note that non-zero constant polynomials consisting of a single element of R have degree 0, and the degree of the zero polynomial is undefined, although some texts take it to be -1 or $-\infty$.

Theorem 12.11.1. *If R is an integral domain, then so is $R[x]$.*

Theorem 12.11.2. *If R is an integral domain, then the units of R and $R[x]$ coincide.*

Note that these properties can fail if R is not an integral domain. For example, $\mathbb{Z}_4$ is not a integral domain as $2 \cdot 2 = 4 \equiv 0$ in $\mathbb{Z}_4$. Then, the polynomial $f = 2x + 1 \in \mathbb{Z}_4[x]$ gives $f \cdot f = 4x^2 + 4x + 1 \equiv 1$, so f is a unit in $\mathbb{Z}_4[x] \setminus \mathbb{Z}_4$.

We can also define polynomial rings in multiple variables. We write $R[x_1, \ldots, x_n]$ for the ring of polynomials in n independent commuting indeterminates $x_1, \ldots, x_n$ with coefficients in R. A *monomial* in this ring is an expression of the form $x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$, where $\alpha_1, \ldots, \alpha_n$ are non-negative integers, and a *polynomial* in this ring is a linear combination of these monomials with coefficients in R.

Note that we can also build up a polynomial ring in multiple variables as a chain of polynomial rings in single variables. For instance, if $S = R[x_1]$, then $R[x_1, x_2] = S[x_2]$, and so on. By induction on the previous 2 theorems, if R is an integral domain, then $R[x_1, \ldots, x_n]$ is an integral domain and the units of R and $R[x_1, \ldots, x_n]$ coincide.

Lemma 12.11.3. *$R[x_1, \ldots, x_n]$ is commutative if and only if R is commutative.*

12.11.1 Polynomial Division

Throughout this section, F will be a field.

Theorem (Polynomial Divison with Remainder)**.** *For any $f, g \in F[x]$ with g non-zero, there exist $q, r \in F[x]$ such that $f = qg + r$, where either $r = 0$ or $\deg(r) < \deg(g)$.*

Theorem (Remainder Theorem)**.** *Let $f = f(x) \in F[x]$. Then, for $a \in F$, $f(a) = 0$ if and only if $(x - a)$ divides f.*

Proof. By the previous proposition,

$$f(x) = g(x)(x-a) + r(x)$$

Since $\deg(x-a) = 1$, $r = 0$ or $\deg(r) < 1$, so $r \in F$ is a constant polynomial. Then,

$$\begin{aligned} f(a) &= g(a)(a-a) + r \\ &= r \end{aligned}$$

■

Corollary 12.11.3.1. *If $f \in F[x]$ is not the zero polynomial, then $f(a) = 0$ for at most $\deg(f)$ distinct values of $a \in F$. That is, a polynomial of degree d has at most d roots.*

Proof. By induction on $\deg(f)$. If $\deg(f) = 0$, then f is a constant non-zero function, so $f(a) \neq 0$. If otherwise $\deg(f) > 0$ and f has no roots, we are done.

Now, suppose $f(a) = 0$ for some $a \in F$, so $f = g(x-a)$ with with $\deg(g) = \deg(f) - 1$. If we then have $f(b) = 0$, then either $a = b$, or $g(b) = 0$, in which case, there are at most $\deg(f) - 1$ such values of b by the inductive hypothesis. ■

Theorem 12.11.4. *Let F be a field. Then, all finite subgroups of the unit group F^* are cyclic.*

Corollary 12.11.4.1. *If p is prime, then the set $\mathbb{Z}_p \setminus \{0\} = \{1,2,\dots,p\}$, under multiplication modulo p, is a cyclic group of order $p-1$.*

12.12 Principal Ideal Domains

The ring R will be an integral domain (and is hence commutative) for this section.

Recall that in a commutative ring, the principal ideals are those of the form $(a) = aR$ for some fixed $a \in R$.

A domain R is a *principal ideal domain* (PID) if every ideal of R is principal.

Theorem 12.12.1. *For every field F, the polynomial ring $F[x]$ is a principal ideal domain.*

Various familiar properties of divisibility that hold in $\mathbb{Z}$ hold in more general PIDs. But first, we need to extend the notion of divisibility to general integral domains.

Let $x,y \in R$. We say that x *divides* y if $y = xr$ for some $r \in R$, and we write $x|y$ to denote this relation.

Lemma 12.12.2. *The following statements are equivalent in an integral domains R:*

(i) $x|y$;

(ii) $y \in (x)$;

(iii) $(y) \subseteq (x)$.

Proof. *(i)* → *(ii)*: If $x|y$, then $y = xr$ for some $r \in R$, so $y \in (x) = \{xt : t \in R\}$.

(ii) → *(iii)*: If $y \in (x)$, then $y = xr$ for some $r \in R$, so

$$\begin{aligned} (y) &= \{yt : t \in R\} \\ &= \{(xr)t : t \in R\} \\ &= \{x(rt) : t \in R\} \\ &\subseteq \{xk : k \in R\} \end{aligned}$$

$$= (x)$$

(iii) → *(i)* $y \in \{yt : t \in R\} \subseteq \{xr : r \in R\}$, so $y = xr$ for some $r \in R$ and $x|y$. ■

Let $x,y \in R$. If both $x|y$ and $y|x$, then x and y are *associate* in R, and we write $x \sim y$.

Lemma 12.12.3. *The following statements are equivalent in an integral domains R:*

(i) $x \sim y$*;*

(ii) $(y) = (x)$*;*

(iii) There exists a unit $q \in R$ *such that* $x = qy$*.*

Example.

- In $\mathbb{Z}$, the only units are ±1, so $x \sim y$ if and only if $|x| = |y|$.
- If F is a field, then the units in $F[x]$ are the non-zero constants, so $x \sim y$ if and only if $x = ay$ for some $a \in F \setminus \{0\}$, so every polynomial is associate to a unique monic polynomial.

△

Let $x,y \in R$. A *greatest common divisor* $\gcd(x,y)$, also called a *highest common factor*, is an element $d \in R$ such that,

(i) $d|x$ and $d|y$;

(ii) if $k|x$ and $k|y$ for some $k \in R$, then $k|d$.

so a greatest common divisor is a maximal element with respect to the partial ordering induced by divisibility.

A *least common multiple* $\mathrm{lcm}(x,y)$ is an element $m \in R$ such that,

(i) $x|m$ and $y|m$;

(ii) if $x|k$ and $y|k$ for some $k \in R$, then $m|k$.

so a least common multiple is a minimal element, as above. Greatest common divisors and least common multiples are dual notions. Note that $\gcd(0,x) = x$ and $\mathrm{lcm}(0,x) = 0$ for any $x \in R$.

Note that a greatest common divisor is not unique. For example, in $\mathbb{Z}$, 2 and -2 are both greatest common divisors of 4 and 6. Any two greatest common divisors must divide each other, and are hence associate. Similar statements hold for least common multiples. So, gcds and lcms and are unique up to the associate relation.

Proving existence of gcds is more difficult. In arbitrary integral domains, they do not always exist, but in PIDs, they do, and in fact, for the PID $\mathbb{Z}$ this is exactly the statement of Bézout's identity.

Theorem 12.12.4. *If R is a PID, then* $\mathrm{lcm}(x,y)$ *and* $\gcd(x,y)$ *exist for all* $x,y \in R$*. Furthermore, there exist* $r,s \in R$ *such that* $\gcd(x,y) = rx + sy$*.*

12.12.1 Prime and Irreducible Elements

There are two different ways to characterise prime numbers, but these definitions lead to distinct notions in arbitrary domains.

Let $r \in R \setminus \{0\}$. Then, r is *irreducible* if,

(i) r is not a unit;

(ii) if $r = ab$, then either a or b is a unit.

Let $r \in R \setminus \{0\}$. Then, r is *prime* if,

(*i*) r is not a unit;

(*ii*) if $r|ab$, then $r|a$ or $r|b$.

Theorem 12.12.5. *If R is a domain, then every prime is also irreducible.*

In general, the converse does not hold in an arbitrary integral domain, but it does in a PID.

Theorem 12.12.6. *If R is a PID, then every irreducible is also prime.*

Together, these theorems show that prime and irreducible elements coincide in PIDs.

An integral domain R is a *factorisation domain* (FD) if each non-unit $x \in R \setminus \{0\}$ admits a factorisation $x = r_1 \cdot r_2 \cdots r_n$, where the r_i are irreducible.

A factorisation domain R is furthermore a *unique factorisation domain* (UFD) if for any two factorisations $\prod_{i=1}^{n} r_i = \prod_{i=1}^{m} s_i = x$ of a non-unit $x \in R \setminus \{0\}$, we have $n = m$, and there exists a permutation $\sigma \in S_n$ such that $r_i \sim s_{\sigma(i)}$ for all i.

Theorem 12.12.7. *If R is a UFD, then every irreducible is also prime.*

So, prime and irreducible elements also coincide in UFDs.

Lemma 12.12.8. *A PID is a FD.*

Theorem 12.12.9. *If R is an FD in which all irreducibles are prime, then R is a UFD. In particular, every PID is a UFD.*

Theorem 12.12.10. *Any finite collection of elements in a UFD has a* gcd *and an* lcm.

12.12.2 Number Fields

An ideal I of a ring R is *maximal* if $I \neq R$, but if J is any ideal of R such that $I \subseteq J \subseteq R$, then $I = J$, or $J = R$.

Theorem 12.12.11. *An ideal I in a commutative ring R is maximal if and only if R/I is a field.*

Theorem 12.12.12. *For $a \neq 0$, the principal ideal (a) in a PID R is maximal if and only if a is irreducible.*

If F is a field, and $f \in F[x]$ has degree $\deg(f) > 0$, then the elements of the quotient ring $F[x]/(f)$ correspond to polynomials in $F[x]$ with degree less than f, where multiplication is done modulo f.

When f is irreducible, the previous two theorems imply that $F[x]/(f)$ is a field. The case $F = \mathbb{Q}$ is particularly important as $\mathbb{Q}[x]/(f)$ is isomorphic to a subfield of $\mathbb{C}\mathbb{C}$.

An element $\alpha \in \mathbb{C}$ is *algebraic* over $\mathbb{Q}$ if it satisfies a polynomial $f(\alpha) = 0$ for some $f \in \mathbb{Q}[x]$ with $\deg(f) > 0$. An element that is not algebraic is called *transcendental.*

Recall that for any $\alpha \in \mathbb{C}$, the evaluation map $\phi_\alpha : \mathbb{Q}[x] \to \mathbb{C}$, defined by $f \mapsto f(\alpha)$, is a ring homomorphism. Here, there are two cases to consider; whether α is algebraic or not.

If α is transcendental, then there are no polynomials $f \in \mathbb{Q}[x]$ such that $f(\alpha) = 0$, so $\ker(\phi_\alpha)$ contains only the zero polynomial, and so, by the first isomorphism theorem, we have $\operatorname{im}(\phi_\alpha) \cong \mathbb{Q}[x]$. If α is algebraic, then there exists a non-zero polynomial $f \in \mathbb{Q}[x]$ such that $f(\alpha) = 0$, so $f \in \ker(\phi_\alpha)$, and since $\ker(\phi_\alpha)$ is an ideal of the PID $F[x]$, $\ker(\phi_\alpha)$ must be a principal ideal, so there is some $m \in F[x]$ such that $\ker(\phi_\alpha) = (m)$.

This polynomial m is not necessarily unique, but any two distinct values must divide each other and thus be associate in $F[x]$. By multiplying by constants, we can assume that m is monic, and this monic polynomial is unique and is called the *minimal polynomial* of α over $\mathbb{Q}$.

Theorem 12.12.13. *If α is algebraic in $\mathbb{C}$, then there is a unique non-zero irreducible monic polynomial $m \in \mathbb{Q}[x]$ such that $m(\alpha) = 0$.*

By the first isomorphism theorem, we then have,

$$\operatorname{im}(\phi_\alpha) \cong \mathbb{Q}[x]/(f)$$

and since f is irreducible, (f) is a maximal ideal, and hence $Q[x]/(f)$ is a field, so $\operatorname{im}(\phi_\alpha)$ is a subfield of $\mathbb{C}$, denoted $\mathbb{Q}(\alpha)$.

Fields of this type are called *number fields.*

12.13 Polynomials

A field F is *algebraically closed* if for every $f(x) \in F[x]$ with degree $\deg(f) > 0$, there exists $a \in F$ such that $f(a) = 0$.

Example.

- $\mathbb{C}$ is an algebraically closed field.
- The subfield $\mathbb{A} = \{a \in \mathbb{C} : \exists f \in \mathbb{Q}[x], f(a) = 0\} \subset \mathbb{C}$ of $\mathbb{C}$ of *algebraic numbers* is also an algebraically closed field.

△

Theorem 12.13.1. *If F is an algebraically closed field, then the irreducibles in $F[x]$ are exactly the polynomials of degree 1, so each irreducible is associate to $(x - a)$ for a unique $a \in F$.*

12.13.1 Eisenstein's Criterion

It is difficult to check polynomials in $\mathbb{Z}[x]$ for irreducibility, but *Eisenstein's criterion* provides an sufficient (but not necessary) condition for irreducibility that is often simpler to use.

Let R be a UFD. Then, note that if a non-constant polynomial $f \in R[x]$ is irreducible, its coefficients need to be jointly coprime, as, if a is a non-unit in R that divides all the coefficients of f, then a is a non-unit in $R[x]$ that divides f.

A non-zero polynomial $f = a_n x^n + \cdots + a_1 x + a_0 \in R[x]$ is *primitive* if $\gcd_{0 \leq i \leq n}(a_i) = 1$.

So, any non-zero $f \in \mathbb{R}[x]$ can be written as af_0 where $a \in R$ is the gcd of the coefficients of f and f_0 is primitive.

Theorem (Eisenstein's Criterion)**.** *Let R be a UFD, and let $f = a_n x^n + \cdots + a_1 x + a_0 \in R[x]$ be a primitive polynomial. If there exists a prime $p \in R$ such that,*

- *$p \nmid a_n$;*
- *$p \mid a_i$ for $0 \leq i < n$;*
- *$p^2 \nmid a_0$,*

or,

- *$p^2 \nmid a_n$;*
- *$p \mid a_i$ for $0 \leq i < n$;*
- *$p \nmid a_0$,*

then f is irreducible in $R[x]$.

Example. $3x^3 + 10x^2 + 12x + 2$ is irreducible in $\mathbb{Z}[x]$ as gcd(3,10,12,1) $= 1$, and Eisenstein's criterion applies with $p = 2$. △

12.13.2 Fields of Fractions

Let R be an integral domain, and define the set,

$$\begin{aligned} W &= R \times (R \setminus \{0\}) \\ &= \{(x,y) \in R \times R : y \neq 0\} \end{aligned}$$

We define an equivalence relation on W by $(a,b) \sim (c,d)$ if and only if $a \cdot d = b \cdot c$. Then, the equivalence classes of an element (a,b) is called a *fraction*, and is denoted $\frac{a}{b}$.

Let $Q(R)$ be the set of equivalence classes of W.

Theorem 12.13.2. *If R is an integral domain, then $Q(R)$ is a field under the operations,*

$$\frac{a}{b} + \frac{c}{d} = \frac{a \cdot d + b \cdot c}{b \cdot d} \qquad \frac{a}{b} \cdot \frac{c}{d} = \frac{a \cdot c}{b \cdot d}$$

and the map $\pi : R \to Q(R)$ defined by $r \mapsto \frac{r}{1}$ is an injective ring homomorphism.

The field $Q(R)$ is called the *field of fractions* of an integral domain R.

Example.

- $Q(\mathbb{Z}) = \mathbb{Q}$
- $Q(F[x])$ is the field of *rational functions* p/q, $p,q \in F[x], q \neq 0$, in one variable x, commonly denoted by $F(x)$.

△

12.13.3 Gauss' Lemma

Lemma 12.13.3. *The product of two primitive polynomials is primitive.*

Proof. Let $f = a_0 + a_1x + \cdots + a_mx^m$ and $g = b_0 + b_1x + \cdots + b_nx^n$ be primitive, and $fg = c_0 + c_1x + \cdots + c_{m+n}x^{m+n}$.

If fg is not primitive, then some non-unit of R divides all the coefficients c_i of fg. This non-unit has a least one irreducible factor, so some irreducible $p \in R$ divides c_i for $0 \leq i \leq m + n$.

Since f is primitive, p cannot divide every a_i, so suppose that $p|a_i$ for $0 \leq i < k$, but $p \nmid a_k$ for some $k \geq 0$. Similarly, choose $\ell \geq 0$ such that $p|b_i$ for $0 \leq i < \ell$, but $p \nmid a_\ell$.

We have $c_{k+\ell} = \sum_{i=0}^{k+\ell} a_ib_{k+\ell-i}$, where we take any undefined coefficients to be 0. Since p is a prime that does not divide a_k or b_ℓ, it does not divide a_kb_ℓ, but p does divide every other term in this sum. So, $p \nmid c_{k+\ell}$, which is a contradiction. ■

Theorem 12.13.4. *Let R be a UFD with a field of fractions $Q = Q(R)$. Then, a primitive polynomial in $R[x]$ is irreducible if and only if it is irreducible in $Q[x]$.*

Proof. Let $f \in R[x]$ be primitive. If $\deg(f) = 0$, then f is a unit in R, and is irreducible in neither $R[x]$ nor $Q[x]$. Otherwise, suppose that $f = gh$ for some non-units $g,h \in R[x]$. By primitivity, $\deg(f) > 0$ and $\deg(h) > 0$, so, if f is reducible in $R[x]$, then it is also reducile in $Q[x]$.

Conversely, suppose that f is primitive and reducible in $Q[x]$, so $f = gh$ for some non-units $g,h \in Q[x]$, and again, we have $\deg(f) > 0$ and $\deg(h) > 0$.

Let a_1 be the least common multiple of all the denominators of the coefficients of $g(x)$, so $a_1 g \in R[x]$. Now, let a_2 be the greatest common divisor of all the coefficients of $a_1 g(x)$, and define $a = \frac{a_1}{a_2}$. Then, $a \in Q$, and $ag \in R[x]$ with ag primitive.

Similarly define $b = \frac{b_1}{b_2} \in Q$ with $bh \in R[x]$ and bh primitive. So, by the previous lemma, $f' = abgh = abf$ is primitive and hence $a_2 b_2 f' = a_1 b_1 f$, with f and f' both primitive.

So, $a_1 b_1$ and $a_2 b_2$ are both equal to the greatest common divisor of the coefficients of $a_1 b_1 f$, and by uniqueness of the greatest common divisor, $a_1 b_1$ and $a_2 b_2$ are associate in R. But, then $u = ab = \frac{a_1 b_1}{a_2 b_2}$ is a unit in R, and $f = (ag)(u^{-1} bh)$ is a factorisation of f in $R[x]$, so f is reducible in $R[x]$. ■

In particular, we have *Gauss' lemma*:

Lemma 12.13.5 (Gauss). *A primitive irreducible polynomial in* $\mathbb{Z}[x]$ *is irreducible in* $\mathbb{Q}[x]$.

Corollary 12.13.5.1. *If R is a UFD, then there are two distinct types of irreducibles in $R[x]$; irreducible elements in R, and primitive elements in $R[x]$ that are irreducible in $Q[x]$.*

Proof. This follows from integral domains sharing units with their polynomial rings (Theorem 12.11.2), Gauss' lemma, and the fact that a polynomial of non-zero degree in $R[x]$ that is not primitive is reducible. ■

Theorem 12.13.6. *If R is a UFD, then so is $R[x]$.*

Proof. Clearly, we can factorise any $f \in R[x]$ into a product of irreducible elements of R and irreducible primitive polynomials in $R[x]$. For uniqueness, suppose that

$$\prod_{i=1}^{k}(p_i) \cdot \prod_{i=1}^{n} f_i = \prod_{i=1}^{\ell}(q_i) \cdot \prod_{i=1}^{m} g_m$$

are two factorisations of the same polynomial in $R[x]$, where the p_i and q_i are irreducible in R, and the f_i and g_i are irreducible primitive polynomials in $R[x]$. By Theorem 12.13.4, the f_i and g_i are also irreducible elements of the field of fractions $Q[x]$ of R, whereas the p_i and q_i are units in $Q[x]$.

$Q[x]$ is a PID, so it is a UFD, and hence $n = m$, and, after permuting the g_i if necessary, f_i and g_i are associates in $Q[x]$ for $1 \leq i \leq n$; that is, for each i, we have $g_i = a_i f_i$ for some $a_i \in Q$. But then, $a_i = \frac{b_i}{c_i}$, with $b_i, c_i \in R$, and $b_i f_i = c_i g_i$.

Since f_i and g_i are primitive, c_i and b_i are both greatest common divisors of the coeficients of $c_i f_i$, so b_i and c_i are associates in R, and hence the a_i are all units in R. We can now cancel the f_i and conclude that

$$\prod_{i=1}^{k} p_i = \prod_{i=1}^{\ell}(q_i) \cdot a$$

where $a = \prod_{i=1}^{n} a_i$ is a unit in R. We then have $k = \ell$ as R is a UFD, so, after permuting the q_i if necessary, p_i and q_i are associates for all i. ■

Example.

- $\mathbb{Z}[x]$ is a UFD, but is not a PID as $(2) + (x)$ is not principal.
- By induction on the previous theorem, $R[x_1, x_2, \ldots, x_n]$ is a UFD for any $n > 0$ for any UFD R.

△

12.14 Exercises

These questions are again in no particular order of subject. Some of the questions are significantly more difficult than others, mostly those at the end, while some can be done in a single sentence – some are solvable just by recalling and stating definitions of algebraic structures.

1. Prove that the empty set cannot form a ring.
2. Prove that the trivial ring is commutative.
3. Prove that the set of 2×2 matrices with,
   (a) integer entries forms a ring.
   (b) integer entries does not form a field.
   (c) integer entries and non-zero determinant, along with the 2×2 zero matrix, still does not form a field.
4. Prove that $\mathbb{Z}$ is a ring under addition and multiplication.
5. Prove that $\mathbb{Z}/4\mathbb{Z}$ is a ring under addition and multiplication modulo 4.
6. Prove that $2\mathbb{Z}$ and $3\mathbb{Z}$ are not isomorphic as rings.
7. Prove that $\mathbb{Z}[i]/(1+i)$ is isomorphic to $\mathbb{Z}/2\mathbb{Z}$.
8. Determine all ring homomorphisms $\mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$. (Find all of them, and prove there are no others.)
9. Determine all ring homomorphisms $\mathbb{Z} \to R$ where R is an arbitrary ring.
10. Let $(R, +, \times)$ be a ring, and suppose the additive identity is 0_R. Prove that $\forall x \in R, 0_R \times x = x \times 0_R = 0_R$. That is, prove that the additive identity is also the zero element for the ring product (hence also justifying the name, "ring zero" for this element).
11. Give an example of a non-commutative ring.
    (a) Give an example of a finite non-commutative ring.
    (b) What is order of the smallest possible non-commutative ring?
12. Prove that complex conjugation is a ring homomorphism $\mathbb{C} \to \mathbb{C}$.
13. Prove that there is no ring homomorphism $\mathbb{Z}/n\mathbb{Z} \to \mathbb{Z}$ for any $n \geq 1$.
14. Suppose that R is a ring such that for all elements $a,b,c \in R$, $ca = cb$ implies $a = b$. Prove that R is commutative.
15. Let R and S be rings, and suppose $\phi : R \to S$ is a surjective ring homomorphism. Prove that the image of an ideal of R under f is an ideal of S.
16. Let R be a ring. Prove that the following statements are equivalent:
    (i) R is a field.
    (ii) The only ideals of R are (0) and R.
    (iii) All ring homomorphisms $\phi : R \to S$ are injective for any ring S.
17. Let R be a commutative ring, and consider the polynomial ring $R[x,y]$. Let (x) be the principal ideal of $R[x,y]$ generated by x. Prove that $R[x,y]/(x)$ is isomorphic to $R[y]$ as a ring.
18. Let I be a non-zero ideal of the Gaussian integers, $\mathbb{Z}[i]$. Prove that $\mathbb{Z}[i]/I$ is finite.

19. Let $\phi : R \to S$ be a ring homomorphism, and let I be a prime ideal of S. Prove that $f^{-1}[I]$ is a prime ideal of R.
20. Let R be an integral domain, and let I be an ideal of R. Prove or disprove the statement that R/I is also an integral domain.
21. Prove that for each positive integer n, the polynomial

    $$\left(\prod_{i=1}^{n}(x-i)\right)-1$$

    is irreducible over $\mathbb{Z}$.
22. Let F be a finite field of characteristic p. Prove that the number of elements of F is p^n for some positive integer n.
23. Prove that the only field automorphism of $\mathbb{R}$ is the identity map.
24. Prove all algebraically closed fields are infinite.
25. Prove that the quotient ring $\mathbb{Z}[i]/(1+i)$ is a field.
26. Let R be a non-trivial ring, and suppose that $a,b \in R$ are elements such that $ab = 1_R$.
    (a) Prove that if a or b is not a zero divisor, then $ba = 1_R$.

    Now suppose that $ba \neq 1_R$.

    (b) Prove that $1 - ba$ is idempotent.
    (c) Prove that $b^n(1 - ba)$ is nilpotent for all positive integer n.
    (d) Prove that R has infinitely many nilpotent elements.
27. Is it possible for the equation $x + x = 1$ to have more than one solution in x in a not necessarily commutative ring R?
28. Find a polynomial of degree 2 over $\mathbb{Z}/4\mathbb{Z}$ that has 4 roots.
29. Compute the characteristics of the following rings:
    (a) $\mathbb{Z}$;
    (b) $\mathbb{Q}$;
    (c) $\mathbb{Z}/n\mathbb{Z}$;
    (d) $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/4\mathbb{Z} \times \mathbb{Z}/10\mathbb{Z}$;
    (e) $\mathbb{Z}[i]/(1+i)$, where i is the imaginary unit;
    (f) $\mathbb{Z}[\omega]/(2-5\omega)$ where ω is a primitive third root of unity.
30. Prove that the characteristic of any field is either zero or prime.
31. Find a ring R that that satisfies $\mathbb{Z} \subseteq R \subseteq \mathbb{Q}$.
32. Let R be any commutative ring.
    (a) Prove that if $f,g \in R[x]$, and g is monic, then there exist $q,r \in R[x]$ with $f = gq + r$, where either $r = 0$ or $\deg(r) < \deg(g)$.
    (b) By means of a counterexample, show that the previous statement does not hold if we do not require that g is monic.

33. Let $\phi : \mathbb{Z}[x] \to \mathbb{R}$ be a ring homomorphism defined by $f \mapsto f\big(\sqrt{3}\big)$. Prove that $\ker(\phi)$ is a principal ideal in $\mathbb{Z}[x]$.
34. Let R be a finite commutative ring. Prove that every prime ideal of R is a maximal ideal of R.
35. Let R be a PID, and let $a \in R$ be a non-zero non-unit element. Prove that the following statements are equivalent:
    (a) The ideal (a) is maximal.
    (b) The ideal (a) is prime.
    (c) The element a is irreducible.
36. Prove that the polynomial $x^n - 2$ is irreducible over $\mathbb{Q}$ using Eisenstein's criterion, and hence deduce that $\sqrt[n]{2}$ is irrational for all integers $n \geq 2$.
37. Prove that $\mathbb{Z}[x]/(n)$ is isomorphic to $(\mathbb{Z}/n\mathbb{Z})[x]$.
38. Let R be a ring, and let x be an nilpotent element of R. Prove that $1_R + x$ and $1_R - x$ are units.
39. Let R be a commutative ring, and define

    $$\mathrm{Nil}(R) = \{r \in R : \exists n \geq 1, r^n = 0_R\}$$

    to be the subset of nilpotent elements of R.
    (a) Prove that $\mathrm{Nil}(R)$ is an ideal of R.
    (b) Show by means of a counterexample that if R is not commutative, then $\mathrm{Nil}(R)$ is not necessarily an ideal of R.
    (c) Prove that if $r \in \mathrm{Nil}(R)$, then $1_R - r$ is invertible in R.
    (d) Prove that

    $$\mathrm{Nil}(R) = \bigcap_{\substack{I \subseteq R \\ I \mathrm{prime}}} I$$

    (e) Prove that $\mathrm{Nil}\big(R/\mathrm{Nil}(R)\big) = \{0_{R/\mathrm{Nil}(R)}\}$.
40. Determine all rings that have equal cardinality and characteristic.

Chapter 13

Group Theory

“One should never try to prove anything that is not almost obvious.”

— Alexander Grothendieck

13.1 Glossary

$H \leq G$	H is a subgroup of G.
$H \trianglelefteq G$	H is a normal subgroup of G.
gH	A coset of H in G; the set $gH = \{gh : g \in G\}$.
G/H	The set of left cosets of H in G; the set $\{gH : g \in G\}$.
G/N	The quotient or factor group of G by N; the set of left cosets of a normal subgroup N in G, equipped with the operation $gN \circ hN = ghN$
$[G : H]$	The index of H in G; the cardinality $\lvert G/H \rvert$; the number of distinct left cosets of H in G.
${}^{g}h$	The conjugation of h by g; the element ghg^{-1}.
${}^{g}H$	The conjugation of a subset $H \subseteq G$ by an element $g \in G$; the set $gHg^{-1} = \{ghg^{-1} : h \in H\}$.
$N_G(H)$	The normaliser of H in G; the set $\{g \in G : gHg^{-1} = H\}$. The normaliser is always a subgroup of G.
$C_G(x)$	The centraliser or commutant of x in G; the set of elements that commute with x; the set $\{g \in G : gx = xg\}$. The centraliser is always a subgroup of G.

$Z(G)$ The centre of G; the set of elements that commute with all elements of G; the set $\{g \in G : \forall h \in G : gh = hg\}$. The centre is always a normal subgroup in G.

$\mathrm{Cl}(x)$, ${}^{G}x$ The conjugacy class of x; the set $\{gxg^{-1} : g \in G\}$.

$\mathrm{Orb}_G(x)$ The orbit of x in G; the set of possible images of x under an action; the set $\{g \cdot x : g \in G\}$.

$\mathrm{Stab}_G(x)$ The stabiliser of x in G; the set of elements that fix x; the set $\{g \in G : g \cdot x = x\}$. The stabiliser is always a subgroup of G.

$\mathrm{fix}_X(g)$ The set of fixed points of g; the set $\{x \in X : g \cdot x = x\}$.

$\mathrm{Syl}_p(G)$ The set of Sylow p-subgroups of G.

$F_p(G)$ The set $\{x \in G : x \neq 1_G \text{ and } |x| \text{ is a power of } p\}$.

$|G|_p$ The highest power of p that divides G; if $|G| = p^n m$, then $|G|_p = p^n$.

$H \ltimes_\phi K$ The semidirect product of H and K; the set $H \times K$ equipped with the multiplication $(h_1,k_1) \cdot (h_2,k_2) := (h_1h_2,\phi_{h_2^{-1}}(k_1)k_2)$, where $\phi : H \to \mathrm{Aut}(K)$ is a homomorphism and $\phi(h) = \phi_h$.

$[g,h]$ The commutator of g and h; the element $ghg^{-1}h^{-1}$.

$[G,G]$ The commutator subgroup of G; the subgroup generated by $\big\langle [g,h] \bigm| g,h \in G \big\rangle$.

$[H,K]$ The commutator subgroup of H and K, given $H,K \leq G$; the subgroup generated by $\big\langle [h,k] \bigm| h \in H,k \in K \big\rangle$.

G^{ab} The abelianisation of G; the abelian quotient group $G/[G,G]$.

$G^{(n)}$ The nth derived subgroup of G, where $G^{(0)} = G$ and $G^{(n)} = \big[G^{(n-1)},G^{(n-1)}\big]$ for $n \in \mathbb{N}$.

13.2 Review

Recall that a *group* is a pair $(G,\circ)$, consisting of an *underlying set* G and a *group operation* $\circ : G\times G \to G$ that satisfies the following properties:

(G1) $\forall a,b,c \in G : (a \circ b) \circ c = a \circ (b \circ c)$ (associativity);

(G2) $\exists 1_G \in G, \forall g \in G : g \circ 1_G = 1_G \circ g = g$ (existence of identity);

(G3) $\forall g \in G, \exists g^{-1} \in G : g \circ g^{-1} = g^{-1} \circ g = 1_G$ (existence of inverses).

The group is furthermore *abelian* if the group operation additionally satisfies

(A) $\forall a,b \in G : a \circ b = b \circ a$ (commutativity).

When the context is clear, we will usually omit the operation and simply say that G is a group.

Sometimes, closure of $\circ$ over the set G is also included as an axiom, but this is implicit in $\circ$ being an operation over G.

It follows from these axioms that the identity element and the inverse of any given element g are unique, so we are justified in calling them *the* identity and *the* inverse of g.

The number of elements in a group G is called the *order* of G, and is denoted by $|G|$. (This coincides with the cardinality of the underlying set, so the notation is meaningful.)

Theorem 13.2.1 (Basic Properties of Groups)**.**

- *If $ga = gb$ or $ag = bg$, then $a = b$ (cancellative property);*
- *The identity element 1_G is unique;*
- *For every element g, the inverse g^{-1} is unique;*
- *If e_ℓ is a left identity (i.e. $e_\ell g = g$ for all $g \in G$), and/or e_r is a right identity, then $e_\ell = 1_G = e_r$;*
- *If ℓ is a left inverse for an element g (i.e. $\ell g = 1_G$), and/or r is a right inverse for g, then $\ell = g^{-1} = r$;*
- *For all $a,b \in G$, $(ab)^{-1} = b^{-1}a^{-1}$;*
- *For all $g \in G$, $(g^{-1})^{-1} = g$.*

13.2.1 Symmetric Groups

Let X be a finite set. We write $\mathrm{Sym}(X)$ for the set of bijections $f : X \to X$. This set has group structure under composition:

(G1) For any functions $f,g,h \in \mathrm{Sym}(X)$ and $x \in X$, $\big((f \circ g) \circ h\big)(x) = f\big(g(h(x))\big) = \big(f \circ (g \circ h)\big)(x)$;

(G2) The identity function id_X is the identity element;

(G3) The inverse function f^{-1} for a function f is also its inverse in the group.

This group is called the *symmetric group* on X, and its elements are called *permutations*.

The symmetric group is abelian if and only if $|X| \leq 2$.

13.2.1.1 Cycle Notation

Let $a_1,a_2,\ldots,a_r$ be distinct elements of a set X. The *cycle* $(a_1,a_2,\ldots,a_r)$ represents the permutation $f \in \mathrm{Sym}(X)$ with

- $f(a_i) = a_{i+1}$ for $1 \leq i < r$;

- $f(a_r) = a_1$;
- $f(b) = b$ for $b \in X \setminus \{a_1,a_2,\dots a_r\}$.

The empty cycle () is a cycle, corresponding to the identity permutation id_X.

Two cycles $(a_1,\dots,a_r)$ and $(b_1,\dots,b_s)$ are *disjoint* if $\{a_1,\dots,a_r\} \cap \{b_1,\dots,b_s\} = \varnothing$.

Note that the representation of a permutation in cycle notation is not unique. For instance, (1,2,3) = (3,1,2) = (2,3,1).

Theorem 13.2.2.

- $\big|\mathrm{Sym}(X)\big| = |X|!$.
- *Every permutation in* $\mathrm{Sym}(X)$ *can be expressed as a product of disjoint cycles.*

  Moreover, this product is unique in the sense that if $f \in \mathrm{Sym}(X)$ *has representations* $f = f_1 \cdots f_m = g_1 \cdots g_n$*, where the* f_i *and* g_i *are disjoint cycles of length greater than* 1*, then* $m = n$ *and* $\{f_1,\dots,f_m\} = \{g_1,\dots,g_n\}$.

13.2.2 General Linear Groups

Let K be a field and n be a positive integer. We define the set $\mathrm{GL}_n(K)$ to be the set of invertible $n \times n$ matrices with entries in K. Under the operation of matrix multiplication, this set forms a group called the *general linear group of dimension n over K*.

Recall that if K is a field (or more generally, a ring), then the *characteristic* of K is the smallest positive number p such that

$$p1_K = \underbrace{1_K + \cdots + 1_K}_{p} = 0_K$$

if such a number exists, and 0 otherwise. In the finite case, such a number will always exist, and moreover, this number is prime. The characteristic also satisfies

$$|K| = p^n$$

for some positive integer n.

Theorem 13.2.3. *Let K be a finite field, and let $q = |K|$. Then,*

$$\big|GL_n(K)\big| = q^{\binom{n}{2}} \prod_{i=1}^{n} (q^i - 1)$$

13.2.3 Orders of Elements

In multiplicative notation, we write g^n to mean the n-fold iteration of the group operation on g. If $n = 0$, then $g^n = 1_G$, and if $n < 0$, then $g^n = (g^{-1})^n$.

Let G be a group, and let $g \in G$. The *order* of g, denoted by $|g|$ is the smallest positive integer n such that $g^n = 1_G$, if such a number exists, and ∞ otherwise:

$$|g| := \begin{cases} \min\{n \in \mathbb{Z}^+ : g^n = 1_G\} & \exists n \in \mathbb{Z}^+ : g^n = 1_G \\ \infty & \text{otherwise} \end{cases}$$

Theorem 13.2.4.

- *The identity element* 1_G *is the unique element of order* 1.
- *For all* $g \in G$*,* $|g| = |g^{-1}|$.

Proof. Clearly, 1_G has order 1. Now suppose an element $e \in G$ also has order 1. Then, $e = e^1 = 1_G$, so $e = 1_G$.

Suppose $|g| = n$. Then, $(g^{-1})^n = (g^n)^{-1} = (1_G)^{-1} = 1_G$, so $|g^{-1}| = n$. ■

Lemma 13.2.5. *Let G be a group and let $a,b \in G$ have finite order. Then,*

(*i*) *If $\ell \in \mathbb{Z}^+$, then $a^\ell = 1_G$ if and only if n divides ℓ;*

(*ii*) *If $m \in \mathbb{Z}^+$, then $|a^m| = |a|/\gcd(|a|,m)$;*

(*iii*) *If a and b commute, then $|ab|$ divides $\operatorname{lcm}(|a|,|b|)$;*

(*iv*) *If a and b commute and $\langle a\rangle \cap \langle b\rangle = \{1_G\}$, then $|ab| = \operatorname{lcm}(|a|,|b|)$.*

Proof.

(*i*) If $n = |a|$ divides ℓ, then $\ell = nr$ for some integer r, so $a^\ell = (a^n)^r = (1_G)^r = 1_G$.

Conversely, if $a^\ell = 1_G$, then by the Euclidean algorithm, $\ell = qn + r$ for some $q \in \mathbb{Z}$ and $0 \le r < n$. Then, $1_G = a^\ell = a^{qn+r} = (a^n)^q a^r = a^r$. Since n is the minimum positive integer such that $a^n = 1_G$, and $r < n$, we have $r = 0$.

(*ii*) If m divides $n = |a|$, then $n = m\ell$ for some integer ℓ, so $a^{m\ell} = 1$. Then, by (*i*), n divides $m\ell$, so $|a|/m$ divides ℓ. Thus, the order of a^m is the least positive integer ℓ such that $|a|/m$ divides ℓ, so $|a^m| = |a|/m$ if m divides $|a|$.

More generally, let $k = \gcd(n,m)$, so for some integer s, we have,

$$m = ks$$
$$\frac{mn}{k} = sn$$

so $(a^m)^{n/k} = (a^n)^s = 1_G$, so $|a^m|$ divides n/k by (*i*).

We also have $k = sn + tm$ for some integers s,t by Bézout's lemma, so $a^k = a^{sn}a^{tm} = a^{tm}$. But,

$$a^{tm|a^m|} = (a^{mn^m})^t = 1_G$$

so $|a^{tm}| = |a^k|$ divides $|a^m|$ by (*i*). We also have $|a^k| = |a|/k$ from above, so $|a|/k$ divides $|a^m|$. It follows that $|a|/k = |a^m|$ as required.

(*iii*) Let $\ell = \operatorname{lcm}(|a|,|b|)$. Then, $(ab)^\ell = a^\ell b^\ell = 1_G$, so ab has finite order, and $|ab|$ divides ℓ by (*i*).

(*iv*) Let $n = |ab|$. Then $1_G = (ab)^n = a^n b^n$ so $a^n = (b^{-1})^n = b^{-n}$, so $a^n \in \langle b\rangle$. But, $a^n \in \langle a\rangle$, so $1_G = a^n \in \langle a\rangle \cap \langle b\rangle = \{1_G\}$. Similarly, $b^n = 1_G$. So, $|a|$ and $|b|$ divide n by (*i*), so $\operatorname{lcm}(|a|,|b|) \le n$, so (*iii*) gives $|ab| = \operatorname{lcm}(|a|,|b|)$.

■

13.2.4 Subgroups

A subset $H \subseteq G$ of a group G is a *subgroup* of $(G,\circ)$ if $(H,\circ)$ is itself a group, and we write $H \le G$ to denote this relation.

Lemma 13.2.6. *Let $H \subseteq G$ be a non-empty subset. Then, $H \le G$ if and only if for all $g,h \in H$, we have $gh^{-1} \in H$.*

Given an element $g \in G$, the (*cyclic*) *subgroup generated by* g is the subgroup defined by

$$\langle g\rangle \coloneqq \{g^i : i \in \mathbb{Z}\}$$

and we say that g is a *generator* of G. Conversely, a group is called *cyclic* if it is in this form.

Lemma 13.2.7. *If $G = \langle g \rangle$ is cyclic, then $|G| = |g|$.*

More generally, given a non-empty subset $S \subseteq G$, the *subgroup generated by* S is the subgroup defined by

$$\langle S \rangle \coloneqq \{s_1^{\epsilon_1} s_2^{\epsilon_2} \cdots s_m^{\epsilon_m} : m \in \mathbb{N}, s_i \in S, \epsilon_i \in \{\pm 1\}\}$$

That is, the subgroup containing all linear combinations of elements in S. If $S = \{s_1, \ldots, s_n\}$, then we also write $\langle S \rangle = \langle s_1, \ldots, s_n \rangle$ for this subgroup.

13.2.4.1 Cosets

Given a subgroup $H \leq G$ of a group G and an element $g \in G$, the *left coset* gH of H in G is the set

$$gH = \{gh : h \in H\} \subseteq G$$

Lemma 13.2.8. *Let G be a group and $H \leq G$ a subgroup. Then, the following are equivalent for all $g, k \in G$:*

(i) $k \in gH$;

(ii) $gH = kH$;

(iii) $g^{-1}k \in H$.

Proof. $(i) \to (ii)$: Note that $hH = H$ for all $h \in H$. Now, if $k \in gH$, then $k = gh$ for some $h \in H$, so $kH = (gh)H = g(hH) = gH$.

$(ii) \to (iii)$: Because H is a subgroup, $1_G \in H$, so $k = k1_G \in kH$. If $kH = gH$, then also $k \in gH$, so for some $h \in H$, $k = gh$, so $g^{-1}k = h \in H$.

$(iii) \to (i)$: If $g^{-1}k = h \in H$, then $k = gh \in gH$. ■

Let G be a group and $H \leq G$ be a subgroup. Define the relation $\sim_H$ on G with $g \sim_H h$ if and only if $gH = hH$.

Corollary 13.2.8.1. *$\sim_H$ is an equivalence relation on G.*

Lemma 13.2.9. *Let G be a group and $H \leq G$ be a subgroup. Then,*

(i) For all $g, h \in G$, either $gH = hH$ or $gH \cap hH = \varnothing$;

(ii) If $\{g_iH\}_{i \in I}$ is the set of $\sim_H$-equivalence classes in G, then

$$G = \bigsqcup_{i \in I} g_i H$$

Proof. Since $\sim_H$ is an equivalence relation, distinct $\sim_H$-equivalence classes are pairwise disjoint and partition G. Both parts follow. ■

Theorem 13.2.10 (Lagrange)**.** *Let G be a finite group and let $H \leq G$ be a subgroup. Then, $|H|$ divides $|G|$. Specifically,*

$$|G| = |G : H||H|$$

Proof. The left cosets of H in G partition G by the previous lemma. Also, each left coset gH is equinumerous to H since $h \mapsto gh$ is a bijection $H \to gH$ (with inverse given by $h \mapsto g^{-1}h$), and the number of left cosets is the index $[G : H]$. The result follows. ■

Let G be a group and $H \leq G$ be a subgroup.

- The set of left cosets of H in G is denoted by $G/H \coloneqq \{gH : g \in G\}$.
- The number of distinct left cosets of H in G (i.e. the cardinality $|G/H|$) is called the *index* of H in G, and is denoted by $[G : H]$. If G is finite, then

$$[G : H] = |G|/|H|$$

Corollary 13.2.10.1. *Let G be a finite group and let $g \in G$. Then $|g|$ divides $|G|$.*

Proof. The subgroup $\langle g\rangle$ has order $|g|$. The result follows from Lagrange's theorem. ∎

13.2.5 Normal Subgroups

Lemma 13.2.11. *Let $H \leq G$ be a subgroup of a group G, and let $g \in G$. Then, ${}^{g}H = gHg^{-1} = \{ghg^{-1} : h \in H\}$ is a subgroup of G.*

Let G be a group and let $H \leq G$ be a subgroup.

- H is *normal* in G if $gHg^{-1} = H$ for all $g \in G$, and we write $H \trianglelefteq G$ to denote this relation.
- The *normaliser* of H in G, is the subgroup of G defined by

$$N_G(H) \coloneqq \{g \in G : gHg^{-1} = H\}$$

Note that H is normal in G if and only if $N_G(H) = G$.

Theorem 13.2.12. *Let G be a group and let $H \leq G$ be a subgroup. Then,*

(*i*) *H is normal in G if and only if $gHg^{-1} \subseteq H$ for all $g \in G$;*

(*ii*) *If $[G : H] = 2$, then H is normal in G;*

(*iii*) *$H \trianglelefteq N_G(H) \leq G$;*

(*iv*) *$G \trianglelefteq G$;*

(*v*) *$\{1_G\} \trianglelefteq G$.*

A non-trivial group G is *simple* if the only normal subgroups of G are $\{1_G\}$ and G.

Given subsets $A,B \subseteq G$ of a group G, we write $AB \coloneqq \{ab : a \in A, b \in B\}$ for the internal product of A and B. In general, this is not a subgroup, even if A and B are both subgroups.

Lemma 13.2.13. *Let N be normal in G, and let $g,h \in G$. Then, $(gN)(hN) = ghN$.*

Let N be normal in G. Then, the binary operation $\circ : G/N \times G/N \to G/N$ defined by $(gN)\circ(hN) = ghN$ is called the *natural binary operation* of G/H.

With the natural binary operation $\circ$, $(G/N,\circ)$ is a group called the *quotient* or *factor* group of G by N.

13.2.6 Group Homomorphisms

Let $(G,\circ)$ and $(H,*)$ be groups.

A map $\phi : G \to H$ is a *group homomorphism* if $\phi(g \circ h) = \phi(g) * \phi(h)$ for all $g,h \in G$.

If ϕ is a homomorphism and has an inverse (or equivalently, is bijective), then φ is an *isomorphism*, and we say that G and H are *isomorphic*, written as $G \cong H$. An isomorphism from a group to itself is also called an *automorphism*.

We define the *kernel* and *image* of a homomorphism ϕ as the sets

$$\begin{aligned}\ker(\phi) &\coloneqq \{g \in G : \phi(g) = 1_G\}\\ \operatorname{im}(\phi) &\coloneqq \{\phi(g) : g \in G\}\end{aligned}$$

Let N be normal in G. A the map $\pi : G \to G/N$ defined by $\pi(g) = gN$ is a surjective homomorphism called the *quotient map* or *natural homomorphism* from G to G/N.

Theorem 13.2.14. *If n and m are coprime, then $C_n \times C_m \cong C_{nm}$.*

Theorem 13.2.15 (First Isomorphism Theorem)**.** *Let G and H be groups, and let $\phi : G \to H$ be a group homomorphism. Then,*

(*i*) $\ker(\phi) \trianglelefteq G$*;*

(*ii*) $\operatorname{im}(\phi) \leq H$*;*

(*iii*) $G/\ker(\phi) \cong \operatorname{im}(\phi)$.

Theorem 13.2.16 (Second Isomorphism Theorem)**.** *Let G be a group, $H \leq G$ a subgroup, and $N \trianglelefteq G$ be normal in G. Then,*

(*i*) $NH = HN \leq G$*;*

(*ii*) $H \cap N \trianglelefteq H$*;*

(*iii*) $H/(H \cap N) \cong NH/N$.

Theorem 13.2.17 (Third Isomorphism Theorem)**.** *Let G be a group, and let $N,K \trianglelefteq G$ be normal in G with $N \subseteq K \subseteq G$. Then,*

(*i*) $K/N \trianglelefteq G/N$*;*

(*ii*) $(G/N)/(K/N) \cong G/K$.

Theorem 13.2.18 (Correspondence Theorem)**.** *Let G be a group, and let $N \trianglelefteq G$ be normal in G. Then, there is a bijection between the subgroups of G containing N and the subgroups of G/N. More precisely, the map*

$$f : \{S : S \leq G/N\} \to \{S : N \leq S \leq G\} : S \mapsto S/N$$

is a bijection, and moreover, this map sends normal subgroups to normal subgroups.

13.3 Permutation Groups

Let X be a set. A subgroup of $\operatorname{Sym}(X)$ is called a *permutation group* on X.

For $g \in \operatorname{Sym}(X)$, the *support* of g is the set

$$\operatorname{supp}(g) \coloneqq \{x \in X : g(x) \neq x\}$$

and for a permutation group G, the *support* of G is the set

$$\operatorname{supp}(G) \coloneqq \{x \in X : g(x) \neq x\}$$

If $G = \langle g \rangle \leq \operatorname{Sym}(X)$, then $\operatorname{supp}(\langle g \rangle) = \operatorname{supp}(g)$. Also note that if

$$g = (a_1, \ldots, a_{m_1}) \cdots (a_{m_{t-1}+1}, \ldots, a_{m_t})$$

is a product of disjoint cycles, then

$$\operatorname{supp}(g) = \{a_1, \ldots, a_{m_1}, a_{m_1+1}, \ldots, a_{m_{t-1}+1}, \ldots, a_{m_t}\}$$

Theorem 13.3.1. *Let X be a finite set. Then,*

(*i*) *Disjoint cycles in* $\operatorname{Sym}(X)$ *commute;*

(*ii*) *If $f = (a_1, \ldots ,a_r) \in \operatorname{Sym}(X)$ is a cycle of length r, then f has order $|f| = r$.*

More generally, if $f = f_1 \cdots f_m$ is a product of disjoint cycles, then f has order

$$|f| = \operatorname{lcm}\bigl(|f_1|, \ldots ,|f_m|\bigr)$$

(*iii*) *Let $f = (a_1, \ldots ,a_r) \in \operatorname{Sym}(X)$ and $g \in \operatorname{Sym}(X)$. Then,*

$$^{g}f = gfg^{-1} = \bigl(g(a_1), \ldots ,g(a_r)\bigr)$$

Let $n \geq 3$ and set $X = \{1, \ldots ,n\}$. Define the permutations $\sigma,\tau \in \operatorname{Sym}(X)$ by

$$\sigma := (1, \ldots ,n)$$
$$\tau := \prod_{i=1}^{\lfloor \frac{n}{2} \rfloor}(i,n - i + 1)$$

Then, the *dihedral group* D_{2n} of order $2n$ is the subgroup of $\operatorname{Sym}(X)$ generated by σ and τ.

Example. If $n = 8$, then

$$D_{16} = \bigl\langle\{(1,2,3,4,5,6,7,8),(1,8)(2,7)(3,6)(4,5)\}\bigr\rangle$$

△

Lemma 13.3.2. *If $H,K \leq G$ with $H = \langle A\rangle$ finite and $K = \langle B\rangle$ for some subsets $A,B \subseteq G$, then $K \subseteq N_G(H)$ if and only if $^{b}a \in H$ for all $a \in A$ and $b \in B$.*

Theorem 13.3.3. *Let $n \geq 3$ and $D_{2n} = \bigl\langle\{\sigma,\tau\}\bigr\rangle$. Then,*

(*i*) $|D_{2n}| = 2n$*;*

(*i*) $\langle\sigma\rangle \trianglelefteq D_{2n}$*, and* $|\langle\sigma\rangle| = n$*. In particular,* D_{2n} *is not simple.*

Let X be a finite set. A permutation $f \in \operatorname{Sym}(X)$ is *even* if it has an even number of cycles of even length in its decomposition into disjoint cycles, and is *odd* otherwise.

Equivalently, a permutation is even if it can be decomposed into an even number of not necessarily disjoint transpositions and odd otherwise.

The set $\operatorname{Alt}(X) := \{f \in \operatorname{Sym}(X) : f \text{ is even}\}$ is the *alternating group* on X, and is a subgroup of $\operatorname{Sym}(X)$ of order $|X|!/2$. That is, $[\operatorname{Sym}(X) : \operatorname{Alt}(X)] = 2$.

Theorem 13.3.4. *If X and Y are finite sets with $|X| = |Y|$, then* $\operatorname{Sym}(X) \cong \operatorname{Sym}(Y)$*.*

Proof. For any bijection $F : Y \to X$, the homomorphism $\phi : \operatorname{Sym}(X) \to \operatorname{Sym}(Y)$ defined by $\phi(f) = F^{-1} \circ f \circ F$ is an isomorphism. ∎

We write S_n for the symmetric group on the set $\{1, \ldots ,n\}$. By the previous theorem, $\operatorname{Sym}(X) \cong S_n$ whenever $|X| = n$.

13.3.1 Group Actions

Let G be a group and X a set. A (*left*) *group action* of G on X is a map $\cdot : G \times X \to X$ such that

(*i*) $(gh) \cdot x = g \cdot (h \cdot x)$ for all $g,h \in G$ and $x \in X$;

(*ii*) $1_G \cdot x = x$ for all $x \in X$.

In this case, we say that G *acts on* X or that X is a G*-set.*

Three important group actions are as follows:

- **Left-multiplication**:

  Let G be a group and take $X = G$. Then, $g \cdot x \coloneqq gx$ defines an action of G on itself:

  (*i*) $(gh) \cdot x = (gh)x = g(hx) = g \cdot (h \cdot x)$;

  (*ii*) $1_G \cdot x = 1_G x = x$.

- **Conjugation**:

  Let G be a group and take $X = G$. Then, $g \cdot x \coloneqq gxg^{-1}$ defines an action of G on itself:

  (*i*) $(gh) \cdot x = (gh)x(gh)^{-1} = ghxh^{-1}g^{-1} = g \cdot (hxh^{-1}) = g \cdot (h \cdot x)$;

  (*ii*) $1_G \cdot x = 1_G x 1_G^{-1} = x$.

- **Action on Cosets**:

  Let G be a group and $H \leq G$ be a subgroup. Take $X = G/H \coloneqq \{gH : g \in G\}$ to be the set of left cosets of H in G. Then, $g \cdot (xH) = (gx)H$ defines a group action on this set of cosets:

  (*i*) $(gh) \cdot xH = g(hxH) = g \cdot (hxH) = g \cdot (h \cdot xH)$;

  (*ii*) $1_G \cdot xH = (1_G x)H = xH$.

Theorem 13.3.5 (Group Action Induces Homomorphism into Symmetric Group)**.** *Let $\cdot$ be an action of a group G on a set X. For $g \in G$, define the map $\phi(g) : X \to X$ by $\phi(g)(x) = g \cdot x$. Then, $\phi(g) \in \mathrm{Sym}(X)$ and $\phi : G \to \mathrm{Sym}(X)$ is a homomorphism.*

Proof. For any $g,h \in G$ and $x \in X$,

$$\begin{aligned} \phi(gh)(x) &= (gh) \cdot x \\ &= g \cdot (h \cdot x) \\ &= \big(\phi(g)\phi(h)\big)(x) \end{aligned}$$

■

Let $\cdot$ be an action of a group G on a set X. The *kernel* of the action $\cdot$, denoted $\ker(G,X,\cdot\,)$, is defined to be the kernel of the homomorphism $\phi : G \to \mathrm{Sym}(X)$ as defined above:

$$\ker(G,X,\cdot\,) \coloneqq \{g \in G : \forall x \in X, g \cdot x = x\} \subseteq G$$

The *image* of the action $\cdot$, denoted $\mathrm{im}(G,X,\cdot\,)$ is the image of ϕ:

$$\mathrm{im}(G,X,\cdot\,) \coloneqq \{\phi(g) : g \in G\} \subseteq \mathrm{Sym}(X)$$

Note that by the first isomorphism theorem, we have

- $\ker(G,X,\cdot\,) \trianglelefteq G$;
- $\mathrm{im}(G,X,\cdot\,) \leq \mathrm{Sym}(X)$.

The action $\cdot$ is *faithful* if the kernel is trivial, $\ker(G,X,\cdot\,) = \{1_G\}$, and *trivial* if the kernel is the entire group, $\ker(G,X,\cdot\,) = G$.

Example.

(*i*) The left-multiplication action of a group on itself is always faithful.

(*ii*) The conjugation action of a group on itself is trivial if and only if $gxg^{-1} = x$ for all $g,x \in G$. That is, if and only if G is abelian.

(*iii*) If G acts on the set G/H of cosets of a subgroup $H \leq G$, then the action is trivial if and only if $gH = H$ for all $g \in G$. That is, if and only if $H = G$.

So, if H is a proper subgroup of G, then $\ker(G,G/H,\cdot\,)$ is a proper normal subgroup of G.

△

Theorem 13.3.6. *If $\cdot$ is a faithful action of G on X, then G is isomorphic to a subgroup of* $\mathrm{Sym}(X)$.

Proof. As $\cdot$ is faithful, we have $G/\ker(G,X,\cdot\,) = G/\{1_G\} \cong G$, so by the first isomorphism theorem,

$$\begin{aligned} G &\cong G/\ker(G,X,\cdot\,) \\ &\cong \mathrm{im}(G,X,\cdot\,) \\ &\leq \mathrm{Sym}(X) \end{aligned}$$

■

Let $\cdot$ be an action of a group G on a set X, and let $x \in X$.

The *orbit* of x in G is the set of possible images of x under the action:

$$\mathrm{Orb}_G(x) \coloneqq \{g \cdot x : g \in G\} \subseteq X$$

The *stabiliser* of x in G is the set of elements of G that fix x:

$$\mathrm{Stab}_G(x) \coloneqq \{g \in G : g \cdot x = x\} \subseteq G$$

The *centraliser* or *commutant* of x in G is the set of elements that commute with x:

$$C_G(x) \coloneqq \{g \in G : gx = xg\}$$

(This notion is independent from group actions.)

Lemma 13.3.7. *The stabiliser and centraliser of any element $g \in G$ are subgroups of G.*

The *centre* of G is the set of elements of G that commute with every element of G:

$$Z(G) = \{g \in G : \forall h \in G : gh = hg\}$$

Note that

$$Z(G) = \bigcap_{g \in G} C_G(g)$$

so, as an intersection of subgroups, the centre is itself a subgroup (and is in fact normal in G).

Example. We compute the orbits and stabilisers of the three group actions from before.

- **Left-multiplication** ($X = G$, $g \cdot x \coloneqq gx$):

  For any $y \in X = G$, we have $y^{-1}x \in G$, so $y = (y^{-1}x) \cdot x$ and $y \in \mathrm{Orb}_G(x)$, so $\mathrm{Orb}_G(x) = X$ for all $x \in X$. Also, $g \cdot x = gx = x$ if and only if $g = 1_G$, so $\mathrm{Stab}_G(x) = \{1_G\}$ for all $x \in G$.

- **Conjugation** ($X = G$, $g \cdot x := gxg^{-1}$):

  The orbit $\mathrm{Orb}_G(x) = \{gxg^{-1} : g \in G\}$ of an element $x \in X$ under conjugation is also called the *conjugacy class* of x in G, also written as $\mathrm{Cl}(x)$ or ${}^G x$.

  For any $g \in G$, $g \cdot x = gxg^{-1} = x$ if and only if $gx = xg$, so $\mathrm{Stab}_G(x) = C_G(x)$ for all $x \in X = G$. Also,

$$\begin{aligned}\ker(G,X,\cdot\,) &= \{g \in G : \forall x \in X : g \cdot x = x\}\\ &= \{g \in G : \forall x \in X : gxg^{-1} = x\}\\ &= Z(G)\end{aligned}$$

- **Action on Cosets** ($X = G/H$, $g \cdot (xH) = (gx)H$):

  The stabiliser of $xH \in X$ is

$$\begin{aligned}\mathrm{Stab}_G(xH) &= \{g \in G : g \cdot xH = xH\}\\ &= \{g \in G : (gx)H = xH\}\\ &= \{g \in G : (x^{-1}gx)H = H\}\\ &= \{g \in G : x^{-1}gx \in H\}\\ &= \{xgx^{-1} \in G : xx^{-1}gxx^{-1} \in H\}\\ &= \{xgx^{-1} \in G : g \in H\}\\ &= xHx^{-1}\\ &= {}^x H\end{aligned}$$

  Also, if $xH,yH \in X$, then $(yx^{-1}) \cdot xH = yH$, so $\mathrm{Orb}_G(xH) = X$ for all $xH \in X$.

△

Theorem 13.3.8. *Let $\cdot$ be an action of a group G on a set X, and let $x \in X$. Then,*

(*i*) $\mathrm{Stab}_G(X) \le G$;

(*ii*) $\bigcap_{x \in X} \mathrm{Stab}_G(x) = \ker(G,X,\cdot\,)$.

Theorem 13.3.9 (Orbit-Stabiliser)**.** *Let G be a group acting on a finite set X and let $x \in X$. Then,*

$$|\mathrm{Orb}_G(x)| = [G : \mathrm{Stab}_G(x)] = \frac{|G|}{|\mathrm{Stab}_G(x)|}$$

Corollary 13.3.9.1. *Let G be a finite group acting on a set X. Then,*

(*i*) *For all $x,y \in X$, either* $\mathrm{Orb}_G(x) = \mathrm{Orb}_G(x)$, *or* $\mathrm{Orb}_G(x) \cap \mathrm{Orb}_G(y) = \varnothing$. *That is, orbits partition X.*

(*ii*) $|\mathrm{Orb}_G(x)|$ *divides* $|G|$.

Proof.

(*i*) Define a relation $\sim$ on X such that $x \sim y$ if and only if $y = g \cdot x$ for some $g \in G$. This relation is reflexive, by taking $g = 1_G$; symmetric, by taking inverses; and transitive, by multiplying the given g values with the group operation.

So, $\sim$ is an equivalence relation. The result then follows immediately from equivalence classes partitioning sets.

(*ii*) Follows immediately from the orbit-stabiliser theorem.

■

Theorem 13.3.10 (Cayley)**.** *Every finite group G is isomorphic to a subgroup of a symmetric group.*

Proof. The kernel of the left-multiplication action of G on itself is the set

$$\ker(G,G,\cdot\,) = \{g \in G : \forall x \in X : gx = x\}$$

For any $g \in G$ such that $gx = x$ for all $x \in G$, we have $g1_G = 1_G$, so $g = 1_G$, and hence the kernel is trivial, so the action is faithful. The result then follows from Theorem 13.3.6. ■

Theorem 13.3.11. *Let G be a finite group with $|G| = p^n$ for a prime p and $n \geq 1$. Then, $|Z(G)| > 1$.*

Proof. By Corollary 13.3.9.1, $|{}^G x| = |\mathrm{Orb}_G(x)|$ divides $|G|$, so $|{}^G x|$ is a power of p.

By definition, $Z(G) = \{x \in G : |{}^G x| = 1\}$. Suppose $|Z(G)| = 1$, so only one conjugacy class has cardinality 1, and the rest have cardinality p^{a_i}. Since orbits partition G, the cardinality of G is equal to the sum of the cardinalities of the orbits:

$$|G| = 1 + p^{a_1} + \cdots + p^{a_k}$$

However, this has residue 1 modulo p, contradicting that $|G| = p^n \equiv 0 \pmod{p}$. ■

Corollary 13.3.11.1. *Let G be a finite group with $|G| = p^n$ for a prime p and natural n. Then,*

(*i*) *If $n = 2$, then G is abelian.*

(*ii*) *If $n = 3$, then either G is abelian, or $|Z(G)| = p$.*

Theorem 13.3.12 (Cauchy)**.** *Let G be a finite group and let p be a prime divisor of $|G|$. Then, G has an element of order p. Moreover, the number of elements of G of order p is congruent to -1 modulo p.*

Theorem 13.3.13. *Let G be a finite group and let $H,K \leq G$. Then,*

$$|HK| = |KH| = \frac{|H||K|}{|H \cap K|}$$

Theorem 13.3.14. *Let G be a finite group and let $H,K \leq G$. Then,*

$$|G : H \cap K| \leq |G : H||G : K|$$

13.3.2 Fixed Points

Let G be a group acting on a set X, and let $g \in G$.

An element $x \in X$ is a *fixed point* if $g \cdot x = x$. The set of all fixed points for a given $g \in G$ is denoted by

$$\mathrm{fix}_X(g) \coloneqq \{x \in X : g \cdot x = x\}$$

An element $g \in G$ is *fixed point free* if $\mathrm{fix}_X(g) = \varnothing$.

Lemma 13.3.15 (Burnside)**.** *Let G be a finite group acting on a finite set X, and let $X/G \coloneqq \{\mathrm{Orb}_G(x) : x \in X\}$ be the set of orbits in G. Then,*

$$|X/G| = \frac{1}{|G|} \sum_{g \in G} |\mathrm{fix}_X(g)|$$

This lemma was stated and proved by Burnside in his 1897 book on finite groups, but attributed it to Frobenius, 1887. However, even before Frobenius, the result was known to Cauchy in 1845. Consequently, this lemma is sometimes called the *lemma that is not Burnside's*, or just *the not-Burnside lemma.*

Proof. First, the sum can be rewritten as

$$\begin{aligned}\sum_{g\in G}|\mathrm{fix}_X(g)| &= \left|\left\{(g,x)\in G\times X : g\cdot x = x\right\}\right| \\ &= \sum_{x\in X}|\mathrm{Stab}_G(x)|\end{aligned}$$

Then, by the orbit-stabiliser theorem,

$$|\mathrm{Stab}_G(x)| = \frac{|G|}{|\mathrm{Orb}_G(x)|}$$

so

$$\begin{aligned}\sum_{x\in X}|\mathrm{Stab}_G(x)| &= \sum_{x\in X}\frac{|G|}{|\mathrm{Orb}_G(x)|} \\ &= |G|\sum_{x\in X}\frac{1}{|\mathrm{Orb}_G(x)|}\end{aligned}$$

Let Y be the set of distinct orbits in X. Note that X is partitioned by its orbits, so,

$$\begin{aligned} &= |G|\sum_{A\in X/G}\sum_{x\in A}\frac{1}{|\mathrm{Orb}_G(x)|} \\ &= |G|\sum_{A\in X/G}\sum_{x\in A}\frac{1}{|A|} \\ &= |G|\sum_{A\in X/G}1 \\ &= |G||X/G|\end{aligned}$$

and the result follows. ∎

The action of G on X is *transitive* if for any two points $x,y \in X$, there exists $g \in G$ such that $g \cdot x = y$. Or equivalently, if G only has one orbit, or $\mathrm{Orb}_G(x) = X$ for all $x \in X$.

Corollary 13.3.15.1. *If a finite group G acts transitively on a finite set X with $|X| > 1$, then G contains a fixed point free element.*

Proof. Suppose G does not contain any fixed point free elements, so $|\mathrm{fix}_X(g)| \geq 1$ for all $g \in G$. Then, G acts transitively, so $|X/G| = 1$, and Burnside's lemma gives

$$\begin{aligned}|G| &= \sum_{g\in G}|\mathrm{fix}_X(g)| \\ &= |\mathrm{fix}_X(y)| + \sum_{g\in G\setminus\{1_G\}}|\mathrm{fix}_X(g)| \\ &= |X| + \sum_{g\in G\setminus\{1_G\}}|\mathrm{fix}_X(g)| \\ &\geq |X| + |G| - 1\end{aligned}$$

so $1 \geq |X|$, contradicting that $1 < |X|$. ∎

13.4 The Sylow Theorems

Lagrange's theorem states that if H is a subgroup of a finite group G, then $|H|$ divides $|G|$. Does the converse hold? That is, if G is a finite group, and r divides $|G|$, then does G contain a subgroup H of order r?

In general, this is not the case. For instance, if G is a non-abelian finite simple group, then G has no subgroup of order $|G|/2$. Such a subgroup H would have index 2 in G and would be a proper normal subgroup of G; also, G is non-abelian, so $|G| > 2$ and $1 < |H| < |G|$, contradicting that G is simple.

We write $|G|_p$ to denote the highest power of p that divides G. That is, if $|G| = p^n m$ with p,m coprime, then $|G|_p = p^n$.

- A subgroup $H \leq G$ is a *p-subgroup* of G if $|H|$ is a power of p.
- Let $P \leq G$ and suppose $|P| = |G|_p$. Then, P is called a *Sylow p-subgroup* of G.
- We write $\mathrm{Syl}_P(G)$ to denote the set of Sylow p-subgroups of G.

Example. Take $G = S_4$. We have $|G| = 4! = 2^3 \cdot 3$, so $|G|_2 = 2^3$ and $|G|_3 = 3$.

1. $P = \{1_G,(1,2,3),(3,2,1)\}$ has order $|P| = 3 = |G|_3$, so P is a Sylow 3-subgroup of G;
2. $K_4 = \{1_G,(1,2)(3,4),(1,3)(1,4),(1,4)(2,3)\}$ has order $|K_4| = 2 \neq |G|_2$, so K_4 is a 2-subgroup of G, but not a Sylow 2-subgroup;
3. $D_8 = \langle \sigma,\tau \rangle$ with $\sigma = (1,2,3,4)$ and $\tau = (1,4)(2,3)$ has order $|D_8| = 8 = |G|_2$, so D_8 is a Sylow 2-subgroup of G.
4. A_4 is not a p-subgroup of G for any prime p.
5. The trivial subgroup $\{1_G\}$ is a Sylow p-subgroup for all prime p.

△

Theorem 13.4.1 (Sylow)**.** *Let G be a finite group with order $|G| = p^n m$ with p,m coprime. Then,*

1. *G has at least one Sylow p-subgroup.*
2. *All Sylow p-subgroups of G are conjugate. That is, if H and K are Sylow p-subgroups of G, then there exists an element $g \in G$ such that $gHg^{-1} = K$.*
3. *Any p-subgroup of G is contained in a Sylow p-subgroup of G.*
4. *The number r of Sylow p-subgroups of G satisfies $r \equiv 1 \pmod p$ and $r \mid m$.*

13.4.1 Applications

By Sylow theorem 2, G acts on $\mathrm{Syl}_p(G)$ by conjugation, and for any $P \in \mathrm{Syl}_p(G)$, $\mathrm{Orb}_G(P) = \mathrm{Syl}_p(G)$. The stabiliser of P under conjugation is then the normaliser:

$$\begin{aligned}
\mathrm{Stab}_G(P) &= \{g \in G : g \cdot P = P\} \\
&= \{g \in G : gPg^{-1} = P\} \\
&= \{g \in G : gP = Pg\} \\
&= N_G(P)
\end{aligned}$$

Corollary 13.4.1.1. *Let G be a finite group, p be a prime divisor of $|G|$, and $P \in \mathrm{Syl}_p(G)$. Then,*

(i) $|\mathrm{Syl}_p(G)| = [G : N_G(P)]$*;*

(ii) $|\mathrm{Syl}_p(G)|$ *divides* $|G|/|G|_p$*;*

(iii) *$P \trianglelefteq G$ if and only if $|\operatorname{Syl}_p(G)| = 1$. That is, unique Sylow p-subgroups are normal.*

Proof.

(i) By the orbit-stabiliser theorem

$$\begin{aligned}|\operatorname{Syl}_p(G)| &= |\operatorname{Orb}_G(P)| \\ &= [G : \operatorname{Stab}_G(P)] \\ &= [G : N_G(P)]\end{aligned}$$

(ii) Since $P \leq N_G(P)$, by Lagrange's theorem, $|N_G(P)| = |P||N_G(P) : P|$. Then,

$$\begin{aligned}|\operatorname{Syl}_p(G)| &= [G : N_G(P)] \\ &= \frac{|G|}{|N_G(P)|} \\ &= \frac{|G|}{|P|[N_G(P) : P]}\end{aligned}$$

which divides $\frac{|G|}{|P|} = \frac{|G|}{|G|_p}$.

(iii) $P \trianglelefteq G$ if and only if $G = N_G(P)$. Then, by the orbit-stabiliser theorem,

$$\begin{aligned}|\operatorname{Orb}_G(P)| &= \frac{|G|}{|\operatorname{Stab}_G(P)|} \\ |\operatorname{Syl}_p(G)| &= \frac{|G|}{|N_G(P)|}\end{aligned}$$

so $G = N_G(P)$ if and only if $|\operatorname{Syl}_P(G)| = 1$.

■

Corollary 13.4.1.2. *Let G be a finite group and let p be a prime divisor of $|G|$. Define the set*

$$F_p(G) \coloneqq \{x \in G : x \neq 1_G \text{ and } |x| \text{ is a power of } p\}$$

Then,

(i)

$$F_p(G) = \bigcup_{P \in \operatorname{Syl}_p(G)} (P \setminus \{1_G\})$$

(ii) *$|F_p(G)| \geq |G|_p - 1$, with equality if and only if $|\operatorname{Syl}_p(G)| = 1$;*

(iii) *If $|G|_p = p$, then $|F_p(G)| = |\operatorname{Syl}_p(G)|(p-1)$, with equality if and only if $|\operatorname{Syl}_p(G)| = 1$.*

13.4.1.1 Proving Groups of a Particular Order are Not Simple

Example. Let G be a group of order $20 = 2^2 \times 5$. Can G be simple?

By Sylow's first theorem, G has Sylow 5-subgroups. By Sylow's fourth theorem, the number r of Sylow 5-subgroups divides 2^2 and satisfies $r \equiv 1 \pmod 5$. It follows that $r = 1$ is the only value that satifies this requirement, so G has a unique Sylow 5-subgroup, which must be normal in G and hence G cannot be simple. △

Example. Let G be a group of order $48 = 2^4 \times 3$. Can G be simple?

By Sylow theorem 1, G has Sylow 2-subgroups and Sylow 3-subgroups. By Sylow's fourth theorem, the number r of Sylow 2 subgroups divides 3 and satisfies $r \equiv 1 \pmod 2$. We must have $r = 1,3$, so G has either 1 or 3 Sylow 2-subgroups.

If there is only 1 Sylow 2-subgroup, then it is normal in G. Otherwise, G has 3 Sylow 2-subgroups and G acts non-trivially (and transitively) on $\mathrm{Syl}_2(G)$ by conjugation. This action induces a non-trivial homomorphism $\phi : G \to S_3$ (as in Theorem 13.3.5).

By the first isomorphism theorem $G/\ker(\phi) \cong \mathrm{im}(\phi)$, so by Lagrange's theorem,

$$\begin{aligned}|G/\ker(\phi)| &= |\mathrm{im}(\phi)|\\ |G|/|\ker(\phi)| &= |\mathrm{im}(\phi)|\\ |G|/|\mathrm{im}(\phi)| &= |\ker(\phi)|\end{aligned}$$

Because ϕ is non-trivial, $1 < |\mathrm{im}(\phi)| \leq |S_3| = 6$, so $\frac{48}{6} \leq |\ker(\phi)| < \frac{48}{1}$ and hence $\ker(\phi)$ is a non-trivial normal subgroup of G. △

Example. Let G be a group of order $2\,552 = 8 \times 11 \times 29$. Can G be simple?

Take $p = 11$, so $|G| = 11 \times (8 \times 29) = 11^1 \times 232$. The number of Sylow 11-subgroups, r, must divide 232 and satisfy $r \equiv 1 \pmod{11}$. Consider the factorisation $232 = 2^3 \times 29$; the factors of 232 are then: 1, 2, 4, 8, $29 \equiv 7$, $58 \equiv 3$, $116 \equiv 6$, and $232 \equiv 1$, so $r = 1{,}232$ are the possible solutions.

Now, if G has more than 1 Sylow 11-subgroup, then it must have 232 Sylow 11-subgroups. As 11 is prime, these subgroups must be cyclic, so every non-identity element generates the group. It follows that these subgroups intersect only at the identity element, so each subgroup contributes 10 elements of order 11, so there must be $232 \times 10 = 2\,320$ elements of order 11 in G.

Now, take $p = 29$, so $|G| = 29 \times (8 \times 11) = 29^1 \times 88$. By identical arguments as before, the number of Sylow 29-subgroups must be 1 or 88, and again, as 29 is prime, each subgroup must be cyclic, so if there is more than 1 Sylow 29-subgroup, then there are $88 \times 28 = 2\,464$ elements of order 28.

Now, by Sylow's first theorem, there exist Sylow 29 and 11-subgroups. If there are more than one of each, then we have 2 320 and 2 464 elements of order 11 and 29, respectively. But these values sum to more than $2\,552 = |G|$, so we cannot simultaneously have more than 1 Sylow 29 and 11-subgroups. But then, any unique Sylow p-subgroup is normal, so G cannot be simple. △

13.4.1.2 Proving a Particular Group is Simple

Corollary 13.4.1.3. *Let G be a finite group and let p be a prime divisor of $|G|$. Define the set*

$$F_p(G) := \{x \in G : x \neq 1_G \text{ and } |x| \text{ is a power of } p\}$$

Then,

(*i*) *Let N be normal in G. If $x \in N$, then ${}^{G}x \subseteq N$.*

(*ii*) *Let N be normal in G and suppose p does not divide $[G : N]$. Then,*

(*a*) $\mathrm{Syl}_p(N) = \mathrm{Syl}_p(G)$;

(*b*) $F_p(G) = F_p(N)$.

Theorem 13.4.2. *A_5 is simple.*

Proof. Suppose for a contradiction that A_5 has a non-trivial proper subgroup N. By Lagrange's theorem, $|N|$ divides $|A_5| = 5!/2 = 60$, so the prime factors of $|N|$ are 2, 3 and 5.

Now, note that

- A_5 has 24 elements of order 5 – these are the 5-cycles, and there are $P_5^5 = \frac{5!}{(5-5)!} = 120$ permutations of 5 elements from {1,2,3,4,5}. Dividing by 5 to account for cyclic shifts, there are $\frac{120}{5} = 24$ such elements.
- A_5 has 20 elements of order 3 – these are the 3-cycles, and there are $P_5^5 = \frac{5!}{(5-5)!} = 120$ permutations of 5 elements from {1,2,3,4,5}
- A_5 has 15 elements of order 2 – are those of the form $(ab)(cd)$ for a,b,c,d distinct elements of {1,2,3,4,5}. There are $P_4^5 = \frac{5!}{(5-4)!}$ permutations of 4 elements from 5, but 2 ways to cyclic shift within each cycle, and 2! ways to permute the cycles themselves, so there are $\frac{120}{2 \cdot 2 \cdot 2!} = 15$ elements of order 2.

Suppose p divides $|N|$ for $p = 3$ or $p = 5$. Then, p does not divide $[G : N]$, so by Corollary 13.4.1.3(i), $F_p(G) = F_p(N)$.

If $p = 3$, then $F_p(N) = F_p(G) = 20$, so $|N| \geq 21$. Since $|N|$ divides 60 and is less than 60, $|N| = 30$. Similarly, if $p = 5$, then $F_p(N) = F_p(G) = 24$, so $|N| \geq 25$. Again, we must have $|N| = 30$.

So, if 3 or 5 divide $|N|$, then $|N| = 30$ and both 3 and 5 divide $|N|$, so $F_3(N) = 20$ and $F_5(N) = 24$. But then, $|N| = 30 > 20 + 24$, which is a contradiction.

Now suppose neither 3 nor 5 divide $|N|$. Then, $|N|$ divides 4 by Lagrange's theorem, so by Cauchy's theorem, there exists $x \in N$ with order 2. By Corollary 13.4.1.3(ii), we then have $4 = |N| \geq |^G x| = 15$. ■

13.4.2 Simplicity of A_n

Lemma 13.4.3.

(i) *Let $n \geq 3$ and let X_n be the set of 3-cycles in S_n. Note that $X_n \subseteq A_n$ since 3-cycles decompose into a pair (i.e. an even number) of transpositions. Then, $A_n = \langle X_n \rangle$.*

(ii) *Let $n \geq 5$. Then, any two 3-cycles are conjugate in A_n.*

Lemma 13.4.4. *For $n \geq 5$, any non-identity permutation $\sigma \in A_n$ has a conjugate σ' such that $\sigma \neq \sigma'$ and $\sigma(i) = \sigma'(i)$ for some $i \in \{1,2,\dots,n\}$.*

Theorem 13.4.5. *A_n is simple for all $n \geq 5$.*

Proof. We induct on n. We already have that A_5 is simple, so assume $n \geq 6$.

A_n acts on the set $X_n = \{1,2,\dots,n\}$ in the natural way. For each $i \in X_n$, define

$$H_i := \mathrm{Stab}_{A_n}(i) \cong A_{n-1}$$

and by the inductive hypothesis, $H_i \cong A_{n-1}$ is simple. Note that H_i contains a 3-cycle containing 3 points of X_n other than i.

Suppose A has a non-trivial proper subgroup $N \triangleleft A_n$. Take any non-identity permutation $\sigma \in N$. By the previous lemma, there exists a conjugate $\sigma' \in N$ such that $\sigma \neq \sigma'$ and $\sigma(i) = \sigma'(i)$ for some $i \in X_n$.

Since normal subgroups are closed under conjugation, $\sigma' \in N$, so $\sigma^{-1}\sigma' \in N$, $\sigma^{-1}\sigma' \neq 1_{A_n}$, and $\sigma^{-1}\sigma'(i) = i$. Thus $\sigma^{-1}\sigma' \in H_i$ and so $N \cap H_i \neq \{1_{A_n}\}$.

Now, $N \triangleleft A_n$ so $N \cap H_i \triangleleft H_i$ by the second isomorphism theorem. But, $H_i \subseteq N$ contains a 3-cycle, so by Lemma 13.4.3(ii), N contains all 3-cycles of A_n. The result then follows from Lemma 13.4.3(i). ■

13.5 Classifying Groups of Small Order

13.5.1 Semidirect Products

Given two groups H and K, their cartesian product $H \times K$ has group structure by applying the group operations pointwise. This group is called the (*external*) *direct product* of H and K.

This extends naturally to any arbitrary collection of groups, with the product operation applied pointwise on each coordinate.

Theorem 13.5.1. *Let H and K be normal subgroups of a group G such that $G = HK$ and $H \cap K = \{1_G\}$. Then,*

(i) $hk = kh$ for all $h \in H$ and $k \in K$, so if H and K are both abelian, then G is abelian;

(ii) $G \cong H \times K$.

Recall that an automorphism of a group G is an isomorphism $G \to G$. The set $\mathrm{Aut}(G)$ of automorphisms of G has group structure under function composition and is called the *automorphism group* of G.

Let H and K be groups, and let $\phi : H \to \mathrm{Aut}(K)$ be a homomorphism. Write ϕ_h for $\phi(h)$ and define a binary operation $\cdot : (H \times K) \times (H \times K) \to H \times K$ by

$$(h_1,k_1) \cdot (h_2,k_2) := (h_1 h_2, \phi_{h_2^{-1}}(k_1)k_2)$$

Then, $(H \times K, \cdot\,)$ has group structure and is called the (*external*) *semidirect product* of H and K with respect to ϕ, denoted by $H \ltimes_\phi K$.

Example. Three important semidirect products are generated by homomorphisms as follows:

- **The trivial homomorphism**:

  Let H and K be any groups. Then, the map $\phi : H \to \mathrm{Aut}(K)$ defined by $\phi(h) = \mathrm{id}_K$ is the trivial homomorphism, and the resulting semidirect product operation is given by

  $$\begin{aligned}(h_1,k_1) \cdot (h_2,k_2) &= (h_1,h_2,\phi_{h_2^{-1}}(k_1)k_2) \\ &= (h_1,h_2,\mathrm{id}_K(k_1)k_2) \\ &= (h_1,h_2,k_1k_2)\end{aligned}$$

  so

  $$H \ltimes_\phi K \cong H \times K$$

- **The inversion homomorphism**:

  Let $H = C_2 = \langle c \rangle$ and let K be any abelian group. Then, the map $\phi : H \to \mathrm{Aut}(K)$ defined by $\phi(1_H) = \mathrm{id}_K$ and $\phi(h) = (k \mapsto k^{-1})$ (i.e. the identity element is sent to the identity automorphism, and every other element is sent to the inversion automorphism) is a homomorphism.

  If $K \cong C_n$, then the resulting semidirect product is isomorphic to the dihedral group of order $2n$:

  $$C_2 \ltimes_\phi C_n \cong D_{2n}$$

- **The conjugation homomorphism**:

  Let G be a group and let $H \leq G$ and $K \trianglelefteq G$. Then, the map $\phi : H \to \mathrm{Aut}(K)$ defined by $\phi(h) = (k \mapsto hkh^{-1})$ is a homomorphism.

  This last homomorphism will be useful with the following lemma:

△

Lemma 13.5.2. *Let G be a group and let $H \leq G$ and $K \trianglelefteq G$. If $G = HK$ and $K \cap H = \{1_G\}$, then*

$$G \cong H \ltimes_\phi K$$

Proof. ■

Example. Let $n \geq 3$ be an integer, and consider the dihedral group $G = D_{2n} = \langle \sigma,\tau \rangle$, where

$$\sigma \coloneqq (1,\ldots,n)$$
$$\tau :- \prod_{i=1}^{\lfloor \frac{n}{2} \rfloor} (i,n-i+1)$$

Let $K = \langle \sigma \rangle = \{1_G,\sigma,\sigma^2,\ldots,\sigma^{n-1}\}$ and $H = \langle \tau \rangle = \{1_G,\tau\}$. Recall that ${}^\tau\sigma = \tau\sigma\tau^{-1} = \sigma^{-1}$, so ${}^\tau k = k^{-1}$ for all $k \in K$.

Since $|\tau| = 2$, $|\sigma| = n$, and $D_{2n} = K \sqcup \tau K$, we have $G = HK$ and $H \cap K = \{1_G\}$, so by the previous lemma, we have $G \cong H \ltimes_\phi K$, where ϕ is the inversion homomorphism. △

Lemma 13.5.3. *Let G be a non-abelian finite group and suppose that*

1. *G has a cyclic subgroup K of order $n \coloneqq |G|/2$;*
2. *$G \setminus K$ contains an element G of order 2;*
3. *If $i \in \{0,1,\ldots,n-1\}$ satisfies $i^2 \equiv 1 \pmod{n}$, then $i \equiv \pm 1 \pmod{n}$.*

Then,

$$G \cong D_{2n}$$

Example. The following are some examples of positive integers n that satisfy the third hypothesis of the previous lemma.

- For $n = 6$, $0^2,1^2,2^2,3^2,4^2,5^2 \equiv 0,1,4,3,4,1 \pmod{6}$, so $i^2 \equiv 1 \pmod{6}$ if and only if $i = 1,5 \equiv \pm 1 \pmod{6}$.
- Let $n = p$ where p is prime. Then,

  $$\begin{aligned} i^2 &= 1 \\ i^2 - 1 &= 0 \\ (i-1)(i+1) &= 0 \end{aligned}$$

  Since $\mathbb{Z}/p\mathbb{Z}$ is a field, it has no zero divisors, so either $i - 1 = 0$ or $i + 1 = 0$, so $i^2 = 1$ if and only if $i = \pm 1$ in $\mathbb{Z}/p\mathbb{Z}$.
- Let $n = p^2$ where p is prime.

  If $p = 2$, we have $0^2,1^2,2^2,3^2 \equiv 0,1,0,1 \pmod{4}$, so $i^2 \equiv 1 \pmod{4}$ if and only if $i = 1,3 \equiv \pm 1 \pmod{4}$.

  Otherwise, suppose p is odd and let $i \in \{0,1,\ldots,p^2-1\}$ such that $i^2 \equiv 1 \mod p^2$. Then, p^2 divides $(i-1)(i+1)$.

  Since p is odd, it divides at most one of the factors, because if it divided both, it would also divide their difference $(i+1)-(i-1) = 2$, contradicting that p is odd. So, p^2 also divides at most one of the factors.

  So, p^2 divides $i - 1$ or $i + 1$. Then, since $0 \leq i \leq p^2 - 1$, the only possibilities are $i = 1,p^2 - 1 \equiv \pm 1 \pmod{p^2}$.

△

13.5.2 Semidirect Products of Abelian and Cyclic Groups

We consider the following special case of semidirect products: let G be a finite group with $|G|/2$ odd, and suppose G has an abelian normal subgroup K of order $|G|/2$.

The *commutator* of two elements $g,h \in G$ is the element $[g,h] := ghg^{-1}h^{-1}$. Similarly, we define the subgroup $[K,x] := \big\langle \{[k,x] : k \in K\} \big\rangle$.

Lemma 13.5.4 (Fitting)**.**

(i) ${}^{x}a = xax^{-1} = a^{-1}$ *for all* $a \in [K,x]$*;*

(ii) $K = C_K(x) \times [K,x]$*;*

(iii) $G \cong \big(H \ltimes_\phi [K,x]\big) \times C_K(x)$*, where* $\phi : H \to \operatorname{Aut}\big([K,x]\big)$ *is the inversion homomorphism.*

13.5.3 Abelian Groups

Theorem 13.5.5 (Fundamental Theorem of Finite Abelian Groups)**.** *Let G be a finite abelian group. Then, there exist divisors $d_1, \ldots ,d_r$ of $|G|$ such that $d_1 \mid d_2 \mid \cdots \mid d_r$ and*

$$G \cong \bigoplus_{i=1}^{r} \mathbb{Z}_{d_i}$$

13.5.4 Groups of order p, p^2, or $2p$, for prime p

Lemma 13.5.6. *If $|G| = p$ with p prime, then $G \cong C_p$.*

Proof. Take any non-identity element $g \in G$. By lagrange's theorem, $|g|$ divides $|G| = p$. Since $g \neq 1_G$, $|g| = p$ so $G = \langle g \rangle$. ■

Lemma 13.5.7. *If $|G| = p^2$ with p prime, then either $G \cong C_{p^2}$ or $G \cong C_p \times C_p$.*

Proof. We have already proved that all groups of order p^2 are abelian (Corollary 13.3.11.1), so G is abelian. The fundamental theorem of finite abelian groups then gives the result. ■

Lemma 13.5.8. *If $|G| = 2p$ with $p \neq 2$ prime, then either $G \cong C_{2p}$ or $G \cong D_{2p}$.*

Proof. If G is abelian, then $G \cong C_2 \times C_p \cong C_{2p}$ by the fundamental theorem of finite abelian groups.

Otherwise, G is non-abelian. Let $P \in \operatorname{Syl}_p(G)$. The number r of Sylow p-subgroups divides 2 and satisfies $r \equiv 1 \pmod p$, so since $p \neq 2$, we must have $r = 1$, so $P \trianglelefteq G$.

Since p is odd, it follows that all elements of G of order 2 lie in $G \setminus P$. Also, since $\mathbb{Z}/p\mathbb{Z}$ is a field, the only solutions of the equation $i^2 - 1 = 0$ are congruent to ± 1 modulo p. Then, Lemma 13.5.3 gives that $G \cong D_{2p}$, as required. ■

13.5.5 Groups of order $2p^2$, for odd prime p

Let $p \neq 2$ be prime, $H = C_2$, and $K = C_p \times C_p$. Let $\phi : H \to \operatorname{Aut}(K)$ be the inversion homomorphism. The group $H \ltimes_\phi K$ is then called the *generalised dihedral group of order* $2p^2$ and is denoted by GD_{2p^2}.

Lemma 13.5.9. *If $|G| = 2p^2$ with $p \neq 2$ prime, then G is isomorphic to one of the following:*

- C_{2p^2}*;*
- $C_p \times C_{2p}$*;*
- $C_p \times D_{2p}$*;*

- D_{2p^2};
- GD_{2p^2}.

13.5.6 Groups of order pq, for prime p,q with $p < q$ and $p \nmid q - 1$

Lemma 13.5.10. *Let $|G| = pq$ with p,q prime, satisfying $p < q$ and $p \nmid q - 1$. Then, $G \cong C_{pq}$.*

Proof. The number r of Sylow p-subgroups divides q and satisfies $r \equiv 1 \pmod{p}$. If $r = q$, then $q \equiv 1 \pmod{p}$, so $q - 1 \equiv 0 \pmod{p}$, contradicting that p does not divide $q - 1$. Thus, $r = 1$.

Similarly, the number s of Sylow q-subgroups divides p and satisfies $s \equiv 1 \pmod{q}$. Since $p < q$, p is already a least residue modulo q, so $s = p$ leads to a contradiction $p \equiv 1 \pmod{q}$, so $s = 1$.

So, G has a normal Sylow p-subgroup, say H, and a normal Sylow q-subgroup, say K. By Lagrange's theorem, $H \cap K = \{1_G\}$. By Theorem 13.3.13,

$$|HK| = \frac{|H||K|}{|H \cap K|} = \frac{pq}{1} = pq = |G|$$

so $G = HK$. Then, by Theorem 13.5.1, $G \cong H \times K$. Note that, being of prime order, H and K are both cyclic. Let $H = \langle h \rangle$ and $K = \langle k \rangle$. These generators commute, so $|hk| = |h||k| = pq = |G|$, so $G = \langle xy \rangle = C_{pq}$, as required. ■

We have now classified all groups of the following orders:

1,2,3,4,5,6,7,9,10,11,13,14,15,17,18

We will not classify groups of order 16, as there are too many, but we will now classify groups of order 8 and 12.

13.5.7 Groups of order 8

We have already seen a non-cyclic group of order 8, namely D_8. We now define another.

The *quaternion group* Q_8 is the group of unit basis quaternions under quaternion multiplication:

$$Q_8 := \{1,i,j,k,-1,-i,-j,-k\}$$

That is,

- $1q = q1 = q$ and $(-1)q = q(-1) = -q$ for all $q \in Q_8$;
- $ij = -ji = k$, $jk = -kj = i$, and $ki = -ik = j$;
- $1^2 = 1$, and $i^2 = j^2 = k^2 = ijk = -1$.

The quaternion group can also be defined as the group with presentation

$$Q_8 := \langle i,j,k \mid i^2 = j^2 = k^2 = ijk \rangle$$

where the identity is denoted 1, the element $i^2 = j^2 = k^2 = ijk$ is denoted -1, and the elements i^3, j^3, and k^3 are denoted $-i$, $-j$, and $-k$, respectively.

Lemma 13.5.11.

(i) $Z(Q_8) = \{\pm 1\}$.

(ii) *G has 1 element of order 2, namely -1, and 6 elements of order 4, namely $\pm i$, $\pm j$, and $\pm k$.*

(iii) $G = \langle i,j \rangle = \langle j,k \rangle = \langle k,i \rangle$.

(iv) $Q_8 \ncong D_8$ *since* D_8 *has* 5 *elements of order* 2 *and* 2 *elements of order* 4.

Lemma 13.5.12. *If* $|G| = 8$*, then* G *is isomorphic to one of the following:*

- $C_2 \times C_2 \times C_2$*;*
- $C_4 \times C_2$*;*
- C_8*;*
- D_8*;*
- Q_8.

13.5.8 Groups of order 12

We have already seen some non-cyclic groups of order 12, namely D_{12} and A_4. We now define another.

Let $H = C_4 = \langle h \rangle$ and $K = C_3$. Define $\phi : H \to \mathrm{Aut}(K)$ by $\phi(h^i) = (k \mapsto k^{(-1)^i})$. The resulting semidirect product $H \ltimes_\phi K$ is called the *dicyclic group* of order 12, denoted by Dic_{12}.

Lemma 13.5.13. *If* $|G| = 12$*, then* G *is isomorphic to one of the following:*

- $C_3 \times C_2 \times C_2 \cong C_6 \times C_2$*;*
- C_{12}*;*
- D_{12}*;*
- A_4*;*
- Dic_{12}.

13.5.9 Unique Simple Group of Order 60

Theorem 13.5.14. *If* $|G| = 60$*, then* $G \cong A_5$.

13.6 Soluble Groups

13.6.1 Composition Series

We write $H < G$ or $H \lneq G$ to mean that H is a proper subgroup of G, and similarly, $H \lhd G$ or $H \ntrianglelefteq G$ to mean that H is a proper normal subgroup of G.

A *composition series* of a group G is a sequence of nested normal subgroups $(G_i)_{i=1}^r$ satisfying

$$\{1_G\} = G_0 \ntrianglelefteq G_1 \ntrianglelefteq G_2 \ntrianglelefteq \cdots \ntrianglelefteq G_r = G$$

such that G_i/G_{i-1} is simple for each $1 \leq i \leq r$, and r is called the *length* of the series.

Example.

1. Let $p \neq 2$ be prime and let $G = D_{2p} = \langle \sigma, \tau \rangle$. Let $G_0 = \{1_G\}$, $G_1 = \langle \sigma \rangle \cong C_p$, and $G_2 = G$. These groups satisfy the normality requirements, and the quotients are given by $G_1/G_0 \cong G_1 \cong C_p$, $G_2/G_1 \cong \langle \tau \rangle \cong C_2$, which are both simple. Thus,

$$\{1_G\} \ntrianglelefteq \langle \sigma \rangle \ntrianglelefteq D_{2p}$$

is a composition series of length 2.

2. Let $n \geq 5$, and let $G = S_n$. Let $G_0 = \{1_G\}$, $G_1 = A_n$, and $G_2 = S_n$. These groups satisfy the normality requirements, and the quotients are given by $G_1/G_0 \cong G_1 \cong A_n$, $G_2/G_1 \cong C_2$, which are both simple. Thus,
$$\{1_G\} \ntrianglelefteq A_n \ntrianglelefteq S_n$$
is a composition series of length 2.

3. Let $G = D_8 = \langle \sigma, \tau \rangle$. Let $G_0 = \{1_G\}$, $G_1 = \langle \sigma^2 \rangle$, $G_2 = \langle \sigma \rangle$, and $G_3 = D_8$. These groups satisfy the normality requirements, and the quotients are all isomorphic to C_2, which is simple, so
$$\{1_G\} \ntrianglelefteq \langle \sigma^2 \rangle \ntrianglelefteq \langle \sigma \rangle \ntrianglelefteq D_8$$
is a composition series of length 3.

△

Note that if G is the trivial group, then the series
$$\{1_G\} = G_0 = G$$
is a composition series of G of length 0.

Theorem 13.6.1. *Every finite group has a composition series.*

Corollary 13.6.1.1. *Let G be a finite group and let $N \trianglelefteq G$. Suppose that*
$$\begin{aligned} \{1_G\} &= N_0 \ntrianglelefteq N_1 \ntrianglelefteq \cdots \ntrianglelefteq N_r = N \\ \{1_G\} &= \frac{X_0}{N} \ntrianglelefteq \frac{X_1}{N} \ntrianglelefteq \cdots \ntrianglelefteq \frac{X_s}{N} = \frac{G}{N} \end{aligned}$$
are composition series for N and G/N, respectively, where each X_i in the second series is a subgroup of G containing N. In particular, $X_0 = N$ and $X_s = G$.

Then,
$$\{1_G\} = N_0 \ntrianglelefteq N_1 \ntrianglelefteq \cdots \ntrianglelefteq N_r = N = X_0 \ntrianglelefteq X_1 \ntrianglelefteq \cdots \ntrianglelefteq X_s = G$$
is a composition series for G of length $r + s$.

13.6.2 Jordan-Hölder Theorem

Two composition series I and II of a group G
$$\begin{aligned} \{1_G\} &= A_0 \ntrianglelefteq A_1 \ntrianglelefteq \cdots \ntrianglelefteq A_r = G \qquad \text{(I)} \\ \{1_G\} &= B_0 \ntrianglelefteq B_1 \ntrianglelefteq \cdots \ntrianglelefteq B_s = G \qquad \text{(II)} \end{aligned}$$
are *equivalent* and write I $\sim$ II if $r = s$ and there is a bijection
$$f : \{A_i/A_{i-1} : 1 \leq i \leq r\} \to \{B_i/B_{i-1} : 1 \leq i \leq s\}$$
such that $A_i/A_{i-1} \cong f(A_i/A_{i-1})$ for each $1 \leq i \leq r$.

Theorem 13.6.2 (Jordan-Hölder)**.** *Let*
$$\begin{aligned} \{1_G\} &= A_0 \ntrianglelefteq A_1 \ntrianglelefteq \cdots \ntrianglelefteq A_r = G \qquad \text{(I)} \\ \{1_G\} &= B_0 \ntrianglelefteq B_1 \ntrianglelefteq \cdots \ntrianglelefteq B_s = G \qquad \text{(II)} \end{aligned}$$
be two composition series of a finite group G. Then, I $\sim$ II.

This theorem implies that, up to isomorphism, the quotients G_i/G_{i-1} and the length r of any composition series of a finite group G are invariants of that group.

Let

$$\{1_G\} = G_0 \ntrianglelefteq G_1 \ntrianglelefteq G_2 \ntrianglelefteq \cdots \ntrianglelefteq G_r = G$$

be a composition series for a finite group G, with uniqueness up to equivalence given by the Jordan-Hölder theorem. Then, the quotient groups G_i/G_{i-1} for $1 \leq i \leq r$ are called the *composition factors* of G, and r is called the *composition length* of G.

A finite group is *soluble* if it is trivial or if its composition factors are all cyclic groups of prime order (or equivalently, simple abelian groups).

Example.

(*i*) Let G be a finite abelian group. Then, any quotient of any subgroup of G is abelian, so any composition factor of G is a simple abelian group, i.e. a cyclic group of prime order. Thus, all abelian groups are soluble.

(*ii*) Let $n \geq 5$ and consider A_n. Then, A_n is a non-abelian simple group, so it has precisely one composition factor, namely itself, which is non-abelian. Thus, A_n is not soluble for any $n \geq 5$.

△

Lemma 13.6.3. *Let G be a finite group and let N be normal in G. Then, G is soluble if and only if both N and G/N are soluble.*

Proof. Write $\mathrm{CF}(G)$ for the (multi)set of composition factors of G. By Corollary 13.6.1.1 and the Jordan-Hölder theorem,

$$\mathrm{CF}(G) = \mathrm{CF}(N) \cup \mathrm{CF}(G/N)$$

Thus, G is soluble if and only if both N and G/N are soluble. ■

Example. Let $G = D_{2n} = \langle \sigma,\tau \rangle$ and let $N = \langle \sigma \rangle \trianglelefteq G$. N is abelian and $|G/N| = 2$, so G/N is abelian, so both are soluble, and hence G is soluble. △

13.6.3 Commutators

Recall that the commutator of two elements $g,h \in G$ is the element $[g,h] := ghg^{-1}h^{-1}$. Note that $[g,h] = 1_G$ if and only if g and h commute.

Example. Consider the alternating group A_5.

$$\begin{aligned}\big[(1,2,4),(1,3,5)\big] &= (1,2,4)(1,3,5)(1,2,4)^{-1}(1,3,5)^{-1} \\ &= (1,2,4)(1,3,5)(4,2,1)(5,3,1) \\ &= (1,2,3)\end{aligned}$$

More generally, if $\{x,a,b,c,d\} = \{1,2,3,4,5\}$,

$$\big[(x,a,b)(x,c,d)\big] = (x,a,b)(x,c,d)(b,a,x)(d,c,x) = (x,a,c)$$

△

The *commutator subgroup* $[G,G]$ is the subgroup of G generated by all of its commutators:

$$[G,G] := \big\langle [g_1,g_2] \bigm| g_1,g_2 \in G \big\rangle$$

More generally, if $H,K \leq G$, we define

$$[H,K] := \big\langle [h,k] \bigm| h \in H, k \in K \big\rangle$$

to be the *commutator subgroup of H and K*.

Example.

1. In any abelian group G, $[g,h] = 1_G$ for all $g,k \in G$, so the commutator subgroup $[G,G] = \langle 1_G \rangle = \{1_G\}$ is trivial.
2. Let $G = A_5$. As seen in the example above, every 3-cycle in A_5 is the commutator of some pair of 3-cycles. But A_5 is generated by 3-cycles, so $[A_5,A_5] = A_5$.

△

The *abelianisation* G^{ab} of a group G is the quotient $G/[G,G]$.

Theorem 13.6.4. *For any group G,*

(i) $[G,G] \trianglelefteq G$*;*

(ii) G^{ab} *is abelian.*

(iii) If N is normal in G and G/N is abelian, then $[G,G] \leq N$

Proof. *(i)* For all $g,h,j \in G$,

$$\begin{aligned} g[h,k]g^{-1} &= ghkh^{-1}k^{-1}g^{-1} \\ &= gh(g^{-1}g)k(g^{-1}g)h^{-1}(g^{-1}g)k^{-1}g^{-1} \\ &= (ghg^{-1})(gkg^{-1})(gh^{-1}g^{-1})(gk^{-1}g^{-1}) \\ &= (ghg^{-1})(gkg^{-1})(ghg^{-1})^{-1}(gkg^{-1})^{-1} \\ &= [ghg^{-1},gkg^{-1}] \\ &\in [G,G] \end{aligned}$$

For a general element $[h_1,k_1][h_2,k_2]\cdots[h_r,k_r] \in [G,G]$, we have,

$$\begin{aligned} g[h_1,k_1][h_2,k_2]\cdots[h_r,k_r]g^{-1} &= g[h_1,k_1](g^{-1}g)[h_2,k_2](g^{-1}g)\cdots(g^{-1}g)[h_r,k_r]g^{-1} \\ &= \big(g[h_1,k_1]g^{-1}\big)\big(g[h_2,k_2]g^{-1}\big)\cdots\big(g[h_r,k_r]g^{-1}\big) \\ &\in [G,G] \end{aligned}$$

so $[G,G] \trianglelefteq G$.

(ii) We prove a more general statement: a quotient group G/N is abelian if and only if every commutator is in N. That is, if and only if $[G,G] \subseteq N$.

Let $g,h \in G$. Then,

$$\begin{aligned} (gN)(hN) &= (hN)(gN) \\ (gN)(hN) &= (hN)(gN)N \\ (gN)^{-1}(hN)^{-1}(gN)(hN) &= N \\ [gN,hN] &= N \\ [g,h]N &= N \\ [g,h] &\in N \end{aligned}$$

where we used that N is the identity in G/N on the second line. So, gH and hN commutes if and only if $[g,h] \in N$, so G/N is abelian if and only if $[g,h] \in N$ for all $g,h \in G$. In particular, if $N = [G,G]$, then every commutator is in N be definition of the commutator subgroup, so $G^{\mathrm{ab}} = G/[G,G]$ is abelian.

(iii) Proved in part *(ii)*.

■

Corollary 13.6.4.1. *A group G is abelian if and only if $[G,G] = \{1_G\}$.*

Given a group G, define $G^(0) \coloneqq G$ and recursively define the *nth derived subgroup* as

$$G^{(n)} \coloneqq \left[G^{(n-1)},G^{(n-1)}\right]$$

for each $n \in \mathbb{N}$. Then, the descending series

$$G^{(0)} \geq G^{(1)} \geq G^{(2)} \geq \cdots \geq G^{(n)} \geq G^{(n+1)} \geq \cdots$$

is called the *derived series* of G.

By definition, we have

- $\left(G^{(n)}\right)^{(m)} = G^{(n+m)}$;
- $H^{(n)} \leq G^{(n)}$ for all $H \leq G$.

Theorem 13.6.5. *Let G be a finite group. Then, G is soluble if and only if $G^{(n)} = \{1_G\}$ for some $n \in \mathbb{N}$.*

Proof. Suppose G is soluble. We induct on $|G|$.

If $|G| = 1$, then G is trivial, as is $G^{(0)}$. Suppose otherwise that $|G| > 1$ and define $N \coloneqq [G,G] \trianglelefteq G$. Then, N is soluble by Lemma 13.6.3, as it is a normal subgroup of a soluble group.

By definition of solubility, G has a composition series

$$\{1_G\} = G_0 \ntrianglelefteq G_1 \ntrianglelefteq \cdots \ntrianglelefteq G_r = G$$

where all the composition factors G_i/G_{i-1} are cyclic with prime order. In particular, G/G_{r-1} is cyclic and hence abelian, so $[G,G] = N \leq G_{r-1}$, giving $|N| < |G|$. So, $N^{(m)} = \{1_G\}$ for some $m \in \mathbb{N}$ by the inductive hypothesis. Since $G^{(n)} = [G,G]^{(n-1)}$ by definition, it follows that $G^{m+1} = \{1_G\}$ as required.

Now, for the reverse implication, suppose that $G^{(n)} = \{1_G\}$ for some $n \in \mathbb{N}$. We induct on $|G|$.

If $|G| = 1$, then G is trivial and hence soluble. Suppose otherwise that $|G| > 1$ and again define $N \coloneqq [G,G] \trianglelefteq G$. If $N = G$, then $G^{(n)} = [G,G]^{(n-1)} = G^{(n-1)} = \cdots = G^{(1)} = G^{(0)} = G$, which contradicts the inductive hypothesis. So, $N \ntrianglelefteq G$.

Since $N^{(n-1)} = [G,G]^{(n-1)} = G^{(n)} = \{1_G\}$, N is soluble by the inductive hypothesis. Also, $G/N = G/[G,G] = G^{\text{ab}}$ is abelian and hence soluble. So, G is also soluble by Lemma 13.6.3. ■

A previous result gave that normal subgroups of a soluble group are soluble, but this theorem implies that *any* subgroup of a soluble group is soluble.

Corollary 13.6.5.1. *If G is a finite soluble group, and $H \leq G$, then H is soluble.*

Proof. Since G is soluble, $G^{(n)} = \{1_G\}$ for some $n \in \mathbb{N}$. Since $H^{(n)} \leq G^{(n)}$, we must have $H^{(n)} = \{1_G\}$, so H is soluble. ■

13.6.4 Examples of Soluble Groups

Theorem 13.6.6. *Let G be a group of order p^n for some prime p and $n \in \mathbb{N}$. Then, G is soluble, and furthermore, all composition factors of G are isomorphic to C_p.*

Proof. We proceed by strong induction on $|G|$.

If $|G| = p^1 = p$, then $G \cong C_p$ is cyclic of prime order, so G is soluble with composition length 1, and its composition factor is C_p.

Assume that $|G| = p^n > p$ and that the result holds for all groups of order less than $|G|$. Then, by Theorem 13.3.11, the centre $Z \coloneqq Z(G)$ is non-trivial. The centre Z is abelian and hence soluble. Also, G/Z is soluble by the inductive hypothesis, so G is soluble by Lemma 13.6.3. ■

Theorem 13.6.7. *Let G_1 and G_2 be finite soluble groups. Then, $G \coloneqq G_1 \times G_2$ is soluble.*

Proof. Consider the projection homomorphism $\pi_1 : G \to G_1$. Define $N \coloneqq \ker(\pi) = \{1_{G_1}\} \times G_2 \cong G_2$, so N is soluble.

Also, $\operatorname{im}(\pi) = G_1$ is soluble, so by the first isomorphism theorem,

$$G/\ker(\pi) \cong \operatorname{im}(\pi)$$
$$G/N \cong G_1$$

and hence G/N is soluble, so G is soluble by Lemma 13.6.3. ■

Corollary 13.6.7.1. *Let $G_1, \ldots, G_t$ be finite soluble groups. Then, $G \coloneqq G_1 \times \cdots \times G_t$ is soluble.*

Proof. Induction on the previous result. ■

Chapter 14

Galois Theory

Chapter 15

Representation Theory

In this chapter, we study groups by their representations as matrices or linear transformations, and their associated modules.

Chapter 16

Symmetric Functions and Integrable Probability

Chapter 17

Geometric Group Theory

Chapter 18

Reflection Groups

> *"We share a philosophy about linear algebra: we think basis-free, we write basis-free, but when the chips are down we close the office door and compute with matrices like fury."*
>
> — Paul Halmos, *Celebrating 50 Years of Mathematics*

18.1 Reflection Groups

Let V be a finite-dimensional vector space over $\mathbb{R}$. A form $\langle \cdot , \cdot \rangle : V \times V \to \mathbb{R}$ is

- *symmetric* if $\langle x,y \rangle = \langle y,x \rangle$ for all $x,y \in V$;
- *bilinear* if $\langle \alpha x + \beta y,z \rangle = \alpha \langle x,z \rangle + \beta \langle y,z \rangle$ and $\langle x,\alpha y + \beta z \rangle = \alpha \langle x,y \rangle + \beta \langle x,z \rangle$ for all $x,y \in V$ and $\alpha,\beta \in \mathbb{R}$;
- *positive definite* if $\langle x,x \rangle \geq 0$ for all $x \in V$, with equality if and only if $x = 0_V$.

All the forms we will consider will be symmetric and bilinear.

A space equipped with a form satisfying the three properties above is called a *Euclidean space*.

Example. $V = \mathbb{R}^2$ with $\big\langle (x_1,x_2),(y_1,y_2) \big\rangle = x_1y_1 - x_1y_2 - x_2y_1 + 2x_2y_2$ is a Euclidean space. △

Example. $V = \mathbb{R}^n$ with $\langle x,y \rangle = x \cdot y = \sum_i x_iy_i$ is the standard Euclidean space, denoted by $\mathbb{E}^n$. △

Via the Gram-Schmidt process, every Euclidean vector space V has an orthonormal basis – that is, a basis $(e_i)_{i=1}^n$ of unit vectors ($\|e_i\| = 1$) that are pairwise orthogonal ($e_i \cdot e_j = 0$ for $i \neq j$). Thus, we can find an isomorphism $V \to \mathbb{E}^n$ preserving the bilinear form.

Example. In the non-standard Euclidean space above,

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

is such a basis. △

The *general linear group* $GL(V)$ of a vector space V is the group of linear automorphisms of V with the operation of composition.

$$\begin{aligned} GL(V) &:= \operatorname{Aut}(V) \\ &= \big\{ (T : V \to V) : T \text{ is linear and bijective} \big\} \end{aligned}$$

If V has dimension $n < \infty$, then this is isomorphic to the group of $n \times n$ invertible matrices with the operation of matrix multiplication.

The *orthogonal group* $O(V)$ of a vector space V is the subgroup of the general linear group consisting of transformations that preserve the bilinear form

$$O(V) \coloneqq \big\{T \in GL(V) : \big\langle T(x) \cdot T(y)\big\rangle = \langle x,y\rangle\big\}$$

Because vector norms are defined in terms of bilinear forms, e.g. $\|x\| = \sqrt{\langle x,x\rangle}$, orthogonal transformations $T \in O(V)$ also preserve vector norms: $\|x\| = \|T(x)\|$ for any $x \in V$.

Let $x \in V$ be non-zero. We define the map $S_x : V \to V$ by

$$S_x(z) \coloneqq z - 2\frac{\langle x,z\rangle}{\langle x,x\rangle}x$$

This is always an element of $O(V)$. Note also that $S_x(x) = -x$.

An element $T \in O(V)$ is a *reflection* if the set of fixed points $V^T \coloneqq \{v \in V : T(v) = v\}$ is an $(n-1)$-dimensional subspace of V.

Example. The map S_x is a reflection, and the fixed subspace is the orthogonal complement $\{x\}^\perp$. △

The next lemma shows that all reflections are of this form.

Lemma 18.1.1. *Let V be a Euclidean space. Let T be a reflection, and let $x \in (V^T)^\perp$ be non-zero. Then, $T(x) = -x$ and $T = S_x$.*

Proof. Let $v \in V^T$, so $v = T(v)$. Then, $\big\langle v,T(x)\big\rangle = \big\langle T(v),T(x)\big\rangle = \langle v,x\rangle = 0$, so v and $T(x)$ are orthogonal, i.e., $T(x) \in (V^T)^\perp$.

Since $\dim\big((V^T)^\perp\big) = \dim(V) - \dim(V^T) = 1$, $T(x) = \alpha x$ for some $\alpha \in \mathbb{R}$. Then, since x is non-zero,

$$\begin{aligned}\langle x,x\rangle &= \langle T(x),T(x)\rangle \\ &= \langle \alpha x,\alpha x\rangle \\ &= \alpha^2\langle x,x\rangle\end{aligned}$$

since $x \neq 0$, $\langle x,x\rangle \neq 0$, so $\alpha^2 = 1$. If $\alpha = 1$, then $T(x) = x$ and $x \in V^T$, contradicting that $x \in (V^T)^\perp$. Hence, $\alpha = -1$, so $T(x) = -x$.

Now, suppose $z \in V$. Then, by linearity in the first argument,

$$\begin{aligned}\left\langle z - \frac{\langle x,z\rangle}{\langle x,x\rangle}x,\, x\right\rangle &= \langle z,x\rangle - \frac{\langle x,z\rangle}{\langle x,x\rangle}\langle x,x\rangle \\ &= \langle z,x\rangle - \langle z,x\rangle \\ &= 0\end{aligned}$$

so $z - \frac{\langle x,z\rangle}{\langle x,x\rangle}x \in \{x\}^\perp = V^T$, and $T\big(z - \frac{\langle x,z\rangle}{\langle x,x\rangle}x\big) = z - \frac{\langle x,z\rangle}{\langle x,x\rangle}x$. So,

$$\begin{aligned}T(z) &= T\left(z - \frac{\langle z,x\rangle}{\langle x,x\rangle}x + \frac{\langle z,x\rangle}{\langle x,x\rangle}x\right) \\ &= T\left(z - \frac{\langle z,x\rangle}{\langle x,x\rangle}x\right) + \frac{\langle z,x\rangle}{\langle x,x\rangle}T(x) \\ &= z - \frac{\langle z,x\rangle}{\langle x,x\rangle}x - \frac{\langle z,x\rangle}{\langle x,x\rangle}x \\ &= z - 2\frac{\langle z,x\rangle}{\langle x,x\rangle}x\end{aligned}$$

$$= S_x(z)$$

■

From this lemma, we deduce

- Every reflection T in a Euclidean space is of the form S_x for some x determined uniquely up to scaling. Such an x is called the *root* of the reflection T.
- For all non-zero $x \in V$, the map S_x is a reflection.
- Every reflection T is involutive, i.e., satisfies $T^2 = \mathrm{id}_V$.

A *finite reflection group* is a pair (G,V) consisting of a Euclidean space V and a finite subgroup $G < O(V)$ generated by reflections, e.g. $G = \big\langle \{S_x : S_x \in G\} \big\rangle$.

Example. The trivial group generated by no reflections forms the trivial reflection group $(0,V)$. △

Example. $\big(\{\mathrm{id}_{\mathbb{R}}\,,f\},\mathbb{R}\big)$ with f defined by $x \mapsto -x$ is a reflection group, with $f = S_1$. △

Two reflection groups (G_1,V_1) and (G_2,V_2) are *equivalent* if there exists an isometry $\varphi : V_1 \to V_2$ such that $\varphi G_1 \varphi^{-1} := \{\varphi T \varphi^{-1} : T \in G_1\} = G_2$, written as $(G_1,V_1) \simeq (G_2,V_2)$.

Example. The reflection group $\big(\{\mathrm{id}_{\mathbb{R}^2}\,,S_{(0,1)}\},\mathbb{R}^2\big)$ generated by the reflection along the x-axis and the reflection group $\big(\{\mathrm{id}_{\mathbb{R}^2}\,,S_{(0,1)}\},\mathbb{R}^2\big)$ generated by the reflection along the y-axis are equivalent, with the isometry given by the rotation

$$\varphi = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

△

Example. Let $x,y \in \mathbb{R}^2$ be non-zero. Consider the group $\langle S_x,S_y \rangle$ generated by the reflections S_x and S_y.

As matrices, we have

$$\begin{aligned} \det(S_x S_y) &= \det(S_x)\det(S_y) \\ &= (-1)^2 \\ &= 1 \end{aligned}$$

so $S_x S_y \in SO(2)$, i.e., is a rotation by some angle α (in fact, α is twice the angle between x and y). If $\frac{\alpha}{2\pi} \notin \mathbb{Q}$, then $S_x S_y$ has infinite order and $\langle S_x,S_y \rangle$ is not finite. Otherwise, if $\frac{\alpha}{2\pi} = \frac{n}{k}$ for coprime integers $n,k \in \mathbb{Z}$, then $S_x S_y$ is a rotation by $\frac{2\pi n}{k}$ and hence has order k. From this, we deduce that $\langle S_x,S_y \rangle$ is isomorphic to the dihedral group $\mathrm{Dih}(k)$ of symmetries on the k-gon of order $2n$ (i.e. as generated by the reflection $\sigma = S_x$ and the rotation $\tau = S_x S_y$). △

We define the group $I_2(k)$ as

$$\begin{aligned} I_2(k) &:= \Big(\big\langle S_{(1,0)}, S_{\left(\cos(\frac{\pi}{k}),\, \sin(\frac{\pi}{k})\right)} \big\rangle, \mathbb{R}^2\Big) \\ &= \big(\mathrm{Dih}(k),\mathbb{R}^2\big) \end{aligned}$$

with the subscript matching the dimension. Note that $|I_2(k)| = 2k$.

Example. The symmetric group $\mathrm{Sym}(n)$ acts on $\{1,\ldots,n\}$. We can extend this action to $\mathbb{R}^n$ as follows.

Let $(e_i)_{i=1}^n$ be a basis of $\mathbb{R}^n$. For each permutation $\sigma \in \mathrm{Sym}(n)$, define $T_\sigma : \mathbb{R}^n \to \mathbb{R}^n$ by $T_\sigma(e_i) = e_{\sigma(i)}$. Then, $\sigma \mapsto T_\sigma$ defines a homomorphism $\mathrm{Sym}(n) \to O(\mathbb{R}^n)$ and hence a subgroup of $O(\mathbb{R}^n)$.

The symmetric group $\mathrm{Sym}(n)$ is generated by the transpositions (i,j), $i \neq j$, and $T_{(i,j)} = S_{e_i - e_j}$ is a reflection, so $T_{(i,j)}(e_i - e_j) = e_j - e_i = -(e_i - e_j)$. Now, any vector $y \in (e_i - e_j)^\perp$ will have equal i and j coordinates, and hence $T_{(i,j)}(y) = (y)$. Thus, this defines a finite reflection group $\big(\mathrm{Sym}(n), \mathbb{R}^n\big)$.

However, note that the vector $x = (1,1\ldots,1)$ is fixed by any T_σ, so the orthogonal complement $x^\perp := \big\{(a_1,\ldots,a_n) : \sum_i a_i = 0\big\}$ is also invariant (i.e. permuting the summands doesn't change its value) and is isomorphic to $\mathbb{R}^{n-1}$.

So, we can reduce $\big(\mathrm{Sym}(n),\mathbb{R}^n\big)$ to $\big(\mathrm{Sym}(n),x^\perp\big) \simeq \big(\mathrm{Sym}(n),\mathbb{R}^{n-1}\big)$ △

We define the group A_{n-1} (unrelated to the alternating group) by:

$$\begin{aligned} A_{n-1} &:= \big(\mathrm{Sym}(n),x^\perp\big) \\ &\simeq \big(\mathrm{Sym}(n),\mathbb{R}^{n-1}\big) \end{aligned}$$

again, with the subscript matching the dimension. Note that $|A_n| = (n+1)!$ and $A_2 = I_2(3)$.

We define the group B_n by modifying A_n as:

$$B_n := \big(\langle \mathrm{Sym}(n), \underset{i=1,\ldots,n}{S_{e_i}} \rangle,\mathbb{R}^n\big)$$

and also the group D_n as:

$$B_n := \big(\langle \mathrm{Sym}(n), \underset{i\neq j}{S_{e_i+e_j}} \rangle,\mathbb{R}^n\big)$$

D_n has index 2 in B_n.

18.2 Root Systems

Lemma 18.2.1. *For any $T \in O(V)$,*

$$TS_xT^{-1} = S_{T(x)}$$

Proof. First, note

$$\begin{aligned} (TS_xT^{-1})\big(T(x)\big) &= (TS_xT^{-1}T)(x) \\ &= (TS_x)(x) \\ &= T(-x) \\ &= -T(x) \end{aligned}$$

More generally,

$$\begin{aligned} (TS_xT^{-1})(z) &= T\big(S_x(T^{-1}z)\big) \\ &= T\left(T^{-1}z - 2\frac{\langle T^{-1}z,x\rangle}{\langle x,x\rangle}x\right) \\ &= T(T^{-1}z) - T\left(2\frac{\langle T^{-1}z,x\rangle}{\langle x,x\rangle}x\right) \\ &= z - 2\frac{\langle T^{-1}z,x\rangle}{\langle x,x\rangle}T(x) \\ &= z - 2\frac{\langle z,Tx\rangle}{\langle Tx,Tx\rangle}T(x) \\ &= S_{T(x)} \end{aligned}$$

where the last line follows from $T \in O(V)$ being orthogonal, i.e. $\langle u,v\rangle = \big\langle T(u),T(v)\big\rangle$. ■

The *root system* of a finite reflection group (G,V) is the set

$$\Phi_{(G,V)} := \big\{x \in V : S_x \in G, \|x\| = 1\big\}$$

Theorem 18.2.2. *The root system of a finite reflection group (G,V) satisfies the following properties:*

(i) *If $x \in \Phi_{(G,V)}$, then $\mathbb{R}x \cap \Phi_{(G,V)} = \{x,-x\}$;*

(ii) *The cardinality of $\Phi_{(G,V)}$ is twice the number of reflections in G.*

(iii) *If $T \in G$ and $x \in \Phi_{(G,V)}$, then $T(x) \in \Phi_{(G,V)}$*

Proof.

(i) If $x \in \Phi_{(G,V)}$, then $\|x\| = 1$, so $\mathbb{R}x \cap \Phi_{(G,V)} = \big\{\alpha x : \alpha \in \mathbb{R}, |a|\|x\| = 1\big\} = \{x,-x\}$.

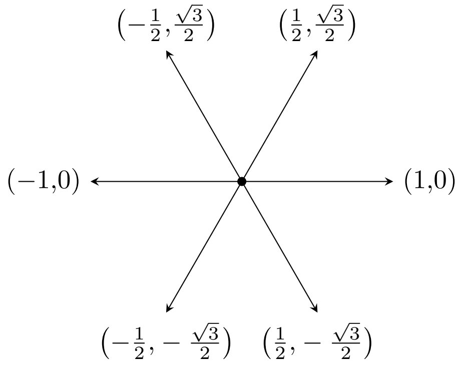
(ii) For each reflection $S_x \in G$, there are two elements $x, -x \in \Phi_{(G,V)}$.

(iii) If $x \in \Phi_{(G,V)}$, then $S_x \in G$. Then, for any $T \in G$, $TS_xT^{-1} = S_{T(x)} \in G$. Also, since $T \in G < O(V)$ is orthogonal, $\|T(x)\| = \|x\| = 1$. Hence, $T(x) \in \Phi_{(G,V)}$.

■

Example. $I_2(3)$ describes the symmetry of an equilateral triangle, so the roots are given by the unit vectors on the lines of symmetry:

$$\Phi_{I_2(3)} = \Big\{(1,0),(-1,0),(\tfrac{1}{2},\tfrac{\sqrt{3}}{2}),(-\tfrac{1}{2},-\tfrac{\sqrt{3}}{2}),(\tfrac{1}{2},-\tfrac{\sqrt{3}}{2}),(-\tfrac{1}{2},\tfrac{\sqrt{3}}{2})\Big\}$$


$(-\frac{1}{2},\frac{\sqrt{3}}{2})$
$(\frac{1}{2},\frac{\sqrt{3}}{2})$
$(-1,0)$
$(1,0)$
$(-\frac{1}{2},-\frac{\sqrt{3}}{2})$
$(\frac{1}{2},-\frac{\sqrt{3}}{2})$

△

18.2.1 Abstract Root Systems

More generally, any set of vectors $\Phi \subseteq V$ is a *root system* if:

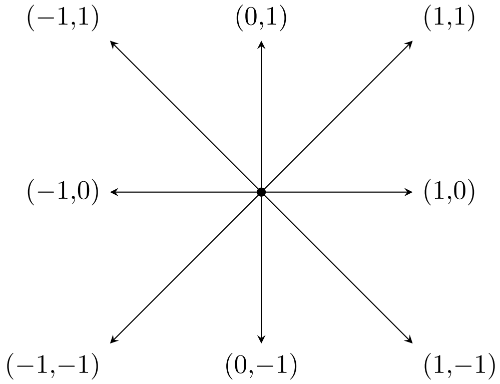
(i) $0_V \notin \Phi$;

(ii) If $x \in \Phi$, then $\mathbb{R}x \cap \Phi = \{x,-x\}$;

(iii) If $x,y \in \Phi$, then $S_y(x) \in \Phi$.

Note that we do not require the vectors in an abstract root system to have unit norm, unlike the root system associated to a finite reflection group.

Example. $\Phi_1 = \{-1,0,1\}^2 \setminus \big\{(0,0)\big\}$ is a root system:



△

Example. The following is a root system in $\mathbb{R}^4$:

$$\Phi_2 = \left\{e_i - e_j : 1 \leq i,j \leq 4, i \neq j\right\}$$

△

Example. The following is a root system in $\mathbb{R}^4$:

$$\Phi_3 = \left\{e_i - e_j, \pm e_i : 1 \leq i,j \leq 4, i \neq j\right\}$$

△

18.2.2 Simple Systems

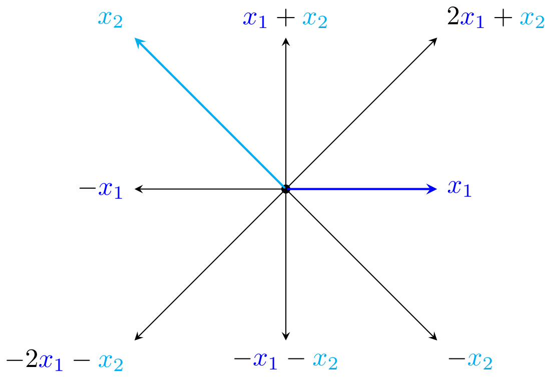
A set $\Pi \subseteq \Phi$ is a *simple system* if

- (i) Π is linearly independent;
- (ii) For every $x \in \Phi$, $x = \sum_i \alpha_i y_i$ for some $y_i \in \Pi$ and $\alpha_i \in \mathbb{R}$ satisfying $\alpha_i \geq 0$ for all i, or $\alpha_i \leq 0$ for all i.

A simple system is similar to a basis in that it is linearly independent and it spans Φ, but with the stronger requirement that these linear combinations have all positive or all negative coefficients.

Example. A simple system for the root system Φ_1 defined in a previous example is given by

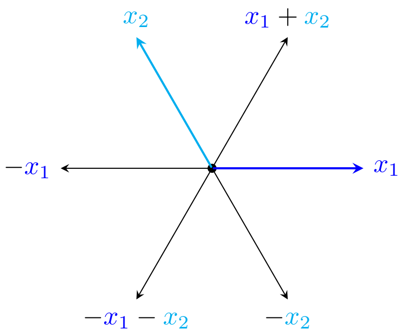
$$\Pi = \left\{x_1 = (1,0), x_2 = (-1,1)\right\}$$



△

Example. A simple system for $\Phi_{I_2(3)}$ is given by

$$\Pi = \{x_1 = (1,0), x_2 = (-\tfrac{1}{2}, \tfrac{\sqrt{3}}{2})\}$$



with △

Example. A simple system for the root system Φ_2 defined in a previous example is given by

$$\Pi_2 = \{e_1 - e_2, e_2 - e_3, e_3 - e_4\}$$

△

Example. A simple system for the root system Φ_3 defined in a previous example is given by

$$\Pi_2 = \{e_1 - e_2, e_2 - e_3, e_3 - e_4, e_4\}$$

△

In the examples above, we note that the angle between the roots in a simple system is obtuse (at least, for the examples we can visualise in $\mathbb{R}^2$). This turns out that this is a necessary condition for a set of roots to be a simple system:

Lemma 18.2.3. *Let Π be a simple system and suppose that $x,y \in \Pi$ are distinct. Then, $\langle x,y \rangle \leq 0$.*

Proof. Suppose otherwise that $\langle x,y \rangle > 0$. Then, $S_x(y) \in \Phi$ is given by $S_x(y) = y - 2\frac{\langle x,y \rangle}{\langle x,x \rangle}x = y - \alpha x$, where $\alpha = 2\frac{\langle x,y \rangle}{\langle x,x \rangle} > 0$.

Since Π is linearly independent, $1y - \alpha x$ is the unique representation of $S_x(y)$ as a linear combination of elements of Π. But then, we have coefficients $1 > 0$ and $-\alpha < 0$, contradicting that Π is simple. ■

This lemma makes it slightly easier to construct simple systems: once the first vector has been chosen, we only have to consider vectors pointing in the opposite direction as candidates to be added to the simple system. For instance, in the previous example, once we have picked x_1, we only need to check the two vectors on the left (obviously excluding $-x_1$).

18.2.3 Ordered Vector Spaces

An *ordered vector space* is a vector space V equipped with a non-strict total ordering $\leq$ compatible with the vector space structure. That is:

- $\forall x,y,z \in V$, if $x \leq y$, then $x + z \leq y + z$ (compatibility with vector addition);
- $\forall x,y \in V, \forall \alpha \in \mathbb{R}$, if $x \leq y$ and $\alpha > 0$, then $\alpha x \leq \alpha y$;
- $\forall x,y \in V, \forall \alpha \in \mathbb{R}$, if $x \leq y$ and $\alpha < 0$ then $\alpha y \leq \alpha x$ (compatibility with scalar multiplication).

We write $x < y$ if $x \leq y$ and $x \neq y$.

The *lexicographical order* on $\mathbb{R}^n$ is given by $(x_1,\ldots,x_n) < (y_1,\ldots,y_n)$ if there exists $k \in \{1,\ldots,n\}$ such that $x_i = y_i$ for $i < k$ and $x_k < y_k$.

That is, compare the first components of the vectors using the usual ordering on $\mathbb{R}$; in case of a tie, compare the second components, and so on. In this way, the lexicographical ordering is a generalisation of dictionary ordering to non-alphabetical symbols.

Example. The vectors in $V = \{-1,0,1\}^2$ are lexicographically ordered as:

$$(-1,-1) < (-1,0) < (-1,1) < (0,-1) < (0,0) < (0,1) < (1,-1) < (1,0) < (1,1)$$

If we label -1 as a, 0 as b, and 1 as c and concatenate the components of each vector together, we have:

$$aa < ab < ac < ba < bb < bc < ca < cb < cc$$

matching the ordinary dictionary ordering of strings. △

Theorem 18.2.4. *Every possible total ordering on $\mathbb{R}^n$ is a lexicographical ordering for some basis.*

If $\leq$ is a total ordering on V, then for each $x \in V \setminus \{\mathbf{0}\}$, either $x < \mathbf{0} < -x$ or $-x < \mathbf{0} < x$, so every ordered vector space V can be partitioned into three sets: namely, the elements strictly less than 0, $V_- := \{x \in V : x < \mathbf{0}\}$, the elements strictly greater than 0, $V_+ := \{x \in V : x > \mathbf{0}\}$, and the singleton containing the zero vector, $\{\mathbf{0}\}$.

A *positive system* in a root system Φ is a subset $\Phi_+ \subset \Phi$ satisfying $\Phi_+ = \Phi \cap V_+$, where V_+ is induced by some total ordering on V. Similarly, a *negative system* is a subset $\Phi_- \subset \Phi$ such that $\Phi_- = \Phi \cap V_-$ for some total ordering on V.

Example. In the previous example, we saw the lexicographical ordering on $V = \{-1,0,1\}^2$. A positive system for the root system $\Phi_1 = V \setminus \{\mathbf{0}\}$ is then given by the elements greater than $\mathbf{0} = (0,0)$:

$$\Phi_+ = \big\{(0,1),(1,-1),(1,0),(1,1)\big\}$$

△

Example. Another positive system is given by

$$\Phi_+ = \big\{(1,-1),(1,0),(1,1),(0,-1)\big\}$$

with the ordering inducing the positive system given by the lexicographic ordering with respect to the basis $\big\{(1,-1),(0,-1)\big\}$. △

18.2.4 Quasisimple Systems

A subset $\Omega \subseteq \Phi_+$ is a *quasisimple system* if

- (i) For each $x \in \Phi_+$, there exists a collection of scalar coefficients $\alpha_i \geq 0$ such that $x = \sum_i \alpha_i y_i$ for $y_i \in \Omega$;
- (ii) Ω is minimal with respect to property (i).

Compared to simple systems, it is relatively easy to construct a quasisimple system:

Example. Consider the positive system

$$\Phi_+ = \big\{z_1 = (0,1), z_2 = (1,-1), z_3 = (1,0), z_4 = (1,1)\big\}$$

from a previous example.

Clearly, the whole set $\{z_1,z_2,z_3,z_4\}$ satisfies property (i).

But, $z_4 = z_1 + z_3$, so z_4 may be replaced in any linear combination with $z_1 + z_3$, and all the coefficients are still positive, so $\{z_1,z_2,z_3\}$ still satisfies (i).

Now, we note that $z_3 = z_1 + z_2$, so again, we may remove z_3 to obtain $\{z_1,z_2\}$.

At this point, we cannot remove any more vectors, so this set is minimal, and $\Omega = \{z_1,z_2\}$ is a quasisimple system. △

Lemma 18.2.5. *Let Ω be a quasisimple system and suppose that $x,y \in \Omega$ are distinct. Then, $\langle x,y\rangle \leq 0$.*

Proof. Suppose $\langle x,y\rangle > 0$. We have $S_x(y) \in \Phi$ and $S_x(y) = y - 2\frac{\langle x,y\rangle}{\langle x,x\rangle}x = y - \alpha x$, with $\alpha > 0$.

Suppose $S_x(y) \in \Phi_+$, so $S_x(y) = y - \alpha x = \sum_{z\in\Omega} \alpha_z z$ for some scalars $\alpha_z \geq 0$. Then,

$$\begin{aligned} y - \alpha x &= \alpha_y y + \sum_{z\in\Omega\setminus\{y\}} \alpha_z z \\ (1-\alpha_y)y &= \alpha x + \sum_{z\in\Omega\setminus\{y\}} \alpha_z z \end{aligned}$$

if $\alpha_y < 1$, then dividing through by $1 - \alpha_y$ gives:

$$\begin{aligned} y &= \frac{1}{1-\alpha_y}\left(\alpha x + \sum_{z\in\Omega\setminus\{y\}} \alpha_z z\right) \\ y &= \frac{\alpha x}{1-\alpha_y} + \sum_{z\in\Omega\setminus\{y\}} \frac{\alpha_z}{1-\alpha_y} z \end{aligned}$$

so we can express y as a linear combination with positive coefficients, so $\Omega \setminus \{y\}$ is a quasisimple system, contradicting the minimality of Ω. So, $\alpha_y \geq 1$. Then,

$$\begin{aligned} y - \alpha_y y &= \alpha x + \sum_{z\in\Omega\setminus\{y\}} \alpha_z z \\ 0 &= (\alpha_y - 1)y + \alpha x + \sum_{z\in\Omega\setminus\{y\}} \alpha_z z \end{aligned}$$

All the coefficients on the right are non-negative, and $\Omega \subseteq \Phi_+$, so the right side is in V_+. Since $\alpha > 0$, αx is non-zero, and hence the right side is non-zero, which is a contradiction.

Otherwise, $S_x(y) \in \Phi_-$, so $S_x(y) = y - \alpha x = \sum_{z\in\Omega} -\alpha_z z$ for some scalars $\alpha_z \geq 0$. Then,

$$\begin{aligned} y - \alpha x &= \sum_{z\in\Omega} -\alpha_z z \\ \alpha x - y &= \sum_{z\in\Omega} \alpha_z z \\ x - \frac{1}{\alpha} y &= \sum_{z\in\Omega} \frac{\alpha_z}{\alpha} z \end{aligned}$$

so the previous argument applies, with the roles of x and y reversed. ■

Theorem 18.2.6. *Every quasisimple system is a simple system.*

Proof. Every root in Φ_+ can be written as a non-negative linear combination L of roots in Ω. But then, every root in Φ_- can be written as the non-negative linear combination $-L$. So Ω satisfies property (ii) of a simple system.

We are left to show that Ω is linearly independent. Suppose there is a linear combination

$$\sum_{z\in\Omega}\alpha_z z = 0$$

with non-negative coefficients α_z not all zero. Define the sets

$$A \coloneqq \{z\in\Omega : \alpha_z > 0\}, \qquad B \coloneqq \{z\in\Omega : \alpha_z < 0\}$$

and define the non-negative scalars $\beta_z \coloneqq -\alpha_z$. Sorting positive and negative coefficients, we have

$$\begin{aligned}\sum_{\substack{z\in\Omega\\ \alpha_z>0}}\alpha_z z + \sum_{\substack{z\in\Omega\\ \alpha_z<0}}\alpha_z z &= 0\\ \sum_{\substack{z\in\Omega\\ \alpha_z>0}}\alpha_z z &= \sum_{\substack{z\in\Omega\\ \alpha_z<0}}-\alpha_z z\\ \sum_{z\in A}\alpha_z z &= \sum_{z\in B}\beta_z z\end{aligned}$$

Define the vector y to be equal to these sums, which is non-negative. Then,

$$\begin{aligned}0 &\le \|y\|^2\\ &= \langle y,y\rangle\\ &= \left\langle \sum_{z\in A}\alpha_z z, \sum_{z\in B}\beta_z z\right\rangle\\ &= \sum_{s\in A}\sum_{t\in B}\alpha_s\beta_t\langle s,t\rangle\end{aligned}$$

The scalars are non-negative, and by the previous lemma, the forms are all non-positive, so the whole sum is non-positive:

$$\le 0$$

so the coefficients must be all zero, contradicting their construction. ■

Theorem 18.2.7. *There is a bijection between positive systems and simple systems. Specifically, every positive system contains a unique simple system, and every simple system is contained within a unique positive system.*

Proof. Let Φ_+ be a positive system. Consider the set of all subsets of Φ_+ satisfying property (i) of a quasisimple system. Note that Φ_+ itself satisfies this property, so this set is non-negative. Now, choose one which is minimal, giving a quasisimple system, which is a simple system.

Suppose this process yields two distinct simple systems $\Pi,\Pi' \subseteq \Phi_+$. Without loss of generality, we can find $x\in\Pi'\setminus\Pi$. Because $x\in\Phi_+$ and Π is a simple system, there exists a linear decomposition

$$x = \sum_{y\in\Pi}\alpha_y y$$

where the coefficients are all non-negative or all non-positive. Since $x\in\Phi_+$ is positive, the coefficients must in fact be all non-negative. Also, at least one of the coefficients, say α_{y_0}, is positive, as x is non-zero.

Because $y\in\Phi_+$ and Π' is also a simple system, we can decompose each y as a linear combination

$$y = \sum_{z\in\Pi'}\beta^y_z z$$

again with all coefficients non-negative, since $y \in \Phi_+$. Again, at least one of the coefficients is positive in each decomposition, since each y is non-zero. Then, we may rewrite the decomposition of x as

$$\begin{aligned} x &= \sum_{y \in \Pi} \alpha_y y \\ &= \sum_{y \in \Pi} \alpha_y \left(\sum_{z \in \Pi'} \beta_z^y z \right) \\ &= \sum_{y \in \Pi} \left(\sum_{z \in \Pi'} \alpha_y \beta_z^y \right) z \\ &= \sum_{z \in \Pi'} \left(\sum_{y \in \Pi} \alpha_y \beta_z^y \right) z \end{aligned}$$

Since $x \in \Pi'$, and Π' is linearly independent, we must have

$$\sum_{y \in \Pi} \alpha_y \beta_z^y = \begin{cases} 0 & z \neq x \\ 1 & z = x \end{cases}$$

As every term in this sum is positive, $\sum_{y \in \Pi} \alpha_y \beta_z^y \geq \alpha_{y_0} \beta_z^{y_0}$, so $\beta_z^{y_0} = 0$ unless $z = x$. Then, we have $y_0 = \beta_x^{y_0} x$. Since $\Phi \cap \mathbb{R}x = \{x, -x\}$, we have that $y_0 = x$, giving a contradiction.

Now, given a simple system Π, we can extend it to a basis B of V. Then, take the lexicographic order on V with respect to B. By construction, $\Pi \subset V_+$, so $\Pi \subset \Phi_+$.

Uniqueness follows since every element of Φ is a sum of elements of Π with either non-negative or non-positive coefficients. The sums with non-negative coefficients are in V_+ and these give exactly Φ_+. ■

Theorem 18.2.8. *Let $\Phi \supseteq \Phi_+ \supseteq \Pi$ be a root system, a positive system, and a simple system, respectively. Then, for all $x \in \Pi$ and all $y \in \Phi_+$:*

- *if $x \neq y$, then $S_x(y) \in \Phi_+$;*
- *if $x = y$, then $S_x(y) = -x \in \Phi_-$.*

Proof. (*ii*) follows from roots being negated under their associated reflections.

Otherwise, assume $x \neq y$, and let $y = \sum_{z \in \Pi} \alpha_z z$. Since $y \in \Phi_+$, $\alpha_z \geq 0$ for all z, and also since y is non-zero, at least one coefficient α_{z_0} is non-zero. Thus,

$$\begin{aligned} S_x(y) &= y - \alpha x \\ &= -\alpha x + \sum_{z \in \Pi} \alpha_z z \\ &= (\alpha_x - \alpha)x + \sum_{z \in \Pi \setminus \{x\}} \alpha_z z \end{aligned}$$

If $\alpha_x - \alpha > 0$, then this is a non-negative decomposition of $S_x(y)$, so $S_x(y) \in \Phi_+$.

Otherwise, $\alpha_x - \alpha < 0$, so this is a decomposition of $S_x(y)$ into a linear combination with both positive coefficients $\{\alpha_z\}_{z \neq x}$, and a negative coefficient $\alpha_x - \alpha$.

But $S_x(y) \in \Phi$, and Π is simple, so there also exists a non-negative or non-positive decomposition. So $S_x(y)$ has two distinct decompositions into linear combinations of vectors in Π, contradicting the linear independence of Π. ■

Intuitively, one might think that applying a reflection would perhaps swap all the vectors in the positive and negative half-spaces V_+ and V_-, or something similar. But, this theorem tells us that only one of these roots ever changes from being positive to negative: a root system captures a lot of information about the set of reflections its roots generate.

Given a root system Φ, we define its associated group G as

$$G = \langle S_x \mid x \in \Phi \rangle \leq O(V)$$

where V is the vector space containing Φ. This group acts on the root system Φ, since every element of G is a composition of reflections, and root systems are closed under reflection.

We claim that this group is finite.

Lemma 18.2.9. *If Φ_+ is a positive system, then $S_x(\Phi_+)$ is a positive system for any $x \in \Pi$.*

Lemma 18.2.10. *Let Π and Π' be simple systems in Φ. Then, there exists $g \in G$ such that $g(\Pi) = \Pi'$.*

Proof. Let Φ_+ and Φ'_+ be the associated unique positive systems for Π and Π, respectively, and let Φ_- and Φ'_- be the corresponding negative systems. We induct on $k := |\Phi_+ \cap \Phi_-|$.

If $k = 0$, then $\Phi_+ = \Phi'_+$, so $\Pi = \Pi' = \mathrm{id}_G(\Pi)$, since each positive system contains a unique simple system.

Assume the result holds for some arbitrary fixed $k \geq 0$. Then, $k + 1 \geq 1$, so $\Phi_+ \neq \Phi'_+$. Now, pick some $x \in \Pi \cap \Phi'_-$. Such an x exists, or else $\Pi \subseteq \Phi \setminus \Phi'_- = \Phi'_+$, so $\Pi = \Pi'$, contradicting the inductive hypothesis.

Now, consider $S_x(\Phi_+)$. By the previous theorem, every root apart from x is invariant under this reflection, and x alone is negated, so

$$S_x(\Phi_+) = \big(\Phi_+ \setminus \{x\}\big) \cup \{-x\}$$

Then,

$$S_x(\Phi_+) \cap \Phi'_- = (\Phi_+ \cap \Phi'_-) \setminus \{x\}$$

has cardinality k. ■

Theorem 18.2.11.

(*i*) *$G = \langle S_x \mid x \in \Pi \rangle$;*

(*ii*) *For all $y \in \Phi$, there exists $x \in \Pi$ and $g \in G$ such that $y = g(x)$.*

The first part of the theorem states that we can reduce the generating set from the entire root system Φ to just a simple system $\Pi \subseteq \Phi$. The second point says that every vector y in a root system is contained within a simple system $g(\Pi)$ for some $g \in G$. So in a way, a simple system contains almost as much information as the entire root system.

Proof. Let $x \in \Phi$, so $x = \sum_{r \in \Pi} \alpha_r r$. The *height* of x with respect to Π is defined as

$$h(x) := \sum_{r \in \Pi} \alpha_r$$

Define $G_0 = \langle S_x \mid x \in \Pi \rangle \leq G$. Then, for some fixed arbitrary $y \in \Phi_+$, define Λ_y to be the intersection of the orbit of y under G_0 with Φ_+:

$$\begin{aligned}\Lambda_y &:= G_0 \cdot y \cap \Phi_+ \\ &= \{g_0(y) : g_0 \in G_0\} \cap \Phi_+\end{aligned}$$

Pick $z \in \Lambda_y$ with minimal height. Since $z \in \Phi_+$, we have

$$z = \sum_{x \in \Pi} \alpha_x x$$

with $\alpha_x \geq 0$ for all x. Then,

$$\begin{aligned} 0 &\leq \|z\|^2 \\ &= \langle z,z\rangle \\ &= \left\langle z, \sum_{x\in\Pi} \alpha_x x \right\rangle \\ &= \sum_{x\in\Pi} \alpha_x \langle z,x\rangle \end{aligned}$$

so $\langle z,x\rangle \geq 0$ for some $x \in \Pi$. Then,

$$\begin{aligned} S_x(z) &= z - 2\frac{\langle x,z\rangle}{\langle x,x\rangle} x \\ S_x(z) &= z - \alpha x \\ h\big(S_x(z)\big) &= h\big(z - \alpha x\big) \\ h\big(S_x(z)\big) &= h\big(z\big) - \alpha \end{aligned}$$

since z has minimal height in Λ_y, $S_x(z) \notin \Lambda_y$. Also, z is in the orbit $G_0 \cdot y$, so also $S_x(z) \in G_0 \cdot y$, and hence $S_x(z) \in \Phi_-$. But, the only root that can change sign under the reflection S_x is x, so $x = z$, and x is also in the orbit $G_0 \cdot y$. That is, there exists $g \in G_0$ such that $x = g \cdot y$, or $y = g^{-1} \cdot x$, proving (ii).

Now, given $y \in \Phi$, let $x \in \Pi$ and $g \in G_0$ be such that $y = g \cdot x$. Then, $gS_xg^{-1} = S_{g(x)} = S_y$, so any reflection with a root in Φ can be expressed as the composition of a reflection S_x in $\Pi \subseteq G_0$ and two reflections $g,g^{-1} \in G_0$. So $G = G_0$. ■

The *length* of an element $g \in G$ is defined as

$$\ell(g) := \min\{n : \exists x_1,x_2,\ldots,x_n \in \Pi : g = S_{x_1}S_{x_2}\cdots S_{x_n}\}$$

That is, the length of an element g is the minimum number of reflections required to compose into g.

- $\ell(g) = 0$ if and only if $g = 1_G$;
- $\ell(S_x) = 1$ for all $x \in \Pi$.

While defined algebraically, this notion of length has geometric meaning, relating to root systems:

Theorem 18.2.12. *For all $g \in G$,*

$$\begin{aligned} \ell(g) &= \big|(g \cdot \Phi_+) \cap \Phi_-\big| \\ &= \big|\{x \in \Phi_+ : g \cdot x \in \Phi_-\}\big| \end{aligned}$$

That is, the length of g is equal to the number of positive roots that become negative when g is applied to them.

Proof. Define $N(g) := (g \cdot \Phi_+) \cap \Phi_-$, and $n(g) := |N(g)|$. The goal is to show that $n(g) = \ell(g)$.

We have:

- $n(g) = n(g^{-1})$, as $x \mapsto -g \cdot x$ is a bijection $N(g)$ to $N(g^{-1})$.

- For each $x \in \Pi$,
$$n(S_x g) = \begin{cases} n(g)+1 & g^{-1}\cdot x \in \Phi_+ \\ n(g)-1 & g^{-1}\cdot x \in \Phi_- \end{cases}$$
and
$$n(S_x g^{-1}) = n(gS_x) = \begin{cases} n(g)+1 & g\cdot x \in \Phi_+ \\ n(g)-1 & g\cdot x \in \Phi_- \end{cases}$$

We show that $n(g) \le \ell(g)$. Let $g = S_{x_1}S_{x_2}\cdots S_{x_k}$ with $k = \ell(g)$. Then,

$$\begin{aligned} n(g) &\le n(S_{x_1}S_{x_2}\cdots S_{x_k}) \\ &\le n(S_{x_1}S_{x_2}\cdots S_{x_{k-1}}) + 1 \\ &\le n(S_{x_1}S_{x_2}\cdots S_{x_{k-2}}) + 2 \\ &\vdots \\ &\le n(S_{x_1}) + (k-1) \\ &\le k \\ &= \ell(g) \end{aligned}$$

Now, suppose $n(g) < k$, so there exists an index $i \le k$ such that $n(S_{x_1}S_{x_2}\cdots S_{x_i}) = n(S_{x_1}S_{x_2}\cdots S_{x_{i-1}}) - 1$, so $S_{x_1}S_{x_2}\cdots S_{x_{i-1}}(x_i) \in \Phi_-$. Pick j to be the maximum index such that $S_{x_j}\cdots S_{x_{i-1}}(x_i) \in \Phi_-$, but $S_{x_{j+1}}\cdots S_{x_{i-1}}(x_i) \in \Phi_+$. Thus,
$$S_{x_{j+1}}\cdots S_{x_{i-1}}(x_i) = x_j$$
Let $h = S_{x_{j+1}}\cdots S_{x_{i-1}}$. Then, $hS_{x_i}h^{-1} = S_{x_j}$, so $S_{x_j}hS_{x_i} = hS_{x_i}h^{-1}hS_{x_i} = h$, so we can remove S_{x_j} and S_{x_i} from the decomposition of g, contradicting that $\ell(g) = k$. ■

Theorem 18.2.13. *Let Π and Π' be simple systems in Φ. Then, there exists a* ***unique*** *$g \in G$ such that $g\Pi = \Pi'$.*

Proof. Existence was proved in a previous theorem. For uniqueness, suppose $g\cdot\Pi = h\cdot\Pi = \Pi'$. Then, $h^{-1}g\cdot\Pi = \Pi$. Now, consider the positive root system Φ_+ associated to Π. Then, $\Phi_+ = h^{-1}g\cdot\Phi_+$, and hence $\ell(h^{-1}g) = 0$, so $h^{-1}g = 1_G$, and $g = h$.

■

Corollary 18.2.13.1. *There is a bijection between G and the set of simple systems in G.*

Proof. Orbit-stabiliser theorem. ■

Corollary 18.2.13.2. *Given a finite root system $\Phi \subseteq V$, $\big(\langle S_x \mid x \in \Phi\rangle, V\big)$ is a finite reflection group with order at most $2^{|\Phi|}$.*

So, we can convert a reflection group (G,V) into a root system $\Phi_{(G,V)}$, and this corollary tells us that we can recover (G,V) from $\Phi_{(G,V)}$.

Conversely, if we start with an abstract root system Φ, we can convert this into a reflection group (G,V), and from there, obtain the root system $\Phi_{(G,V)}$, which will be equal to the set of roots in Φ, each normalised to unit length.

18.3 Presentations of Groups

18.3.1 Free Groups

Given a set X, a *word* $w = x_1x_2\cdots x_n$ on X is a finite sequence of *letters* $(x_i)_{i=1}^n \subseteq X \cup X^{-1}$, where a letter is either an element of X or the formal inverse of an element of X, and we say that n is the *length* of the word. Note that the empty sequence of length 0 is a word, denoted by $\varnothing$.

The *concatenation* of two words $x_1x_2\cdots x_m$ and $y_1y_2\cdots y_n$ is the word $x_1\cdots x_my_1\ldots y_n$.

A word w' is an *elementary contraction* of a word w if $w = y_1xx^{-1}y_2$ and $w' = y_1y_2$, where y_1,y_2 are (possibly empty) words and $x \in X \cup X^{-1}$, and we write $w \searrow w'$. We also say that w is an *elementary expansion* of w' and write $w' \nearrow w$.

Two words a and b are *equivalent* if there are words $w_1,\ldots,w_n$ such that $a = w_1$ and $b = w_n$ and for each i, either $w_i \nearrow w_{i+1}$ or $w_i \searrow w_{i+1}$.

The *free group* $F(X)$ on the set X is the set of equivalence classes of words in X. The group operation is given by $[w]\cdot[w'] = [ww']$, and the identity element is given by $[\varnothing]$, also denoted by ε or e. The inverse of the element $[x_1x_2\cdots x_n]$ is given by $[x_m^{-1}\cdots x_2^{-1}x_1^{-1}]$.

A word is *reduced* if it does not admit an elementary contraction.

Theorem 18.3.1. *Every element of $F(X)$ is represented by a unique reduced word.*

18.3.2 Presentations

The free group satisfies a universal property in the category of groups, namely, given any function $f : X \to G$ from a set X to a group G, there is a unique homomorphism $\varphi : F(X) \to G$ $\varphi\big([x]\big) = f(x)$. That is, such that

$$\begin{tikzcd}
& F(X) \arrow[dd, dashed, "\varphi"] \\
X \arrow[ur, hook, "\iota"] \arrow[dr, "f"'] & \\
& G
\end{tikzcd}$$

commutes. That is, homomorphisms $F(X) \to G$ uniquely correspond to functions $X \to G$.

Because of this, given a group G, we can always find a set X and a surjection $F(X) \rightarrowtail G$.

Let G be a group and $B \subseteq G$ be a subset. The *normal subgroup generated by* B, denoted $\langle\langle B\rangle\rangle$, is smallest normal subgroup of G containing b. Or equivalently,

$$\langle\langle B\rangle\rangle := \bigcap_{B\subseteq N\trianglelefteq G} N$$

Because the intersection of normal subgroups is a normal subgroup, $\langle\langle B\rangle\rangle$ is itself normal in G.

Lemma 18.3.2.

$$\langle\langle B\rangle\rangle = \left\{\prod_{i=1}^{n} g_i b^{\pm 1} g_i^{-1} : n \in \mathbb{N}, b_i \in B, g_i \in G\right\}$$

If N is a normal subgroup of G containing B, then it certainly contains all the conjugates $g_ib_i^{\pm1}g_i^{-1}$, so N is a subset of this set. Conversely, this set contains B, as when $n = 0$ and $g_i = 1_G$, the product is just $b_i \in B$, and it can also be verified that this set is normal in G.

Let X be a set and $R \subseteq F(X)$. The group with *presentation* $\langle X \mid R \rangle$ is defined as

$$\langle X \mid R \rangle \coloneqq F(X)/\langle\langle R \rangle\rangle$$

Elements of the set R are called *relations.* Intuitively, the presentation $\langle X \mid R \rangle$ is the group on X that is as free as possible, subject to the constraint that every relation in R is identified with the identity.

Example.

- $\langle X \mid \varnothing \rangle \cong F(X)$
- $\langle t \mid t^n \rangle \cong \mathbb{Z}/n\mathbb{Z}$
- $\langle x,y \mid xyx^{-1}y^{-1} \rangle \cong \mathbb{Z}^2$

△

Example. The dihedral group of order $2n$ has presentation

$$\langle \sigma,\tau \mid \sigma^n,\tau^2,\tau\sigma\tau^{-1}\sigma \rangle$$

△

Because the relations are just specifying which elements are identified with the identity, we sometimes write equalities on the right side of a presentation to identify two expressions in the presentation. For instance, the more common presentation of the dihedral group of order $2n$ is given by

$$\langle \sigma,\tau \mid \sigma^n,\tau^2,\tau\sigma\tau^{-1} = \sigma^{-1} \rangle$$

Here, $\tau\sigma\tau^{-1} = \sigma^{-1}$ is called a *relator*, as it is not an element of R.

Example. The group with presentation

$$\langle a,b \mid ba^2b^{-1} = a^3,ab^2a^{-1} = b^3 \rangle$$

is the trivial group. △

As seen by this example, it is not immediately obvious what group any given presentation represents.

In fact, the *word problem* for a finitely generated group is the decision problem of determining whether two words in generators represent the same element. It turns out that the word problem is undecidable, so there is no algorithm to determine whether any given word is non-trivial.

Lemma 18.3.3. *Let $G = \langle X \mid R \rangle$. If w and w' are two words in X, then $[w] = [w']$ if and only if one can be obtained from the other by a finite sequence of applications of:*

- *elementary contractions/expansions;*
- *inserting any relation $r \in R$, or its inverse, into one of the words.*

Obviously, adding or removing gg^{-1} or $g^{-1}g$ into a word does not change its equivalence class, as it still represents the same reduced word. Similarly, elements of G are words in X modulo relations, so adding relations into a word also does not change its equivalence class.

The useful property of group presentations is that it is easy to determine when a map is a homomorphism by using the universal property of the free group.

Lemma 18.3.4. *Let $G = \langle X \mid R \rangle$ and H be groups. Let $f : X \to H$ be a set function, and let φ be its unique extension from the universal property of the free group. Then, f descends to a homomorphism*

$\bar{\varphi} : G \to H$ if and only if $\varphi(r) = 1_H$ for all $r \in R$.

$$\begin{tikzcd}
 & F(X) \arrow[dr, "q"] \arrow[dd, dashed, "\exists!\varphi"] & \\
X \arrow[ur, hook, "\iota"] \arrow[dr, "f"'] & & F(X)/\langle\langle R\rangle\rangle \arrow[dl, dashed, "\exists!\bar{\varphi}"] \\
 & H &
\end{tikzcd}$$

Proof. If $\bar{\varphi}$ is a homomorphism, then it must send the identity to the identity, so every every element of $\langle\langle R\rangle\rangle \supseteq R$ must be sent to the identity.

Conversely, suppose $\varphi(r) = 1_H$ for all $r \in R$. Any element $s \in \langle\langle R\rangle\rangle$ can be expressed as

$$\begin{aligned}
s &= \prod_{i=1}^{n} s_i r_i^{\pm 1} s_i^{-1} \\
\varphi(s) &= \varphi\left(\prod_{i=1}^{n} s_i r_i^{\pm 1} s_i^{-1}\right) \\
\varphi(s) &= \prod_{i=1}^{n} \varphi(s_i)\varphi(r_i^{\pm 1})\varphi(s_i^{-1}) \\
\varphi(s) &= \prod_{i=1}^{n} \varphi(s_i)1_H^{\pm 1}\varphi(s_i^{-1}) \\
\varphi(s) &= \prod_{i=1}^{n} \varphi(s_i)\varphi(s_i)^{-1} \\
\varphi(s) &= 1_H
\end{aligned}$$

so $s \in \ker(\varphi)$, and hence $\langle\langle R\rangle\rangle \subseteq \ker(\varphi)$, so this is a well-defined homomorphism from G to H. ■

Example. Consider the two presentations

$$G_1 = \langle x,y \mid xyx^{-1}y^{-1}\rangle, \qquad G_2 = \langle \sigma,\tau \mid \sigma^{2n},\tau^2,\tau\sigma\tau^{-1} = \sigma^{-1}\rangle$$

and define the set function $f : G_1 \to G_2$ on generators by

$$\begin{aligned}
f(x) &= \tau \\
f(y) &= \sigma^n
\end{aligned}$$

Noormally, to verify that this defines a homomorphism, we would need to check that $f(\hat{x}\hat{y}) = f(\hat{x})f(\hat{y})$ for all words $\hat{x},\hat{y} \in G_1$. Because G_1 is an infinite group, this is difficult to do. However, the previous lemma tells us that we only need to verify that the relations are in the kernel:

$$\begin{aligned}
f(xyx^{-1}y^{-1}) &= \tau\sigma^n\tau^{-1}\sigma^{-n} \\
&= \tau\big[\sigma\cdots\sigma\big]\tau^{-1}\sigma^{-n} \\
&= \tau\big[\sigma(\tau^{-1}\tau)\sigma(\tau^{-1}\tau)\cdots(\tau^{-1}\tau)\sigma(\tau\tau^{-1})\sigma\big]\tau^{-1}\sigma^{-n} \\
&= (\tau\sigma\tau^{-1})(\tau\sigma\tau^{-1})\cdots(\tau\sigma\tau^{-1})(\tau\sigma\tau^{-1})\sigma^{-n} \\
&= (\tau\sigma\tau^{-1})^n\sigma^{-n} \\
&= a^{-n}a^{-n} \\
&= a^{-2n} \\
&= 1_H
\end{aligned}$$

△

18.4 Coxeter Groups

Coxeter groups are defined via either graphs or matrices, as defined here.

A *Coxeter matrix* $M = (m_{ij})$ is on a set X is an $|X| \times |X|$ matrix satisfying:

- $m_{ij} \in \mathbb{N}_{\geq 1} \cup \{\infty\}$;
- $m_{ij} = m_{ji}$;
- $m_{ij} = 1$ if and only if $i = j$.

That is, it is a symmetric matrix with 1 along the diagonal, and integers at least 2 or infinity in all other entries.

A *Coxeter graph* Γ is an undirected finite simple graph with edges labelled by elements of $\mathbb{N}_{\geq 3} \cup \{\infty\}$.

Theorem 18.4.1. *There is a one-to-one correspondence between Coxeter graphs and Coxeter matrices.*

Proof. Given a Coxeter matrix M on a set X, we construct the Coxeter graph on $n = |X|$ vertices labelled $\{1, \ldots, n\}$, where two vertices i,j are adjacent if and only if $m_{ij} \geq 3$.

Conversely, given a Coxeter graph G with vertex set V, we construct the $|V| \times |V|$ Coxeter matrix M by

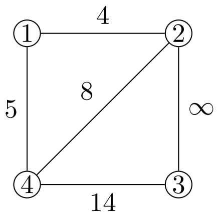
$$m_{ij} = \begin{cases} 1 & i = j \\ w(i,j) & \text{if edge } (i,j) \text{ exists} \\ 2 & \text{else} \end{cases}$$

■

Example. Given the Coxeter matrix

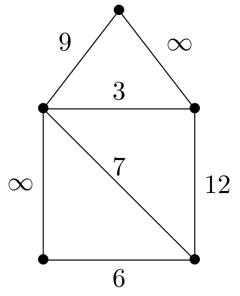
$$M = \begin{bmatrix} 1 & 4 & 2 & 5 \\ 4 & 1 & \infty & 8 \\ 2 & \infty & 1 & 14 \\ 5 & 8 & 14 & 1 \end{bmatrix}$$

the associated Coxeter graph is then

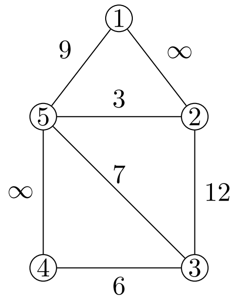


△

Example. Given the Coxeter graph



we first pick an ordering of the vertices:



then fill in the 5×5 matrix, with 1 on the diagonal; copying the edge weight of each edge; and filling in 2 otherwise:

$$M = \begin{bmatrix} 1 & \infty & 2 & 2 & 9 \\ \infty & 1 & 12 & 2 & 3 \\ 2 & 12 & 1 & 6 & 7 \\ 2 & 2 & 6 & 1 & \infty \\ 9 & 3 & 7 & \infty & 1 \end{bmatrix}$$

△

The matrix representation is more useful if the graph is very large, since lots of edges are hard to visualise. Conversely, if the graph is very sparse, and the matrix will be full of 2s and will be hard to read.

To simplify Coxeter graphs, it is convention to omit the label for edges with weight 3, since these edges will occur very frequently.

Given an $n \times n$ Coxeter matrix $M = (m_{ij})$ over a set X, the *Coxeter group* W_Γ is the group given by the presentation

$$W_\Gamma := \langle X \mid \forall i,j \leq n : (x_i x_j)^{m_{ij}} = 1 \rangle$$

- If $m_{ij} = \infty$, then there is no relation.
- If $i = j$, we have the relation $(x_i^2)^1 = x_i^2 = 1$ for every generator.
- If $m_{ij} = 2$, then $(x_i x_j)^2 = 1$, so

$$\begin{aligned} x_i x_j x_i x_j &= 1 \\ x_i x_j x_i x_j^2 &= x_j \\ x_i x_j x_i^2 &= x_j x_i \\ x_i x_j &= x_j x_i \end{aligned}$$

  so x_i and x_j commute.

Equivalently, given a Coxeter graph $\Gamma = (V,E)$ with associated Coxeter matrix $M = (m_{ij})$,the Coxeter group W_Γ defined by Γ is given by the presentation

$$W_\Gamma \coloneqq \big\langle V \mid \forall i,j \in V : i^2,(ij)^{m_{ij}} \big\rangle$$

Lemma 18.4.2. *Let $G = \langle a_1, \ldots ,a_n\rangle$ be a group generated by elements a_i all of order $|a_i| = 2$. Then, G is a quotient of the Coxeter group given by the Coxeter matrix with entries $m_{ij} = |a_i a_j|$.*

Proof. The Coxeter group given by this matrix has presentation

$$W = \big\langle x_1, \ldots , x_n \mid (x_i x_j)^{|a_i a_j|} = 1\big\rangle$$

Define the function $\varphi : W \to G$ on generators x_i by

$$\varphi(x_i) = a_i$$

We check that the relations are in the kernel of this map:

$$\begin{aligned}\varphi(x_i x_j) &= (a_i a_j)^{|a_i a_j|}\\ &= 1_G\end{aligned}$$

so φ defines a group homomorphism $W \to G$. ■

Theorem 18.4.3. *Let (G,V) be a finite reflection group with $G = \langle S_x \mid x \in \Pi\rangle$, and let W be the Coxeter group*

$$W = \big\langle x_1, \ldots , x_n \mid (x_i x_j)^{|S_i S_j|} = 1\big\rangle$$

as defined in the previous proof. Then, the homomorphism $\varphi : W \to G$

$$\varphi(x_i) = S_i$$

as defined in the previous proof is an isomorphism.

So, not only can we reduce the generating set of a finite reflection group from an entire root system Φ to only a simple system Π, this theorem then says further that the only relations that are relevant are the orders of *pairs* of reflections. That is, there are no relations of the form $S_i S_j S_k \ldots = 1$.

Lemma 18.4.4 (Deletion Condition)**.** *Let (G,V) be a finite reflection group, and let $\Pi \subseteq \Phi_{(G,V)}$ be a simple system. Suppose $g = S_{x_1} S_{x_2} \cdots S_{x_n}$ for some roots $x_1, \ldots ,x_n \in \Pi$, and $\ell(g) < n$. Then, there exist indices $1 \le i < j \le n$ such that*

$$g = S_{x_i} \cdots S_{x_{i-1}} S_{x_{i+1}} \cdots S_{x_{j-1}} S_{x_{j+1}} \cdots S_{x_n}$$

Example. There is a single Coxeter group on 1 generator, given by the presentation

$$\langle a \mid a^2\rangle \cong \mathbb{Z}/2\mathbb{Z}$$

△

Example. For two generators a and b, there are two options, depending on the value of m_{ab}.

- If $m_{ab} = m < \infty$, then
$$\langle a,b \mid a^2,b^2,(ab)^m\rangle \cong \mathrm{Dih}(n)$$
is dihedral group of order $2n$.
- If $m_{ab} = \infty$, then
$$\langle a,b \mid a^2,b^2\rangle \cong \mathrm{Dih}(\infty)$$
is the infinite dihedral group, which can be interpreted as the

△

Example. For three generators, a, b, and c, we have

$$W = \big\langle a,b,c \mid a^2,b^2,c^2,(ab)^k,(bc)^\ell,(ac)^m \big\rangle$$

If $k,\ell,m < \infty$, we have 3 cases:

- if $\frac{1}{k} + \frac{1}{\ell} + \frac{1}{m} = 1$, then this group describes the isometries of tilings of Euclidean 2-space, where each generator is a reflection;
- if $\frac{1}{k} + \frac{1}{\ell} + \frac{1}{m} < 1$, then this group describes the isometries of platonic solids, or the suspension of regular n-gons;
- if $\frac{1}{k} + \frac{1}{\ell} + \frac{1}{m} > 1$, then this group describes the isometries of tilings of hyperbolic 2-space.

where all the reflections meet at angles $\frac{\pi}{k}$, $\frac{\pi}{\ell}$, and $\frac{\pi}{m}$. △

18.4.1 Geometric Representations of Coxeter Groups

Lemma 18.4.5. *Let $U \subseteq V$ be a finite-dimensional subspace. If $U \cap U^\perp = \{0_V\}$, then $V = U \oplus U^\perp$.*

Proof. Let $e_1, \dots ,e_n$ be a basis for U, and let $v \in V$. The goal is to find scalars x_i such that $v - \sum_{i=1}^n x_i e_i \in V^\perp$. Or equivalently, such that

$$\left\langle v - \sum_{i=1}^n x_i e_i,e_j \right\rangle = 0$$

for all $1 \leq j \leq n$. This is same as solving the system of linear equations $\langle v,e_j\rangle = \sum_{i=1}^n x_i \langle e_i,e_j\rangle$. ■

Corollary 18.4.5.1. *If $x \in V$ satisfies $\langle x,x\rangle \neq 0$, then $V = \mathbb{R}x \oplus \{x\}^\perp$*

Given the above, we can define a *generalised reflection* as follows. Let $x \in V$ be such that $\langle x,x\rangle \neq 0$. We define $S_x : V \to V$ by

$$S_x(y) = y - 2\frac{\langle x,y\rangle}{\langle x,x\rangle}x$$

This satisfies similar properties to a reflection:

- $S_x(x) = -x$;
- $S_x\big|_{x^\perp} = \mathrm{id}_{x^\perp}$;
- $S_x^2 = \mathrm{id}_V$;
- $S_x \in O\big(V,\langle\,\cdot\,,\,\cdot\,\rangle\big)$;

Given a Coxeter group W_Γ associated to a Coxeter graph $\Gamma = (V,E)$, the goal is to find a group homomorphism $\rho_\Gamma : W_\Gamma \to O\big(V_\Gamma,\langle\,\cdot\,,\,\cdot\,\rangle\big)$, where V_Γ is the $\mathbb{R}$-vector space with basis $\{e_i : i \in V\}$ given by vertex set V of Γ.

We define the symmetric bilinear form $\langle\,\cdot\,,\,\cdot\,\rangle_\Gamma$ on V_Γ by

$$e_i,e_j = -\cos\left(\frac{\pi}{m_{i,j}}\right)$$

where $m_{ij} = w(i,j)$ is the weight of the edge (i,j) in Γ.

Note that

$$\langle e_i,e_i\rangle = -\cos\left(\frac{\pi}{1}\right) = 1$$

If $m_{ij} = 2$, then

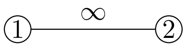
$$\langle e_i, e_j \rangle = -\cos\left(\frac{\pi}{2}\right) = 0$$

so e_i and e_j are orthogonal.

If $m_{ij} = \infty$, then

$$\langle e_i, e_j \rangle = -\cos\left(\frac{\pi}{\infty}\right) = -1$$

Consider the Coxeter graph



Then,

$$\begin{aligned}\langle e_1 + e_2, e_1 + e_2 \rangle &= \langle e_1, e_1 \rangle + 2\langle e_1, e_2 \rangle + \langle e_2, e_2 \rangle \\ &= 1 - 2 + 1 \\ &= 0\end{aligned}$$

So, we have $\langle x, x \rangle = 0$ for $x \neq 0$, so this form is *not* positive definite, and hence $\big(V_\Gamma, \langle\, \cdot\,,\, \cdot\, \rangle\big)$ is not necessarily a Euclidean space.

One effect of this is that orthogonal complements do not behave as in Euclidean spaces. For instance, if $W \subseteq V$, then:

- $W^\perp \cap W = \{0_V\}$ does not hold;
- $W = (W^\perp)^\perp$ does not hold, but $W \subseteq (W^\perp)^\perp$ does;
- $V = W \oplus W^\perp$ does not hold.

Theorem 18.4.6. *Let Γ be a Coxeter graph with two vertices x and y. Then, V_Γ is Euclidean if and only if $m_{xy} = w(x,y) < \infty$.*

Theorem 18.4.7. *Let $v = \alpha e_x + \beta e_y \in V_\Gamma$. Then,*

$$\begin{aligned}\langle v, v \rangle &= \alpha^2 \langle e_x, e_x \rangle + 2\alpha\beta \langle e_x, e_y \rangle + \beta^2 \langle e_y, e_y \rangle \\ &= \alpha^2 + 2\alpha\beta \cos\left(\frac{\pi}{m_{xy}}\right) + \beta^2 \\ &\geq (\alpha - \beta)^2 \\ &\geq 0\end{aligned}$$

with equality if and only if $\alpha = \beta = 0$, or if $m_{xy} = \infty$ and $\alpha = \beta$.

We define a map $\rho_\Gamma : W_\Gamma \to O\big(V_\Gamma, \langle\, \cdot\,,\, \cdot\, \rangle\big)$ on generators by

$$\rho_\Gamma(x) = S_{e_x}$$

Theorem 18.4.8. *The composition $S_{e_x} S_{e_y}$ has order m_{xy}.*

Proof. Suppose $m_{xy} < \infty$. Then, by the previous theorem, we have that the form restricted to $U = \mathbb{R}e_x \oplus \mathbb{R}e_y$ is Euclidean and hence non-degenerate. In particular, $U \cap U^\perp = \{0\}$, so $V_\Gamma = U \oplus U^\perp$. Since $U^\perp \subset e_x^\perp$, we have $S_{e_x}\big|_{e_x^\perp} = \text{id}$, and similarly, $S_{e_y}\big|_{e_y^\perp} = \text{id}$.

So, on U, this composition is a rotation by $2\pi/m_{xy}$ and hence has order m_{xy}. Now suppose $m_{xy} = \infty$. Both S_{e_x} and S_{e_y} preserve U, so we can compute matrices for $S_{e_x}\big|_U$ and $S_{e_y}\big|_U$ as:

$$S_{e_x} = \begin{bmatrix} -1 & 2 \\ 0 & 1 \end{bmatrix} \qquad S_{e_y} = \begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix}$$

Their product is

$$\begin{bmatrix} 3 & -2 \\ 2 & -1 \end{bmatrix}$$

which has infinite order, since its Jordan canonical form is

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

■

Corollary 18.4.8.1. *ρ_Γ defines a group homomorphism.*

Corollary 18.4.8.2. *The element xy in W_Γ has order exactly m_{xy}.*

Proof. Since we have the relation $(xy)^{m_{xy}}$, the order of xy divides m_{xy}, and is therefore at most m_{xy}. However, its image under ρ_Γ has order m_{xy}, so it must have order at least m_{xy}. So $|xy| = m_{xy}$. ■

Recall that if a and b are vertices of Γ that are not connected by an edge, then $m_{ab} = 2$, so a and b commute in W_Γ.

Lemma 18.4.9. *Suppose Γ is disconnected, so $\Gamma = \Gamma_1 \sqcup \Gamma_2$. Then,*

$$\begin{aligned} W_\Gamma &\cong W_{\Gamma_1} \times W_{\Gamma_2} \\ V_\Gamma &\cong V_{\Gamma_1} \oplus V_{\Gamma_2} \\ \rho_\Gamma &\cong \rho_{\Gamma_1} \oplus \rho_{\Gamma_2} \end{aligned}$$

and V_{Γ_1} is orthogonal to V_{Γ_2}.

Lemma 18.4.10. *If Γ is a connected graph, then any W_Γ-invariant proper subspace is contained in $V_\Gamma^\perp$.*

Proof. Suppose $U \subseteq V_\Gamma$ is preserved by W_Γ. We claim that for each $x \in \Gamma$, we have either $e_x \in U$, or $U \subseteq e_x^\perp$.

Suppose $U \not\subseteq e_x^\perp$ so there exists $u \in U$ such that $\langle u,e_x \rangle \neq 0$. Then,

$$\begin{aligned} S_{e_x}(u) &= u - 2\frac{\langle u,e_x \rangle}{\langle e_x,e_x \rangle} e_x \\ \frac{1}{2\langle u,e_x \rangle}\big(S_{e_x}(u) - u\big) &= e_x \end{aligned}$$

Since U is preserved by W_Γ, $S_{e_x}(u) \in U$, and also $u \in U$, so $e_x \in U$ as it is a linear combination of vectors in U.

This partitions the vertices of Γ into the sets

$$S_1 \coloneqq \{x \in \Gamma : e_x \in U\} \qquad S_2 \coloneqq \{x \in \Gamma : U \subseteq e_y^\perp\}$$

However, for all $x \in S_1$ and $y \in S_2$, we have $\langle e_x,e_y \rangle = 0$, so $m_{xy} = 2$, but Γ is connected. It follows that the vertices of Γ are contained entirely within one of the sets. If it is S_1, then $U = V_\Gamma$. Otherwise, if $U = S_2$ ■

Corollary 18.4.10.1. *If Γ is connected, and V_Γ is Euclidean, then ρ_Γ is irreducible. That is, there are no proper W_Γ-invariant subspaces.*

Let $\rho : G \to GL_n(\mathbb{R})$ be a group homomorphism. Then, ρ is *completely reducible* if there exist $\rho(G)$-invariant subspaces $V_1, \dots, V_k$ such that $\mathbb{R}^n \cong V_1 \oplus \cdots \oplus V_k$, and G acting on V_i is irreducible.

In other words, there exists a basis of $\mathbb{R}^n$ such that the image of ρ is a block matrix with blocks along the diagonal and zero elsewhere.

Lemma 18.4.11. *If G is a finite group, then any representation $\rho : G \to GL_n(\mathbb{R})$ is completely reducible.*

Lemma 18.4.12. *If W_Γ is finite and Γ is connected, then ρ_Γ is irreducible.*

Suppose we have a group G acting on vector spaces X and Y. A linear map $f : X \to Y$ is *G-equivariant* if $f(g \cdot x) = g \cdot f(x)$. We denote the set of G-equivariant linear maps from X to Y by $\mathrm{Map}_G(X,Y)$. If $X = Y$, then these are G-equivariant endomorphisms and are denoted $\mathrm{End}_G(X)$.

Lemma 18.4.13. *Let Γ be connected and let W_Γ be finite. Then,* $\mathrm{End}_{W_\Gamma}(V_\Gamma) = \{k \, \mathrm{id}_{V_\Gamma} : k \in \mathbb{R}\} \cong \mathbb{R}$.

Theorem 18.4.14. *Suppose W_Γ is a finite group. Then $\big(V_\Gamma, \langle \cdot\,,\cdot \rangle_\Gamma\big)$ is Euclidean.*

Corollary 18.4.14.1. *If W_Γ is finite, then $\big(\rho_\Gamma(W_\Gamma), V_\Gamma\big)$ is a finite reflection group.*

So far, given a finite reflection group (G,V), we can find a Coxeter group W_Γ which is isomorphic to G.

Given a Coxeter group W_Γ, we have a representation that induces a finite reflection group $\big(\rho_\Gamma(W_\Gamma), V_\Gamma\big)$.

We will show that this representation ρ_Γ is faithful.

To do this we redefine some concepts for general Coxeter groups.

Let $g \in W_\Gamma$. Then, the *length* of g is

$$\ell(g) \coloneqq \min\{n : \exists x_1, \dots, x_n \in \Gamma : g = x_1 x_2 \cdots x_n\}$$

This satisfies similar properties to lengths for finite reflection groups:

- $\ell(g) = 0$ if and only if $g = 1_{W_\Gamma}$;
- $\ell(gh) \le \ell(g) + \ell(h)$;
- $\ell(gh) \ge \ell(g) - \ell(h)$;
- for all $g \in W_\Gamma$ and $x \in \Gamma$, $\ell(gx) = \ell(g) + 1$ or $\ell(gx) = \ell(g) - 1$.

We abbreviate $\rho_\Gamma(x)$ to ρ_x.

Let W_Γ be a Coxeter group. The *root system* associated to W_Γ is defined by

$$\Phi_\Gamma \coloneqq \big\{\rho_w(e_x) : w \in W_\Gamma, x \in \Gamma\big\}$$

That is, Φ_Γ is the union of the orbits of the basis vectors e_x.

A root is *positive* if it can be written as a non-negative linear combination of the e_x, and is *negative* if it can be written as a non-positive linear combination of the e_x.

Given a subset I of the vertices of Γ, the *parabolic subgroup W_I* of W_Γ corresponding to I is the subgroup of W_Γ generated by I. For $w \in W_I$, let $\ell_I(w)$ denote the length of w in the generating set I.

Theorem 18.4.15. *Let $g \in W_\Gamma$ and let x be a vertex of Γ. If $\ell(gx) > \ell(w)$, then $g \cdot e_x$ is a positive root. Similarly, if $\ell(gx) < \ell(w)$, then $g \cdot e_x$ is a negative root.*

Corollary 18.4.15.1. *Ever root in Φ is positive or negative.*

Theorem 18.4.16. *The representation ρ_Γ is faithful. That is,* $\ker(\rho_\Gamma) = \{1\}$.

Proof. If not, then let $g \in \ker(\rho_\Gamma)$ such that $\ell(w) > 1$. Then, there exists $x \in \Gamma$ such that $\ell(gx) < \ell(g)$. But then $e_x = g \cdot e_x$ must be a negative root ■

Theorem 18.4.17. *The standard parabolic subgroup of W_Γ corresponding to I is isomorphic to the Coxeter group with vertex set I and labels coming from Γ.*

Corollary 18.4.17.1. *If W_Γ is finite, then $\Pi = \{e_x : x \in \Gamma\}$ is a simple system in Φ_Γ.*

A reflection group (G,V) is *essential* if V is the span of $\Phi_{(G,V)}$.

Theorem 18.4.18. *For any finite reflection group (G,V),*

$$(G,V) \simeq \big(G, \operatorname{span}(\Phi_{(G,V)}) \oplus U\big)$$

where G acts on U trivially.

Theorem 18.4.19. *The map $\Gamma \mapsto \big(\rho_\Gamma(W_\Gamma),V_\Gamma\big)$ is a bijection from the set of finite Coxeter graphs up to labelled graph isomorphisms, to the set of essential finite reflection groups.*

Note that

$$\Phi_{\left(\rho_\Gamma(W_\Gamma),V_\Gamma\right)} = \Phi_{W_\Gamma}$$

18.5 The Finiteness Criterion

The topology on $GL(V_\Gamma)$ comes from a norm on the set $\operatorname{End}(V_\Gamma)$ of all linear endomorphisms $T : V_\Gamma \to V_\Gamma$ as follows.

Let $\|\cdot\|$ be any norm on V_Γ, and for $T : V_\Gamma \to V_\Gamma$, define the operator norm by

$$\begin{aligned}\|T\| &:= \sup_{\|x\|=1} \|T(v)\| \\ &= \sup_{\|x\|\neq 0} \frac{\|T(v)\|}{\|v\|}\end{aligned}$$

The operator norm satisfies:

- $\|T(x)\| = \|T\|\|x\|$;
- $\|T\| = 0$ if and only if $T = 0$;
- $\|T + S\| \leq \|T\| + \|S\|$;
- if $T \in O(V_\Gamma)$, then $\|T\| = 1$;
- $O(V_\Gamma)$ is a closed, bounded, and compact subset of $\operatorname{End}(V_\Gamma)$.

Theorem 18.5.1. *Suppose that V_Γ is a Euclidean space. Then, W_Γ is a finite group.*

We have already proved the converse of this statement, so we have:

Corollary 18.5.1.1. *The Coxeter group W_Γ is finite if and only if V_Γ is Euclidean.*

A Coxeter graph Γ is *positive definite* if V_Γ is Euclidean. Everything we have done has allowed us to reduce to the case of understanding connected, positive definite Coxeter graphs.

A symmetric matrix is *positive definite* if the associated bilinear form defined by $\langle x,y\rangle = x^\top Ay$ is positive definite.

Lemma 18.5.2. *A symmetric matrix A is positive definite if and only if all of its eigenvalues are positive.*

Lemma 18.5.3. *Let $A = (a_{ij})_{1\leq i,j\leq n}$ be a symmetric matrix. Then, the associated form is positive definite if and only if for each $k \in \{1,\ldots,n\}$, the upper left k-submatrix $A_k = (a_{ij})_{1\leq i,j\leq k}$ has positive determinant.*

Given a Coxeter graph $\Gamma = (V,E)$, and an induced Coxeter subgraph Λ with vertex set $I \subseteq V$, there is a natural inclusion $W_\Lambda \hookrightarrow W_\Gamma$, and the image of this map is the parabolic subgroup W_I, and we have $W_I \cong W_\Lambda$.

$$\begin{tikzcd}
W_\Lambda \arrow[r, hook] \arrow[d, "\rho_\Lambda"'] & W_\Gamma \arrow[d, "\rho_\Gamma"] \\
GL(V_\Lambda) \arrow[r, hook] & GL(V_\Gamma)
\end{tikzcd}$$

Theorem 18.5.4. *If Λ is a Coxeter subgraph of Γ and:*

- *W_Γ is finite, then W_Λ is also finite;*
- *V_Γ is Euclidean, then V_Λ is also Euclidean;*
- *Γ is positive definite, then Λ is also positive definite.*

Let C_Γ be the matrix associated to the bilinear form on V_Γ. That is,

$$C_\Gamma = \left(-2\cos\left(\frac{\pi}{m_{xy}}\right)\right)_{x,y\in\Gamma}$$

We define $d(\Gamma) = \det(C_\Gamma)$.

Lemma 18.5.5. *Suppose Γ is a graph with a leaf node whose unique edge has label 3. Let Γ_1 be the graph obtained by deleting this node from Γ, and let Γ_2 be the graph obtained by deleting both endpoints of this edge. Then,*

$$d(\Gamma) = 2d(\Gamma_1) - d(\Gamma_2)$$

Proof. Order the vertices of the graph such that the last row in the Coxeter matrix for Γ corresponds to the leaf node, and the $(n-1)$ row corresponds to the other endpoint of the leaf node's edge:

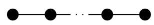
$$M = \begin{bmatrix} & & & * & 0 \\ & C_{\Gamma_2} & & \vdots & \vdots \\ & & & * & 0 \\ * & \cdots & * & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{bmatrix}$$

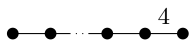
Laplacian expansion along the last row gives:

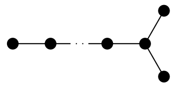
$$\begin{aligned} d(\Gamma) &= 2d(\Gamma_1) + (-1)^{(n+n-1)}(-1)\det\begin{bmatrix} & & & 0 \\ & C_{\Gamma_2} & & \vdots \\ & & & 0 \\ * & \cdots & * & -1 \end{bmatrix} \\ &= 2d(\Gamma_1) - d(\Gamma_2) \end{aligned}$$

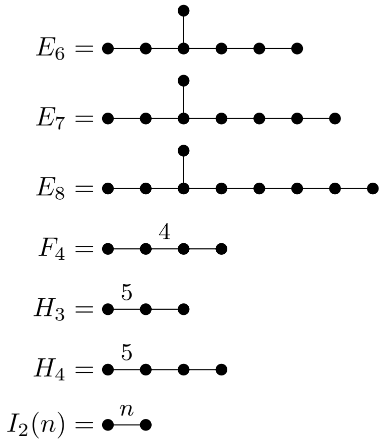
■

Theorem 18.5.6. *The following graphs have $d(\Gamma) > 0$:*

$A_n =$ 

$B_n =$ 

$D_n =$ 



where the subscript denotes the number of vertices in the Coxeter graph.

Theorem 18.5.7. *Suppose that* Γ *is a connected positive definite Coxeter graph. Then,* Γ *is one of the graphs above.*

18.6 The Exchange and Deletion Conditions

Let (W,S) be a pair consisting of a group W and a generating set S of elements of order 2. We say that (W,S) satisfies the *deletion condition* if, whenever $g = s_1 s_2 \cdots s_n$ for some $s_1,\ldots,s_n \in S$ with $\ell(g) < n$, there exists $1 \leq i < j \leq n$ such that

$$g = s_1 \cdots s_{i-1} s_{i+1} \cdots s_{j-1} s_{j+1} \cdots s_n$$

Note that the deletion condition depends on both the group and the generating set. For instance,

$$\big(\mathbb{Z}/2 \times \mathbb{Z}/2, \{(1,0),(0,1)\}\big)$$

satisfies the deletion condition, but

$$\big(\mathbb{Z}/2 \times \mathbb{Z}/2, \{(1,0),(0,1),(1,1)\}\big)$$

does not.

We have already seen that finite reflection groups satisfy this condition. To prove that this result also holds for Coxeter groups, we have to use the root system associated to the Coxeter group.

Recall that the root system for Γ is defined as $\Phi_\Gamma \coloneqq \{\rho_w(e_x) : w \in W_\Gamma, x \in \Gamma\}$. That is, the union of the orbits of the basis vectors e_x.

A root is *positive* if it can be written as a non-negative linear combination of the e_x, and is *negative* if it can be written as a non-positive linear combination of the e_x.

Lemma 18.6.1. *Let* $x \in \Gamma$. *Then,* $\rho_x(\Phi_+) \cap \Phi_- = \{e_x\}$.

Theorem 18.6.2. *Let* $\Gamma = (V,E)$ *be a Coxeter graph. Then,* (W_Γ, V) *satisfies the deletion condition.*

Let (W,S) be a pair consisting of a group W and a generating set S of elements of order 2. We say that (W,S) satisfies the *exchange condition* if: whenever $s_1 \cdots s_r = t_1 \cdots t_r$ are two words in S representing the same element $w \in W$ with $\ell(w) = r$ and $s_1 \neq t_1$, then there is an index $i \in \{2,\ldots,r\}$ such that $w = s_1 t_1 \cdots t_{i-1} t_{i+1} \cdots t_r$.

Theorem 18.6.3. *Suppose* (W,S) *satisfies the deletion condition. Then,* (W,S) *satisfies the exchange condition.*

Corollary 18.6.3.1. *Let W_Γ be a Coxeter group. Then, the pair (W_Γ,Γ) satisfies the exchange condition.*

Given a pair (W,S) satisfying the exchange condition, we can construct a Coxeter graph with vertex set S and edge weights $m_{st} = |st|$ to be the order of the word st. This yields a Coxeter group that surjects onto W. Let $M = (m_{ij})$ be the Coxeter matrix associated with this Coxeter group. Then, an *M-elementary reduction* of a word $w \in W$ is one of the following operations:

- Delete a subword of the form ss for $s \in S$;
- Replace a subword $sts\cdots$ with $tst\cdots$ where each of the words has exactly $m_{st} = |st|$ letters.

Theorem 18.6.4. *Let (W,S) be a pair satisfying the deletion condition, and let M be the associated Coxeter matrix. Let $w = s_1 \cdots s_k$ be a word with length $\ell(w) = k$. Then, given any other decomposition $w = t_1 \cdots t_m$, we can obtain $s_1 \cdots s_k$ from $t_1 \cdots t_m$ using M-elementary reductions.*

Corollary 18.6.4.1. *If (W,S) is a pair satisfying the deletion condition, then W is a Coxeter group.*

18.7 The Davis Complex

If W_Γ is finite, then (W_Γ,V_Γ) is a finite reflection group, so it has a nice group action on Euclidean space. Furthermore, this action preserves the unit sphere, and this restricts to a nice action on S^n.

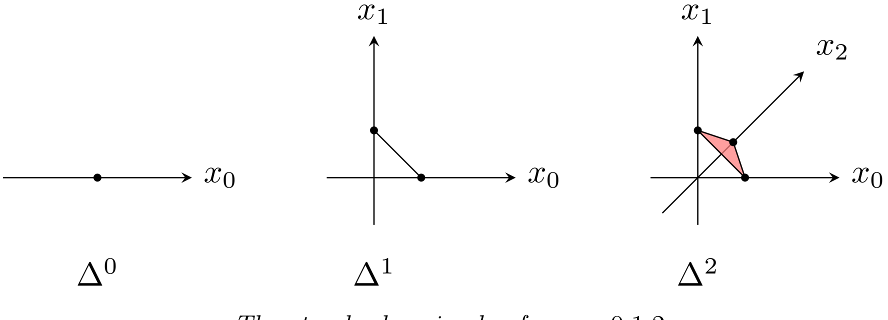
For infinite Coxeter groups, the geometric representation gives a faithful action on the inner product space $\big(V_\Gamma,\langle\,\cdot\,,\cdot\,\rangle_\Gamma\big)$. For the infinite case, there is a "nicer" space upon which the Coxeter group acts, and this is known as the *Davis complex*.

18.7.1 Simplicial Complexes

The *standard n-simplex* $\Delta^n \subseteq \mathbb{R}^{n+1}$ is the subspace

$$\Delta^n := \left\{ \mathbf{x} \in \mathbb{R}^{n+1} : x_i \geq 0, \sum_{i=0}^{n} x_i = 1 \right\}$$

whose vertices $v_0,v_1,\ldots,v_n$ are the unit vectors along the coordinate axes.



The standard n-simplex for $n = 0,1,2$

The *vertice set* $V(\Delta^n)$ of the n-simplex is the set of points where $x_i = 1$ for some i, and we denote the vertex corresponding to $x_i = 1$ by v_i. Each dimension adds an additional vertex, so $|V(\Delta^n)| = n + 1$. Also note that $V(\Delta^n)$ forms a basis for $\mathbb{R}^{n+1}$.

For each non-empty set $A \subseteq \{0,\ldots,n\}$, we define a *face* Δ_A of Δ^n to be the subspace

$$\Delta_A := \big\{ (x_0,\ldots,x_n) \in \Delta^n : \forall i \notin A, x_i = 0 \big\}$$

Note that we consider Δ^n to be a face of itself.

We define the *interior* $\mathring{\Delta}^n$ of the n-simplex to be the subspace of points where $x_i > 0$ for all i. Note that for $n = 0$, we have $\mathring{\Delta}^0 = \Delta^0$.

Suppose $m \leq n$ and suppose we have an injection $f : \{0,\dots,m\} \to \{0,\dots,n\}$. Then, this map extends to a map $f_* : \mathbb{R}^{m+1} \to \mathbb{R}^{n+1}$ given by

$$f_*(v_i) = v_{f(i)}$$

for the bases $V(\Delta^m)$ and $V(\Delta^n)$. This induces a continuous map $\Delta^m \to \Delta^n$, which we call a *face inclusion.*

An *abstract simplicial complex* is a pair $K = (V,\Sigma)$, where V is a set containing the *vertices* of K, and Σ is a set of finite subsets of V containing the *simplices* of K, satisfying:

- for each $v \in V$, $\{v\} \in \Sigma$;
- if $\sigma \in \Sigma$ and $\tau \subseteq \sigma$, then $\tau \in \Sigma$ (transitive);

We can associate to each abstract simplicial complex a topological space.

The *topological realisation* $|K|$ of an abstract simplicial complex $K = (V,\Sigma)$ is obtained as follows:

1. For each $\sigma \in \Sigma$, take a copy Δ^n_σ of the standard n-simplex, where $n = |\sigma| - 1$, and pick a bijection $V(\Delta^n_\sigma) \to \sigma$.
2. Whenever $\tau \subset \sigma$, using the above bijections, obtain an injection $V(\Delta_\tau) \to V(\Delta_\sigma)$, inducing a face inclusion $f_{\tau\sigma} : \Delta_\tau \to \Delta_\sigma$.
3. Define
$$|K| = \left(\bigsqcup_{\sigma \in \Sigma} \Delta_\sigma \right) / \sim$$
where $x \sim f_{\tau\sigma}(x)$ for all $x \in \Delta_\tau$ and all $\tau \subset \sigma$.

That is, $|K|$ is the disjoint union of the simplices modulo the face inclusions.

Example. If $K_X = \big(X,\mathcal{P}(X)\big)$, then $|K_X| \cong \Delta^{|X|-1}$. △

For each $\sigma \in \Sigma$, there is a natural map $\Delta_\sigma \to |K|$ that identifies Δ_σ with a subspace of $|K|$. Given $\tau \neq \sigma$, we see that $\mathring{\Delta}_\sigma \cap \mathring{\Delta}_\tau = \varnothing$. Thus, for each point $x \in |K|$, there is a unique $\sigma \in \Sigma$ such that $x \in \mathring{\Delta}_\sigma$.

18.7.2 Geometric Realisations of Posets

Recall that a *partially ordered set* or *poset* is a set equipped with a reflexive, symmetric, and transitive relation, or a *partial ordering*, $\leq$.

Given a poset $(X, \leq)$, we can associate a simplicial complex $K_X = (X,\Sigma)$, where Σ consists of the subsets $\{x_0,\dots,x_n\}$ such that $x_i \leq x_j$ if $i \leq j$. We call $|K_X|$ the *geometric realisation* of X.

Given a Coxeter graph Γ, we can consider the set $P_\Gamma \coloneqq \{I \subseteq \Gamma : W_I \text{ is finite}\}$. This is a poset when ordered by inclusion. Note that $\varnothing \in P_\Gamma$ since $W_\varnothing = \{1\}$.

Denote the geometric realisation of P_Γ by K_Γ. As before, for each point $x \in |K|$, there is a unique $\sigma \in \Sigma$ such that $x \in \mathring{\Delta}_\sigma$. Since a simplex σ corresponds to a chain in P_Γ, there is a minimal element in σ, which we will denote by I_x. We define the *point stabiliser* of x to be W_{I_x}, also denoted by W_x when no confusion will arise.

The *Davis complex* Σ_Γ associated to Γ is the space $K_\Gamma \times W_\Gamma / \sim$, where $(x,w) \sim (x',w')$ if $x = x'$ and $w^{-1}w' \in W_x$.

Lemma 18.7.1. *The Davis complex Σ_Γ has an action of W_Γ given by $w \cdot [(x,z)] = [(x,wz)]$.*

Lemma 18.7.2. *The stabiliser of the point $[(x,z)]$ is given by zW_xz^{-1}.*

Chapter 19

Lie Groups

Chapter 20

Lie Algebras

A *Lie algebra* is a vector space equipped with an additional multiplication operation that is typically non-associative. Lie algebras are closely related to Lie groups, which are groups that are also smooth manifolds; every Lie group induces a Lie algebra as the tangent space at the identity, in which case, the Lie bracket measures the failure of commutativity for the Lie group. Conversely, to any finite-dimensional Lie algebra over the $\mathbb{R}$ or $\mathbb{C}$, there is a corresponding connected Lie group. This correspondence allows us to study the structure and classification of Lie groups in terms of Lie algebras.

Lie groups and Lie algebras find extensive applications in physics – in particular, quantum and particle mechanics – where Lie groups arise as symmetry groups of physical systems and their Lie algebras may be interpreted as infinitesimal symmetry motions of those systems.

20.1 Lie Algebras

All of the vector spaces we consider will be finite dimensional over a field K.

A *Lie bracket* on a vector space L is a bilinear map $[-,-] : V \times V \to V$ with the additional properties:

(L1) (*Alternation*) For all $x \in L$, $[x,x] = 0$;

(L2) (*Jacobi identity*) For all $x,y,z \in L$, $\big[x,[y,z]\big] + \big[z,[x,y]\big] + \big[y,[x,z]\big] = 0$.

The pair $\big(L,[-,-]\big)$ is then called a *Lie algebra* over K. We often suppress the Lie bracket and the field, and refer to a Lie algebra by the underlying vector space L.

The *dimension* of a Lie algebra L is the dimenion of L as a vector space.

Lemma 20.1.1 (Anticommutativity)**.** *Let L be a Lie algebra. Then, for all $x,y \in L$,*

$$[x,y] = -[y,x]$$

Proof. By (L1), for all $x,y \in L$, $[x+y,x+y] = 0$, so by bilinearity, $[x,x]+[x,y]+[y,x]+[y,y] = 0$. Again by (L1), $[x,x] = 0 = [y,y]$, so $[x,y] + [y,x] = 0$, and hence $[x,y] = -[y,x]$. ■

Lemma 20.1.2. *If* $\operatorname{char}(K) \neq 2$*, then the alternating property is equivalent to anticommutativity.*

Proof. The forward implication is shown in the previous lemma. Conversely, suppose L satisfies (L2) and anticommutativity. Then, $[x,x] = -[x,x]$, so $2[x,x] = 0$. Since the characteristic of K is not 2, 2 is invertible, so $[x,x] = 0$. ■

Example.

(*i*) Let V be any vector space and define $[-,-] : V \times V \to V$ to be the constant zero vector map. This bracket trivially satisfies the Lie bracket axioms, so $\big(V,[-,-]\big)$ is a Lie algebra, called an *abelian* Lie algebra.

Every 1-dimensional Lie algebra is necessarily abelian since if e is the basis element, then $[a,b] = [\alpha e,\beta e] = \alpha\beta[e,e] = 0$.

(*ii*) Let $L = \mathbb{R}^3$ be a vector space over $\mathbb{R}$. The cross product satisfies the Lie bracket axioms, so $\mathbb{R}^3$ is a Lie algebra over $\mathbb{R}$.

(*iii*) Consider the set $L = M_n(K)$ of $n \times n$ matrices with entries in K as a n^2-dimensional vector space over K. Define the bracket $[-,-] : L \times L \to L$ by

$$[A,B] = AB - BA$$

This is linear in the first argument:

$$[\lambda A + \mu B,C] = (\lambda A + \mu B)C - C(\lambda A + \mu B) = \lambda(AC - CA) + \mu(BC - CB) = \lambda[A,C] + \mu[B,C]$$

and since $[A,B] = AB - BA = -(BA - AB) = -[B,A]$, we also have linearity in the second argument. The bracket is also alternating since $[A,A] = AA - AA = \mathbf{0}$. We also have:

$$\begin{aligned}
\big[A,[B,C]\big] &= [A,BC - CB] \\
&= A(BC - CB) - (BC - CB)A \\
&= ABC - ACB - BCA + CBA \\
\big[C,[A,B]\big] &= CAB - CBA - ABC + BAC \\
\big[B,[C,A]\big] &= BCA - BAC - CAB + ACB
\end{aligned}$$

The 12 terms are the positive and negatives of the permutations of A, B, and C, so adding these together, we obtain $\mathbf{0}$, and the Jacobi identity holds. So $\big(M_n(K),[-,-]\big)$ is a Lie algebra.

This Lie algebra is also denoted by $\mathfrak{gl}_n(K)$ (since it is the Lie algebra of the Lie group $GL_n(K)$).

(*iv*) Let V be any vector space and consider the endomorphism space $\text{End}(V)$ of V. We similarly define the bracket $[-,-] : \text{End}(V) \times \text{End}(V) \to \text{End}(V)$ by $[S,T] = S \circ T - T \circ S$. This defines a Lie algebra, denoted by $\mathfrak{gl}(V)$.

(*v*) Consider the linear subspace $L = \big\{A \in M_n(K) : \text{tr}(A) = 0\big\} \subseteq M_n(K)$ of matrices with zero trace. Since the trace is linear and satisfies $\text{tr}(AB) = \text{tr}(BA)$, the restriction of the Lie bracket from $\mathfrak{gl}_n(K)$ is closed on L since $\text{tr}\big([A,B]\big) = \text{tr}(AB - BA) = \text{tr}(AB) - \text{tr}(BA) = \text{tr}(AB) - \text{tr}(AB) = 0$. Thus, L is again a Lie algebra, denoted by $\mathfrak{sl}_n(K)$ (again, since it is the Lie algebra of the Lie group $SL_n(K)$).

(*vi*) Let $L = \big\{A \in M_n(K) : a_{ij} = 0 \text{ for all } i > j\big\}$ be the set of **non-strictly** upper triangular matricies over K. Again, the Lie bracket from $\mathfrak{gl}_n(K)$ is closed on L since the product and sum of upper triangular matricies is again upper triangular, so L is again a Lie algebra, denoted by $\mathfrak{b}_n(K)$.

(*vii*) Let $L = \big\{A \in M_n(K) : a_{ij} = 0 \text{ for all } i \geq j\big\}$ be the set of **strictly** upper triangular matricies over K. Again, L is a Lie algebra, denoted by $\mathfrak{u}_n(K)$.

△

20.1.1 Structure Constants

Let L be a Lie algebra, and let $e_1, \dots, e_n$ be a basis of L. Then, for $a,b \in L$, we may express them as linear combinations of the basis elements and use the linearity of the Lie bracket to obtain:

$$\begin{aligned}[a,b] &= \left[\sum_i \alpha_i e_j, \sum_j \beta_j e_j\right] \\ &= \sum_{i,j} \alpha_i \beta_j [e_i,e_j]\end{aligned}$$

Removing the diagonal elements and combining the antisymmmetric combinations, this simplifies to:

$$= \sum_{i,j} (\alpha_i \beta_j - \beta_i \alpha_j)[e_i,e_j]$$

If we compute the Lie brackets $[e_i,e_j]$ of the basis elements, then we can compute any other Lie bracket $[a,b]$ using this formula.

Example. Consider the space $\mathfrak{gl}_2(\mathbb{R})$ with basis

$$e_1 = E_{11} = \begin{bmatrix}1 & 0\\0 & 0\end{bmatrix} \qquad e_2 = E_{12} = \begin{bmatrix}0 & 1\\0 & 0\end{bmatrix} \qquad e_3 = E_{21} = \begin{bmatrix}0 & 0\\1 & 0\end{bmatrix} \qquad e_4 = E_{22} = \begin{bmatrix}0 & 0\\0 & 1\end{bmatrix}$$

Applying the Lie bracket to pairs of these basis elements, we have:

$$\begin{array}{llll} [e_1,e_1] = \mathbf{0} & [e_1,e_2] = e_2 & [e_1,e_3] = -e_3 & [e_1,e_4] = \mathbf{0} \\ [e_2,e_1] = -e_2 & [e_2,e_2] = \mathbf{0} & [e_2,e_3] = e_1 - e_4 & [e_2,e_4] = e_2 \\ [e_3,e_1] = e_3 & [e_3,e_2] = e_4 - e_1 & [e_3,e_3] = \mathbf{0} & [e_3,e_4] = -e_3 \\ [e_4,e_1] = \mathbf{0} & [e_4,e_2] = -e_2 & [e_4,e_3] = e_3 & [e_4,e_4] = \mathbf{0} \end{array}$$

(Since the Lie bracket is anticommutative and alternating, we only really need to compute the 6 entries above the diagonal.)

Let

$$\begin{aligned} A &= \begin{bmatrix}1 & 2\\3 & 4\end{bmatrix} & \qquad B &= \begin{bmatrix}5 & -6\\-7 & 8\end{bmatrix} \\ &= 1e_1 + 2e_2 + 3e_3 + 4e_4 & &= 5e_1 - 6e_2 - 7e_3 + 8e_4 \end{aligned}$$

By direct computation, the Lie bracket $[A,B]$ is given by:

$$\begin{aligned}[A,B] &= AB - BA \\ &= \begin{bmatrix}-9 & 10\\-13 & 14\end{bmatrix} - \begin{bmatrix}-13 & -14\\17 & 18\end{bmatrix} \\ &= \begin{bmatrix}4 & 24\\-30 & -4\end{bmatrix}\end{aligned}$$

Alternatively, the formula yields:

$$\begin{aligned}[A,B] &= \big(1\cdot(-6) - 5\cdot 2\big)[e_1,e_2] + (1\cdot(-7) - 5\cdot 3)[e_1,e_3] + (1\cdot 8 - 5\cdot 4)[e_1,e_4] \\ &\quad + \big(2\cdot(-7) - (-6)\cdot 3\big)[e_2,e_3] + \big(2\cdot 8 - (-6)\cdot 4\big)[e_2,e_4] + \big(3\cdot 8 - (-7)\cdot 4\big)[e_3,e_4] \\ &= -16[e_1,e_2] - 22[e_1,e_3] - 12[e_1,e_4] + 4[e_2,e_3] + 40[e_2,e_4] + 52[e_3,e_4] \\ &= -16e_2 + 22e_3 + 4(e_1 - e_4) + 40e_2 - 52e_3 \\ &= 4e_1 + 24e_2 - 30e_3 - 4e_4 \\ &= \begin{bmatrix}4 & 24\\-30 & -4\end{bmatrix}\end{aligned}$$

△

Since $[e_i{,}e_j] \in L$, these brackets themselves can also be expressed in the basis as:

$$[e_i{,}e_j] = \sum_k c_{ij}^k e_k$$

The coefficients c_{ij}^k are called the *structure constants* of L with respect to the basis $e_1,\ldots{,}e_n$.

Example. In the above example, the Lie brackets are:

$$\begin{aligned} &[e_1{,}e_2] = e_2, && [e_1{,}e_3] = -e_3 && [e_1{,}e_4] = \mathbf{0} \\ &[e_2{,}e_3] = e_1 - e_4, && [e_2{,}e_4] = e_2 && [e_3{,}e_4] = -e_3 \end{aligned}$$

The corresponding structure constants are thus given by:

$$\begin{aligned} &c_{12}^1 = 0, && c_{12}^2 = 1, && c_{12}^3 = 0, && c_{12}^4 = 0, \\ &c_{13}^1 = 0, && c_{13}^2 = 0, && c_{13}^3 = -1, && c_{13}^4 = 0, \\ &c_{14}^1 = 0, && c_{14}^2 = 0, && c_{14}^3 = 0, && c_{14}^4 = 0, \\ &c_{23}^1 = 1, && c_{23}^2 = 0, && c_{23}^3 = 0, && c_{23}^4 = -1, \\ &c_{24}^1 = 0, && c_{24}^2 = 1, && c_{24}^3 = 0, && c_{24}^4 = 0, \\ &c_{34}^1 = 0, && c_{34}^2 = 0, && c_{34}^3 = -1, && c_{34}^4 = 0, \end{aligned}$$

In more detail, $[e_1{,}e_2] = 0e_1 + 1e_2 + 0e_3 + 0e_4$, so the four corresponding structure constants (the first row) are the coefficients 0, 1, 0, and 0. △

As a shortcut, the Lie bracket of elementary matrices may be computed as:

$$[E_{ij}{,}E_{k\ell}] = \delta_{jk}E_{i\ell} - \delta_{\ell i}E_{kj}$$

20.1.2 Homomorphisms

Let L_1 and L_2 be Lie algebras over a field K. A function $\phi : L_1 \to L_2$ is a *Lie algebra homomorphism* if:

(*i*) ϕ is K-linear;

(*ii*) $\forall x,y \in L_1, \phi\big([x,y]_{L_1}\big) = \big[\phi(x),\phi(y)\big]_{L_2}$.

That is, a Lie algebra homomorphism is a homomorphism of the underlying vector space that preserves the Lie bracket.

Given a basis $\{e_i\}_{i=1}^n$ of L_1, it follows from the previous formula $[a,b] = \sum_{i,j} \alpha_i\beta_j[e_i,e_j]$ that, to show property (*ii*), it suffices to verify that ϕ preserves the Lie brackets $[e_i,e_j]$ of the basis elements.

If ϕ is furthermore bijective (or equivalently, invertible), then ϕ is a *Lie algebra isomorphism.* If there exists a Lie algebra isomorphism between L_1 and L_2, we say that L_1 and L_2 are isomorphic, and denote this relation by $L_1 \cong L_2$.

Lemma 20.1.3. *Let L_1 and L_2 be Lie algebras over a field K. Then, $L_1 \cong L_2$ if and only if there exist bases $\mathcal{B}_1$ of L_1 and $\mathbb{B}_2$ of L_2 such that the structure constants c_{ij}^k of L_1 with respect to $\mathcal{B}_1$ are the same as the structure constants of d_{ij}^k of L_2 with respect to $\mathcal{B}_2$.*

Proof. For the forward implication, let $\mathcal{B}_1 = (e_i)_{i=1}^n$ be a basis of L_1 and let $\phi : L_1 \to L_2$ be a Lie algebra isomorphism. Transport the basis $\mathcal{B}_1$ along ϕ to a basis $\mathcal{B}_2 = (f_i)_{i=1}^n = \big(\phi(e_i)\big)_{i=1}^n$ of L_2.

Then,

$$\begin{aligned}\phi\big([e_i,e_j]\big) &= \phi\left(\sum_k c_{ij}^k e_k\right)\\ &= \sum_k c_{ij}^k \phi(e_k)\\ &= \sum_k c_{ij}^k f_k\end{aligned}$$

We also have

$$\begin{aligned}\phi\big([e_i,e_j]\big) &= \big[\phi(e_i),\phi(e_j)\big]\\ &= [f_i,f_j]\\ &= \sum_k d_{ij}^k f_k\end{aligned}$$

Comparing coefficients, we have $c_{ij}^k = d_{ij}^k$ for all i,j,k.

For the reverse implication, suppose there exist bases $\mathcal{B}_1 = (e_i)_{i=1}^n$ of L_1 and $\mathcal{B}_2 = (f_i)_{i=1}^n$ of L_2 such that $c_{ij}^k = d_{ij}^k$ for all i,j,k.

Define a linear map $\phi : L_1 \to L_2$ on basis elements by $e_i \mapsto f_i$ and linearly extending. This is a K-linear isomorphism since $\mathcal{B}_1$ and $\mathcal{B}_2$ are bases. It remains to check that ϕ is a Lie algebra homomorphism.

$$\begin{aligned}\phi\big([e_i,e_j]\big) &= \phi\left(\sum_k c_{ij}^k e_k\right)\\ &= \sum_k c_{ij}^k \phi(e_k)\\ &= \sum_k c_{ij}^k f_k\end{aligned}$$

$$\begin{aligned}&= \sum_k d_{ij}^k f_k \\ &= [f_i,f_j] \\ &= [\phi(e_i),\phi(e_j)]\end{aligned}$$

■

Example. For any Lie algebra L, the identity map $\mathrm{id}_L : L \to L$ is trivially a Lie algebra isomorphism. △

Example. For any field K, the trace map $\mathrm{tr} : \mathfrak{gl}_n(K) \to K$ is a Lie algebra homomorphism, where K is equipped with the identically zero Lie bracket (i.e. is abelian):

(i) trace is linear;

(ii) for all $A,B \in \mathfrak{gl}_n(K)$, we have $\mathrm{tr}\big([A,B]\big) = 0$ by basic properties of the trace, while $[\mathrm{tr}(A), \mathrm{tr}(B)] = 0$ since K is abelian. So the trace preserves the Lie bracket.

△

20.1.3 Subalgebras

Let L be a Lie algebra. A *Lie subalgebra* K of L is a subset $K \subseteq L$ such that:

(i) K is a linear subspace of L;

(ii) K is closed under the Lie bracket: $\forall a,b \in K, [a,b] \in K$.

That is, K is a subset of L that is also a Lie algebra under (the restriction of) the Lie bracket of L.

Example. $\mathfrak{sl}_n(K)$ (zero-trace matrices) is a Lie subalgebra of $\mathfrak{gl}_n(K)$ (all matrices). Similarly, $\mathfrak{b}_n(K)$ (upper triangular) and $\mathfrak{u}_n(K)$ (strictly upper triangular) are subalgebras of $\mathfrak{gl}_n(K)$.

Moreover, any strictly upper triangular matrix has zero trace, so $\mathfrak{u}_n(K) \subseteq \mathfrak{sl}_n(K)$, and every strictly upper triangular matrix is upper triangular, so also $\mathfrak{u}_n(K) \subseteq \mathfrak{b}_n(K)$. △

Example. Consider the space $\langle e_2,e_3\rangle$ in $\mathfrak{gl}_2(K)$. This is a linear subspace of $\mathfrak{gl}_2(K)$, but is not a Lie subalgebra, since it is not closed under the Lie bracket: $[e_2,e_3] = e_1 - e_4 \notin \langle e_2,e_3\rangle$. △

20.1.4 Ideals

Let L be a Lie algebra. An *ideal* I of L is a subset $I \subseteq L$ such that:

(i) I is a linear subspace of L;

(ii) I absorbs Lie brackets with any element of L: $\forall x \in L\, \forall i \in I, [x,i] \in I$.

Clearly, every ideal is a subalgebra, but the converse generally fails. Also, unlike for rings, there is no distinction between left, right, and two-sided ideals, since if $[x,i] \in I$, then $[i,x] = -[x,i] \in I$, as I is a linear subspace.

Example.

(i) $\mathfrak{sl}_n(K)$ is an ideal of $\mathfrak{gl}_n(K)$, since the trace of a Lie bracket is always zero.

(ii) Neither $\mathfrak{b}_n(K)$ nor $\mathfrak{u}_n(K)$ are ideals of $\mathfrak{gl}_n(K)$:

Let M_{ij} be the elementary matrix with $m_{ij} = 1$ and zero elsewhere.

For the former, let $A = M_{21} \in \mathfrak{gl}_n(K)$ and $B = M_{11} \in \mathfrak{b}_n(K)$. Then, $[A,B] = M_{21}M_{11} - M_{11}M_{21} = M_{21} - \mathbf{0} = M_{21}$ is not upper triangular.

For the latter, let $A = M_{21} \in \mathfrak{gl}_n(K)$ and $B = M_{112} \in \mathfrak{u}_n(K)$. Then, $[A,B] = M_{21}M_{12} - M_{12}M_{21} = M_{11} - M_{22}$ is diagonal, and not strictly upper triangular.

(iii) The previous counterexample for $\mathfrak{u}_n(K)$ also shows that $\mathfrak{u}_n(K)$ is not an ideal of $\mathfrak{sl}_n(K)$, since M_{21} in particular has zero trace.

(iv) $\mathfrak{u}_n(K)$ is an ideal of $\mathfrak{b}_n(K)$ since if $A \in \mathfrak{u}_n(K)$ and $B \in \mathfrak{b}_n(K)$, then the diagonals of AB and BA are zero, so $[A,B] \in \mathfrak{u}_n(K)$.

$\triangle$

Lemma 20.1.4. *For any Lie algebra L,*

(i) *L is an ideal of L;*

(ii) *$\{0_L\}$ is an ideal of L;*

(iii) *The centre $Z(L) = \{z \in L : \forall x \in L, [z,x] = 0\}$ is an ideal of L.*

Proof.

(i) This is trivial since the Lie bracket is closed on L by definition.

(ii) $\{0_L\}$ is a linear subspace of L, and for any $x \in L$, $[x,0] = [x,0+0] = [x,0] + [x,0]$, so $[x,0] = 0$.

(iii) The centre is a linear subspace of L since if $\alpha \in K$: $0_L \in Z(L)$ since $[0,x] = 0$; if $z_1,z_2 \in Z(L)$, then $[z_1 + z_2,x] = [z_1,x] + [z_2,x] = 0 + 0 = 0$; and if $z \in Z(L)$ and $\lambda \in K$, $[\lambda z,x] = \lambda[z,x] = \lambda 0 = 0$.

Now, if $z \in Z(L)$ and $x \in L$, $[z,x] = 0$ so $[z,x] \in Z(L)$, as required.

$\blacksquare$

Lemma 20.1.5. *Let $\phi : L_1 \to L_2$ be a Lie algebra homomorphism. Then,*

(i) *$\mathrm{im}(\phi)$ is a Lie subalgebra of L_2;*

(ii) *$\ker(\phi)$ is an ideal of L_1.*

Proof. From basic linear algebra, $\mathrm{im}(\phi)$ is a subspace of L_2 and $\ker(\phi)$ is a subspace of L_1.

(i) For any $x,y \in \mathrm{im}(\phi)$, there exist $x',y' \in L_2$ such that $x = \phi(x')$ and $y = \phi(y')$. Then, $[x,y] = \big[\phi(x'),\phi(y')\big] = \phi\big([x',y']\big) \in \mathrm{im}(\phi)$, so $\mathrm{im}(\phi)$ is closed under the Lie bracket and is hence a Lie subalgebra.

(ii) For any $z \in L_1$ and $x \in \ker(\phi)$, $\phi\big([z,x]\big) = \big[\phi(z),\phi(x)\big] = \big[\phi(z),0\big] = 0$, so $[z,x] \in \ker(\phi)$, as required.

$\blacksquare$

Example. Consider the Lie algebra homomorphism $\mathrm{tr} : \mathfrak{gl}_n(K) \to K$. The image is all of K, since for $\alpha \in K$, $\alpha E_{1,1}$ has trace α; and the kernel is, by definition, $\mathfrak{sl}_n(K)$. So, by the previous lemma, $\mathfrak{sl}_n(K)$ is an ideal of $\mathfrak{gl}_n(K)$ (as we have already verified before). $\triangle$

Let L be a Lie algebra and $I,J \subseteq L$ be ideals of L. We define the *ideal sum* as the pointwise sum:

$$I + J \coloneqq \{i + j : i \in I, j \in J\}$$

and the *ideal Lie bracket* as the subspace generated by all of the Lie brackets:

$$[I,J] \coloneqq \big\langle [i,j] : i \in I, j \in J \big\rangle$$

Lemma 20.1.6. *Let L be a Lie algebra and $I,J \subseteq L$ be ideals of L. Then, $I \cap J$, $I + J$, and $[I,J]$ are ideals of L.*

Proof. From basic linear algebra, the three sets are linear subspaces of L.

(*i*) If $x \in L$ and $i \in I \cap J$, then $[x,i] \in I$ since $i \in I$, and $[x,i] \in J$ since $i \in J$. So $[x,i] \in I \cap J$.

(*ii*) If $x \in L$ and $i \in I+J$, then $i = i' + j'$ for some $i' \in I, j' \in J$. Then, $[x,i] = [x,i'+j'] = [x,i'] + [x,j']$. Since I is an ideal, $[x,i'] \in I$, and similarly, $[x,j'] \in J$. So $[x,i] \in I + J$.

(*iii*) If $x \in L$ and $i \in [I,J]$, then i is some linear combination of Lie brackets $[i',j']$, so

$$\begin{aligned}[x,i] &= \left[x, \sum_{k=1}^{n} c_k [i_k,j_k]\right] \\ &= \sum_{k=1}^{n} c_k \big[x,[i_k,j_k]\big]\end{aligned}$$

It remains to show that the $\big[x,[i_k,j_k]\big]$ are in $[I,J]$. By the Jacobi identity, for any $x \in L$, $i \in I$, and $j \in J$,

$$\begin{aligned}\big[x,[i,j]\big] &= -\big[j,[x,i]\big] - \big[i,[j,x]\big] \\ &= \big[[x,i],j\big] - \big[i,[j,x]\big]\end{aligned}$$

Since I is an ideal, $[x,i] \in I$, and similarly, $[j,x] \in J$. So $\big[x,[i,j]\big] \in [I,J]$.

■

Example. For any Lie algebra L, $[L,L]$ is an ideal of L called the *derived subalgebra* of L. △

20.1.5 Adjoint Homomorphism

Let L be a Lie algebra. In particular, L is a vector space, so we may also consider the Lie algebra $\mathfrak{gl}_n(L)$ of K-linear maps $L \to L$. We define the *adjoint homomorphism* $\operatorname{ad} : L \to \mathfrak{gl}_n(L)$ by

$$\operatorname{ad}(x) = [x,-]$$

Note that $\operatorname{ad}(x)$ is an element of $\mathfrak{gl}_n(L)$ since $\operatorname{ad}(x)(y) = [x,y] \in L$, and $\operatorname{ad}(x)$ is linear:

$$\begin{aligned}\operatorname{ad}(x)(\lambda y_1 + \mu y_2) &= [x,\lambda y_1 + \mu y_2] \\ &= \lambda[x,y_1] + \mu[x,y_2] \\ &= \lambda \operatorname{ad}(x)(y_1) + \mu \operatorname{ad}(x)(y_2)\end{aligned}$$

Lemma 20.1.7. *The adjoint homomorphism is a Lie algebra homomorphism.*

Proof. First, ad is linear:

$$\begin{aligned}\operatorname{ad}(\alpha x + \beta y)(z) &= [\alpha x + \beta y,z] \\ &= \alpha[x,z] + \beta[y,z] \\ &= \big(\alpha \operatorname{ad}(x) + \beta \operatorname{ad}(y)\big)(z)\end{aligned}$$

and ad also preserves the Lie bracket:

$$\operatorname{ad}\big([x,y]\big)(z) = \big[[x,y],z\big]$$

$$\begin{aligned}
&= -\big[z,[x,y]\big] \\
&= \big[x,[y,z]\big] + \big[y,[z,x]\big] \\
&= \big[x,[y,z]\big] - \big[y,[x,z]\big] \\
&= \mathrm{ad}(x)\big(\mathrm{ad}(y)(z)\big) - \mathrm{ad}(y)\big(\mathrm{ad}(x)(z)\big) \\
&= \big(\mathrm{ad}(x) \circ \mathrm{ad}(y) - \mathrm{ad}(y) \circ \mathrm{ad}(x)\big)(z) \\
&= \big[\mathrm{ad}(x), \mathrm{ad}(y)\big](z)
\end{aligned}$$

■

Lemma 20.1.8. *The kernel of the adjoint homomorphism is the centre of L:*

$$\ker(\mathrm{ad}) = Z(L)$$

Proof. By definition, the centre is the collection of all $z \in L$ such that $[z,x] = 0$ for all $x \in L$. That is, all the $z \in L$ such that $\mathrm{ad}(z)$ is the zero map, which is precisely the kernel of ad. ■

20.1.6 Quotient Algebras

The next standard algebraic construction is the quotient. As with rings, we will only obtain a Lie algebra when quotienting by an ideal.

Let V be a vector space, and $W \subseteq V$ a linear subspace. Given $v \in V$, we define the associated *coset* of W in V by:

$$v + W := \{v + w : w \in W\}$$

Lemma 20.1.9.

(*i*) *$x + W = y + W$ if and only if $x - y \in W$;*

(*ii*) *Any two cosets are either disjoint or equal.*

Proof.

(*i*) For forward implication, suppose $x+W = y+W$. Since $x = x+0 \in x+W = y+W$, $x = y+w$ for some $w \in W$. So $x - y = w \in W$. For the reverse implication, suppose $x - y = w_0 \in W$. Then, for any $x + w \in x + W$, $x + w = (y + w_0) + w = y + (w_0 + w) \in y + W$, so $x + W \subseteq y + W$. Similarly, for any $y + w \in y + W$, $y + w = (x - w_0) + w = x + (w - w_0) \in x + W$, so $y + W \subseteq x + W$. So $x + W = y + W$.

(*ii*) If the cosets are not disjoint, then there exists $z \in (x + W) \cap (y + W)$, so $y = x + w_1 = y + w_2$. Then, $x - y = w_2 - w_1 \in W$, so by the previous part, $x + W = y + W$.

■

We define the quotient V/W to be the set of cosets of V in W:

$$V/W := \{v + W : v \in V, w \in W\}$$

We define an addition $\oplus$ on V/W by:

$$(x + W) \oplus (y + W) := (x + y) + W$$

and a scalar multiplication by:

$$\lambda(x + W) := (\lambda x) + W$$

These are well-defined, as different representatives of the cosets differ only by an element of W, which is absorbed into the result. Under these operations, V/W inherits a vector space structure, with zero element $0_{V/W} = 0_V + W = W$.

Let L be a Lie algebra and I an ideal of L. We have that V/I is a vector space, but we can endow it with a bracket operation as follows:

$$[x + I,y + I] \coloneqq [x,y] + I$$

This is well-defined since, if $x + I = x' + I$ and $y + I = y' + I$, then $x = x' + i_1$ and $y = y' + i_2$, and:

$$\begin{aligned}[x + I,y + I] &= [x,y] + I\\ &= [x' + i_1,y' + i_2] + I\\ &= [x',y'] + [x',i_2] + [i_1,x'] + [i_1,i_2] + I\end{aligned}$$

Since I is an ideal, the three bracket on the right are absorbed into I:

$$\begin{aligned}&= [x',y'] + I\\ &= [x' + I,y + I]\end{aligned}$$

Theorem 20.1.10. *This operation defines a Lie bracket on L/I.*

Proof.

(*i*) The bracket in L is bilinear, so the bracket in L/I is defined on representatives and thus inherits bilinearity:

$$\begin{aligned}\big[(\alpha x + I) \oplus (\beta y + I),z + I\big] &= \big[(\alpha x + \beta y) + I,z + I\big]\\ &= [\alpha x + \beta y,z] + I\\ &= \big(\alpha[x,z] + \beta[y,z]\big) + I\\ &= \alpha\big([x,z] + I\big) \oplus \beta\big([y,z] + I\big)\\ &= \alpha[x + I,z + I] \oplus \beta[y + I,z + I]\end{aligned}$$

and similarly in the second argument.

(*ii*) The bracket similarly inherits the alternating property:

$$\begin{aligned}[x + I,x + I] &= [x,x] + I\\ &= 0_L + I\\ &= 0_{L/I}\end{aligned}$$

(*iii*) The Jacobi identity also descends from the Lie bracket on L:

$$\begin{aligned}\big[x + I,[y + I,z + I]\big] &= [x,[y,z]] + I\\ \big[z + I,[x + I,y + I]\big] &= [z,[x,y]] + I\\ \big[y + I,[z + I,x + I]\big] &= [y,[z,x]] + I\end{aligned}$$

By the Jacobi identity in L, $[x,[y,z]] + [z,[x,y]] + [y,[z,x]] = 0_L$, so the sum of the three terms above reduces to $0_L + I = 0_{L/I}$.

■

Let $\pi : L \to L/I$ be the natural quotient map $x \mapsto x + I$. Then,

$$\begin{aligned}\pi(x+y) &= (x+y)+I\\ &= (x+I)\oplus(y+I)\\ &= \pi(x)\oplus\pi(y)\end{aligned}$$

and

$$\begin{aligned}\pi(\lambda x) &= (\lambda x)+I\\ &= \lambda(x+I)\\ &= \lambda\pi(x)\end{aligned}$$

so π is linear; *pi* also preserves the Lie bracket:

$$\begin{aligned}\pi\big([x,y]\big) &= [x,y]+I\\ &= [x+I,y+I]\\ &= \big[\pi(x),\pi(y)\big]\end{aligned}$$

so π is a Lie algebra homomorphism.

Theorem 20.1.11 (First Isomorphism Theorem). *Let $\phi : L_1 \to L_2$ be a Lie algebra homomorphism. Then,*

(*i*) $\mathrm{im}(\phi)$ *is a Lie subalgebra of* L_2*;*

(*ii*) $\ker(\phi)$ *is an ideal of* L_1*;*

(*iii*) $L_1/\ker(\phi) \cong \mathrm{im}(\phi)$

Proof. Parts (*i*) and (*ii*) were proved in Lemma 20.1.5.

For (*iii*), let $I = \ker(\phi)$ and define the map $f : L_1/I \to L_2$ by $f(x+I) = \phi(x)$. This is well-defined since if $x + I = y + I$, then $x - y \in I = \ker(\phi)$

$$\begin{aligned}f(x+I) &= \phi(x)\\ &= \phi(x-y+y)\\ &= \phi(x-y)+\phi(y)\\ &= 0+\phi(y)\\ &= \phi(y)\\ &= f(y+I)\end{aligned}$$

f is also linear, since

$$\begin{aligned}f\big((x+I)\oplus(y+I)\big) &= f\big((x+y)+I\big)\\ &= \phi(x+y)\\ &= \phi(x)+\phi(y)\\ &= f(x+I)+f(y+I)\end{aligned}$$

and

$$\begin{aligned}f\big(\lambda(x+I)\big) &= f\big((\lambda x)+I\big)\\ &= \phi(\lambda x)\\ &= \lambda\phi(x)\end{aligned}$$

$$= \lambda f(x+I)$$

f also preserves the Lie bracket:

$$\begin{aligned} f\big([x+I,y+I]\big) &= f\big([x,y]+I\big) \\ &= \phi\big([x,y]\big) \\ &= \big[\phi(x),\phi(y)\big] \\ &= \big[f(x+I),f(y+I)\big] \end{aligned}$$

So f is a Lie algebra homomorphism.

Furthermore, f surjects onto the image of ϕ, since for any $\phi(x) \in \operatorname{im} f$, $f(x+I) = \phi(x)$, and f is injective since

$$\begin{aligned} \ker(f) &= \{x+I \in L/I : f(x+I) = 0\} \\ &= \{x+I \in L/I : \phi(x) = 0\} \\ &= \{x+I \in L/I : x \in \ker(\phi)\} \\ &= \{x+I \in L/I : x \in I\} \\ &= \{I\} \\ &= \{0_{L/I}\} \end{aligned}$$

So f witnesses the isomorphism $L_1/\ker(\phi) \cong \operatorname{im}(\phi)$. ■

Example. △

The other standard isomorphism theorems also hold for Lie algebras, their proofs being similar to the corresponding proofs for rings:

Theorem 20.1.12 (Second Isomorphism Theorem)**.** *Let L be a Lie algebra, and $I,J \subseteq L$ be ideals of L. Then,*

$$\frac{I+J}{J} \cong \frac{I}{I\cap J}$$

Theorem 20.1.13 (Third Isomorphism Theorem)**.** *Let L be a Lie algebra, and $I,J \subseteq L$ be ideals of L. Then, J/I is an ideal of L/I, and,*

$$\frac{L/I}{J/I} \cong L/J$$

Theorem 20.1.14 (Correspondence Theorem)**.** *Let L be a Lie algebra, and $I \subseteq L$ be an ideal of L. Then, there is a bijection*

$$\{J : J \subseteq I \textit{ is an ideal of } L\} \cong \{K : K \textit{ is an ideal of } L/I\}$$

20.1.7 Direct Sums

Let L_1 and L_2 be Lie algebras, and consider the cartesian product of the underlying sets:

$$L_1 \times L_2 = \big\{(x,y) : x \in L_1, y \in L_2\big\}$$

The operations on L_1 and L_2 naturally descend pointwise to this product:

$$\begin{aligned} (x,y) + (x',y') &\coloneqq (x+x',y+y') \\ \lambda(x,y) &\coloneqq (\lambda x,\lambda y) \\ \big[(x,y),(x',y')\big] &\coloneqq \big([x,y],[x',y']\big) \end{aligned}$$

Under these operations, this set is a Lie algebra, denoted by $L_1 \oplus L_2$ called the *direct sum* of L_1 and L_2.

Lemma 20.1.15. *Let L_1 and L_2 be Lie algebras. Then,*

(i) $[L_1 \oplus L_2, L_1 \oplus L_2] = [L_1, L_1] \oplus [L_2, L_2]$*;*

(ii) $Z(L_1 \oplus L_2) = Z(L_1) \oplus Z(L_2)$*;*

(iii) $\{(x,0) : x \in L_1\}$ *is an ideal of* $L_1 \oplus L_2$*, isomorphic to* L_1*;*

(iv) $\{(0,y) : y \in L_2\}$ *is an ideal of* $L_1 \oplus L_2$*, isomorphic to* L_2*;*

(v) The projections $\pi_i : L_1 \oplus L_2 \to L_i$ *are Lie algebra homomorphisms.*

This algebra $L_1 \oplus L_2$ is also called the *external* direct sum, since we have formed a new algebra from two unrelated algebras L_1 and L_2. In constrast, the *internal* direct sum is defined as follows:

Let $L_1, L_2 \subseteq L$ be subalgebras of a Lie algebra L such that:

(i) $L_1 \cap L_2 = \{0_L\}$;

(ii) $[L_1, L_2] = \{0_L\}$.

Then, the linear subspace $L_1 + L_2$ is naturally a Lie subalgebra of L since

$$\begin{aligned}[x + y, x' + y'] &= [x,x'] + [x,y'] + [x',y] + [y,y'] \\ &= [x,x'] + [y,y'] \\ &\in L_1 + L_2\end{aligned}$$

Lemma 20.1.16. *The internal direct sum* $L_1 + L_2$ *is isomorphic to the external direct sum* $L_1 \oplus L_2$*.*

Proof. Define the map $\phi : L_1 \oplus L_2 \to L_1 + L_2$ by $(x,y) \mapsto x + y$. Linearity is clear, and for Lie brackets, we have:

$$\begin{aligned}\phi\big([(x,y),(x',y')]\big) &= \phi\big([x,y],[x',y']\big) \\ &= [x,y] + [x',y']\end{aligned}$$

and

$$\begin{aligned}\big[\phi(x,y), \phi(x',y')\big] &= [x + y, x' + y'] \\ &= [x,x'] + [x,y'] + [x',y] + [y,y']\end{aligned}$$

Since $[L_1, L_2] = \{0_L\}$, the mixed brackets vanish and the two expressions are equal.

Then, ϕ is injective since if $\phi(x,y) = x + y = 0$, then $x = -y \in L_1 \cap L_2 = \{0_L\}$, so $x = y = 0$, and ϕ has trivial kernel. We also have that ϕ is surjective, since every element of $x + y \in L_1 + L_2$ has preimage $(x,y) \in L_1 \oplus L_2$. So ϕ is an isomorphism. ■

20.2 Representations

Let L be a Lie algebra over K. A *representation* of L is a Lie algebra homomorphism

$$\phi : L \to \mathfrak{gl}(V)$$

where V is a vector space over K. If $\ker(\phi)$ is trivial, then ϕ is called *faithful.*

Example.

(i) Every matrix Lie algebra is "really" a faithful representation of the underlying abstract Lie algebra. For instance, the abstract Lie algebra $\mathfrak{sl}_2(\mathbb{C})$ has basis elements e,h,f and Lie brackets $[e,h] = -2e$, $[e,f] = h$, and $[f,h] = 2f$, which we often represent as matrices.

(ii) If L is a Lie subalgebra of $\mathfrak{gl}(V)$, then the inclusion $\iota : L \to \mathfrak{gl}(V)$ is a representation called the *natural representation* of L.

(iii) The zero homomorphism $\phi : L \to \mathfrak{gl}(V)$ is the *trivial representation* of L.

(iv) The adjoint homomorphism $\operatorname{ad} : L \to \mathfrak{gl}(L)$ is a representation called the *adjoint representation* of L. This representation is faithful if and only if $Z(L) = \{0_L\}$.

△

20.3 Soluble and Nilpotent Lie Algebras

20.3.1 Solubility

Lemma 20.3.1. *Let L be a Lie algebra and I an ideal of L. Then, L/I is abelian if and only if $[L,L] \subseteq I$.*

Proof. By definition, L/I is abelian if $[x+I,y+I] = 0_{L/I}$ for all $x,y \in L$, or equivalently, $[x,y] + I = 0_{L/I} = I$. This holds if and only if $[x,y] \in I$, and since $[L,L] = \big\langle [x,y] : x,y \in L \big\rangle$, this is equivalent to $[L,L] \subseteq I$. ■

Corollary 20.3.1.1. *The ideal $I = [L,L]$ is the smallest ideal of L such that L/I is abelian.*

The *derived series* of L is the sequence $L^{(0)},L^{(1)},L^{(2)},\ldots$, defined inductively as follows:

(i) $L^{(0)} = L$;

(ii) $L^{(k+1)} = \big[L^{(k)},L^{(k)}\big]$.

Lemma 20.3.2. *For any $k \in \mathbb{N}$, $L^{(k)}$ is an ideal of L, and $L^{(0)} \supseteq L^{(1)} \supseteq L^{(2)} \supseteq \cdots$.*

Proof. We have already seen that $[L,L]$ is an ideal of L, so the chain of containments follows by induction. The Lie bracket of ideals is also an ideal, so $L^{(k+1)} = \big[L^{(k)},L^{(k)}\big]$ is an ideal of $L^{(0)} = L$ by induction. ■

A Lie algebra L is *soluble* if there exists $n \in \mathbb{N}$ such that

$$L^{(n)} = \{0_L\}$$

Example.

(i) If L is abelian, then L is soluble. This is immediate since $L^{(1)} = [L,L] = \big\langle [x,y] : x,y \in L \big\rangle = \langle 0 \rangle = \{0\}$.

(ii) $L = \mathfrak{b}_n(\mathbb{C})$ is soluble for all $n \in \mathbb{N}$. The Lie bracket of any two upper tringular matrices is strictly upper triangular. Continuing to take Lie brackets, the matrices gain an additional zero diagonal at each step, and thus eventually all become the zero matrix after at most n iterations. So $L^{(n)} = \{0\}$, and $\mathfrak{b}_n(\mathbb{C})$ is soluble.

(iii) Let $L = \mathfrak{gl}_n(\mathbb{C})$. Since $\operatorname{tr}(AB) = \operatorname{tr}(BA)$, $[A,B] \in \mathfrak{sl}_n(\mathbb{C})$, so $L^{(1)} \subseteq \mathfrak{sl}_n(\mathbb{C})$. Conversely, $\mathfrak{sl}_n(\mathbb{C})$ is generated by the brackets $[E_{ij},E_{jk}] = E_{ik}$ (the off-diagonal elements) and $[E_{ij},E_{ji}] = E_{ii} - E_{jj}$ (the diagonal elements preserving zero trace), so $\mathfrak{sl}_n(C) \subseteq L^{(1)}$.

By the same argument, $[\mathfrak{sl}_n(\mathbb{C}),\mathfrak{sl}_n(\mathbb{C})] = \mathfrak{sl}_n(\mathbb{C})$, so $L^{(k)} = L^{(1)} = \mathfrak{sl}_n(\mathbb{C})$ for all $k \in \mathbb{N}$, and $\mathfrak{sl}_n(\mathbb{C}) \neq \{0\}$ for $n \geq 2$. Thus, $\mathfrak{gl}_n(\mathbb{C})$ and $\mathfrak{sl}_n(\mathbb{C})$ are not soluble for $n \geq 2$.

△

Lemma 20.3.3. *Let $\phi : L \to L'$ be a Lie algebra homomorphism. Then, for all $n \in \mathbb{N}$,*

$$\phi(L^{(n)}) = \phi(L)^{(n)}$$

Proof. We induct on n. For $n = 0$, $\phi(L^{(0)}) = \phi(L) = \phi(L)^{(0)}$. Now suppose the result holds for some fixed arbitrary n. Then,

$$\begin{aligned}\phi(L^{(n+1)}) &= \phi\big([L^{(n)},L^{(n)}]\big)\\ &= \big[\phi(L^{(n)}),\phi(L^{(n)})\big]\\ &= [\phi(L)^{(n)},\phi(L)^{(n)}]\\ &= \phi(L)^{(n+1)}\end{aligned}$$

■

Lemma 20.3.4.

(*i*) *If L is soluble, then every Lie subalgebra of L is soluble.*

(*ii*) *If L is soluble, then every homomorphic image of L is soluble.*

(*iii*) *If I is an ideal of L such that L/I and I are soluble, then L is soluble.*

(*iv*) *If I and J are soluble ideals of L, then $I + J$ is also soluble.*

Proof.

(*i*) If L is a Lie subalgebra of L, then $M^{(i)} \subseteq L^{(i)}$ for all $i \in \mathbb{N}$, so if L is soluble, there exists $n \in \mathbb{N}$ such that $L^{(n)}$ is trivial, so $M^{(n)} \subseteq L^{(n)}$ must also be trivial.

(*ii*) Let $\phi : L \to M$ be a Lie algebra homomorphism. Since $\phi(L^{(i)}) = \phi(L)^{(i)}$. Since L is soluble, $L^{(n)} = \{0_L\}$ for some $n \in \mathbb{N}$, so $\phi(L)^{(n)} = \phi(L^{(n)}) = \phi\big(\{0_L\}\big) = \{0_M\}$ is also soluble.

(*iii*) Since L/I is soluble, $(L/I)^{(n)} = \{0\}$ for some $n \in \mathbb{N}$. Let $\pi : L \to L/I$ be the natural quotient homomorphism. Then, $\{0\} = (L/I)^{(n)} = \pi(L)^{(n)} = \pi(L^{(n)})$, so $L^{(n)} \subseteq \ker(\pi) = I$. Then, I is soluble, so $I^{(m)} = \{0\}$ for some $m \in \mathbb{N}$. So $(L^{(n)})^{(m)} \subseteq \{0\}$, and L is soluble.

(*iv*) By the, second isomorphism theorem,

$$\frac{I + J}{J} \cong \frac{I}{I \cap J}$$

Since $I/(I \cap J) = \pi(I)$, it is soluble by (*ii*). So $(I + J)/J$ is also soluble, and hence $I + J$ is soluble by (*iii*).

■

20.3.2 Simple and Semisimple Lie Algebras

A non-abelian Lie algebra L is *simple* if it has no proper non-zero ideals. That is, the only ideals of L are $\{0\}$ and L.

Example.

(*i*) $\mathfrak{sl}_n(\mathbb{C})$ is an proper non-zero ideal of $\mathfrak{gl}_n(\mathbb{C})$ for $n \geq 2$, so $\mathfrak{gl}_n(\mathbb{C})$ is not simple for $n \geq 2$. Also, for $n = 1$, $\mathfrak{gl}_1(\mathbb{C}) \cong \mathbb{C}$ is 1-dimensional over $\mathbb{C}$ and is hence abelian (and thus also non-simple).

(*ii*) $\mathfrak{sl}_2(\mathbb{C})$ is simple.

Suppose otherwise that $\mathfrak{sl}_2(\mathbb{C})$ has a non-zero proper ideal I. Let

$$e_1 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad e_2 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \qquad e_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

be a basis of $\mathfrak{sl}_2(\mathbb{C})$, with brackets $[e_1,e_2] = e_3$, $[e_1,e_3] = -2e_1$, and $[e_2,e_3] = 2e_2$.

Suppose $e_1 \in I$. Then, $[e_1,e_2] = e_3 \in I$, and $[e_2,e_3] = 2e_2 \in I$, so $e_2 \in I$. So $I = \mathfrak{sl}_2(\mathbb{C})$. Through similar considerations, if $e_2 \in I$, then necessarily $e_1,e_3 \in I$, and if $e_3 \in I$, then $e_1,e_2 \in I$. So any non-zero ideal must be equal to $\mathfrak{sl}_2(\mathbb{C})$, so $\mathfrak{sl}_2(\mathbb{C})$ is simple.

△

Let L be a Lie algebra, and consider the set $\mathcal{R}$ of soluble ideals of L:

$$\mathcal{R} = \{I : I \text{ is a soluble ideal of } L\}$$

Let $R \in \mathcal{R}$ be an ideal of maximum dimension, which exists as L is finite-dimensional. For any $I \in \mathcal{R}$, $I + R$ is a soluble ideal of L, and since $R \subseteq I + R$, $\dim(R) \leq \dim(I + R)$. But the dimension of R is maximal, so $\dim(I + R) \leq \dim(R)$, so $\dim(I + R) = \dim(R)$, which holds if and only if $I \subseteq R$. So R contains every soluble ideal of L.

Moreover, R is unique, since if $R' \in \mathcal{R}$ is another soluble ideal containing every soluble ideal of L, then $R' \subseteq R$, since R' is soluble, and $R \subseteq R'$, since R is soluble.

Thus, for any Lie algebra L, there exists a unique soluble ideal of L that contains every soluble ideal of L. This ideal is denoted by $\operatorname{Rad}(L)$, and is called the *radical* of L. Since the solubility of L implies the solubility of every Lie subalgebra of L, L is soluble if and only if $\operatorname{Rad}(L) = L$.

A Lie algebra L is *semisimple* if $\operatorname{Rad}(L) = \{0_L\}$. That is, if it has no non-zero *soluble* ideals.

Example.

(*i*) $\mathfrak{sl}_2(\mathbb{C})$ is semisimple.

(*ii*) $\mathfrak{gl}_2(\mathbb{C})$ is not semisimple.

△

Lemma 20.3.5. *For any Lie algebra L, $L/\operatorname{Rad}(L)$ is semisimple.*

Proof. Let K be a soluble ideal of $L/\operatorname{Rad}(L)$. By the correspondence theorem, there is a corresponding ideal I of L with $\operatorname{Rad}(L) \subseteq I$ and $I/\operatorname{Rad}(L) = K$. Since both K and $\operatorname{Rad}(L)$ are soluble, so is I. But by the definition of a radical, $I \subseteq \operatorname{Rad}(I)$, so $I = \operatorname{Rad}(L)$. So $K = \{0\}$, as required. ■

Lemma 20.3.6. *If L is a complex simple Lie algebra, then L is semisimple.*

Proof. Suppose otherwise that L is not semisimple, so $\operatorname{Rad}(L) \neq \{0\}$. Since L is simple, it has no proper non-zero ideals, and hence $\operatorname{Rad}(L) = L$. So L is soluble, and $L^{(1)} = [L,L] = \{0\}$ (since L has no proper non-zero ideals, and $[L,L] = L$ contradicts solubility). But then, L is abelian, which contradicts that L is simple. ■

20.3.3 Nilpotent Lie Algebras

The *lower central series* of L is the sequence $L^0,L^1,L^2,\ldots$, defined inductively as follows:

(*i*) $L^0 = L$;

(*ii*) $L^{(k+1)} = \big[L,L^k\big]$.

Lemma 20.3.7. *For any $k \in \mathbb{N}$, L^k is an ideal of L, and $L^0 \subseteq L^1 \subseteq L^2 \subseteq \cdots$.*

Proof. Identical to Lemma 20.3.2. ■

A Lie algebra L is *nilpotent* if there exists $n \in \mathbb{N}$ such that

$$L^n = \{0_L\}$$

The connection to ordinary nilpotency of operators will be made explored later.

Example.

(*i*) If L is abelian, then L is nilpotent.

(*ii*) As seen earlier, if $L = \mathfrak{sl}_n(\mathbb{C})$, $[L,L] = L$, so $L^k = L$ and $\mathfrak{sl}_n(\mathbb{C})$ is not nilpotent.

(*iii*) The Lie algebra $L = \mathfrak{u}_3(\mathbb{C})$, called the *Heisenberg* Lie algebra, is nilpotent. $L = \langle E_{12},E_{13},E_{23}\rangle$, and we have $[E_{12},E_{13}] = 0$, $[E_{12},E_{23}] = E_{13}$, $[E_{13},E_{23}] = 0$. Then, using the structure constants above, $[A,B] = \alpha E_{13}$, so $L^1 = \langle E_{13}\rangle_{\mathbb{C}}$. Then, using the structure constants, we also have that $[A,E_{13}] = 0$ for any $A \in L$, so $L^2 = [L,L^1] = \big[L,\langle E_{13}\rangle\big] = \{0\}$.

More generally, $L = \mathfrak{u}_n(\mathbb{C})$ is nilpotent for all $n \in \mathbb{N}$.

△

Lemma 20.3.8.

(*i*) *If L is nilpotent, then every Lie subalgebra of L is nilpotent.*

(*ii*) *If L is nilpotent and non-trivial, then $Z(L)$ is non-trivial.*

(*iii*) *If $L/Z(L)$ is nilpotent, then L is nilpotent.*

Theorem 20.3.9. *Let L be a Lie algebra. Then, for all $n \in \mathbb{N}$,*

$$L^{(n)} \subseteq L^n$$

Proof. We induct on n. For $n = 0$, $L^{(0)} = L \subseteq L = L^0$. Now, suppose the inclusion holds for some arbitrary fixed n.

Since $L^{(n)} \subseteq L$, every Lie bracket $[x,y] \in [L^{(n)},L^{(n)}]$ also lies in $[L,L^{(n)}]$, so

$$\begin{aligned} L^{(n+1)} &= [L^{(n)},L^{(n)}] \\ &\subseteq [L,L^{(n)}] \\ &\subseteq [L,L^n] \\ &= L^{n+1} \end{aligned}$$

■

Corollary 20.3.9.1. *Every nilpotent Lie algebra is soluble.*

Proof. If L is nilpotent, then $L^n = \{0\}$ for some $n \in \mathbb{N}$. Then, $L^{(n)} \subseteq L^n = \{0\}$, and L is soluble. ■

20.3.4 Weights

Let V be a vector space over a field K. Recall that a non-zero vector $v \in V$ is an eigenvector for a linear map $T : V \to V$ if there exists an eigenvalue $\lambda \in K$ such that $T(v) = \lambda v$.

Let V be a vector space and H be a Lie subalgebra of $\mathfrak{gl}(V)$. A non-zero vector $v \in V$ is an *eigenvector* for H if v is an eigenvector for every $T \in H$.

That is, $v \in V$ is an eigenvector for H if for every $T \in H$, there exists $\lambda_T \in K$ such that $T(v) = \lambda_T v$. This induces a function $\lambda : H \to K$ that sends each transformation $T \in H$ to its eigenvalue:

$$\lambda(T) = \lambda_T$$

So equivalently, a non-zero vector $v \in V$ is an eigenvector for H if there exists a function $\lambda : H \to K$ such that $T(v) = \lambda(T)v$ for all $T \in H$.

Now, given such a function λ, consider the set V_λ of all eigenvectors of H consistent with this function:

$$V_\lambda := \{w \in V : \forall T \in H, T(w) = \lambda(T)w\}$$

By the construction of λ, we have that $v \in V$, so V is non-empty. Then, for all $T \in H$, $\alpha \in K$, and x,$y \in V_\lambda$,

$$\begin{aligned} T(x+y) &= T(x) + T(y) \\ &= \lambda(T)x + \lambda(T)y \\ &= \lambda(T)(x+y) \end{aligned}$$

so $x + y \in V_\lambda$, and

$$T(\alpha x) = \lambda(T)\alpha x$$

so $\alpha x \in V_\lambda$. So, V_λ is closed under vector addition and scaling. Also, $0 \in V_\lambda$ since the zero vector is preserved under any linear transformation T and annihilates any scalar, so V_λ is a linear subspace of V.

Moreover, for any S,$T \in H$, α,$\beta \in K$, and $w \in V_\lambda$,

$$\begin{aligned} \lambda(\alpha T + \beta S)w &= (\alpha T + \beta S)(w) \\ &= \alpha T(w) + \beta S(w) \\ &= \alpha\lambda(T)w + \beta\lambda(S)w \\ &= (\alpha\lambda(T) + \beta\lambda(T))w \end{aligned}$$

so $\lambda : H \to K$ is linear.

Let V be a vector space over a field K and H be a Lie subalgebra of $\mathfrak{gl}(V)$. A *weight* of H is a linear function $\lambda : H \to K$ such that the *weight space* V_λ is a non-trivial linear subspace of V.

Example. Consider $L = \mathfrak{gl}_3(\mathbb{C})$, and let $H = \mathfrak{b}_3(\mathbb{C}) \subseteq L$. A general element $A \in H$ has the form

$$A = \begin{bmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{bmatrix}$$

Then, the basis vector e_1 is an eigenvector of this matrix:

$$\begin{bmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} a \\ 0 \\ 0 \end{bmatrix} = a \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

so e_1 is an eigenvector for H. The associated weight $\lambda : H \to K$ is then given by

$$\lambda\left(\begin{bmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{bmatrix}\right) = a$$

and the weight space V_λ is given by

$$V_\lambda = \{ae_1 : a \in K\} = \langle e_1 \rangle$$

△

Example. Let $L = \mathfrak{sl}_3(\mathbb{C})$, and $H = \langle h_1, h_2 \rangle \subseteq L$, where

$$h_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad h_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

We have:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x \\ -y \\ 0 \end{bmatrix}$$
$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ y \\ -z \end{bmatrix}$$

By inspection,

- e_1 is an eigenvector for
  - h_1 with eigenvalue 1;
  - h_2 with eigenvalue 0;
- e_2 is an eigenvector for
  - h_1 with eigenvalue -1;
  - h_2 with eigenvalue 1;
- e_2 is an eigenvector for
  - h_1 with eigenvalue 0;
  - h_2 with eigenvalue -1.

So, we may define three weights for each eigenvector by mapping h_1 and h_2 to the corresponding eigenvalues:

$$\begin{aligned} \lambda_1(h_1) &= 1; & \lambda_1(h_1) &= -1; & \lambda_1(h_1) &= 0; \\ \lambda_1(h_2) &= 0; & \lambda_1(h_2) &= 1; & \lambda_1(h_2) &= -1; \\ V_{\lambda_1} &= \langle e_1 \rangle & V_{\lambda_2} &= \langle e_2 \rangle & V_{\lambda_3} &= \langle e_3 \rangle \end{aligned}$$

△

Lemma 20.3.10. *Let V be a vector space over a field K of characteristic $\mathrm{char}(K) = 0$, L be a Lie subalgebra of $\mathfrak{gl}(V)$, and I be an ideal of L. Then, for any weight $\lambda : I \to K$, V_λ is L-invariant. That is, for every $T \in L$,*

$$T(V_\lambda) \subseteq V_\lambda$$

20.3.5 Engel's Theorem

Let L be a Lie algebra. An element $x \in L$ is *ad-nilpotent* if there exists $n \in \mathbb{N}$ such that the n-fold iteration of $\mathrm{ad}(x)$ is the zero map:

$$\mathrm{ad}(x)^n = \mathbf{0}$$

That is, $\mathrm{ad}(x)$ is a nilpotent operator in the ordinary sense.

Lemma 20.3.11. *If L is a nilpotent Lie algebra, then every element of L is ad-nilpotent.*

Proof. Since L is nilpotent, there exists $n \in \mathbb{N}$ such that $L^n = \{0\}$, so for every $x,y \in L$, applying the Lie bracket n times yields 0:

$$\mathrm{ad}(x)^n(y) = [\underbrace{x,[x,\cdots[x}_{n},y]\cdots]] = 0$$

■

The question is, does the converse hold?

Lemma 20.3.12. *Let V be a vector space, and L a Lie subalgebra of $\mathfrak{gl}(V)$. If $x \in L$ is nilpotent as a linear map, then it is ad-nilpotent.*

Proof. Since x is nilpotent, there exists n such that $x^n = \mathbf{0}$. For any $y \in L$, consider the $2n$-fold iteration of $\mathrm{ad}(x)$. By induction, we can prove that:

$$\mathrm{ad}(x)^{2n}(y) = \sum_{i=1}^{2n} \alpha_i x^i y x^{2n-i}$$

for some coefficients $\alpha_i \in K$. The factor x^{2n-i} vanishes on the first half of the sum, while x^i vanishes over the second, so the entire sum vanishes, and x is ad-nilpotent. ■

Lemma 20.3.13. *Let V be a vector space, and L be a Lie subalgebra of $\mathfrak{gl}(V)$ such that every element of L is nilpotent. Then, there exists a non-zero vector v such that v on all of L:*

$$\forall x \in L, x(v) = 0$$

Theorem 20.3.14 (Engel's Theorem for Subalgebras of $\mathfrak{gl}(V)$)**.** *Let V be a vector space, and L be a Lie subalgebra of $\mathfrak{gl}(V)$ such that every element of L is nilpotent. Then, there exists a basis of V such that the matrix of every element of L is strictly upper triangular. In particular, L is nilpotent.*

Theorem 20.3.15 (Engel's Theorem)**.** *Let L be a Lie algebra. If every element of L is ad-nilpotent, then L is nilpotent.*

Proof. Consider the map $\mathrm{ad} : L \to \mathrm{ad}(L) \subseteq \mathfrak{gl}(L)$. For every $x \in L$, $\mathrm{ad}(x)$ is nilpotent, so by the previous theorem, $\mathrm{ad}(L)$ is nilpotent. Then, $\mathrm{ad}(L) \cong L/\ker(\mathrm{ad}) = L/Z(L)$ is nilpotent, so L is nilpotent. ■

20.3.6 Lie's Theorem

Lemma 20.3.16. *Let $L = \mathfrak{b}_n(\mathbb{C})$. Then, $[L,L] = \mathfrak{u}_n(\mathbb{C})$.*

Theorem 20.3.17. *Let V be a vector space over $\mathbb{C}$ and L be a Lie subalgebra of $\mathfrak{gl}(V)$. If L is soluble, there exists an eigenvector for L. That is, there exists a non-zero $v \in V$ such that for every $x \in L$, there exists λ_x such that $x(v) = \lambda_x v$.*

Theorem 20.3.18 (Lie's Theorem)**.** *Let V be a vector space over $\mathbb{C}$, and L be a soluble Lie subalgebra of $\mathfrak{gl}(V)$. Then, there exists a basis of V such that the matrix of every element of L is upper triangular.*

Corollary 20.3.18.1. *Let V be a vector space over $\mathbb{C}$, and L be a soluble Lie subalgebra of $\mathfrak{gl}(V)$. If $x \in [L,L]$, then x is nilpotent.*

Corollary 20.3.18.2. *Let L be a Lie algebra over $\mathbb{C}$. Then, L is soluble if and only if $[L,L]$ is nilpotent.*

20.4 The Killing Form and Cartan's Criteria

20.4.1 Jordan Decomposition

In this section, V is a vector space over $\mathbb{C}$.

Let $x : V \to V$ be a linear map. Then, there exist linear maps $d : V \to V$ and $n : V \to V$ such that

(*i*) $x = d + n$;

(*ii*) d is diagonalisable and n is nilpotent;

(*iii*) $dn = nd$.

Such a decomposition $x = d + n$ is called a *Jordan decomposition* of x.

Lemma 20.4.1. *The Jordan decomposition of $x : V \to V$ is unique. Moreover, there exist polynomials $p,q \in \mathbb{C}[t]$ without constant terms such that $p(x) = d$ and $q(x) = n$.*

Let $\mathcal{B}$ be a basis of V such that the matrix $D = [d]_{\mathcal{B}}$ of d in $\mathcal{B}$ is diagonal. Let $\overline{d}$ be the linear map whose matrix with respect to $\mathcal{B}$ is the complex conjugate $\overline{D}$ of D. Then, there exists $\tilde{p} \in \mathbb{C}[t]$ such that $\tilde{p}(x) = \overline{d}$.

Lemma 20.4.2. *Let $x \in \mathfrak{gl}(V)$. If $x = d + n$ is its Jordan decomposition, then*

$$\operatorname{ad}(x) = \operatorname{ad}(d) + \operatorname{ad}(n)$$

Lemma 20.4.3. *For any $A,B,C \in \mathfrak{gl}(V)$,*

$$\operatorname{tr}\big([A,B]C\big) = \operatorname{tr}\big(A[B,C]\big)$$

Theorem 20.4.4. *Let L be a Lie subalgebra of $\mathfrak{gl}(V)$. If $\operatorname{tr}(x \circ y) = 0$ for all $x,y \in L$, then L is soluble.*

Corollary 20.4.4.1. *Let L be a complex Lie algebra. Then, L is soluble if and only if for all $x \in L$ and $y \in [L,L]$,*

$$\operatorname{tr}\big(\operatorname{ad}(x) \circ \operatorname{ad}(y)\big) = 0$$

20.4.2 The Killing Form

Let L be a complex Lie algebra. The *Killing form* on L is the map $k : L \times L \to \mathbb{C}$ defined by

$$k(x,y) \coloneqq \operatorname{tr}\big(\operatorname{ad}(x) \circ \operatorname{ad}(y)\big)$$

Lemma 20.4.5. *The Killing form is a symmetric bilinear form. Moreover,*

$$k\big([x,y],z\big) = k\big(x,[y,z]\big)$$

Theorem 20.4.6 (Cartan's First Criterion)**.** *Let L be a complex Lie algebra. Then, L is soluble if and only if $k(x,y) = 0$ for all $x \in L$ and $y \in [L,L]$.*

Lemma 20.4.7. *Let L be a complex Lie algebra and I be an ideal of L. Then, the restriction of the Killing form k of L to I is the Killing form k_I on I (i.e. the Killing form on I when I is considered as a complex Lie algebra itself); for all $x,y \in I$,*

$$k(x,y) = k_I(x,y)$$

Let τ be a symmetric bilinear form on a vector space V and let W be a linear subspace of V. We define the *orthogonal complement* $W^\perp$ of W in V to be:

$$W^\perp := \big\{v \in V : \forall w \in W, \tau(v,w) = 0\big\}$$

Lemma 20.4.8. *The set $W^\perp$ is a linear subspace of V.*

In particular, the subspace $V^\perp$ is called the *radical* of τ.

The form τ is *non-degenerate* if its radical $V^\perp = \{0_V\}$ is trivial. Note that for a symmetric bilinear form, positive-definiteness ($\tau(x,x) \neq 0$ whenever $x \neq 0$) implies non-degeneracy.

Recall that for a fixed basis $\mathcal{B} = e_1,\dots,e_n$, a bilinear form τ is uniquely determined by the matrix $[\tau]_{\mathcal{B}} = \big(\tau(e_i,e_j)\big)$, and vice versa. When τ is symmetric matrix, this matrix is symmetric, and τ is non-degenerate if and only if $\det\big([\tau]_{\mathcal{B}}\big) \neq 0$.

A basis $\mathcal{B}$ of V is *orthonormal* if $\tau(e_i,e_i) = 1$ and $\tau(e_i,e_j)$ for all $i \neq j$.

Lemma 20.4.9. *For an non-degenerate symmetric bilinear form τ and a linear subspace W of V,*

$$\dim(V) = \dim(W) + \dim(W^\perp)$$

Moreover, if $W \cap W^\perp = \{0\}$, then $V = W \oplus W^\perp$, $(W^\perp)^\perp = W$, and the restrictions of τ to W and $W^\perp$ are non-degenerate.

We are interested in the orthogonal complements of ideals with respect to the Killing form.

Lemma 20.4.10. *Let L be a complex Lie algebra, and I an ideal of L. Then, the ortheogonal complement*

$$I^\perp = \{x \in L : \forall i \in I, k(x,i) = 0\}$$

is an ideal of L.

Proof. The orthogonal complement is always a linear subspace, so it remains to verify that $I^\perp$ absorbs Lie brackets with any element of L.

Let $x \in I^\perp$, $y \in L$, and $z \in I$. Since I is an ideal, $[y,z] \in I$, so

$$k\big([x,y],z\big) = k\big(x,[y,z]\big) = 0$$

so $[x,y]$ is orthogonal to $z \in I$, and is hence in $I^\perp$. ■

Theorem 20.4.11 (Cartan's Second Criterion)**.** *Let L be a complex Lie algebra. Then, L is semisimple if and only if its Killing form k is non-degenerate.*

Lemma 20.4.12. *Let L be a semisemple complex Lie algebra and I be an ideal of L. Then,*

(i) $I \cap I^\perp = \{0\}$*;*

(ii) $L = I \oplus I^\perp$ *as Lie algebras;*

(iii) I *and* $I^\perp$ *are semisimple as complex Lie algebras.*

Theorem 20.4.13. *Let L be a complex Lie algebra. Then, L is semisimple if and only if there exist simple ideals $L_1,L_2,\dots,L_k \subseteq L$ such that $L = \bigoplus_{i=1}^k L_i$.*

20.4.3 Derivations

Given a field K, a K-algebra A is a vector space over K endowed with an additional bilinear multiplication operation $A \times A \to A$. Bilinearity is equivalent to multiplication distributing over addition and compatibility with scalar multiplication in the vector space.

Example. Lie algebras are K-algebras, with multiplication given by the Lie bracket. △

Let U be a K-algebra. A linear map $\delta : U \to U$ is a *derivation* if it satisfies the *Leibniz law*:

$$\delta(ab) = a\delta(b) + \delta(a)b$$

for all $a,b \in U$. That is, δ satisfies an analogue of the product rule of differentiation.

We denote by $\mathrm{Der}(U)$ the set of derivations on U.

Lemma 20.4.14. *The set* $\mathrm{Der}(U)$ *is a linear subspace of* $\mathrm{End}(V)$*, and in particular, is a vector space over K.*

Lemma 20.4.15. *Let L be a Lie algebra. Then,* $\mathrm{Der}(L)$ *is a LIe subalgebra of* $\mathfrak{gl}(L)$*. In particular,* $\mathrm{Der}(L)$ *is a Lie algebra with Lie bracket* $[a,b] = ab - ba$*.*

Example. The adjoint homomorphism is a derivation: for any $x,a,b \in L$,

$$\begin{aligned}\mathrm{ad}(x)\big([a,b]\big) &= \big[x,[a,b]\big]\\ &= -\big[a,[b,x]\big] - \big[b,[x,a]\big]\\ &= \big[a,[x,b]\big] + \big[[x,a],b\big]\\ &= \big[a,\mathrm{ad}(x)(b)\big] + \big[\mathrm{ad}(x)(a),b\big]\end{aligned}$$

△

For $x \in L$, we call $\mathrm{ad}(x)$ an *inner derivation* of L, and any other derivation an *outer derivation.*

Theorem 20.4.16 (Primary Decomposition)**.** *Let $x \in \mathfrak{gl}(V)$, and suppose the minimal polynomial of x factorises as*

$$(X - \lambda_1)^{a_1} \cdots (X - \lambda_r)^{a_r}$$

where the eigenvalues λ_i are distinct, and $a_i \geq 1$. Then, V decomposes as a direct sum of x-invariant subspaces V_i,

$$V = \bigoplus_{i=1}^{r} V_i$$

where $V_i = \ker(x - \lambda_i 1_V)^{a_i}$ is the generalised eigenspace of x with respect to λ_i.

Theorem 20.4.17. *Let L be a complex semisimple Lie algebra. Then, all derivations of L are inner. That is,*

$$\mathrm{ad}(L) = \mathrm{Der}(L)$$

Lemma 20.4.18. *Let L be a complex Lie algebra, and $\delta \in \mathrm{Der}(L)$ a derivation with Jordan decomposition $\delta = d_\delta + n_\delta$ in $\mathfrak{gl}(L)$. Then, $d_\delta, n_\delta \in \mathrm{Der}(L)$.*

Corollary 20.4.18.1. *Let L be a complex semisimple Lie algebra. Then, for each $x \in L$, there exists unique elements $d,n \in L$ such that:*

(i) *$x = d + n$;*

(ii) *$\mathrm{ad}(d)$ is diagonalisable and $\mathrm{ad}(n)$ is nilpotent;*

(iii) *$[d,n] = 0$.*

Let L be a complex semisimple Lie algebra. Then, the decomposition of an $x \in L$ into $x = d+n$ as above is called the *abstract Jordan decomposition* of x, d is the *semisimple part* of x, and n is the *nilpotent part* of x. If $n = 0$, then $x = d$ is *semisimple*, and if $d = 0$, then $x = n$ is *nilpotent.*

Note that if L is a semisimple Lie subalgebra of $\mathfrak{gl}(V)$, for V a complex vector space, then there is a potential ambiguity, in that every element of $\mathfrak{gl}(V)$ has its original Jordan decomposition as well as this abstract Jordan decomposition. However, these actually coincide:

Theorem 20.4.19. *Let L be a semisimple complex Lie algebra, and $\phi : L \to \mathfrak{gl}(V)$ a representation. Let $x = d+n$ be the abstract Jordan decomposition of $x \in L$. Then, the Jordan decomposition of $\phi(x) \in \mathfrak{gl}(V)$ is $\phi(x) = \phi(d) + \phi(n)$.*

20.5 Root Space Decompositions

20.5.1 Cartan Subalgebras

In this section, all the Lie algebras are complex semisimple.

Lemma 20.5.1. *Suppose $x_1, \ldots ,x_n \in \mathfrak{gl}(V)$ are diagonalisable. Then, there exists a basis of V such that $x_1, \ldots ,x_n$ are diagonal if and only if they pairwise commute.*

Let L be a Lie algebra. A *Cartan subalgebra H* is a Lie subalgebra of H such that

(*i*) H is abelian.

(*ii*) Every element $h \in H$ is semisimple.

(*iii*) H is maximal with respect to (*i*) and (*ii*);

The existence of such a subalgebra is guaranteed, since $\{0\}$ satisfies (*i*) and (*ii*). However, the following lemma shows that we are guarantees more interesting Cartan subalgebras:

Lemma 20.5.2. *A semisimple complex Lie algebra L contains a non-zero Cartan subalgebra.*

At this point there is an obvious question - are Cartan subalgebras unique? The answer is no, but they do all have the same dimension.

Given $y \in L$, we define the *centraliser* of y as the set:

$$C_L(y) = \big\{x \in L : [x,y] = 0\big\}$$

More generally, the centraliser of a subset $Y \subseteq L$ is the set:

$$C_L(A) = \big\{x \in L : \forall y \in Y, [x,y] = 0\big\}$$

Lemma 20.5.3. *For any $y \in L$ and $Y \subseteq L$, the centralisers $C_L(y)$ and $C_L(Y)$ are Lie subalgebras of L.*

Proof. By construction, $C_L(y) = \ker\big(\mathrm{ad}(y)\big)$ and is thus a linear subspace of L. Now, suppose $a,b \in C_L(y)$. Then,

$$\begin{aligned}\big[y,[a,b]\big] &= -\big[b,[y,a]\big] - \big[a,[b,y]\big] \\ &= -[b,0] - [a,0] \\ &= 0\end{aligned}$$

so $[a,b] \in C_L(y)$. $C_L(Y)$ is closed under Lie brackets under the same reasoning, and is a linear subspace, since $C_L(Y) = \bigcap_{y\in Y} C_L(y)$ is an intersection of linear subspaces. ■

Lemma 20.5.4. *Let H be a Cartan subalgebra of L, and let $h_0 \in H$. Then,*

$$H \subseteq C_L(H) \subseteq C_L(h_0)$$

Proof. Since H is abelian, $H \subseteq C_L(H)$. Moreover,

$$\begin{aligned} C_L(H) &= \bigcap_{h \in H} C_L(h) \\ &\subseteq C_L(h_0) \end{aligned}$$

■

Lemma 20.5.5. *Let H be a Cartan subalgebra of L, and $h_0 \in H$ satisfying*

$$\dim C_L(h_0) \leq \dim C_L(h)$$

for all $h \in H$. Then, $C_L(h_0) = C_L(H)$.

20.5.2 Dual Spaces

Given a vector space V over a field K, the *dual space* V^* of V is the set of all linear functionals $V \to K$ equipped with the natural vector space structure of addition and scalar multiplication of linear maps.

Example. The weights of a subalgeba M are elements of M^*. △

Given a basis $e_1, \ldots, e_n$ of V, the *dual basis* $f_1, \ldots, f_n : V \to K$ of V^* is defined as:

$$f_i(e_j) = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

For each $v \in V$, the *evaluation map* $\epsilon_v : V^* \to K$ is defined by

$$\epsilon_v(f) = f(v)$$

This map is linear and thus belongs to V^{**}. The map $\epsilon : V \to V^{**} : v \mapsto \epsilon_v$ is an isomorphism and shows that $V \cong V^{**}$.

Given a bilinear form $\tau : V \times V \to K$, we may define a linear map $\Phi_\tau : V \to V^*$ by $\phi(v) = \tau(v,-)$, with linearity of Φ_τ following from the bilinearity of τ. Conversely, given a linear map $\Phi : V \to V^*$, we may define a bilinear form $\tau_\Phi : V \times V \to K$ by $\tau_\Phi(u,v) = \Phi(u)(v)$. These operations are inverse, in that $\Phi = \Phi_{\tau_\Phi}$ and $\tau = \tau_{\Phi_\tau}$. Finally, if Φ is an isomorphism, then it has trivial kernel and τ_Φ is non-degenerate, and vice versa; if τ is non-degenerate, then Φ_τ is an isomorphism.

20.5.3 Roots of L Relative to a Cartan Subalgebra H

Let L be a complex semisimple Lie algebra, and suppose that H is a Cartan subalgebra of L. Then, H act on L via the adjoint map

$$\operatorname{ad}(h) : L \to L$$

for each $h \in H$.

Since H is abelian and consists of semisimple elements (i.e. $\operatorname{ad}(h)$ is diagonalisable for all $h \in H$), there exists a basis of L consisting of common eigenvectors for all elements of H (rather, for the elements $\operatorname{ad}(h)$, so this is an abuse of notation) If v is such a common eigenvector, then for each $h \in H$, there exists $\alpha(h) \in \mathbb{C}$ such that

$$\operatorname{ad}(h)(v) = \alpha(h)v$$

By definition, $\alpha : H \to \mathbb{C}$ is a weight of H (really of $\text{ad}(H) \cong H$, with the isomorphism coming from the fact that L is semisimple), and so $\alpha \in H^*$. Let L_α be the corresponding weight space of H, which is

$$L_\alpha = \big\{x \in L : \forall h \in H, [h,x] = \alpha(h)x\big\} \neq \{0\}$$

In particular, if $\alpha = 0$,

$$L_0 = \big\{x \in L : \forall h \in H, [h,x] = 0\big\} = C_L(H)$$

Let $\Phi = \big\{\alpha \in H^* : \alpha \neq 0, L_\alpha \neq \{0\}\big\}$. Then, Φ is a *set of roots* of L *relative to* H, and for each $\alpha \in \Phi$, the corresponding *root space* is L_α.

By the primary decomposition theorem,

$$L = L_0 \oplus \bigoplus_{\alpha \in \Phi} L_\alpha$$

Note that $|\Phi|$ is finite, since L is finite dimensional.

Lemma 20.5.6. *Let $\alpha,\beta \in H^*$. Then,*

(*i*) $[L_\alpha,L_\beta] \subseteq L_{\alpha+\beta}$;

(*ii*) *If $\alpha + \beta \neq 0$, then $k(L,\alpha,L_\beta) = 0$;*

(*iii*) *$L_0 \cap L_0^\perp = \{0\}$, and so $k\big|_{L_0}$ is non-degenerate.*

Theorem 20.5.7. *Let L be a complex semisimple Lie algebra and H be a Cartan subalgebra of L. Then, $H = C_L(H)$.*

Corollary 20.5.7.1. *The root space decomposition of L relative H is*

$$L = H \oplus \bigoplus_{\alpha \in \Phi} L_\alpha$$

20.5.4 Sets of Roots Relative to H

Let L be a complex semisimple Lie algebra and H be a Cartan subalgebra of L.

Lemma 20.5.8. *For each non-zero $h \in H$, there exists $\alpha \in \Phi$ with $\alpha(h) \neq 0$.*

Corollary 20.5.8.1. *The set of roots span the dual space H^*.*

Proof. Suppose otherwise. Then, there exists a non-zero $h \in H$ such that $\alpha(h) = 0$ for all $\alpha \in \Phi$, contradicting the previous lemma. ■

Lemma 20.5.9. *If $\alpha \in \Phi$, then $-\alpha \in \Phi$.*

Proof. Suppose otherwise there exists $\alpha \in \Phi$ such that $-\alpha \notin \Phi$. Then, for each $\beta \in \Phi \cup \{0\}$, $k(L_\alpha,L_\beta) = 0$ by Lemma 20.5.6. Thus $L_\alpha \subseteq L^\perp = \{0\}$, so $L_\alpha = \{0\}$, which is a contradiction. ■

Lemma 20.5.10. *For each $\alpha \in \Phi$, there exists a non-zero $t_\alpha \in H$ such that for all $x \in L_\alpha$ and $y \in L_{-\alpha}$,*

$$[x,y] = k(x,y)t_\alpha$$

Moreover, $k(t_\alpha,h) = \alpha(h)$ for all $h \in H$.

Corollary 20.5.10.1. *If $\alpha \in \Phi$, then $L_\alpha,L_{-\alpha} = \langle t_\alpha \rangle_{\mathbb{C}}$.*

Fix a root $\alpha \in \Phi$ and fix non-zero $x \in L_\alpha$ and non-zero $y \in L_{-\alpha}$ such that $[x,y] \neq 0$. We define the set

$$M_\alpha = \big\langle x,y,[x,y]\big\rangle_{\mathbb{C}}$$

The previous two results show that $[x,y] = \lambda t_\alpha$, where $\lambda = k(x,y) \neq 0$.

Lemma 20.5.11. *M_α is a Lie subalgebra of L of dimension* $\dim(M_\alpha) = 3$.

Lemma 20.5.12. *For $\alpha \in \Phi$, $\alpha(t_\alpha) \neq 0$.*

Since $\alpha(t_\alpha) \neq 0$, we have that $k(t_\alpha,t_\alpha) = \alpha(t_\alpha) \neq 0$, so we define:

$$e_\alpha \coloneqq x, \qquad h_\alpha \coloneqq \frac{2}{k(t_\alpha,t_\alpha)}t_\alpha, \qquad f_\alpha = \frac{2}{k(t_\alpha,t_\alpha)k(x,y)}y$$

so

$$M_\alpha = \langle e_\alpha,h_\alpha,f_\alpha\rangle_{\mathbb{C}}$$

Lemma 20.5.13. *For every root $\alpha \in \Phi$, $M_\alpha \cong \mathfrak{sl}_2(\mathbb{C})$.*

Lemma 20.5.14. *For $\alpha \in \Phi$, $t_\alpha = -t_{-\alpha}$, $h_\alpha = -h_{-\alpha}$, and $\alpha(h_\alpha) = 2$.*

20.6 Representations

20.6.1 Modules

Let L be a Lie algebra over K. An *L-module* is a vector space V over K equipped with a bilinear map

$$\cdot\, : L \times V \to V$$

compatible with the Lie bracket in that

$$[x,y] \cdot v = x \cdot (y \cdot v) - y \cdot (x \cdot v)$$

Such a map is said to be an *action* of L on V. We often drop the $\cdot$ and write the action as multiplication.

An L-module V is *trivial* if $x \cdot v = 0$ for all $x \in L$ and $v \in V$.

If V is a vector space and L is a Lie subalgebra of $\mathfrak{gl}(V)$, then the map $x \cdot v \coloneqq x(v)$ defines an action of L on V.

Example. Let $V = \mathbb{C}^3$ and $L = \mathfrak{b}_3(\mathbb{C}) \subseteq \mathfrak{gl}(\mathbb{C}^3)$. We define a bilinear map $L \times V \to V$ by $A \cdot v = Av$. Then,

$$\begin{aligned}[A,B] \cdot v &= (AB - BA)v \\ &= ABv - BAv \\ &= A(Bv) - B(Av) \\ &= A \cdot (B \cdot v) - B \cdot (A \cdot v)\end{aligned}$$

so this map is an action of L on V. △

Lemma 20.6.1. *Let L be a Lie algebra and $\phi : L \to \mathfrak{gl}(V)$ be a representation of L. Then, V is an L-module under the action defined by $x \cdot v \coloneqq \phi(x)(v)$.*

Conversely, if V is an L-module, then there is a corresponding representation $\phi : L \to \mathfrak{gl}(V)$ defined by $\phi(x)(v) \coloneqq x \cdot v$.

Let L be a Lie algebra and V be an L-module. An *L-submodule* of V is a linear subspace W of V which is also an L-module under the same action $x \cdot v$ as for V.

To check that W is a submodule of an L-module V, it suffices to check that W is L-invariant.

Example. Let $V = \mathbb{C}^3$ and $L = \mathfrak{b}_3(\mathbb{C}) \subseteq \mathfrak{gl}(\mathbb{C}^3)$ as above, and consider the linear subspace $W = \langle e_1 \rangle \subseteq V$. Then, for any $A \in L$ and $v \in W$,

$$Av = \begin{bmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{bmatrix} \begin{bmatrix} x \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} ax \\ 0 \\ 0 \end{bmatrix} \in W$$

so W is an L-submodule of V. △

Example. Let V and L be as above, and let $U = \langle e_1, e_2 \rangle \subseteq V$. Then, for any $A \in L$ and $v \in U$,

$$Av = \begin{bmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{bmatrix} \begin{bmatrix} x \\ y \\ 0 \end{bmatrix} = \begin{bmatrix} ax + by \\ dy \\ 0 \end{bmatrix} \in U$$

so U is also an L-submodule of V. △

Let L be a Lie algebra and V,W be L-modules. An *L-module homomorphism* from V to W is a linear map $\phi : V \to W$ such that

$$x \cdot \phi(v) = \phi(x \cdot v)$$

for all $x \in L$ and $v \in V$. If ϕ is further an isomorphism of vector spaces, then it is an *L-module isomorphism.* As usual, we say that V and W are isomorphic if there exists an L-module isomorphism between them, and we write $V \cong W$ to denote for this relation.

Let L be a Lie algebra and V an L-module. Then, V is the *direct sum* of U and W if $V = U \oplus W$ as vector spaces, and both U and W are L-submodules of V.

One may verify that if we define an external direct sum of two L-modules U and W by the action $x \cdot (u + w) = x \cdot u + x \cdot w$ on the vector space $U \oplus W$, we obtain an L-module isomorphic to the internal direct sum.

A L-module V is *irreducible* or *simple* if V is non-trivial and has no proper non-zero submodules. That is, the only submodules of V are $\{0\}$ and V.

An L-module V is *completely reducible* if for any L-submodule W of V, there exists an L-submodule W' of V such that $V = W \oplus W'$.

A module V is *indecomposable* if it cannot be expressed as the direct sum of two non-zero proper L-submodules of V.

Lemma 20.6.2. *For any L-module V, irreducibility implies indecomposability, but not the converse in general.*

Theorem 20.6.3 (Weyl's Theorem). *A non-zero module of a semisimple complex Lie algebra is completely reducible.*

20.6.2 Representation Theory of $\mathfrak{sl}_2(\mathbb{C})$

In this section, we classify the irreducible $\mathfrak{sl}_2(\mathbb{C})$-modules. As usual, let $e = E_{12}$,$h = E_{11} - E_{22}$,$f = E_{21}$ be the standard basis of $\mathfrak{sl}_2(\mathbb{C})$.

Consider the vector space $\mathbb{C}[X,Y]$ of polynomials in two indeterminates X and Y. For each $d \geq 0$, define W_d to be the linear subspace of homogeneous degree-d polynomials in X and Y. A basis for W_d is then given by X^d,$X^{d-1}Y, \dots XY^{d-1}$,Y^d, so $\dim(W_d) = d + 1$.

We define an action of $\mathfrak{sl}_2(\mathbb{C})$ on W_d as to make W_d into an $\mathfrak{sl}_2(\mathbb{C})$-module. To do this, it is sufficient to define the action of e, h, and f on $p \in W_d$:

$$e \cdot p = X\frac{\partial p}{\partial Y}, \qquad h \cdot p = X\frac{\partial p}{\partial X} - Y\frac{\partial p}{\partial Y}, \qquad f \cdot p = Y\frac{\partial p}{\partial X}$$

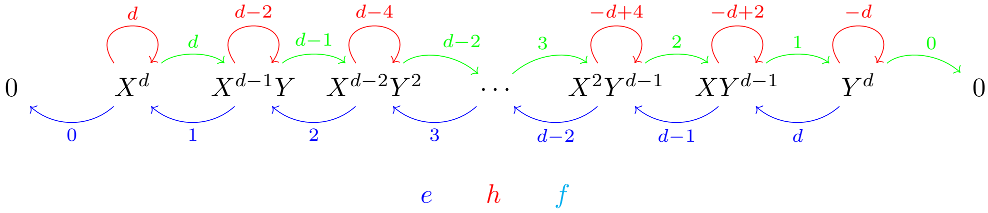
So, on monomials, the actions are:

$$e \cdot X^aY^b = bX^{a+1}Y^{b-1}, \qquad h \cdot X^aY^b = (a-b)X^aY^b, \qquad f \cdot X^aY^b = bX^{a-1}Y^{b+1}$$

Theorem 20.6.4. *W_d is an $\mathfrak{sl}_2(\mathbb{C})$ module with this action.*

Lemma 20.6.5. *For any two basis vectors $v_1 = X^aY^{d-a}$ and $v_2 = X^bY^{d-b}$, there exists a sequence of elements $\ell_1, \ldots ,\ell_n \in \mathfrak{sl}_2(\mathbb{C})$ such that $\ell_1 \cdot (\ell_2 \cdot (\cdots (\ell_n \cdot v_1))) = v_2$.*

Proof. It is sufficient to show that we may reach, starting from X, to $X^{d-1}Y$, to ... to Y^d, and vice versa. $f \cdot X = dX^{d-1}Y$, so the element $\frac{1}{d}f$ maps X^d to $X^{d-1}Y$. In general, to obtain $X^{d-a-1}Y^{a+1}$ from $X^{d-a}Y^a$, $\frac{1}{d-a}f$ will do. Similarly, applying scaled copies of e moves from Y^d to X^d:



■

Theorem 20.6.6. *Each W_d for $d \geq 0$ is irreducible.*

We now want to show that any other irreducible $\mathfrak{sl}_2(\mathbb{C})$-module is isomorphic to W_d for some $d \geq 0$. This completes the classification, since W_d is isomorphic to $W_{d'}$ if and only if $d = d'$, since they have different dimensions otherwise.

Lemma 20.6.7. *Let V be an $\mathfrak{sl}_2(\mathbb{C})$-module, and let $v \in V$ be an eigenvalue of h with eigenvalue λ. Then,*

(i) Either $h \cdot (e^n \cdot v) = (\lambda + 2n)(e^n \cdot v)$ or $e^n \cdot v = 0$;

(ii) Either $h \cdot (f^n \cdot v) = (\lambda - 2n)(f^n \cdot v)$ or $f^n \cdot v = 0$.

We write $x^n \cdot v$ for

$$x^n \cdot v := \underbrace{x \cdot (x \cdot (\cdots (x}_{n} \cdot v) \cdots))$$

Lemma 20.6.8. *Let V be an $\mathfrak{sl}_2(\mathbb{C})$-module. Then, V contains an eigenvector w for h such that $e \cdot w = 0$ and $f^d \cdot w \neq 0$ but $f^{d+1} \cdot w = 0$ for some $d \geq 0$.*

Theorem 20.6.9. *Let V be an irreducible $\mathfrak{sl}_2(\mathbb{C})$-module. Then, V is isomorphic to W_d for some $d \geq 0$.*

20.6.3 The Importance of $\mathfrak{sl}_2(\mathbb{C})$ for Semisimple Complex Lie Algebras

In this section, let L be a complex semisimple Lie algebra, H be a Cartan subalgebra of L, and Φ be the set of roots relative to H. Recall that we have the decomposition of L:

$$L = H \oplus \bigoplus_{\alpha \in \Phi} L_\alpha$$

We also defined $M_\alpha = \langle e_\alpha, h_\alpha, f_\alpha \rangle$, a subalgebra of L isomorphic to $\mathfrak{sl}_2(\mathbb{C})$ for each root $\alpha \in \Phi$. L can be viewed as an L-module using the adjoint representation $x \cdot y = [x,y]$, and since M_α is a subalgebra of L, me may study L as an M_α-module by restricting ad_L to M_α (so the action is the same for $x \in M_\alpha$, $y \in L$).

Lemma 20.6.10. *If V is an M_α-submodule of L, then the eigenvalues of h_α acting on V are integers.*

We have already seen that $-\alpha \in \Phi$ if $\alpha \in \Phi$. It turns out that these are the only multiples of α in Φ:

Theorem 20.6.11. *For each $\alpha \in \Phi$, if $c\alpha \in \Phi$ for some $c \in \mathbb{C}$, then $c = \pm 1$.*

We define the M_α-submodules of L

$$U_\alpha \coloneqq \langle H, L_{c\alpha} \mid c\alpha \in \Phi \rangle_{\mathbb{C}} \subseteq L$$

and

$$K_\alpha \coloneqq \ker(\alpha) \subseteq L$$

Corollary 20.6.11.1. *For any $\alpha \in \Phi$, $U_\alpha = K_\alpha \oplus M_\alpha$.*

Corollary 20.6.11.2. *For any $\alpha \in \Phi$, $\dim(L_\alpha) = 1$.*

If $\beta \in \Phi \cup \{0\}$, the *$\alpha$-root string through β* is the space

$$S \coloneqq \bigoplus_c L_{\beta + c\alpha}$$

where the sum is taken over all $c \in \mathbb{Z}$ such that $\beta + c\alpha \in \Phi$. Since $[L_\gamma, L_\delta] \subseteq L_{\gamma+\delta}$ for any roots $\gamma, \delta \in \Phi$, it follows that S is an M_α-submodule of L.

Strictly speaking, the proper definition of a root string should have the sum range over all $c \in \mathbb{C}$ such that $\beta + c\alpha \in \Phi$, but it turns out that these give the same submodule of L, and we will only need to work with the above definition.

Lemma 20.6.12. *Let $\alpha, \beta \in \Phi$ such that $\alpha \neq \beta$. Then,*

(i) $\beta(h_\alpha) \in \mathbb{Z}$;

(ii) *There exist integers $q, r \geq 0$ such that, given an integer $k \in \mathbb{Z}$, $\beta + k\alpha \in \Phi$ if and only if $-r \leq k \leq q$. Moreover,*

$$r - q = \beta(h_\alpha)$$

(iii) $\beta - \beta(h_\alpha)\alpha \in \Phi$.

Lemma 20.6.13. *For $\alpha, \beta \in \Phi$, $k(h_\alpha, h_\beta) \in \mathbb{Z}$, and $k(t_\alpha, t_\beta) \in \mathbb{Q}$.*

20.7 Root Systems and Classifications

20.7.1 Roots of L

Recall that we have an explicit isomorphism between H and H^*, given by

$$h \mapsto k(h, -)$$

Furthermore, for every $\alpha \in \Phi$, there exists $t_\alpha \in H$ such that $\alpha(-) = k(t_\alpha, -)$. Now, let $\phi \in H^*$, and denote by t_ϕ the element of H satisfying

$$t_\phi \mapsto k(t_\phi, -) = \phi(-)$$

We define a bilinear form $(-,-) : H^* \times H^* \to \mathbb{C}$ by

$$(\theta, \phi) = k(t_\theta, t_\phi)$$

Since k is a symmetric bilinear form on H, $(-,-)$ is a symmetric bilinear form on H^*. In particular, for $\alpha, \beta \in \Phi$, $(\alpha, \beta) = k(t_\alpha, t_\beta) \in \mathbb{Q}$.

Since $H^* = \langle \Phi \rangle_{\mathbb{C}}$, there exist $\alpha_1, \ldots, \alpha_\ell \in \Phi$ that form a basis of H^*. We define the real subspace E of H^* by

$$E \coloneqq \mathbb{R}[\alpha_1, \ldots, \alpha_\ell]$$

Clearly, $\phi \subseteq E$, and we may restrict $(-,-)$ to E, so

$$(-,-) : E \times E \to \mathbb{R}$$

is also a symmetric bilinear form. Then, there exists $t_\theta \in H$ such that

$$\begin{aligned} (\theta,\theta) &= k(t_\theta,t_\theta) \\ &= \operatorname{tr}\big(\operatorname{ad}(t_\theta)^2\big) \\ &= \sum_{\gamma \in \Phi} \gamma(t_\theta)^2 \\ &= \sum_{\gamma \in \Phi} k(t_\gamma,t_\theta)^2 \\ &= \sum_{\gamma \in \Phi} (\gamma,\theta)^2 \end{aligned}$$

Since $(\gamma,\theta) \in \mathbb{Q} \subseteq \mathbb{R}$, and $(\theta,\theta) \geq 0$, $(\theta,\theta) = 0$ if and only if $(\gamma,\theta) = \gamma(t_\theta) = 0$ for all $\gamma \in \Phi$, which means that $\theta = 0$. So, E is in fact an inner product space, and in particular, a Euclidean space, since it is finite-dimensional and real-valued.

Lemma 20.7.1. *Let L be a semisimple complex Lie algebra with roots Φ. Then,*

(i) E is a vector space over R with a real-valued inner product;

(ii) $\langle \Phi \rangle_{\mathbb{R}} = E$, and $0 \notin \Phi$;

(iii) If $\alpha \in \Phi$, then $-\alpha \in \Phi$;

(iv) If $r\alpha \in \Phi$ for some $r \in \mathbb{R}$, then $r = \pm 1$;

(v) For $\alpha,\beta \in \Phi$,

$$\begin{aligned} 2\frac{(\beta,\alpha)}{(\alpha,\alpha)} &= k\left(t_\beta, \frac{2}{k(t_\alpha,t_\alpha)} t_\alpha\right) \\ &= k(t_\beta,h_\alpha) \\ &= \beta(h_\alpha) \\ &\in \mathbb{Z} \end{aligned}$$

and

$$\beta - 2\frac{(\beta,\alpha)}{(\alpha,\alpha)}\alpha \in \Phi$$

20.7.2 Root Systems

Let E be a finite-dimensional vector space over $\mathbb{R}$, and let $(-,-) : E \times E \to \mathbb{R}$ be an inner product.

For every non-zero vector $v \in E$, we define the map $\sigma_v : E \to E$ by

$$\sigma_v(x) = x - 2\frac{(x,v)}{(v,v)}v$$

Geometrically, this is the reflection through the hyperplane orthogonal to v, since $\sigma_v(v) = v - 2v = -v$ and if x is orthogonal to v, then $\sigma_v(x) = x - 0v = x$.

For $u,v \in E$, we abbreviate

$$\langle x,v\rangle := 2\frac{(x,v)}{(v,v)}$$

Geometrically, $\langle \alpha,\beta\rangle = 2\frac{\|\alpha\|}{\|\beta\|}\cos(\theta)$, where θ is the angle between α and β, and can thus be interpreted as a normalised/rescaled projection of α onto β.

Also note that this mapping is linear in the first argument, but *not* the second.

Lemma 20.7.2. *For $x,y,v \in E$, $\big(\sigma_v(x),\sigma_v(y)\big) = (x,y)$.*

A subset $R \subseteq E$ is a *root system* in E if:

(R1) R is finite, $0 \notin R$, and $\langle R\rangle_{\mathbb{R}} = E$;

(R2) If $\alpha \in R$, then $c\alpha \in R$ if and only if $c = \pm 1$;

(R3) If $\alpha \in R$, then $\sigma_\alpha(R) \subseteq R$ (that is, R is closed under reflections through roots);

(R4) If $\alpha,\beta \in R$, then $\langle \alpha,\beta\rangle = 2\frac{(\alpha,\beta)}{(\beta,\beta)} \in \mathbb{Z}$;

Let L be a semisimple Lie algebra over $\mathbb{C}$ with H a Cartan subalgebra of L, and let Φ be the set of roots of L relative to H. As before, let $E = \mathbb{R}[\Phi]$, which is a real vector space with inner product induced by the Killing form k.

For the rest of this section, let R be a root system in E.

Lemma 20.7.3. *For $\alpha,\beta \in R$ with $\alpha \neq \pm\beta$,*

$$\langle \alpha,\beta\rangle\langle \beta,\alpha\rangle \in \{0,1,2,3\}$$

Corollary 20.7.3.1. *Let $\alpha,\beta \in R$. Then,*

(i) *$\langle \alpha,\beta\rangle = 0$ if and only if a and b are orthogonal;*

(ii) *$\langle \alpha,\beta\rangle > 0$ if and only if $\langle \beta,\alpha\rangle > 0$.*

Let $\alpha,\beta \in R$, and without loss of generality, suppose $(\beta,\beta) \geq (\alpha,\alpha)$. Then,

$$\big|\langle \beta,\alpha\rangle\big| = 2\frac{\big|(\beta,\alpha)\big|}{(\alpha,\alpha)} \geq 2\frac{\big|(\beta,\alpha)\big|}{(\beta,\beta)} = \big|\langle \alpha,\beta\rangle\big|$$

We can now use the previous lemma to classify all the possible values of $\langle \beta,\alpha\rangle$:

$\langle \alpha,\beta\rangle$	$\langle \beta,\alpha\rangle$	θ	$\frac{(\beta,\beta)}{(\alpha,\alpha)} = \frac{\|\beta\|^2}{\|\alpha\|^2}$
0	0	$\frac{\pi}{2}$	undefined
1	1	$\frac{\pi}{3}$	1
-1	-1	$\frac{2\pi}{3}$	1
1	2	$\frac{\pi}{4}$	2
-1	-2	$\frac{3\pi}{4}$	2
1	3	$\frac{\pi}{6}$	3
-1	-3	$\frac{5\pi}{6}$	3

Lemma 20.7.4. *Let $\alpha,\beta \in R$ be such that $(\beta,\beta) \geq (\alpha,\alpha)$, and let θ be the angle between α and β. Then,*

(i) *If $\frac{\pi}{2} < \theta < \pi$, then $\alpha + \beta \in R$;*

(ii) *If* $0 < \theta < \frac{\pi}{2}$*, then* $\alpha - \beta \in R$.

Proof. $\sigma_\beta(\alpha) = \alpha - \langle \alpha,\beta \rangle \beta \in R$. If $\frac{\pi}{2} < \theta < \pi$, then from the table above, $\langle \alpha,\beta \rangle = -1$, so $\sigma_\beta(\alpha) = \alpha + \beta \in R$. Similarly, if $0 < \theta < \frac{\pi}{2}$, then $\langle \alpha,\beta \rangle = 1$, and $\sigma_\beta(\alpha) = \alpha - \beta \in R$. ■

Example. WIP △

A root system R is *irreducible* if R cannot be expressed as a disjoint union of two root systems R_1 and R_2 satisfying $(r_1,r_2) = 0$ for $r_1 \in R_1, r_2 \in R_2$.

Example.

(i) $A_1 \times A_1$ is not irreducible since it is the union of two root system of type A_1.

(ii) The root systems of type A_2, B_2, and G_2 are all irreducible. Indeed, the only 1-dimensional root system is of type A_1, so the only reducible root system in $\mathbb{R}^2$ is $A_1 \times A_1$.

△

Lemma 20.7.5. *Let R be a root system in E. Then, there exist non-empty subsets $R_1, \ldots, R_\ell$ of R such that*

(i) $R = \bigsqcup_{i=1}^{\ell} R_i$*;*

(ii) R_i *is an irreducible root system in* $E_i = \langle R_i \rangle_{\mathbb{R}}$*;*

(iii) $E = \bigoplus_{i=1}^{\ell} E_i$*, with* E_i *and* E_j *orthogonal for* $1 \leq i \neq j \leq \ell$.

Let R and R' be root systems of E and E' respectively. Then, R and R' are *isomorphic* if there exists an isomorphict $\phi : E \to E'$ such that

(i) $\phi(R) = R'$;

(ii) For all $\alpha,\beta \in R$, $\big\langle \phi(\alpha),\phi(\beta) \big\rangle = \langle \alpha,\beta \rangle$.

20.7.3 Bases of Root Systems

Let R be a root system. A subset $\mathcal{B} \subseteq R$ is a *base* of R if:

(B1) $\mathcal{B}$ is a basis for E;

(B2) We may express any β as a $\mathbb{Z}$-linear combination of element of $\mathcal{B}$:

$$\beta = \sum_{\alpha \in \mathcal{B}} c_\alpha \alpha$$

where $c_\alpha \in \mathbb{Z}$ for all $\alpha \in \mathcal{B}$. Moreover, either the coefficients c_α are all non-negative, or all non-positive.

The roots in a base $\mathcal{B}$ are then called *simple*.

Given (B1), the expression of β in (B2) is unique.

Let β be a root, and $\beta = \sum_{\alpha \in \mathcal{B}} c_\alpha \alpha$ be the unique expression of β in terms of the base $\mathcal{B}$. Then, the *height* of β is the sum of the coefficients of the expression:

$$\sum_{\alpha \in \mathcal{B}} c_\alpha$$

If the height of β is positive, then we say that β is a *positive root*, and similarly, if the height of β is negative, then β is a *negative root*. We denote by R^+ and R^- the sets of positive and negative roots, respectively.

Note that a root cannot be simultaneously positive and negative, since that would require that every coefficient c_α is zero, in which case the root is zero, which is disallowed in the definition of a root system. So $R^+ \cap R^- = \varnothing$.

Lemma 20.7.6. *Let $\alpha_1,\alpha_2 \in \mathcal{B}$ be distinct simple roots. Then, the angle between them is at least $\frac{\pi}{2}$.*

Proof. Suppose otherwise. Then, $\alpha_1 - \alpha_2 \in R$. This is a $\mathbb{Z}$-linear combination with both positive and negative coefficients, contradicting (B2). ■

To find a base, we can pick any hyperplane in $E = \mathbb{R}^n$ which does not contain any roots. Then, label one side of the hyperplane as positive, and the other as negative. Then, the n nearest roots to the hyperplane form a base.

Theorem 20.7.7. *Every root system has a base.*

20.7.4 The Weyl Group of a Root System

By the definition of a root system, for each root $\alpha \in R$, the corresponding reflection σ_α is an element of $GL(E)$, the group of invertible linear transformations on E.

The *Weyl group* of a root system R is the group

$$W = W(R) := \langle \sigma_\alpha \mid \alpha \in R \rangle \leq GL(E)$$

For each $\alpha \in R$, we have that $\sigma_\alpha(R) \subseteq R$, and since $\subseteq_\alpha$ is a reflection, it is an automorphism of E and thus $\sigma_\alpha(R)$ is a permutation of the roots in R. So, there exists a group homomorphism

$$f : W \to \mathrm{Sym}_{|R|}$$

sending each $w \in W$ to its action on R, viewed as a permutation.

Lemma 20.7.8. *The Weyl group is a subgroup of $\mathrm{Sym}_{|R|}$. In particular, W is finite.*

Proof. It suffices to show that f is injective. If $w \in \ker(f)$, then $f(w) = \mathrm{id}$, so w must have been the identity on R. But since R spans E, w is also the identity on E. So $\ker(f)$ is trivial, and f is injective. ■

Lemma 20.7.9. *If $\alpha \in \mathcal{B}$, then the reflection σ_α permutes the set of positive roots apart from α.*

Proof. Suppose $\beta \in R^+$, and $\alpha \neq \beta$. Then,

$$\beta = \sum_{\gamma \in \mathcal{B}} c_\gamma \gamma$$

with every c_γ non-negative. Since $\alpha \neq \beta$, there must exist $\gamma \in \mathcal{B} \setminus \{\alpha\}$ such that $\alpha_\gamma > 0$ (otherwise, β is a positive multiple of α, and $\pm\alpha$ are the only multiples of α in R).

Now, $\sigma_\alpha(\beta) = \beta - \langle \beta,\alpha \rangle \alpha$, so the expansion of $\sigma_\alpha(\beta)$ differs from the expansion of β only in c_α (by precisely $-\langle \beta,\alpha \rangle$). In particular, c_γ is still positive, so every coefficient is still positive, and $\sigma_\alpha(\beta) \in R^+$. ■

Lemma 20.7.10. *Given $\alpha \in R$, there exists $g \in W_0 := \langle \sigma_\alpha \mid \alpha \in \mathcal{B} \rangle$ and $\beta \in \mathcal{B}$ such that $\alpha = g(\beta)$.*

Lemma 20.7.11. *Suppose $\alpha \in R$ and $g \in W$. Then,*

$$g\sigma_\alpha g^{-1} = \sigma_{g(\alpha)}$$

Theorem 20.7.12. *Let R be a root system, $\mathcal{B}$ be a base of R, and W be its Weyl group. Then,* *(i)* $W = \langle \sigma_\alpha \mid \alpha \in \mathcal{B} \rangle$*;*

(ii) For each $\alpha \in R$*, there exist* $w \in W$ *and* $\alpha_i \in \mathcal{B}$ *such that* $w(\alpha_i) = \alpha$*.*

(iii) If $\mathcal{B}'$ *is another base of* R*, then there exists* $g \in W$ *such that* $\mathcal{B}' = \big\{g(\alpha_i) : \alpha_i \in \mathcal{B}\big\}$*.*

20.7.5 Cartan Matrices and Dynkin Diagrams

Let R be a root system in E with a base $\mathcal{B} = \{\alpha_1, \ldots, \alpha_\ell\}$. The *Cartan matrix* of R is the $\ell \times \ell$ matrix

$$\big(\langle \alpha_i, \alpha_j \rangle\big)_{1 \le i,j \le \ell} = \begin{bmatrix} 2 & \langle \alpha_1, \alpha_2 \rangle & \langle \alpha_1, \alpha_3 \rangle & \cdots & \langle \alpha_1, \alpha_\ell \rangle \\ \langle \alpha_2, \alpha_1 \rangle & 2 & \langle \alpha_2, \alpha_3 \rangle & \cdots & \langle \alpha_2, \alpha_\ell \rangle \\ \langle \alpha_3, \alpha_1 \rangle & \langle \alpha_3, \alpha_2 \rangle & 2 & \cdots & \langle \alpha_3, \alpha_\ell \rangle \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \langle \alpha_\ell, \alpha_1 \rangle & \langle \alpha_\ell, \alpha_2 \rangle & \langle \alpha_\ell, \alpha_3 \rangle & \cdots & 2 \end{bmatrix}$$

(The diagonal is 2 since $\langle \alpha_i, \alpha_i \rangle = 2\frac{(\alpha_i, \alpha_i)}{(\alpha_i, \alpha_i)} = 2$ for all i.)

Lemma 20.7.13. *For any* $\alpha, \beta \in R$ *and* $g \in W$*,*

$$\big\langle g(\alpha), g(\beta) \big\rangle = \langle \alpha, \beta \rangle$$

Proof. For any $\alpha, \beta, \gamma \in R$, since orthogonal transformations like reflections preserve the inner product,

$$\begin{aligned} \big\langle \sigma_\gamma(\alpha), \sigma_\gamma(\beta) \big\rangle &= 2\frac{\big(\sigma_\gamma(\alpha), \sigma_\gamma(\beta)\big)}{\big(\sigma_\gamma(\beta), \sigma_\gamma(\beta)\big)} \\ &= 2\frac{(\alpha, \beta)}{(\beta, \beta)} \\ &= \langle \alpha, \beta \rangle \end{aligned}$$

Since W is generated by the reflections σ_β for $\beta \in R$, any $g \in W$ is a composition of such reflections, so $\big\langle g(\alpha), g(\beta) \big\rangle = \langle \alpha, \beta \rangle$. ■

Lemma 20.7.14. *The Cartan matrix of a root system is unique up to reordering.*

Example. *(i)* For the root system A_2, we can take the base $\{\alpha, \beta\}$. Since the angle between them is $\frac{\pi}{2}$, $\langle \beta, \alpha \rangle = \langle \alpha, \beta \rangle = -1$. So, the Cartan matrix is

$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

(ii) For the root system B_2

(iii)

(iv)

(v)

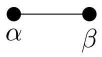
△

We can encode the information of the Cartan matrix in a graph as follows.

The *Dynkin diagram* of a root system R it the graph $\Delta = \Delta(R)$ with vertex set given by the base $\mathcal{B}$, and the number of edges between $\alpha, \beta \in \mathcal{B}$ given by $d_{\alpha,\beta} \coloneqq \langle \alpha, \beta \rangle \langle \beta, \alpha \rangle$. If $d_{\alpha,\beta} > 1$, then the two roots α and β must have different lengths, and we notate the edges with an arrow pointing from the longer root to the shorter one.

Example.

(*i*) From the Cartan matrix of A_2, the Dynkin diagram is



(*ii*) From the Cartan

(*iii*)

(*iv*)

(*v*)

△

Note that the Dynkin diagram of a root system R can be constructed just from the Cartan matrix of R, and conversely, the Cartan matrix can be reconstructed from the Dynkin diagram, since $d_{\alpha,\beta}$ determines $\langle\alpha,\beta\rangle$ and $\langle\beta,\alpha\rangle$ for all $\alpha,\beta \in \mathcal{B}$.

Theorem 20.7.15. *Let R and R' be root systems of E and E' respectively. Then, $R \cong R'$ if and only if $\Delta(R) = \Delta(R')$.*

Lemma 20.7.16. *A root system R is irreducible if and only if $\Delta(R)$ is connected.*

Theorem 20.7.17. *Let R be an irreducible root system with Dynkin diagram $\Delta(R)$. Then, $\Delta(R)$ is one of the following types:*

(*i*) *A_n, $n \geq 1$;*

(*ii*) *B_n, $n \geq 2$;*

(*iii*) *C_n, $n \geq 3$;*

(*iv*) *D_n, $n \geq 4$;*

(*v*) *G_2;*

(*vi*) *F_4;*

(*vii*) *E_6;*

(*viii*) *E_7;*

(*ix*) *E_8.*

Conversely, each such type occurs as the Dynkin diagram of a root system.

WIP

20.7.6 The Classification of Semisimple Complex Lie Algebras

Theorem 20.7.18. *Let L be a complex semisimple Lie algebra. If Φ_1 and Φ_2 are root system associated to two Cartan subalgebras of L, then $\Phi_1 \cong \Phi_2$.*

Corollary 20.7.18.1. *If L_1 and L_2 are complex semisimple Lie algebras with root systems Φ_1 and Φ_2 respectively, then $\Phi_1 \ncong \Phi_2$ implies $L_1 \ncong L_2$.*

WIP

Chapter 21

Commutative Algebra

Chapter 22

Ring Theory

In this chapter, we will develop the general theory of rings and modules, with a focus on non-commutative theory. Commutative algebra is studied in more detail in §21. Some module theory is also covered at the end of §33, with *finitely generated abelian groups*.

Chapter 23

Algebraic Geometry

23.1 Review of Commutative Algebra

A *ring* $(R,+,\cdot,0_R,1_R)$, consists of a set R, two binary operations $+,\cdot : R \times R \to R$, and two distinguished elements $0_R,1_R \in R$ such that $(R,+)$ is an abelian group with identity 0_R; $(R,\cdot)$ is a monoid with identity 1_R; and multiplication distributes over addition.

All the rings we will consider will be commutative, unital, and non-trivial.

A function $f : R \to S$ between rings R and S is a *ring homomorphism* if for all $a,b \in R$,

(i) $f(a+b) = f(a) + f(b)$;

(ii) $f(ab) = f(a)f(b)$;

(iii) $f(1_R) = 1_S$.

An *ideal* I of a ring R is an additive subgroup that absorbs multiplication from the left (or equivalently for commutative rings, the right, or both sides), and we write $I \trianglelefteq R$ to denote this relation.

Example. The set I of polynomials with zero constant term is an ideal of $R = \mathbb{C}[x,y]$. Adding such polynomials has group structure since the coefficients all have additive inverses, and addition of polynomials is associative, commutative, and closed on I. Multiplying any polynomial in R by a polynomial in I yields another polynomial with zero constant term, so I also absorbs multiplication. △

Every element $x \in R$ generates an ideal $\langle x \rangle = xR = \{xr : r \in R\}$. An ideal of this form is called a *principal* ideal.

The *unit ideal* is the entire ring $R = \langle 1_R \rangle$, and the *zero* or *trivial ideal* is the set $\{0_R\} = \langle 0_R \rangle$. An ideal is *proper* if it is a proper subset of the ring; that is, it is not equal to the whole ring.

The intersection of arbitrary ideals is also an ideal, so we may define an ideal *generated by* a set $S \subseteq R$ by

$$\langle S \rangle = \bigcap_{\substack{S \subseteq I \subseteq R \\ I \text{ is an ideal}}} I$$

That is, the ideal $\langle S \rangle$ is then the smallest ideal containing S. We can also think of the ideal $\langle S \rangle$ as the collection of all finite R-linear combinations of elements of S.

An ideal I is *finitely generated* if there is a finite set S such that $I = \langle S \rangle = \langle s_1, \dots, s_n \rangle$.

Example. The elements of the ideal $I \trianglelefteq \mathbb{C}[x,y]$ of polynomials with zero constant term are of the form $xp(x,y) + yq(x,y)$, where $p,q \in \mathbb{C}[x,y]$. That is, every element is the $\mathbb{C}[x,y]$-linear combination of x and y, so $I = \langle x,y \rangle$. △

The preimage of an ideal under a ring homomorphism ϕ is an ideal. In particular, the *kernel* $\ker(\phi) = \phi^{-1}\big[\{0\}\big]$ is an ideal.

23.1.1 Special Elements, Rings, and Ideals

Let R be a commutative ring.

(*i*) An element $x \in R$ is a *unit* if $xy = 1$ for some $y \in R$ – in this case, y is uniquely determined by x and is also denoted by x^{-1};

(*ii*) An element $x \in R$ is a *zero-divisor* if $xy = 0$ for some $y \neq 0$;

(*iii*) An element $x \in R$ is *nilpotent* if $x^n = 0$ for some $n \geq 1$. (This also implies that x is zero-divisor, unless R is trivial.)

(*i*) R is a *field* if R is non-trivial and every non-zero element is a unit. In a field, the only ideals are the zero ideal and the unit ideal;

(*ii*) R is an *integral domain* if R is non-trivial and has no zero-divisors;

(*iii*) R is *reduced* if zero is the only nilpotent element.

(*i*) An ideal $\mathfrak{m} \subset R$ is *maximal* if the only ideal strictly containing it is the unit ideal R;

(*ii*) An ideal $\mathfrak{p} \subset R$ is *prime* if whenever $fg \in \mathfrak{p}$, we have $f \in \mathfrak{p}$ or $g \in \mathfrak{p}$;

(*iii*) An ideal $I \subset R$ is *radical* if whenever $x^n \in I$, $x \in I$.

The *radical* of an ideal I is the ideal

$$\sqrt{I} := \{x \in R : \exists n > 0, x^n \in I\}$$

Equivalently, I is radical if $I = \sqrt{I}$.

Lemma 23.1.1. *Every maximal ideal is prime, and every prime ideal is radical.*

Theorem 23.1.2. *An ideal $I \trianglelefteq R$ is*

(*i*) *maximal;*

(*ii*) *prime;*

(*iii*) *radical,*

if and only if R/I is, respectively,

(*i*) *a field;*

(*ii*) *a integral domain;*

(*iii*) *a reduced ring.*

Given $I \trianglelefteq R$, the *quotient* R/I is the set of *cosets* $x+I = \{x+i : i \in I\}$. This quotient has ring structure under the addition and multiplication defined by $(x+I)+(y+I) = (x+y)+I$ (i.e. the inherited quotient group addition) and $(x + I)(y + I) = xy + I$. The quotient map $\pi : R \to R/I : x \mapsto x + I$ is then a surjective ring homomorphism with kernel I.

Because the quotient map π is a homomorphism, the preimage $\pi^{-1}[J]$ of any ideal $J \subseteq R/I$ is an ideal of R containing I. Conversely, π maps every ideal $K \subseteq R$ containing I onto an ideal in the quotient ring. Therefore, the ideals of R/I are in bijection with the ideals of R containing I:

$$\{\text{ideals of } R/I\} \cong \{\text{ideals of } R \text{ containing } I\}$$

This bijection carries maximal, prime, and radical ideals to maximal, prime, and radical ideals, respectively.

Theorem (First Isomorphism Theorem)**.** *Let $\phi : R \to S$ be a homomorphism with kernel I. Then, $R/I \cong \operatorname{im}(\phi)$. More precisely, the isomorphism $\bar{\phi} : R/I \to \operatorname{im}(\phi)$ is given by $\bar{\phi}(x + I) = \phi(x)$ for all $x \in R$.*

23.2 Affine Subvarieties

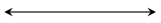
Let k be any field. Later, we will often assume that $k = \mathbb{C}$ since we will want to work over an algebraically closed field, but for now, we could also have $k = \mathbb{R}$,$\mathbb{Q}$,$\mathbb{Z}/p\mathbb{Z}$, etc. (In particular, the case where k is finite or p-adic field is of utility in number theory.)

The set $k^n = \{(x_1,x_2,\ldots,x_n) : x_i \in k\}$ is called *affine n-space* (*over k*), also denoted $\mathbb{A}^n_k$, or even just $\mathbb{A}^n$ if the field is clear or unimportant. We also sometimes write things like $\mathbb{A}^2_{x,y}$ to indicate the indeterminates.

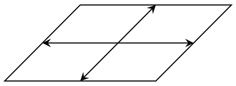
Note that $\mathbb{A}^n_k$ is just k^n as a *set*; it is customary to use different notation since k^n is also a vector space over k, a ring, a topological space with the standard Euclidean topology, etc. We will write $\mathbb{A}^n_k$ whenever we wish to ignore this additional structure, or use an alternative (i.e. we will soon put a topology on $\mathbb{A}^n_k$ distinct from the standard topology on k^n).

Example.

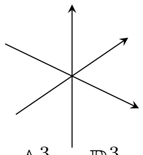
$k = \mathbb{R}$



$\mathbb{A}^1_{\mathbb{R}} = \mathbb{R}^1$

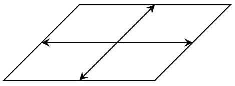


$\mathbb{A}^2_{\mathbb{R}} = \mathbb{R}^2$



$\mathbb{A}^3_{\mathbb{R}} = \mathbb{R}^3$

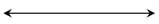
$k = \mathbb{C}$



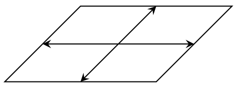
$\mathbb{A}^1_{\mathbb{C}} = \mathbb{C}$

△

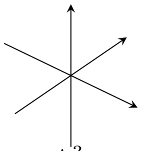
We can't draw $\mathbb{A}^4_{\mathbb{R}}$ or $\mathbb{A}^2_{\mathbb{C}}$ convincingly as they are 4-dimensional over $\mathbb{R}$, so we stop there. Later, we will define a notion of dimension specific to algebraic geometry where $\mathbb{A}^n$ is n-dimensional. Thus, we will suggestively choose to draw $\mathbb{A}^n_k$ as:



$\mathbb{A}^1_k$



$\mathbb{A}^2_k$



$\mathbb{A}^3_k$

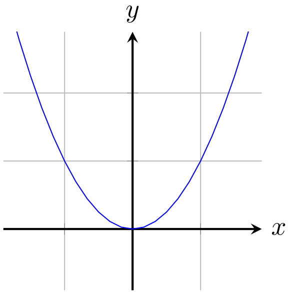
even if $k = \mathbb{C}$. In light of this, the set $\mathbb{A}^1_{\mathbb{C}} = \mathbb{C}^1$ is then called the *complex line*, while $\mathbb{A}^2_{\mathbb{C}} = \mathbb{C}^2$ is called the *complex plane*. (Contrast with analytic contexts, where "complex plane" often refers to $\mathbb{C}^1$, while the term "complex *coordinate* plane" is used for $\mathbb{C}^2$.)

Let $p \in k[x_1, \ldots ,x_n]$ be a polynomial. Then, the *vanishing locus* or *zero locus* of p is the set of points upon which p vanishes:

$$\mathbb{V}(p) \coloneqq \{x \in \mathbb{A}^n : p(x) = 0\}$$

Example. Consider the polynomial $y - x^2 \in \mathbb{C}[x,y]$. Then, the vanishing locus is the set:

$$\mathbb{V}(y - x^2) = \big\{(x,y) \in \mathbb{C}^2 : y = x^2\big\}$$



Note that this picture really depicts $\mathbb{V}(y - x^2) \cap \mathbb{R}^2$. △

When drawing a sketch of a vanishing locus V in $\mathbb{C}^n$, we will only draw its real points $V \cap \mathbb{R}^n$.

More generally, the vanishing locus of a set $S = \{f_i\}_{i\in I} \subseteq k[x_1, \ldots ,x_n]$ of polynomials is the set of points upon which all the polynomials in S vanish:

$$\mathbb{V}(S) \coloneqq \{x \in \mathbb{A}^n : \forall p \in S,p(x) = 0\}$$

If $S = \{f_1, \ldots ,f_k\}$ is finite, then we also write $\mathbb{V}(S) = \mathbb{V}\big(\{f_1, \ldots ,f_k\}\big)$ as $\mathbb{V}(f_1, \ldots ,f_k)$.

Example. $\mathbb{V}(x,y) \subseteq \mathbb{C}^3$ is the complex line in $\mathbb{C}^3$ consisting of the z-axis. △

Theorem 23.2.1.

(i) *For any* $S_1,S_2 \subseteq k[x_1, \ldots ,x_n]$,

$$\mathbb{V}\big(S_1\big) \cup \mathbb{V}\big(S_2\big) = \mathbb{V}\big(S_1S_2\big)$$

where $S_1S_2 = \{fg : f \in S_1,g \in S_2\}$

(ii) *If* I *is any set indexing a collection of sets* $S_i \subseteq k[x_1, \ldots ,x_n]$ *of polynomials, then*

$$\bigcap_{i\in I} \mathbb{V}(S_i) = \mathbb{V}\left(\bigcup_{i\in I} S_i\right)$$

Proof.

(i) If $x \in \mathbb{V}(S_1) \cup \mathbb{V}(S_2)$, then $f(x) = 0$ for all $f \in S_1$, or $g(x) = 0$ for all $g \in S_2$. In either case, $(fg)(x) = f(x)g(x) = 0$ for all $f \in S_1$ and $g \in S_2$, so $x \in \mathbb{V}(S_1S_2)$.

For the reverse containment, suppose that $x \notin \mathbb{V}(S_1) \cup \mathbb{V}(S_2)$, so there exist $f \in S_1$ and $g \in S_2$ such that $f(x) \neq 0$ and $g(x) \neq 0$. Then, $(fg)(x) \neq 0$, so $x \notin \mathbb{V}(S_1S_2)$, proving the claim by contraposition.

(ii) We have $x \in \bigcap \mathbb{V}(S_i)$ if and only if x vanishes on every S_i. But this holds if and only if x vanishes on the union of the S_i, i.e., $x \in \mathbb{V}\left(\bigcup_{i\in I} S_i\right)$.

■

An *affine algebraic set* in $\mathbb{A}_k^n$ is the common vanishing locus of some collection $\{F_i\}_{i\in I}$ of polynomials in $k[x_1,\ldots,x_n]$. Note that the indexing set I may not necessarily be finite or even countable.

If k is algebraically closed, then an affine algebraic set of $\mathbb{A}_k^n$ is an (*affine*) *subvariety of* $\mathbb{A}_k^n$.

Example.

(*i*) The entire space $k^n = \mathbb{V}(0)$ is itself an affine algebraic set.

(*ii*) The empty set $\varnothing = \mathbb{V}(1)$ is an affine algebraic set.

(*iii*) Any point $a = (a_1,\ldots,a_n) \in \mathbb{A}^n$ is an affine algebraic set since $\{a\} = \mathbb{V}(x_1 - a_1,x_2 - a_2,\ldots,x_n - a_n) = \mathbb{V}\big(\{x_i - a_i\}_{i=1}^n\big)$.

(*iv*) Any finite subset $S \subseteq \mathbb{A}^n$ is also an affine algebraic set:

$$S = \bigcup_{s\in S}\{s\} = \bigcup_{s\in S}\mathbb{V}(x_i - s_i) = \mathbb{V}\left(\prod_{s\in S}(x_i - s_i)\right)$$

△

In fact, for $\mathbb{A}_k^1$, these are the only affine algebraic sets possible:

Lemma 23.2.2. *The affine algebraic sets of* $\mathbb{A}_k^1$ *are precisely* $\mathbb{A}_k^1$*,* $\varnothing$*, and all finite subsets.*

Proof. If f is the zero polynomial, then $\mathbb{V}(f) = \mathbb{A}^1$. Otherwise, f is some polynomial of degree d. Then, f has at most d roots, so $\mathbb{V}(f) = \{\text{roots of } f\}$ has cardinality at most d and is, in particular, finite. Then, for any collection $\{f_i\}_{i\in I}$,

$$\mathbb{V}\big(\{f_i\}_{i\in I}\big) = \bigcap_{i\in I}\mathbb{V}(f_i)$$

is the intersection of sets that are either all of $\mathbb{A}^1$, finite, or empty and is thus itself either all of $\mathbb{A}^1$, finite, or empty.

The converse statement that $\mathbb{A}^1$, $\varnothing$, and every finite subset is an affine algebraic set is shown in the previous example. ■

Example. Consider the set $S^1 = \big\{\cos(t) + i\sin(t) : t \in \mathbb{R}\big\} \subseteq \mathbb{A}_{\mathbb{C}}^1$. This set is infinite, but is not all of $\mathbb{A}_{\mathbb{C}}^1$, and is thus not a subvariety of $\mathbb{A}_{\mathbb{C}}^1$. △

A *hypersurface* is the vanishing locus $\mathbb{V}(f)$ of a single polynomial in $\mathbb{A}^n$. If $n = 2$, then such a vanishing locus is also called an *affine plane curve.*

Lemma 23.2.3. *The countably infinite union of affine algebraic sets is not necessarily an affine algebraic set.*

Proof. Consider $\mathbb{A}_{\mathbb{C}}^1$. For each integer $a \in \mathbb{Z}$, the singleton $\{a\} = \mathbb{V}(x - a)$ is a subvariety, but the countably infinite union

$$\mathbb{Z} = \bigcup_{a\in\mathbb{Z}}\{a\}$$

is infinite but not all of $\mathbb{A}_{\mathbb{C}}^1$, and is thus not a subvariety. ■

23.2.1 The Zariski Topology

From now on, we assume that k is algebraically closed unless specified otherwise.

Recall that a topology on a set X is a set $T \subseteq \mathcal{P}(X)$ of *open sets* such that

(T1) X is open and $\varnothing$ is open;

(T2) The arbitrary union of open sets is open;

(T3) The finite intersection of open sets is open.

The complement of an open set is called *closed.*

For our purposes, it will be helpful to characterise topologies in terms of closed sets instead. By De Morgan's laws, a topology on X is equivalently a set $T' \subseteq \mathcal{P}(X)$ of closed sets such that

(T1) X is closed and $\varnothing$ is closed;

(T2) The arbitrary intersection of closed sets is closed;

(T3) The finite union of closed sets is closed.

Now, we have seen that $\mathbb{A}^n$ and $\varnothing$ are both subvarieties, and that the arbitrary intersection and finite unions (by induction on binary unions) of subvarieties are subvarieties.

So, the collection of subvarieties of $\mathbb{A}^n$ defines a topology of closed sets on $\mathbb{A}^n$ called the *Zariski topology.*

Compared to the standard topology, non-empty Zariski-open sets are very "large". While the standard topology has a basis consisting of open balls of arbitrarily small radius, every non-empty Zariski-open set is unbounded in the standard topology, and in fact dense in both the Zariski and standard topology. Furthermore, any two non-empty Zariski-open subsets have non-empty intersection, so the Zariski topology is strongly non-Hausdorff.

Note that this definition is also satisfied by affine algebraic sets when k is not algebraically closed, but we will generally be interested in the case of topologies of subvarieties.

Lemma 23.2.4. *Any subset of $\mathbb{R}^n$ or $\mathbb{C}^n$ that is closed in the Zariski topology is also closed in the standard topology.*

Proof. Let S be closed in the Zariski topology, so $S = \mathbb{V}\big(\{f_i\}_{i\in I}\big) = \bigcap_{i\in I} \mathbb{V}(f_i)$. Since polynomials are continuous with respect to the standard topology and $\{0\}$ is closed in the standard topology, $\mathbb{V}(f_i) = f_i^{-1}\big[\{0\}\big]$ is also closed, and hence the intersection S is also closed. ■

Recall that if X is a topological space and $Y \subseteq X$ is a subset, then Y is naturally a topological space under the *subspace topology*, where the open and closed sets of Y are the open and closed sets of X intersected with Y.

In particular, if Y is a subvariety of $\mathbb{A}^n$, then the Zariski-closed subsets of Y are subvarieties of $\mathbb{A}^n$ intersected with Y. Since intersections of subvarieties are subvarieties, the closed subsets of a subvariety Y are precisely the subvarieties of $\mathbb{A}^n$ that are contained in Y.

If Y is a subvariety of $\mathbb{A}^n$, then a *subvariety of* Y is a Zariski-closed subset of Y, or equivalently, a subvariety of $\mathbb{A}^n$ that is contained in Y.

Theorem 23.2.5. *Subvarieties are compact in the Zariski topology.*

23.2.2 Regular Maps

A map $f : \mathbb{A}^n \to \mathbb{A}^m$ is *regular* or is a *morphism of affine space* if every component is a polynomial. That is, there exist $f_1, \ldots, f_m \in k[x_1, \ldots, x_n]$ such that

$$f(x_1, \ldots, x_n) = \big(f_1(x_1, \ldots, x_n), \ldots, f_m(x_1, \ldots, x_n)\big)$$

Example. The projection map $f : \mathbb{A}^2_{x,y} \to \mathbb{A}^2_{x,y}$ defined by $(x,y) \mapsto x$ is a regular map, since the only component x is a polynomial. △

Example. The map $h : \mathbb{A}^1_t \to \mathbb{A}^2_{x,y}$ defined by $t \mapsto (t^2, t^3)$ is a regular map, since the two components t^2 and t^3 are polynomials. △

This definition naturally extends to subvarieties. If $V \subseteq \mathbb{A}^n$ and $W \subseteq \mathbb{A}^n$ are subvarieties, then a map of sets $f : V \to W$ is *regular* or is a *morphism of subvarieties* if it can be expressed as the restriction of a regular map $\mathbb{A}^n \to \mathbb{A}^m$.

Note that the extension $\mathbb{A}^n \to \mathbb{A}^m$ is not necessarily unique.

Example. Let $V = \mathbb{V}(y - x^2) \subseteq \mathbb{A}^2$ and $W = \mathbb{A}^1$. The map $f : V \to W$ defined by $(x,y) \mapsto x$ is a morphism, since it is the restriction of the regular map $\mathbb{A}^2_{x,y} \to \mathbb{A}^1_t : (x,y) \mapsto x$.

It is also the restriction of the map $(x,y) \mapsto x + y - x^2$, since $y - x^2 = 0$ on V. △

Lemma 23.2.6. *Let $F : \mathbb{A}^n_{x_1,\ldots,x_n} \to \mathbb{A}^m_{y_1,\ldots,y_m}$ be regular. Then, for any $g \in k[y_1, \ldots, y_m]$,*

(i) *$g \circ F \in k[x_1, \ldots, x_n]$ is a polynomial;*

(ii) *$F^{-1}\big[\mathbb{V}(g)\big] = \mathbb{V}(g \circ F)$;*

(iii) *$F^{-1}\big[\mathbb{V}\big(\{g_i\}_{i\in I}\big)\big] = \mathbb{V}\big(\{g_i \circ F\}_{i\in I}\big)$.*

Proof.

(i) Since F is regular, its components are polynomials $F_1, \ldots, F_m \in k[x_1, \ldots, x_n]$. Then, the composition is defined by the substitution:

$$g \circ F = g\big(F_1(x_1, \ldots, x_n), \ldots, F_m(x_1, \ldots, x_n)\big)$$

But a polynomial combination of polynomials is again a polynomial, so $g \circ F \in k[x_1, \ldots, x_n]$.

(ii) By definition, $x \in F^{-1}\big[\mathbb{V}(g)\big]$ if and only if $F(x) \in \mathbb{V}(g)$ if and only if $g\big(F(x)\big)$, or equivalently, $x \in \mathbb{V}(g \circ F)$.

(iii) Similar to the previous, $x \in F^{-1}\big[\mathbb{V}\big(\{g_i\}_{i\in I}\big)\big]$ if and only if $F(x) \in \mathbb{V}\big(\{g_i\}_{i\in I}\big)$ if and only if $g_i\big(F(x)\big)$ for all $i \in I$, or equivalently, $x \in \mathbb{V}\big(\{g_i \circ F\}_{i\in I}\big)$.

■

Corollary 23.2.6.1. *The regular preimage of a subvariety of $\mathbb{A}^m$ is a subvariety of $\mathbb{A}^n$.*

Lemma 23.2.7. *Let $X \subseteq \mathbb{A}^n$ and $Y \subseteq \mathbb{A}^m$ be subvarieties, and $F : X \to Y$ be regular. Then, for any subvariety $W \subseteq Y$, $F^{-1}[W]$ is also a subvariety of X.*

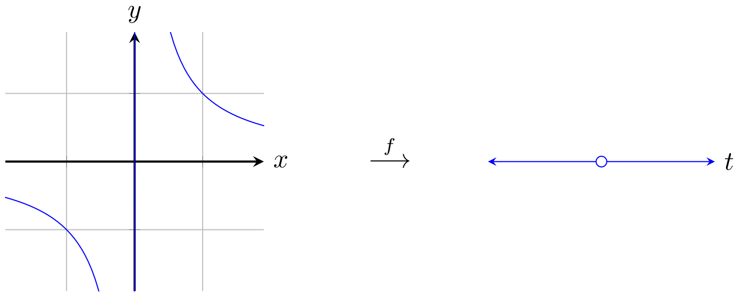
Proof. Since W is a subvariety of Y, it is given by $W = \mathbb{V}\big(\{f_i\}_{i\in I}\big)$. Let $F : \mathbb{A}^n \to \mathbb{A}^m$ be a regular map extending $F : X \to Y$. Then,

$$\begin{aligned} F^{-1}[W] &= \big\{x \in X : F(x) \in \mathbb{V}\big(\{f_i\}_{i\in I}\big)\big\} \\ &= \big\{x \in X : x \in \mathbb{V}\big(\{f_i \circ F\}_{i\in I}\big)\big\} \\ &= X \cap \mathbb{V}\big(\{f_i \circ F\}_{i\in I}\big) \end{aligned}$$

so $F^{-1}[W]$ is a subvariety of X. ■

Note, however, that the regular direct image of a subvariety is not necessarily a subvariety. That is, a regular map need not be a closed map.

Example. Let $V = \mathbb{V}(xy-1)$ be a subvariety of $\mathbb{A}^2_{x,y}$, and let $f : \mathbb{A}^2_{x,y} \to \mathbb{A}^1_t$ be the regular map $(x,y) \mapsto x$. Then, $f(V) = \mathbb{A}^1 \setminus \{0\}$, which is not a subvariety.

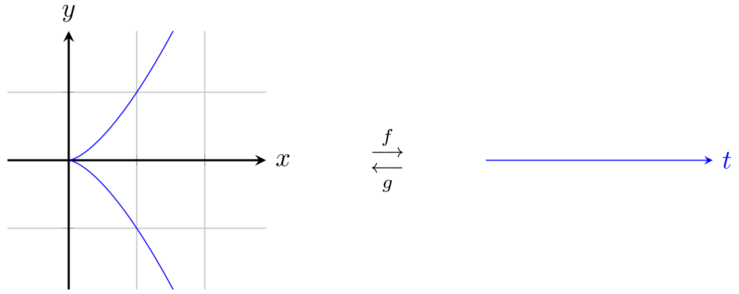


△

A regular map $V \to W$ is furthermore an *isomorphism* (*of subvarieties*) if it has a regular inverse. Two subvarieties are *isomorphic* if there exists an isomorphism between them, and we denote this relation as usual by $V \cong W$.

Example. Consider the subvarieties $V = \mathbb{V}(y - x^2) \subseteq \mathbb{A}^2_{x,y}$ and $W = \mathbb{A}^1_t$. Then, the regular map $f : \mathbb{A}^2_{x,y} \to \mathbb{A}^1_t : (x,y) \mapsto x$ has inverse given by the restriction of the regular map $g : \mathbb{A}^1_t \to \mathbb{A}^2_{x,y} : t \mapsto (t,t^2)$. △

Example. Consider the subvarieties $V = \mathbb{V}(y^2 - x^3) \subseteq \mathbb{A}^2_{x,y}$ and $W = \mathbb{A}^1_t$. The regular map $g : \mathbb{A}^1_t \to \mathbb{A}^2_{x,y}$ defined by $t \mapsto (t^2,t^3)$ is a bijection, and is moreover a homeomorphism in the Zariski topology, but is not an isomorphism of subvarieties.



Since $x = t^2$ and $y = t^3$, the inverse is given by either $\sqrt{x}$, $\sqrt[3]{y}$, or $\frac{y}{x}$; none of which are polynomial. △

23.2.3 Irreducibilty

A topological space X is *reducible* if it is the union of two distinct closed proper subsets, and is *irreducible* otherwise.

Example. The interval $[0,1]$ is reducible in the standard topology, since $[0,1] = [0,\frac{1}{2}],[\frac{1}{2},1]$. △

Example. Any one-point space is irreducible, since any proper subset is necessarily empty. △

By convention, the empty set is considered irreducible.

Lemma 23.2.8. *$\mathbb{A}^1_k$ is irreducible in the Zariski topology for any infinite field k.*

Proof. Any proper closed subset of $\mathbb{A}^1_k$ is either finite or empty, but $\mathbb{A}^1_k$ is infinite, and thus cannot be the union of two such subsets. ■

Theorem 23.2.9. *Every subvariety V of $\mathbb{A}^n$ can be uniquely expressed as the union of finitely many irreducible subvarieties.*

The irreducible subvarieties in this decomposition are called the *irreducible components* of V.

Lemma 23.2.10. *If V is a subvariety and $W \subseteq V$ is an irreducible subvariety, then W is contained in one of the irreducible components of V.*

Proof. If $V = V_1 \cup V_2$ are proper closed subsets and $W \subseteq V$ is closed, then $W = (W \cap V_1) \cup (V \cap V_2)$ are both closed. So W is irreducible only if it is contained entirely within V_1 or V_2. ■

If k is not algebraically closed, then qualitatively different polynomials can have the same vanishing loci. For instance, over $\mathbb{R}$ as *sets*, $\mathbb{V}(x^2 + y^2) = \mathbb{V}(x,y) = \big\{(0,0)\big\} \subseteq \mathbb{A}^2_{\mathbb{R}}$, but we would really like to be able to distinguish these as *subvarieties.*

The problem is that, over $\mathbb{C}$, $\mathbb{V}(x^2 + y^2) \subseteq \mathbb{A}^2_{\mathbb{C}}$ is reducible as $\mathbb{V}(x^2 + y^2) = \mathbb{V}(x + iy) \cup \mathbb{V}(x - iy)$. The regular maps $\mathbb{V}(x^2 + y^2) \to \mathbb{A}^1$ defined by $(x,y) \mapsto 0$ and $(x,y) \mapsto x$ agree as maps of sets over $\mathbb{R}$, but not over $\mathbb{C}$. Even over $\mathbb{R}$, we say that these two maps are distinct *as morphisms.*

The point is that, over an algebraically closed field, the set-theoretic picture faithfully captures the algebro-geometric situation, while over $\mathbb{R}$, it is not enough.

Lemma 23.2.11. *The continuous image of an irreducible space is irreducible.*

Proof. Let X is irreducible and $f : X \to Y$ is continuous and surjective. Suppose that $Y = Y_1 \cup Y_2$ are both closed. Then, since f is continuous, $X = f^{-1}[Y_1] \cup F^{-1}[Y_2]$ are both closed. Since X is irreducible, (at least) one of these must be equal to X, say $f^{-1}[Y_1]$. But then, by the surjectivity of f, $Y = f(X) = Y_1$, and Y is irreducible. ■

Theorem 23.2.12. *Let X be a topological space, and $V \subseteq X$ be a subspace. Then, V is irreducible if and only if $\overline{V}$ is irreducible.*

Let X be a topological space and $V \subseteq X$ be a subspace. Recall that V is *dense* in X if $\overline{V} = X$.

A map $f : X \to Y$ of topological spaces is *dominant* if $f(X)$ is dense in Y.

Example. Consider the map $f : \mathbb{A}^2 \to \mathbb{A}^2$ defined by $(x,y) \mapsto (xy,y)$. Then, $f(\mathbb{A}^2) = \big(\mathbb{A}^2 \setminus \{x\text{-axis}\}\big) \cup \big\{(0,0)\big\}$. This is neither open nor closed, but is dense in $\mathbb{A}^2$, so f is dominant. △

23.2.4 Dimension

Consider the subvariety of $\mathbb{A}^3$ defined by $\mathbb{V}(x^2 + y^2 + z^2 - 1)$. It sees reasonable that the "dimension" of this subvariety should be 2, since it can be thought of as a complex 2-sphere.

What about the subvariety $V = \mathbb{V}(xy,xz) = \mathbb{V}(x) \cup \mathbb{V}(y,z) = \{yz\text{-plane}\} \cup \{x\text{-axis}\}$ of $\mathbb{A}^3$? This variety has two components: the yz-plane, which has dimension 2; and the x-axis which has dimension one. We adopt the convention that the subvariety V should have dimension two.

The *Krull dimension* $\dim(V)$ of a subvariety $V \subseteq \mathbb{A}^n$ is the length d of the longest possible chain $V_0 \subset V_1 \subset \cdots V_{d-1} \subset V_d$ of non-empty irreducible subvarieties of V.

Example. $\mathbb{A}^1$ is 1-dimensional, since its only proper irreducible subvarieties of $\mathbb{A}^1$ are singletons, $\{\text{pt}\} \subseteq \{\text{line}\}$. △

Theorem 23.2.13. *$\mathbb{A}^n$ is n-dimensional.*

Note that if $V_0 \subset \cdots V_d$ is such a maximal chain, then necessarily $\dim(V_k) = k$ and $V_0 = \{\text{pt}\}$. If V is irreducible, then we also have $V_d = V$.

Lemma 23.2.14. *The dimension of a subvariety of $\mathbb{A}^n$ is the maximum dimension of its irreducible components.*

Proof. Let $V = V_1 \cup \cdots \cup V_n$ all be irreducible subvarieties with no containments between any of the V_i. By Lemma 23.2.10, every irreducible subvariety of V lies in one of the V_i, so any chain of irreducible subvarieties of V is also a chain in that V_i. So the dimension of V is at most that of the maximum dimension of the V_i. Conversely, any chain in a V_i is also a chain in V, so the dimension of V is also at least that of the maximum dimension of the V_i. ■

A subvariety is *equidimensional* if all of its irreducible components have the same dimension.

Example. The subvariety $\mathbb{V}(xy,xz) = \mathbb{V}(x) \cup \mathbb{V}(y,z)$ is not equidimensional since $\mathbb{V}(x)$ is 2-dimensional, while $\mathbb{V}(y,z)$ is 1-dimensional. △

Lemma 23.2.15. *If W is a subvariety of V, then* $\dim(W) \leq \dim(V)$.

Proof. Any chain in W is also a chain in V. ■

Lemma 23.2.16. *If $f : X \to Y$ is a surjective regular map of subvarieties, then* $\dim(X) \geq \dim(Y)$.

That is, the dimension of the image is at most the dimension of the source: there are no space-filling curves in algebraic geometry. In fact, we can weaken surjectivity to dominance, and the result still holds:

Theorem 23.2.17. *If $f : X \to Y$ is a dominant regular map of subvarieties, then* $\dim(X) \geq \dim(Y)$.

23.3 Algebraic Foundations

A $\mathbb{C}$-*algebra* is a commutative ring that contains $\mathbb{C}$ as a subring.

Example. Any polynomial ring $R = \mathbb{C}[x_1, \ldots, x_n]$ over $\mathbb{C}$ is a $\mathbb{C}$ algebra since the subspace of contant polynomials in R is isomorphic to $\mathbb{C}$. △

Every $\mathbb{C}$-algebra R is naturally a $\mathbb{C}$-vector space, where the addition of vectors is defined by the addition in R and the multiplication of a scalar in $\mathbb{C}$ by a vector in R is defined by the multiplication in R.

We can define concepts for $\mathbb{C}$-algebras that are analogous to that of rings and ideals:

- The $\mathbb{C}$-subalgebra *generated by* a subset S of a $\mathbb{C}$-algebra R is the set

$$\langle S \rangle = \bigcap_{\substack{S \subseteq A \subseteq R \\ A \text{ is a } \mathbb{C}\text{-algebra}}} A$$

  That is, the $\mathbb{C}$-algebra $\langle S \rangle$ is the smallest $\mathbb{C}$-algebra containing S, or equivalently, the collection of all finite *polynomial* combinations of elements of S with coefficients in R.

- A $\mathbb{C}$-algebra is *finitely generated* if there is a finite set S such that $A = \langle S \rangle$.

  For instance, the polynomial ring $\mathbb{C}[x,y]$ is finitely generated as a $\mathbb{C}$-algebra by the elements x and y.

- A ring homomorphism $\phi : R \to S$ between $\mathbb{C}$-algebras R and S is a $\mathbb{C}$-*algebra homomorphism* if it is additionally $\mathbb{C}$-linear.

Example. The complex conjugate map $z \mapsto \bar{z}$ is a ring homomorphism $\mathbb{C} \to \mathbb{C}$, but is not a $\mathbb{C}$-algebra homomorphism since it is not $\mathbb{C}$-linear. △

If R is a $\mathbb{C}$-algebra and $I \subseteq R$ is an ideal, then R/I is a $\mathbb{C}$-algebra and the quotient map $R \to R/I$ is a $\mathbb{C}$-algebra homomorphism.

Theorem (Universal Property of Polynomial Rings)**.** *Suppose R is a $\mathbb{C}$-algebra and $a_1,\dots,a_n \in R$. Then, there exists a unique $\mathbb{C}$-algebra homomorphism $\phi : \mathbb{C}[x_1,\dots,x_n] \to R$ such that $\phi(x_i) = a_i$.*

Example. Let $R = \mathbb{C}[t]$, and pick $t^2,t^3 \in \mathbb{C}[t]$. Then, there is a unique $\mathbb{C}$-algebra homomorphism $\phi : \mathbb{C}[x,y] \to \mathbb{C}[t]$ such that $\phi(x) = t^2$ and $\phi(y) = t^3$. △

Lemma 23.3.1. *Every finitely generated $\mathbb{C}$-algebra R is the quotient of a polynomial ring.*

Proof. Pick generators $a_1,\dots,a_k$ of R. Then, by the universal property of polynomial rings, there is a unique $\mathbb{C}$-algebra homomorphism $\phi : \mathbb{C}[x_1,\dots,x_k] \to R$ with $\phi(x_i) = a_i$.

Since the a_i generate R, ϕ is surjective, so by the first isomorphism theorem,

$$R = \operatorname{im}\phi \cong \mathbb{C}[x_1,\dots,x_k]/\ker(\phi)$$

■

23.3.1 Hilbert's Basis Theorem

Although the definition of an affine subvariety allows for arbitrarily many polynomials in the vanishing locus, it turns out that every affine subvariety can be expressed as the vanishing locus of only finitely many polynomials. This follows from the *Noetherian* property of polynomials rings.

A ring is *Noetherian* if any of the following equivalent conditions hold:

- Every strictly ascending chain of ideals
$$I_0 \subset I_1 \subset \cdots$$
is finite.
- Every weakly ascending chain of ideals stabilises. That is, for every chain of ideals
$$I_0 \subseteq I_1 \subseteq \cdots$$
there exists $n > 0$ such that
$$I_n = I_{n+1} = I_{n+2} = \cdots$$
- Every ideal is finitely generated.

Lemma 23.3.2. *Every field is Noetherian.*

Lemma 23.3.3. *Let R be Noetherian and $I \subseteq R$ be an ideal. Then, every generating set for I contains a finite generating subset.*

Theorem (Hilbert's Basis Theorem)**.** *If R is Noetherian, then $R[x]$ is Noetherian.*

Corollary 23.3.3.1. *If R is Noetherian, then $R[x_1,\dots,x_n]$ is Noetherian.*

Corollary 23.3.3.2. *$\mathbb{C}[x_1,\dots,x_n]$ is Noetherian.*

23.3.2 Hilbert's Nullstellensatz

The set of polynomials that define a subvariety is not unique. Suppose that f and g vanish on a subvariety V of $\mathbb{A}^n$, and let h be any polynomial in $k[x_1,\dots,x_n]$. Then, $f+g$ and hf also vanish on X. In particular, if $V = \mathbb{V}(S)$, then adding $f+g$ and hf for any polynomial h to S does not change its zero locus. In other words,

Thus, we always have $\mathbb{V}\big(\langle S\rangle\big) = \mathbb{V}(S)$, where $\langle S\rangle$ is the ideal of $k[x_1,\dots,x_n]$ generated by S (that is, the set of all linear combinations of elements in S).

Let V be a subvariety of $\mathbb{A}^n$. Then, the set $\mathbb{I}(V)$ of polynomials that vanish on V is an ideal of $k[x_1,\dots,x_n]$ called the *vanishing ideal* by the same reasoning as above.

$$\mathbb{I}(V) = \big\{f \in k[x_1,\dots,x_n] : \forall x \in V, f(x) = 0\big\}$$

Lemma 23.3.4. *For any subvariety V of $\mathbb{A}^n$, the vanishing ideal $\mathbb{I}(V)$ is a radical ideal of $k[x_1,\dots,x_n]$.*

Proof. Let $f \in k[x_1,\dots,x_n]$ be such that $f^n \in \mathbb{I}(V)$ for some $n > 0$, so $f^n(x) = 0$ for all $x \in V$. Since k is a field, it has no zero divisors, so $f^n(x) = f(x)^n = 0$ if and only if $f(x) = 0$ for all $x \in V$. So $f \in \mathbb{I}(V)$. ■

Theorem 23.3.5. *Every subvariety V is the vanishing locus of finitely many polynomials.*

Proof. Since $V = \mathbb{V}\big(\mathbb{I}(V)\big)$ and $k[x_1,\dots,x_n]$ is Noetherian, $\mathbb{I}(V) = \langle f_1,\dots,f_k\rangle$ is finitely generated, and hence

$$V = \mathbb{V}\big(\mathbb{I}(V)\big) = \mathbb{V}\big(\langle f_1,\dots,f_k\rangle\big) = \mathbb{V}(f_1,\dots,f_k)$$

■

Theorem 23.3.6. *Every subvariety V of $\mathbb{A}^n$ is the intersection of finitely many hypersurfaces.*

Proof. $V = \mathbb{V}\big(\{f_i\}_{i\in I}\big) = \mathbb{V}\big(\langle f_i\rangle_{i\in I}\big)$. By Noetherianness, this ideal is finitely generated, so $\mathbb{V}\big(\langle f_i\rangle_{i\in I}\big) = \mathbb{V}\big(\langle f_1,\dots,f_k\rangle\big) = \bigcap_{i=1}^k \mathbb{V}\big(\langle f_i\rangle\big)$ is a finite intersection of hypersurfaces. ■

Theorem 23.3.7. *For any subvariety V,*

$$\mathbb{V}\big(\mathbb{I}(V)\big) = V$$

Proof. By definition, $V \subseteq \mathbb{V}\big(\mathbb{I}(V)\big)$. Conversely, since V is a subvariety, $V = \mathbb{V}\big(\{f_i\}_{i\in I}\big)$, and by the definition of a vanishing ideal, $f_i \in \mathbb{I}(V)$ for each $i \in I$. Now, for any $x \in \mathbb{V}\big(\mathbb{I}(V)\big)$, x vanishes for each $f \in \mathbb{I}(V)$, and in particular, for each f_i, so $x \in \mathbb{V}\big(\{f_i\}_{i\in I}\big) = V$. ■

So $\mathbb{V}$ is a left inverse to $\mathbb{I}$. What about the other order?

Theorem (Hilbert's Nullstellensatz)**.** *For any ideal $I \subseteq k[x_1,\dots,x_n]$,*

$$\mathbb{I}\big(\mathbb{V}(I)\big) = \sqrt{I}$$

With the previous result, $\mathbb{V}$ and $\mathbb{I}$ are inverse maps when restricted to radical ideals. In this way, Hilbert's Nullstellensatz implies a bijection

$$\big\{\text{affine subvarieties of } \mathbb{A}^n\big\} \underset{\mathbb{V}}{\overset{\mathbb{I}}{\rightleftarrows}} \big\{\text{radical ideals of } k[x_1,\dots,x_n]\big\}$$

So every radical ideal is in fact a vanishing ideal.

If V is a subvariety of W, then the functions vanishing on W also vanish on V. So $\mathbb{I}(V) \subseteq \mathbb{I}(W)$, so this correspondence is order-reversing. More generally, $\mathbb{V}$ is also order-reversing on all ideals, not necessarily radical.

Moreover, this order-reversing correspondence implies that every maximal ideal in $k[x_1,\dots,x_n]$ is the ideal of functions vanishing at a single point $(a_1,\dots,a_n) \in \mathbb{A}^n$. In particular, every maximal ideal has the form $\mathfrak{m}_a = \langle x_1 - a_1, \dots, x_n - a_n\rangle$, and the corresponding subvariety is the singleton $\mathbb{V}(\mathfrak{m}_a) = \{a\} =$

$\big\{(a_1,\dots,a_n)\big\} \subseteq \mathbb{A}^n$. That is, under the above correspondence, the set of maximal ideals of $k[x_1,\dots,x_n]$ is identified with the points of affine n-space $\mathbb{A}^n$.

$$\big\{\text{points of } \mathbb{A}^n\big\} \underset{\mathbb{V}}{\overset{\mathbb{I}}{\rightleftarrows}} \big\{\text{maximal ideals of } k[x_1,\dots,x_n]\big\}$$

Similarly, prime ideals are identified with irreducible subvarieties under this correspondence, and radical ideals correspond to all subvarieties of $\mathbb{A}^n$.

$$\big\{\text{irreducible affine subvarieties of } \mathbb{A}^n\big\} \underset{\mathbb{V}}{\overset{\mathbb{I}}{\rightleftarrows}} \big\{\text{prime ideals of } k[x_1,\dots,x_n]\big\}$$

Lemma 23.3.8. *For any subset $S \subseteq \mathbb{A}^n$,*

$$\mathbb{V}\big(\mathbb{I}(S)\big) = \overline{S}$$

Proof. By definition, every polynomial in $\mathbb{I}(S)$ vanishes everywhere on S, so every point of S vanishes under $\mathbb{I}(S)$, i.e. $S \subseteq \mathbb{V}\big(\mathbb{I}(S)\big)$. Since $\mathbb{V}\big(\mathbb{I}(S)\big)$ is a closed set and $\overline{S}$ is the smallest closed set containing S, $\overline{S} \subseteq \mathbb{V}\big(\mathbb{I}(S)\big)$.

Conversely, let $T = \mathbb{V}\big(\{f_i\}_{i\in I}\big)$ be a closed set containing S, so $f_i \in \mathbb{I}(S)$ for all $i \in I$. Then, $\mathbb{V}\big(\mathbb{I}(S)\big) \subseteq \mathbb{V}\big(\{f_i\}_{i\in I}\big) = T$. Since $\mathbb{V}\big(\mathbb{I}(S)\big)$ is a subset of every closed set T containing S, it is a subset of the closure $\overline{S}$. ■

Corollary 23.3.8.1. *If $V_1 \neq V_2$ are closed, then $\mathbb{I}(V_1) \neq \mathbb{I}(V_2)$.*

That is, subvarieties can be determined by their vanishing ideals.

Corollary 23.3.8.2. $\mathbb{I}(S_1) = \mathbb{I}(S_2)$ *if and only if* $\overline{S_1} = \overline{S_2}$.

Theorem 23.3.9. *Any strictly descending chain of subvarieties of $\mathbb{A}^n_k$ is finite.*

Proof. Let

$$V_1 \supset V_2 \supset V_3 \supset \cdots$$

a descending chain of subvarieties. Then,

$$\mathbb{I}(V_1) \subset \mathbb{I}(V_2) \subset \mathbb{I}(V_3) \subset \cdots$$

is an ascending chain of ideals. Since $k[x_1,\dots,x_n]$ is Noetherian, this chain must be finite, and since $V_i = \mathbb{V}\big(\mathbb{I}(V_i)\big)$, the chain of subvarieties is finite. ■

Lemma 23.3.10. *Given $\{f_i\} \subseteq k[x_1,\dots,x_n]$,*

$$\mathbb{I}\big(\mathbb{V}(\{f_i\}_{i\in I})\big) \supseteq \sqrt{\langle f_i\rangle}$$

Example. Let $k = \mathbb{R}$, and $I = \langle x^2+1\rangle \subseteq \mathbb{R}[x]$. I is a radical ideal, and $\mathbb{V}(x^2+1) = \varnothing$, so $\mathbb{I}\big(\mathbb{V}(x^2+1)\big) = \mathbb{I}(\varnothing) = \mathbb{R}[x] \supseteq \sqrt{I} = I$. △

23.4 The Coordinate Ring

We have seen a correspondence between the geometry of affine n-space $\mathbb{A}^n$ and various ideals of the $\mathbb{C}$-algebra $\mathbb{C}[x_1,\dots,x_n]$. For a subvariety V, what $\mathbb{C}$-algebra describes the subvarieties of V?

Often, to understand an object, we instead study natural classes of functions defined on them. In topology, we study continuous functions on topological spaces; in differential geometry, we study smooth maps on smooth manifolds; and in complex geometry, we study holomorphic maps on complex manifolds. In algebraic geometry, the maps of choice are polynomials.

Let $V \subseteq \mathbb{A}^n$ be an affine subvariety. Given any polynomial $p \in \mathbb{C}[x_1,\dots,x_n]$, the restriction $p\big|_V$ defines a function $V \to \mathbb{C}$. Under the usual pointwise addition and multiplication operations, the set of these functions naturally form a $\mathbb{C}$-algebra $\mathbb{C}[V]$ called the *coordinate ring* of V. In particular, the coordinate ring of the whole affine space $\mathbb{A}^n$ is $\mathbb{C}[\mathbb{A}^n] = \mathbb{C}[x_1,\dots,x_n]$, as expected.

The elements of $\mathbb{C}[V]$ are restrictions of polynomials on $\mathbb{A}^n$, but we usually denote them by the original polynomials. This may be slightly confusing, as two ostensibly different polynomials may agree when restricted to V.

Example. Consider the variety $V = \mathbb{V}(y-x)$ of $\mathbb{A}^n$. Then, the polynomials $xy+1$, x^2+1, and y^2+1 are all the same polynomial on V, since $y = x$ on V. △

Restriction defines a surjective ring homomorphism $\mathbb{C}[x_1,\dots,x_n] \to \mathbb{C}[V]$. By definition, the kernel of this homomorphism is precisely the vanishing ideal $\mathbb{I}(V)$, so by the first isomorphism theorem,

$$\mathbb{C}[V] \cong \frac{\mathbb{C}[x_1,\dots,x_n]}{\mathbb{I}(V)}$$

Which $\mathbb{C}$-algebras are coordinate rings?

Recall that a ring R is *reduced* if for all elements $x \in R$, whenever $x^n = 0$ for some $n > 0$, then $x = 0$. That is, 0 is the only nilpotent element in R.

Lemma 23.4.1. *An ideal I is radical if and only if R/I is reduced.*

Theorem 23.4.2. *R is a finitely generated reduced $\mathbb{C}$-algebra if and only if R is the coordinate ring of some affine subvariety V of affine space.*

Proof. As the quotient of a finitely generated $\mathbb{C}$-algebra $\mathbb{C}[x_1,\dots,x_n]$ by a radical ideal $\mathbb{I}(V)$, coordinate rings are finitely generated reduced $\mathbb{C}$-algebras.

Now, suppose R is a finitely generated reduced $\mathbb{C}$-algebra. Pick generators $a_1,\dots,a_n \in R$. Then, there exists a unique $\mathbb{C}$-algebra homomorphism $\Phi : \mathbb{C}[x_1,\dots,x_n] \to R$ with $\Phi(x_i) = a_i$. Because the a_i generate R, Φ is surjective, so by the first isomorphism theorem,

$$R \cong \frac{\mathbb{C}[x_1,\dots,x_n]}{\ker(\Phi)}$$

Since R is reduced, $\ker(\Phi)$ is radical and hence determines a subvariety $V = \mathbb{V}\big(\ker(\Phi)\big)$ of $\mathbb{A}^n$. So, $\ker(\Phi) = \mathbb{I}(V)$ and hence R is the coordinate ring $\mathbb{C}[V]$. ■

Specifically, the relations between the generators $a_1,\dots,a_n$ determine the ideal $\mathbb{I}(V)$. Selecting a different set of generators $b_1,\dots,b_m \in R$ would then yield a different subvariety W of $\mathbb{A}^m$ with $R \cong \mathbb{C}[W]$. We will soon define a notion of isomorphism between subvarieties that identify W and V.

Example. Let $R = \mathbb{C}[t]$ and consider the generators $x = t$ and $y = 1$. These generators satisfy the relation $y - 1 = 0$, so

$$R \cong \frac{\mathbb{C}[x,y]}{\langle y-1\rangle}$$

This choice of generators corresponds to the subvariety $V = \mathbb{V}(y-1) \subseteq \mathbb{A}^2_{x,y}$.

If we instead choose the generators $a = t^2$, $b = 2t$, and $c = 1-t$, the relations are then $a - b - c^2 + 1 = 0$, so

$$R \cong \frac{\mathbb{C}[a,b,c]}{\langle a-b-c^2+1\rangle}$$

This choice of generators corresponds to the subvariety $W = \mathbb{V}(a-b-c^2+1) \subseteq \mathbb{A}^3_{a,b,c}$. △

So far, we have been viewing subvarieties as certain subsets of affine n-space $\mathbb{A}^n$. However, as we have seen above, we cannot precisely recover a subvariety V from the coordinate ring $\mathbb{C}[V]$, and can only do so up to isomorphism.

The subvarieties V and W are all isomorphic and embed the same subvariety $\mathbb{A}^1$ into $\mathbb{A}^2$ and $\mathbb{A}^3$ respectively. So, we want a definition that captures the notion of a subvariety is an entity by itself that does not depend on an ambient embedding.

Before, we have seen a correspondence between certain geometric features of $\mathbb{A}^n$ and the polynomial ring $\mathbb{C}[x_1,\dots,x_n] = \mathbb{C}[\mathbb{A}^n]$. We will now establish an analogous correspondence between geometric features of an arbitrary affine subvariety V and its coordinate ring $\mathbb{C}[V]$.

Recall that, given an ideal I of R, the quotient map $\pi : R \to R/I$ induces a bijection

$$\begin{aligned} \{\text{ideals of } R/I\} &\xrightarrow{\cong} \{\text{ideals of } R \text{ containing } I\} \\ J &\mapsto \pi^{-1}[J] \end{aligned}$$

Since $\frac{R/I}{J} \cong \frac{R}{\pi^{-1}[J]}$, this bijection sends prime, maximal, and radical ideals to prime, maximal, and radical ideals, respectively.

$$\begin{aligned}
\big\{\text{ideals of } \mathbb{C}[V]\big\} &\cong \begin{Bmatrix}\text{ideals of} \\ \mathbb{C}[x_1,\dots,x_n] \text{ containing } I\end{Bmatrix} \\
\begin{Bmatrix}\text{radical} \\ \text{ideals of } \mathbb{C}[V]\end{Bmatrix} &\cong \begin{Bmatrix}\text{radical ideals of} \\ \mathbb{C}[x_1,\dots,x_n] \text{ containing } I\end{Bmatrix} \cong \begin{Bmatrix}\text{all} \\ \text{subvarieties } W \subseteq V\end{Bmatrix} \\
\begin{Bmatrix}\text{prime} \\ \text{ideals of } \mathbb{C}[V]\end{Bmatrix} &\cong \begin{Bmatrix}\text{prime ideals of} \\ \mathbb{C}[x_1,\dots,x_n] \text{ containing } I\end{Bmatrix} \cong \begin{Bmatrix}\text{irreducible} \\ \text{subvarieties } W \subseteq V\end{Bmatrix} \\
\begin{Bmatrix}\text{maximal} \\ \text{ideals of } \mathbb{C}[V]\end{Bmatrix} &\cong \begin{Bmatrix}\text{maximal ideals of} \\ \mathbb{C}[x_1,\dots,x_n] \text{ containing } I\end{Bmatrix} \cong \begin{Bmatrix}\text{points} \\ (a_1,\dots,a_n) \in V\end{Bmatrix}
\end{aligned}$$

Thus, the Zariski topology on V, the points of V, and the subvarieties of V are all encoded in the $\mathbb{C}$-algebra structure of $\mathbb{C}[V]$. This gives us a way to think of a subvariety V independently from its original construction as a subset of $\mathbb{A}^n$.

An *affine variety* is a subvariety of $\mathbb{A}^n$ together with its Zariski topology and coordinate ring.

23.4.1 The Pullback Homomorphism

Just as each affine variety induces a unique $\mathbb{C}$-algebra as its coordinate ring, every morphism of affine varieties determines a unique $\mathbb{C}$-algebra homomorphism between their coordinate rings *in reverse direction.*

Given any morphism $F : V \to W$, there is a map of coordinate rings $\mathbb{C}[W] \to \mathbb{C}[V]$ defined by precomposition by F:

$$\begin{aligned} \mathbb{C}[W] &\to \mathbb{C}[V] \\ g &\mapsto g \circ F \end{aligned}$$

called the *pullback* of F, denoted by $F^\sharp$.

Lemma 23.4.3. *For any morphism $F : V \to W$, the pullback map $F^\sharp : \mathbb{C}[W] \to \mathbb{C}[V]$ is a $\mathbb{C}$-algebra homomorphism.*

Example. Consider the morphism F of affine varieties defined by

$$\begin{aligned} \mathbb{A}^3_{x,y,z} &\to \mathbb{A}^2_{u,v} \\ (x,y,z) &\mapsto (x^2y,x-z) \end{aligned}$$

The pullback $F^\sharp$ is then defined by

$$\begin{aligned} \mathbb{C}[u,v] &\to \mathbb{C}[x,y,z] \\ u &\mapsto x^2y \\ v &\mapsto x-z \end{aligned}$$

△

Lemma 23.4.4. *Given distinct points $p,q \in V$, there exists a polynomial $g \in \mathbb{C}[V]$ such that $g(p) \neq g(q)$.*

Proof. Since $V \subseteq \mathbb{A}^n_{x_1,\ldots,x_n}$, if $p \neq q$, then they must differ in some coordinate x_i. Then, the polynomial x_i will do. ■

Theorem 23.4.5. *Given morphisms $F,G : V \to W$, $F = G$ if and only if $F^\sharp = G^\sharp$.*

Proof. The forward direction is obvious. Conversely, suppose $F \neq G$, so there exists $x \in V$ such that $F(x) \neq G(x)$. By the previous lemma, there is a polynomial $g \in \mathbb{C}[W]$ such that

$$F^\sharp(g)(x) = g\big(F(x)\big) \neq g\big(G(x)\big) = G^\sharp(g)(x)$$

so $F^\sharp(g) \neq G^\sharp(g)$, and hence $F^\sharp \neq G^\sharp$. ■

23.4.2 The Equivalence of Algebra and Geometry

Lemma 23.4.6. *Given $F : V \to W$ and $G : W \to X$, $(G \circ F)^\sharp = F^\sharp \circ G^\sharp$.*

Theorem 23.4.7. *Every homomorphism $\sigma : S \to R$ of finitely generated reduced $\mathbb{C}$-algebras can be realised essentially uniquely as the pullback of a morphism $F : V \to W$ of affine varieties.*

That is, there exist affine varieties V and W with identifications $\mathbb{C}[V] \cong C$ and $\mathbb{C}[W] \cong S$ and a morphism $F : V \to W$ such that under these identifications, $F^\sharp = \sigma$:

$$\begin{tikzcd}
\mathbb{C}[V] \arrow[dr, leftrightarrow, "\cong"] & & & \mathbb{C}[W] \arrow[lll, "F^\sharp"'] \arrow[dl, leftrightarrow, "\cong"'] \\
& S & R \arrow[l, "\sigma"'] &
\end{tikzcd}$$

Furthermore, the choices of V, W, and F are unique up to unique isomorphism, so if there exist V' and W' with identifications $\mathbb{C}[V'] \cong R$ and $\mathbb{C}[W] \cong S$, and a morphism $F' : V' \to W'$ such that under these identifications $(F')^\sharp = \sigma$, then there exist unique isomorphisms $\alpha : V \to V'$ and $\beta : W \to W'$ such that the following diagram commutes:

$$\begin{tikzcd}
\mathbb{C}[V] \arrow[dr, leftrightarrow, "\cong"] & & & \mathbb{C}[W] \arrow[lll, "F^\sharp"'] \arrow[dl, leftrightarrow, "\cong"'] \\
& S & R \arrow[l, "\sigma"'] & \\
\mathbb{C}[V'] \arrow[uu, dashed, "\alpha^\sharp"] \arrow[ur, leftrightarrow, "\cong"'] & & & \mathbb{C}[W'] \arrow[lll, "(F')^\sharp"] \arrow[uu, "\beta^\sharp"'] \arrow[ul, leftrightarrow, "\cong"]
\end{tikzcd}$$

Corollary 23.4.7.1. *$V \cong W$ if and only if $\mathbb{C}[V] \cong \mathbb{C}[W]$.*

In summary, just as geometry determines algebra, algebra also determines geometry; every finitely generated reduced $\mathbb{C}$-algebra is equivalent to an affine variety V via $R \cong \mathbb{C}[V]$, and every homomorphism $\phi : S \to R$ of $\mathbb{C}$-algebras is equivalent to a morphism $F : V \to W$ of affine varieties by pullback.

23.5 The Spectrum of a Ring

We have seen how $\mathbb{C}[V]$ determines V up to isomorphism; by picking n generators for $\mathbb{C}[V]$, we obtain an embedding $V \hookrightarrow \mathbb{A}^n$. However, we can reconstruct V from $\mathbb{C}[V]$ "abstractly", not as a subset of $\mathbb{A}^n$.

The *maximal spectrum* of a commutative ring R is the set of maximal ideals of R:

$$\operatorname{maxSpec}(R) \coloneqq \{\mathfrak{m} \subset R : \mathfrak{m} \text{ is a maximal ideal}\}$$

So, by Hilbert's Nullstellensatz, there is a bijection

$$\begin{aligned} V &\xrightarrow{\cong} \operatorname{maxSpec}\big(\mathbb{C}[V]\big) \\ a &\mapsto \mathfrak{m}_x = \mathbb{I}\big(\{a\}\big) \end{aligned}$$

We can transport the Zariski topology on V to a topology on $\operatorname{maxSpec}\big(\mathbb{C}[V]\big)$ as follows. The points of a Zariski-closed set $W \subseteq V$ correspond to the set of maximal ideals in $\mathbb{C}[V]$ that contain $\mathbb{I}[W]$, so the closed sets of $\operatorname{maxSpec}\big(\mathbb{C}[V]\big)$ are sets of maximal ideals of $\mathbb{C}[V]$ containing some given ideal of $\mathbb{C}[V]$.

Given any commutative ring R and an ideal I of R, we define a notion of a vanishing locus on the maximal spectrum as:

$$\mathbb{V}^{\text{ms}}(I) \coloneqq \big\{\mathfrak{m} \in \operatorname{maxSpec}(R) : \mathfrak{m} \supseteq I\big\} \subseteq \operatorname{maxSpec}(R)$$

The Zariski topology on $\operatorname{maxSpec}(R)$ is the topology whose closed sets are precisely the sets of the form $\mathbb{V}^{\text{ms}}(I)$ for I an ideal of R.

This gives us a way to define affine varieties without reference to any embeddings in affine space; for any finitely generated reduced $\mathbb{C}$-algebra R, $\operatorname{maxSpec}(R)$ is the corresponding affine variety.

Note, however, that the definition of a maximal spectrum and vanishing locus apply to any commutative ring, and not only finitely generated reduced $\mathbb{C}$-algebras.

Example. Let $R = \mathbb{Z}$. Then,

$$\operatorname{maxSpec}(R) = \big\{\langle p\rangle : p \text{ is prime}\big\}$$

Since $\mathbb{Z}$ is a principle ideal domain, every ideal is of the form $I = \langle n\rangle$, so

$$\begin{aligned} \mathbb{V}^{\text{ms}}(I) &= \big\{\langle p\rangle : \langle m\rangle \subseteq \langle p\rangle\big\} \\ &= \big\{\langle p\rangle : p \mid m\big\} \end{aligned}$$

△

We have seen that every morphism $V \to W$ induces a $\mathbb{C}$-algebra homomorphism $\mathbb{C}[W] \to \mathbb{C}[V]$ by pullback. Does every morphism also induce a morphism $\operatorname{maxSpec}\big(\mathbb{C}[V]\big) \to \operatorname{maxSpec}\big(\mathbb{C}[W]\big)$ of maximal spectra?

Lemma 23.5.1. *Let $F : V \to W$ be a morphism, and let $x \in V$. Then,*

$$(F^\sharp)^{-1}(\mathfrak{m}_x) = \mathfrak{m}_{F(x)}$$

Given a $\mathbb{C}$-algebra homomorphism $\sigma : \mathbb{C}[W] \to \mathbb{C}[V]$, we define the map of maximal spectra

$$\begin{aligned} \sigma^\flat : \operatorname{maxSpec}\big(\mathbb{C}[V]\big) &\to \operatorname{maxSpec}\big(\mathbb{C}[W]\big) \\ \mathfrak{m} &\mapsto \sigma^{-1}[\mathfrak{m}] \end{aligned}$$

Then, for any morphisms $F : V \to W$, the following diagram commutes:

$$\begin{tikzcd} \operatorname{maxSpec}\big(\mathbb{C}[V]\big) \arrow[r, "(F^\sharp)^\flat"] & \operatorname{maxSpec}\big(\mathbb{C}[W]\big) \\ V \arrow[u, "\cong"] \arrow[r, "F"'] & W \arrow[u, "\cong"'] \end{tikzcd}$$

So, morphisms of affine varieties are recoverable purely algebraically from maximal spectra of finitely generated reduced $\mathbb{C}$-algebras.

It seems that affine varieties are well-described by maximal spectra, and we may think to generalise this theory to maximal spectra of arbitrary commutative rings. Unfortunately, while the maximal spectrum construction still yields a topological space, we run into trouble when constructing maps between maximal spectra.

For instance, given a ring homomorphism $\sigma : R \to S$, we would like

$$\begin{aligned} \sigma^\flat : \operatorname{maxSpec}(S) &\to \operatorname{maxSpec}(R) \\ \mathfrak{m} &\mapsto \sigma^{-1}[\mathfrak{m}] \end{aligned}$$

to be a well-defined continuous map of topological spaces. However, the preimage of a maximal ideal under an arbitrary ring homomorphism is not necessarily a maximal ideal.

Example. Let $\sigma : \mathbb{Z} \hookrightarrow \mathbb{Q}$ be the inclusion map. As a field, the trivial ideal is maximal, but the preimage of the trivial ideal in $\mathbb{Q}$ is the trivial ideal in $\mathbb{Z}$, which is not maximal. △

Theorem 23.5.2. *The preimage of a prime ideal under a ring homomorphism is prime.*

Proof. Let $\phi : R \to S$ be a ring homomorphism, I be an ideal of S, and $J = \phi^{-1}[I]$. Suppose that $J = R$. Then, $1_R \in J$, so $\phi(1_R) = 1_S \in I$, so $I = S$, contradicting that I is prime. So J is a proper ideal.

Now, suppose $ab \in J$, so $\phi(ab) = \phi(a)\phi(b) \in I$. Since I is prime, $\phi(a) \in I$ or $\phi(b) \in I$, so $a \in J$ or $b \in J$. So J is prime. ■

The *spectrum* of a commutative ring is the set of *prime* ideals of R:

$$\operatorname{Spec}(R) \coloneqq \{\mathfrak{p} \subset R : \mathfrak{p} \text{ is a prime ideal}\}$$

Again, we define a notion of a vanishing locus on the spectrum as:

$$\mathbb{V}^{\mathrm{s}}(I) \coloneqq \big\{\mathfrak{p} \in \operatorname{Spec}(R) : \mathfrak{p} \supseteq I\big\} \subseteq \operatorname{Spec}(R)$$

and the Zariski topology on $\operatorname{Spec}(R)$ is again the topology whose closed sets are precisely the sets of the form $\mathbb{V}^{\mathrm{s}}(I)$ for I an ideal of R.

The spectrum of a ring, equipped with its Zariski topology, is what Grothendieck called an *affine scheme*.

While the maximal spectrum $\operatorname{maxSpec}\big(\mathbb{C}[V]\big)$ of a coordinate ring $\mathbb{C}[V]$ is canonically isomorphic to V, the spectrum $\operatorname{Spec}\big(\mathbb{C}[V]\big)$ contains more information and is canonically isomorphic to the set of irreducible subvarieties of V.

We write V^{sch} to abbreviate $\operatorname{Spec}\big(\mathbb{C}[V]\big)$.

Example. TODO △

23.6 Morphisms of Affine Schemes

Since primeness of ideals is preserved under ring homomorphism preimages, given a ring homomorphism $\sigma : S \to R$, we can again define a map of spectra:

$$\begin{aligned} \sigma^\flat : \operatorname{Spec}(R) &\to \operatorname{Spec}(S) \\ \mathfrak{p} &\mapsto \sigma^{-1}[\mathfrak{p}] \end{aligned}$$

Lemma 23.6.1. *$\sigma^\flat$ is continuous with respect to the Zariski topology.*

A *morphism of affine schemes* is the data of a ring homomorphism $\sigma : S \to R$ inducing a map $\sigma^\flat : \operatorname{Spec}(R) \to \operatorname{Spec}(S)$

An *affine scheme over* $\mathbb{C}$ is a spectrum $\operatorname{Spec}(R)$ of a not necessarily reduced or finitely generated $\mathbb{C}$-algebra R.

23.7 Projective Varieties

Let k be any field. Then, *n-dimensional projective space over k*, denoted $\mathbb{P}^n_k$ is the set of 1-dimensional subspaces of k^{n+1}. We will write $\mathbb{P}^n$ for n-dimensional complex projective space $\mathbb{P}^n_{\mathbb{C}}$.

Projective n-space can also be interpreted as the quotient

$$\mathbb{P}^n = \frac{\mathbb{C}^{n+1} \setminus \big\{(0,\dots,0)\big\}}{\sim}$$

where $\sim$ identifies two points that lie on the same line through the origin. That is, $(x_0,\dots,x_n) \sim (y_0,\dots,y_n)$ if and only if there exists a non-zero scalar $\lambda \in \mathbb{C}$ such that $(y_0,\dots,y_n) = \lambda(x_0,\dots,x_n)$.

A point in this space can then an equivalence class

$$\big[(z_0,\dots,z_n)\big] = \big\{(\lambda z_0,\dots,\lambda z_n) : \lambda \in \mathbb{C}\big\}$$

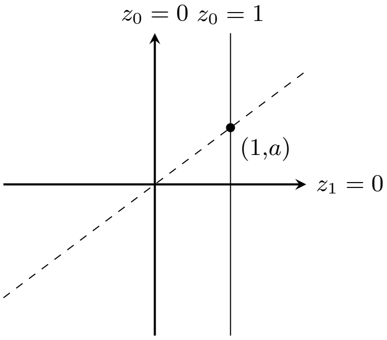
where at least one of the coordinates $z_0,\dots,z_n$ must be non-zero. We denote a representative of the equivalence class of a point $(z_0,\dots,z_n)$ by $[z_0 : \dots : z_n]$, called a *homogeneous coordinate*. This notation emphasises that homogeneous coordinates are really just ratios of coordinates, and are defined only up to non-zero scaling: $[z_0 : \dots : z_n] = [\lambda z_0 : \dots : \lambda z_n]$ for any non-zero $\lambda \in \mathbb{C}$.

$$\mathbb{P}^n = \big\{[z_0 : \ldots : z_n] : \exists i, z_i \neq 0\big\}$$

Example. $[1 : 2] = [\frac{1}{2} : 1] = [i : 2i] \in \mathbb{P}^1$. These all represent the line $\big\{(z_0,z_1) : z_1 = 2z_0\big\} \subseteq \mathbb{C}^2$. △

0-dimensional projective space $\mathbb{P}^0$ is the set of all complex lines through the origin in $\mathbb{C}^1$, of which there is only $\mathbb{C}^1$ itself, so $\mathbb{P}^0 = \{\mathbb{C}^1\}$ is a singleton.

1-dimensional projective space $\mathbb{P}^1$ is the set of all complex lines through the origin in $\mathbb{C}^2$. By fixing a reference line – any complex line not through the origin, say $z_0 = 1$ – we can choose a representative for almost every point as the unique point on the reference line where the line through the origin intersects the reference line. Only one point in $\mathbb{P}^1$ will fail to have a representative under this scheme, namely the unique line through the origin parallel to the reference line, called the *point at infinity.*



Each line L_a of slope a can be represented by homogeneous coordinate:

$$\begin{aligned} L_a &= \big\{(z_0,az_0)\big\} \\ &= [z_0 : az_0] \\ &= [1 : a] \end{aligned}$$

while the vertical line $z_0 = 0$ has coordinate:

$$\begin{aligned} L_\infty &= \{z_0 = 0\} \\ &= [0 : z_0] \\ &= [0 : 1] \end{aligned}$$

So, we have a bijection

$$\begin{aligned} \mathbb{P}^1 \setminus \big\{[0 : 1]\big\} &\cong \{\text{coordinates with } z_0 = 1\} \\ L_a = [z_0,az_0] &\mapsto [1,a] \end{aligned}$$

and by discarding the first coordinate, we have a further isomorphism (as affine varieties):

$$\begin{aligned} \{\text{coordinates with } z_0 = 1\} &\cong \mathbb{A}^1_{\mathbb{C}} \\ [1,a] &\mapsto a \end{aligned}$$

or directly,

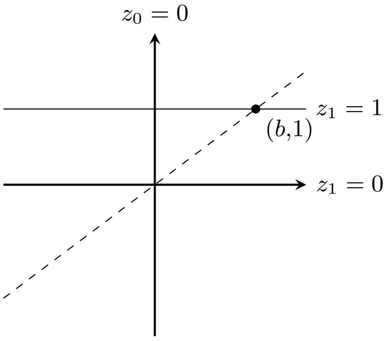
$$\begin{aligned} \mathbb{P}^1 \setminus \big\{[0 : 1]\big\} &\cong \mathbb{A}^1_{\mathbb{C}} \\ [z_0,z_1] &\mapsto \tfrac{z_1}{z_0} \end{aligned}$$

Adding in the point at infinity, this identifies $\mathbb{P}^1$ with the Riemann sphere

$$\mathbb{P}^1 \cong \mathbb{C} \cup \{\infty\}$$

$$[z_0,z_1] \mapsto \begin{cases} \frac{z_1}{z_0} & z_0 \neq 0 \\ \infty & z_0 = 0 \end{cases}$$

We could have also fixed the horizontal line $z_1 = 1$:



This time, every line L_b with slope $\frac{1}{b}$ through the origin is of the form $[b : 1]$, while the horizontal line $z_1 = 0$ has coordinate $[1 : 0]$.

In $\mathbb{P}^2$, we can similarly select a reference plane not passing through the origin, and identify lines through the origin with their intersection with the reference plane. The exceptions will be the lines through the origin parallel to the reference plane – that is, a copy of $\mathbb{P}^1$.

$$\mathbb{P}^2 \cong \mathbb{C}^2 \cup \mathbb{P}^1 = \mathbb{C}^2 \cup \mathbb{C} \cup \{\infty\}$$

For instance, if we have coordinates z_0,z_1,z_2 for $\mathbb{C}^3$, and select the reference plane $z_0 = 1$, the identification sends the homogeneous coordinate $[z_0 : z_1 : z_2] \in \mathbb{P}^2$ to $(\frac{z_1}{z_0}, \frac{z_2}{z_0}) \in \mathbb{C}^2$ whenever $z_0 \neq 0$, and to $[z_1 : z_2] \in \mathbb{P}^1$ when $x_0 = 0$.

Generalising this to arbitrary dimensions,

$$\mathbb{P}^n \cong \mathbb{C}^n \cup \mathbb{P}^{n-1}$$
$$[z_0 : z_1 : \cdots : z_n] \mapsto \begin{cases} (\frac{z_1}{z_0}, \ldots, \frac{z_n}{z_0}) & z_0 \neq 0 \\ [z_1 : \cdots : z_n] & z_0 = 0 \end{cases}$$

We define the set $\mathcal{U}_0$ to be the set of points for which z_0 is non-zero. Under the above mapping, U_0 is identified with the hyperplane $z_0 = 1$ in $\mathbb{C}^{n+1}$, which can be identified with $\mathbb{C}^n$:

$$[z_0 : z_1 : \cdots : z_n] = \left[1 : \frac{z_1}{z_0} : \cdots : \frac{z_n}{z_0}\right] \mapsto \left(1, \frac{z_1}{z_0}, \ldots, \frac{z_n}{z_0}\right) \cong \left(\frac{z_1}{z_0}, \ldots, \frac{z_n}{z_0}\right)$$

The remaining points for which $z_0 = 0$ are then the points at infinity; the lines through the origin in $\mathbb{C}^{n+1}$ parallel to the hyperplane $z_0 = 1$, isomorphic to $\mathbb{P}^{n-1}$.

The choice of z_0 in the above is arbitrary. We define the set $\mathcal{U}_i$ to be the subset of $\mathbb{P}^n$ of points with $z_i \neq 0$

$$\mathcal{U}_i = \big\{[z_0 : \cdots : z_n] \in \mathbb{P}^n : z_i \neq 0\big\} = \mathbb{P}^n \setminus \{\text{1-dimensional subspaces of } \mathbb{V}(z_0)\}$$

isomorphic to $\mathbb{A}^n$ by dividing by and discarding the ith component:

$$\psi_i : \mathcal{U}_i \to \mathbb{A}^n$$
$$[z_0 : \cdots : z_n] \mapsto \left(\frac{z_0}{z_i}, \ldots, \frac{\widehat{z_i}}{z_i}, \ldots \frac{z_n}{z_i}\right)$$

This isomorphism is called the *ith affine chart on* $\mathbb{P}^n$.

Example. In $\mathbb{P}^1$, we have

$$\mathcal{U}_0 = \big\{[z_0 : z_1] : z_0 \neq 0\big\} = \big\{[1 : a] : a \in \mathbb{C}\big\} \cong \mathbb{A}^1_a$$
$$\mathcal{U}_1 = \big\{[z_0 : z_1] : z_1 \neq 0\big\} = \big\{[b : 1] : b \in \mathbb{C}\big\} \cong \mathbb{A}^1_b$$

△

The collection of all the affine charts yields a cover of $\mathbb{P}^n$ by $n + 1$ copies of $\mathbb{A}^n$:

$$\mathbb{P}^n = \bigcup_{i=0}^{n} \mathcal{U}_i$$

Since $\mathbb{P}^n$ is the quotient of $\mathbb{C}^{n+1} \setminus \{0\}$, this $\mathbb{P}^n$ inherits a standard quotient topology from $\mathbb{C}^{n+1}$. In this topology, the affine charts are open.

The open cover $\{\mathcal{U}_i\}$ of $\mathbb{P}^n$ defines an atlas making projective space a complex n-dimensional manifold; we can move between charts via the *transition functions*

$$\psi_j \circ \psi_i^{-1} : \psi_i(\mathcal{U}_i \cap \mathcal{U}_j) \to \psi_j(\mathcal{U}_i \cap \mathcal{U}_j)$$

and these are not only holomorphic, but in fact rational. For instance,

$$\psi_n \circ \psi_0^{-1}(a_1,\ldots,a_n) = \psi_n\big([1 : a_1 : \cdots : a_n]\big) = \left(\frac{1}{a_n},\frac{a_1}{a_n},\ldots,\frac{a_{n-1}}{a_n}\right)$$

Example. On the intersection $\mathcal{U}_0 \cap \mathcal{U}_1 \subseteq \mathbb{P}^1$, both components of a homogeneous coordinate are non-zero, and we can translate between the two as:

$$\mathbb{A}^1 \cong \mathcal{U}_0 \ni a \mapsto [1 : a] = [\tfrac{1}{a} : 1] = \tfrac{1}{a} \in \mathcal{U}_1 \cong \mathbb{A}^1$$

so $b = \frac{1}{a}$, and hence $\mathcal{U}_0 \cap \mathcal{U}_1 \cong \mathbb{A}^1 \setminus \{0\}$. △

23.7.1 Projective Varieties

An element $[z_0 : \cdots : z_n] \in \mathbb{P}^n$ has many representations, given by scaling every component by some $\lambda \neq 0$, so something like:

$$\big\{[z_0 : \cdots : z_n] : z_0 = 5\big\}$$

is *not* well-defined. However, if $z_i = 0$ then $\lambda z_i = 0$ for any λ, so this is well-defined, like the affine charts.

Similarly, given any non-constant polynomial $f \in \mathbb{C}[z_0,\ldots,z_n]$, the value of $f(z_0,\ldots,z_n)$ and $f(\lambda z_1,\ldots,\lambda z_n)$ may not agree, so f does not define a function $\mathbb{P}^n \to \mathbb{C}$ since its value depends on the choice of homogeneous coordinates.

A polynomial is *homogeneous* if all of its terms have the same total degree.

Example. The polynomial $x^2y + 2z^3 \in \mathbb{C}[x,y,z]$ is homegeneous of degree 3. △

Lemma 23.7.1. *If $f(z_0,\ldots,z_n)$ is homogeneous of degree d, then*

$$f(\lambda z_0,\ldots,\lambda z_n) = \lambda^d f(z_0,\ldots,z_n)$$

Example. If $f(x,y,z) = x^2y + z^3$, then

$$\begin{aligned} f(\lambda x,\lambda y,\lambda z) &= (\lambda x)^2(\lambda y) + (\lambda z)^3 \\ &= \lambda^3 x^2 y + \lambda z^3 \\ &= \lambda^3 f(x,y,z) \end{aligned}$$

△

Corollary 23.7.1.1. *If f is homogeneous and $p \in \mathbb{P}^n$, then either for all choices of representatives $[z_0 : \cdots : z_n]$ of p, $f(z_0,\ldots,z_n) = 0$; or for all choices of representatives $[z_0 : \cdots : z_n]$ of p, $f(z_0,\ldots,z_n) \neq 0$.*

That is, a homogeneous polynomials only vanish along lines through the origin; if a homegeneous polynomial vanishes at a point, it must vanish along the entire line through the origin containing that point. Thus, the set of zeros in $\mathbb{C}^{n+1}$ of a homogeneous polynomial is the union of complex lines through the origin. So, while a homogeneous polynomial in $n+1$ variables does not define a function on $\mathbb{P}^n$, it still makes sense to talk about its vanishing locus in $\mathbb{P}^n$.

If $f \in \mathbb{C}[z_0,\ldots,z_n]$ is a homogeneous polynomial, then we define its projective vanishing locus to be

$$\mathbb{V}(f) = \big\{[z_0 : \ldots : z_n] : f(z_0,\ldots,z_n) = 0\big\} \subseteq \mathbb{P}^n$$

A *subvariety of* $\mathbb{P}^n$ is the vanishing locus $\mathbb{V}\big(\{f_i\}_{i\in I}\big)$ of some collection $\{f_i\}_{i\in I} \subseteq \mathbb{C}[z_0,\ldots,z_n]$ of homogeneous polynomials in $n+1$ variables.

A *projective variety* is a closed subvariety of some $\mathbb{P}^n$.

The Zariski topology on $\mathbb{P}^n$ is then the topology in which the closed subsets are precisely the subvarieties of $\mathbb{P}^n$.

Example. Consider the projective subvariety $V = \mathbb{V}(x^2 + y^2 - z^2) \subseteq \mathbb{P}^2$. We can write it as the union of its coordinate charts:

$$V = (V \cap \mathcal{U}_x) \cup (V \cap \mathcal{U}_y) \cup (V \cap \mathcal{U}_z)$$

On the chart $\mathcal{U}_z$ defined by $z \neq 0$, the variety looks like a complex circle; identifying $\mathcal{U}_z$ with $\mathbb{C}^2$, the curve in $\mathcal{U}_z$ is defined by the vanishing locus of $x^2 + y^2 - 1$, while on the charts $\mathcal{U}_x$ and $\mathcal{U}_y$, the curves are given by $1 + y^2 - z^2 = 0$ and $x^2 + 1 - z^2 = 0$, respectively. △

As in the example, the intersection of any projective variety V with one of the affine charts of $\mathbb{P}^n$ is an affine variety. Specifically, if $\mathcal{U}_i$ is an open set of $\mathbb{P}^n$ where the component z_i is non-zero, isomorphic to $\mathbb{A}^n$, then setting the variable z_i to 1 in the defining polynomials for V yields a set of defining polynomials for $V \cap \mathcal{U}_i$. So, just as projective space is covered by affine charts, we can think of a projective variety as being covered by affine varieties.

Another way to visualise a projective variety in $\mathbb{P}^n$ is to imagine a cone-shaped variety in $\mathbb{C}^{n+1}$, but then to identify all points lying on the same line through the origin. The variety in $\mathbb{C}^{n+1}$ defined by a collection of homogeneous polynomials is then called the *affine cone* over the projective variety in $\mathbb{P}^n$ defined by the same homogeneous polynomials.

Given a not-necessarily homogeneous polynomial $f \in \mathbb{C}[z_0, \dots, z_n]$, we say that f *vanishes* at a point $p \in \mathbb{P}^n$ or write $f(p) = 0$, if $f(z_0, \dots, z_n)$ for all choices of representative $[z_0 : \cdots : z_n] = p$, or equivalently, if the line $L_p \subseteq \mathbb{C}^{n+1}$ corresponding to p is entirely contained within the affine vanishing locus $\mathbb{V}(f) \subseteq \mathbb{C}^{n+1}$.

Any polynomial $f \in \mathbb{C}[z_0, \dots, z_n]$ can be expressed uniquely as a sum

$$f = f_0 + f_1 + \cdots + f_d$$

where f_i is homogeneous of degree i. The polynomial f_i is called the *ith homogeneous component* of f.

Theorem 23.7.2. *Let $f \in \mathbb{C}[z_0, \dots, z_n]$ and $p \in \mathbb{P}^n$ such that $f(p) = 0$. Then, for each homogeneous component f_i of f, $f_i(p) = 0$*

Proof. Pick a representative $p = [p_0 : \cdots : p_n]$, so

$$\begin{aligned} 0 &= f(p_0, \dots, p_n) \\ &= f_0(p_0, \dots, p_n) + \cdots + f_d(p_0, \dots, p_n) \end{aligned}$$

Since $f(p) = 0$, f also vanishes at $\lambda[p_0 : \cdots : p_n]$ for any non-zero $\lambda \in \mathbb{C}$:

$$\begin{aligned} 0 &= f(\lambda p_0, \dots, \lambda p_n) \\ &= \lambda^0 f_0(p_0, \dots, p_n) + \cdots + \lambda^d f_d(p_0, \dots, p_n) \end{aligned}$$

This is a polynomial in λ that vanishes for all non-zero λ, and must therefore be the zero polynomial. So the coefficients $f_i(p_0, \dots, p_n)$ must all also vanish. Since the choice of representative was arbitrary, $f_i(p) = 0$. ■

An ideal $I \subseteq \mathbb{C}[z_0, \dots, z_n]$ is *homogeneous* if it can be generated by homogeneous polynomials, or equivalently, whenever $f \in I$, each homogeneous component of f is also in I.

Suppose $V \subseteq \mathbb{P}^n$ is a projective variety. Then, the set

$$\{f \in \mathbb{C}[z_0, \dots, z_n] : \forall p \in V, f(p) = 0\}$$

is called the *homogeneous vanishing ideal of* V and is denoted by $\mathbb{I}(V)$.

Lemma 23.7.3. *$\mathbb{I}(V)$ is a homogeneous radical ideal of $\mathbb{C}[z_0, \dots, z_n]$.*

Theorem 23.7.4 (Projective Nullstellensatz). *There is an inclusion-reversing bijective correspondence*

$$\begin{aligned}\{\textit{projective varieties in } \mathbb{P}^n\} &\cong \{\textit{radical homogeneous ideals in } \mathbb{C}[z_0,\dots,z_n]\} \setminus \{\langle z_0,\dots,z_n\rangle\} \\ V &\mapsto \mathbb{I}(V) \\ \mathbb{V}(I) &\leftarrow\!\shortmid I\end{aligned}$$

Given a projective variety $V \subseteq \mathbb{P}^n$, its *homogeneous coordinate ring* is the quotient

$$\frac{\mathbb{C}[z_0,\dots,z_n]}{\mathbb{I}(V)}$$

Note that this is equal to the *affine* coordinate ring of the affine cone over V. Elements of this ring are also *not* functions on V, and are instead functions on the affine cone over V. This ring also depends on the embedding of V in $\mathbb{P}^n$, and not just the isomorphism class of V.

Given any subset $X \subseteq \mathbb{P}^n$, X inherits the Zariski topology as the subspace topology, where closed subsets of X are sets of the form $X \cap V$ for $V \subseteq \mathbb{P}^n$ a projective variety.

23.7.2 Homogenisation

Given any polynomial $f \in \mathbb{C}[x_1,\dots,x_n]$ of degree d in n variables, we can *homogenise* the polynomial into a homogeneous polynomial $f^+ \in \mathbb{C}[z_0,\dots,z_n]$ of degree d in $n+1$ variables by "padding" lower degree components with a new variable.

Decomposing f into its homogeneous components,

$$f = f_0 + f_1 + \cdots + f_{d-1} + f_d$$

the homogenisation f^+ is given by multiplying each f_i by z_0^{d-i}, and replacing each x_i with z_i:

$$f^+ = z_0^d f_0 + z_0^{d-1} f_1 + \cdots + z_0 f_{d-1} + f_d$$

Example. The polynomial $2 + 3x_1 + 4x_1^2x_2 + 5x_2^3$ is degree 3. The homogenisation is given by

$$2z_0^3 + 3z_0^2z_1 + 4z_1^2z_2 + 5z_2^3$$

△

Equivalently, we can replace each x_i with $\frac{z_i}{z_0}$, then multiply through by z_0^d.

We can also *dehomogenise* a homogeneous polynomial $f \in \mathbb{C}[z_0,\dots,z_n]$ with respect to a variable z_i to obtain a polynomial in $f^\circ \in \mathbb{C}[x_1,\dots,x_n]$ by evaluating $f(z_0,\dots,z_n)$ at $(x_1,\dots,z_i = 1,\dots,x_n)$. This corresponds to restricting the homogenised polynomial to the affine chart $\mathcal{U}_i$.

The dehomogenisation of a degree d homogeneous polynomial f (with respect to z_0 for example) can be seen as two steps:

$$f(z_0,\dots,z_n) \overset{\substack{\text{divide}\\ \text{by } z_0^d}}{\mapsto} \frac{f(z_0,\dots,z_n)}{z_0^d} \overset{x_i=\frac{z_i}{z_0}}{\mapsto} \text{polynomial in } x_1,\dots,x_n$$

Thus, the following are equivalent:

- $[z_0 : \cdots : z_n] \in \mathbb{V}(f) \cap \mathcal{U}_0$;
- $(\frac{z_1}{z_0},\dots,\frac{z_n}{z_0}) = (x_1,\dots,z_n) \in \mathbb{A}^n$;
- $f(z_0,\dots,z_n) = 0$ and $z_0 \neq 0$;

- $\frac{1}{z_0^d}f(z_0,\dots,z_n) = 0$;
- $f(1,\frac{z_1}{z_0},\dots,\frac{z_n}{z_0}) = 0$;
- $f(1,x_1,\dots,x_n) = 0$;
- $f^\circ(x_1,\dots,x_n) = 0$.

So, there is a bijection

$$\mathbb{V}(f) \cap \mathcal{U}_0 \underset{\cong}{\xrightarrow{\psi_0}} \mathbb{V}(\text{dehomogenisation of } f \text{ with respect to } z_0)$$

Lemma 23.7.5. *If $\{F_\alpha\}$ is a set of homogeneous polynomials in $z_0,\dots,z_n$ and f_α is the dehomogenisation of F_α with respect to z_i, then*

$$\psi_i : \mathbb{V}\big(\{F_i\}\big) \cap \mathcal{U}_i \to \mathbb{V}\big(\{f_i\}\big)$$

is a bijection.

Corollary 23.7.5.1. *If $V \subseteq \mathbb{P}^n$ is a projective subvariety, then $V \cap \mathcal{U}_i \subseteq \mathcal{U} \cong \mathbb{A}^n$ is an affine subvariety of $\mathbb{A}^n$ under the identification $\psi_i : \mathcal{U}_i \xrightarrow{\cong} \mathbb{A}^n$.*

Lemma 23.7.6. *For any $f \in \mathbb{C}[x_1,\dots,x_n]$,*

$$(f^+)^\circ = f$$

where the dehomogenisation is with respect to z_0.

Corollary 23.7.6.1. *Given $\{f_\alpha\} \subseteq \mathbb{C}[x_1,\dots,x_n]$ and $W = \mathbb{V}\big(\{f_\alpha\}\big) \subseteq \mathbb{A}^n$, define $V = \mathbb{V}\big(\{f_\alpha^+\}\big) \subseteq \mathbb{P}^n$. Then,*

$$\psi_0(V \cap \mathcal{U}_0) = W$$

So, the affine subvarieties W of $\mathcal{U}_0 \cong \mathbb{A}^n$ are precisely the sets $V \cap \mathcal{U}_0$, where $V \subseteq \mathbb{P}^n$ is a projective subvariety. In other words, the Zariski topology on $\mathcal{U}_0$ as a subspace topology of $\mathbb{P}^n$ is the same as the Zariski topology on $\mathbb{U}_0 \cong \mathbb{A}^n$.

Note that the dehomogenisation map $\{\text{homogeneous polynomials in} z_0,\dots,z_n\} \to \mathbb{C}[x_1,\dots,x_n]$ is not injective, since

$$(z_i^k F)^\circ = F^\circ$$

where the dehomogenisation is with respect to z_i; extra factors of the new variable are discarded under dehomogenisation.

Lemma 23.7.7. *Let $F \in \mathbb{C}[z_0,\dots,z_n]$ be a homogeneous polynomial, and suppose $F = z_0^k G$ where $z_0 \nmid G$. Then,*

$$(F^\circ)^+ = G$$

Corollary 23.7.7.1. *Two homogeneous functions F_1 and F_2 have equal dehomogenisations with respect to z_0 if an only if there exists a polynomial G such that $z_0 \nmid G$ and $F_1 = z_0^k G$ and $F_2 = z_0^\ell G$ for some $k,\ell \in \mathbb{N}$.*

23.7.3 Projective Closures

Let V be an affine variety, with a fixed embedding $V \subseteq \mathbb{A}^n \subseteq \mathbb{P}^n$. The *projective closure* $\overline{V}$ of V is the closure of V in $\mathbb{P}^n$. The closure may be computed in either the Zariski or standard topology on $\mathbb{P}^n$; the result will be the same.

Given an affine variety $V = \mathbb{V}(F_1,\dots,F_r) \subseteq \mathbb{A}^n$, we'd might think that the projective closure $\overline{V}$ of V in $\mathbb{P}^n$ might be defined by the ideal obtained by replacing each of the polynomials F_i with its homogenisation F^+.

Example. Consider the parabola $V = \mathbb{V}(y - x^2) \subseteq \mathbb{A}^2 \subseteq \mathbb{P}^2$. The variables x and y are the affine coordinates for V in $\mathbb{A}^2$, while in $\mathbb{P}^2$, we use homogeneous coordinates x, y, and z, and identify $\mathbb{A}^2$ with the open affine chart $\mathcal{U}_z$ where z is non-zero (say, $z = 1$).

The points of the parabola in $\mathbb{P}^2$ are then the lines through the origin in $\mathbb{C}^3$ connecting to the points on the parabola in the plane $z = 1$, i.e. picture the affine cone modulo scaling. There is a line "missing" from this cone – namely the y-axis where the two branches of the parabola asymptotically converge together.

As a projective variety, the parabola is described by $yz - x^2$ in $\mathbb{P}^2$, so the projective closure of $y - x^2$ is $yz - x^2$. △

In this case, the closure is indeed given by the homogenisation. However, this does not work in general.

Example. Let $T = \mathbb{V}(y - x^2, z - xy) \subseteq \mathbb{A}^n$. We have $z - xy = z - x^3$, so the points of this variety are of the form (x, x^2, x^3), so T is the *twisted cubic*, i.e. the image of the map $\mathbb{A}^1_t \to \mathbb{A}^3_{x,y,z} : t \mapsto (t, t^2, t^3)$. What is the projective closure of T?

The homogenisations of the polynomials are given $wy-x^2$ and $wz-xy$. Let $V = \mathbb{V}(wy-x^2, wz-xy) \subseteq \mathbb{P}^3$. By construction, $V \cap \mathcal{U}_w = T$ (i.e. set $w = 1$). What about $V \setminus T$?

First,

$$V \setminus T = V \setminus (V \cap \mathcal{U}_w) = V \setminus \mathcal{U}_w = V \cap \mathbb{V}(w)$$

so, setting $w = 0$ in the defining polynomials for V, we have $wy - x^2 = -x^2 = 0$ and $wz - xy = -xy = 0$, so $x = 0$.

So, V also contains the point $[0 : y : z : 0]$ where $x = 0 = w$. However, the set $W = T \cup \{[0 : 0 : 1 : 0]\}$ is a closed set strictly contained in V, so V is not the minimal closed set containing T, i.e., $V \neq \overline{T}$. △

Theorem 23.7.8. *Let $V \subseteq \mathbb{A}^n \subseteq \mathbb{P}^n$ be an affine variety, and let $I = \mathbb{I}(V) \subseteq \mathbb{C}[x_1, \ldots, x_n]$ be the radical ideal of all polynomials vanishing on V. Then, the ideal $J = \langle f^+ \mid f \in I\rangle$ of $\mathbb{C}[z_0, \ldots, z_n]$ generated by the homogenisations of all the elements of I is the radical homogeneous ideal of polynomials vanishing on the projective closure $\overline{V}$ in $\mathbb{P}^n$.*

The ideal J is called the *homogenisation* of the ideal I.

The problem in the previous example is that we only homogenised the polynomials $y - x^2$ and $z - xy$. The previous theorem says that if we instead homogenise all the polynomials in the ideal closure of $y - x^2$ and $z - xy$, the generated ideal would then correspond to the projective closure.

23.7.4 Morphisms of Projective Varieties

Consider the map $f : \mathbb{P}^1_{[s:t]} \to \mathbb{P}^2_{[x:y:z]}$ defined by $[s : t] \mapsto [s^2 : st : t^2]$.

This map is well-defined since

$$[s : t] = [\lambda s : \lambda t] \mapsto [\lambda^2 s^2 : \lambda^2 st : \lambda^2 t^2] = [s^2 : st : t^2]$$

and since $[s : t] \in \mathbb{P}^1_{[s:t]}$, s and t cannot simultaneously vanish, so the first and last coordinate of $[s^2 : st : t^2]$ cannot simultaneously vanish, so f does not map onto the origin. More generally, any map between projective spaces is well-defined if it is given in coordinates by homogeneous polynomials of the same degree with empty common vanishing loci.

Since in the image of f, we have $x = s^2$, $y = st$, and $z = t^2$, the coordinates satisfy the relation $xz = y^2$, so the image of f lies on the curve $C = \mathbb{V}(xz - y^2)$ in $\mathbb{P}^2$. Let us examine f on affine charts.

If $s \neq 0$, then $x = s^2 \neq 0$, so $f|_{\mathcal{U}_s} \subseteq \mathcal{U}_x$. Similarly, $f|_{\mathcal{U}_t} \subseteq \mathcal{U}_z$.

On $\mathcal{U}_s$, we have $s \neq 0$, so $x = s^2 \neq 0$ and $f|_{\mathcal{U}_s} \subseteq \mathcal{U}_x$:

$$f|_{\mathcal{U}_s} : \mathcal{U}_s \to \mathcal{U}_x$$

$$[s:t] = [1:\tfrac{t}{s}] \mapsto [1:\tfrac{t}{s},\tfrac{t^2}{s^2}]$$

Identifying $\mathcal{U}_s$ with $\mathbb{A}^1_a$ and $\mathcal{U}_x$ with $\mathbb{A}^2_{u_1,u_2}$, this map is:

$$\begin{aligned} f|_{\mathcal{U}_s} : \mathbb{A}^1_a &\to \mathbb{A}^2_{u_1,u_2} \\ a &\mapsto (a,a^2) \end{aligned}$$

This is a morphism of affine varieties, whose image is the parabola $\mathbb{V}(u_2 - u_1^2) \cong C \cap \mathcal{U}_x$ in the plane.

Similarly, on $\mathcal{U}_t$, $f|_{\mathcal{U}_t} : \mathbb{A}^1_b \to \mathbb{A}^2_{v_1,v_2}$ is described by $b \mapsto (b^2,b)$. Again, the image is a parabola $\mathbb{V}(v_2^2 - v_1)$ in the plane.

Thus, $f : \mathbb{P}^1 \to C$ restricts localy on the coordinate charts covering $\mathbb{P}^1$ to a morphism of affine varieties. This motivates the following definition.

Let $V \subseteq \mathbb{P}^n$ and $W \subseteq \mathbb{P}^m$ be projective varieties. A map of sets $F : V \to W$ is a *morphism of projective varieties* if F is locally a polynomial map at every point of V. That is, for each $p \in V$, there exists an open neighbourhood $U \subseteq V$ of p and homogeneous polynomials $F_0,\ldots,F_m \in \mathbb{C}[z_0,\ldots,z_n]$ such that

- The F_i do not simultaneously vanish on U;
- The restriction $F|_U : U \to W$ agrees with the map $U \to \mathbb{P}^m$ defined by:

$$[z_0 : \cdots : z_n] \mapsto \big[F_0(z_0,\ldots,z_n) : F_1(z_0,\ldots,z_n) : \cdots : F_m(z_0,\ldots,z_n)\big]$$

A morphism of projective varieties is, as usual, an *isomorphism* if it has an map that is also a morphism. Two projective varieties are *isomorphic* if there exists an isomorphism between them.

This definition is compatible with the definition of morphisms for affine varieties.

Example. The simplest example of a isomorphism is given by a change of coordinates in $\mathbb{P}^n$. Let $A = (A_{ij})$ be a full-rank $(n+1) \times (n+1)$ matrix. Then, $A : \mathbb{C}^{n+1} \to \mathbb{C}^{n+1}$ is a linear automorphism, and permutes the 1-dimensional subspaces of $\mathbb{C}^{n+1}$, thus inducing an automorphism $\mathbb{P}^n \to \mathbb{P}^n$:

$$[z_0 : \ldots : z_n] \mapsto \left[\sum_j A_{0j} z_j : \cdots : \sum_j A_{nj} z_j\right]$$

△

Theorem 23.7.9. *Every automorphism of $\mathbb{P}^n$ arises this way, i.e. is a linear automorphism.*

Corollary 23.7.9.1. *If $\lambda \in \mathbb{C}$ is a non-zero scalar, then A and λA induce the same automorphism of $\mathbb{P}^n$.*

Two projective varieties $V,W \subseteq \mathbb{P}^n$ are *projectively equivalent* if there exists an automorphism of $\mathbb{P}^n$ that restricts to an isomorphism $V \to W$.

Theorem 23.7.10. *If V and W are projectively equivalent, then their homogeneous coordinate rings are isomorphic.*

Theorem 23.7.11. *If $F,G \in \mathbb{C}[x,y,z]$ are irreducible homogeneous polynomials of degree 2, then $\mathbb{V}(F) \subseteq \mathbb{P}^2$ and $\mathbb{V}(G) \subseteq \mathbb{P}^2$ are projectively equivalent.*

Let $A = (A_{ij})$ be an $(m+1) \times (n+1)$ matrix with trivial nullspace (so $m \geq n$), representing an injective linear map $\mathbb{C}^{n+1} \to \mathbb{C}^{m+1}$. This induces a morphism $\mathbb{P}^n \to \mathbb{P}^m$ that is linear in homogeneous coordinates.

This map is an embedding, i.e., is an isomorphism on to a closed subvariety of $\mathbb{P}^m$, and the image is a linear subvariety, i.e. is cut out by homogeneous polynomials of degree 1.

Theorem 23.7.12. *If $n > m$, then there do not exist any non-constant morphisms $\mathbb{P}^n \to \mathbb{P}^m$. If $n \leq m$, then there exist non-linear morphisms $\mathbb{P}^n \to \mathbb{P}^m$.*

23.8 Quasiprojective Varieties

A *locally closed* subset of a topological space X is the intersection of an open and closed subset, or equivalently, a closed subset of an open subset.

A *quasiprojective variety* is a locally closed subset of $\mathbb{P}^n$. A quasiprojective variety inherits the Zariski topology from $\mathbb{P}^n$.

Example. The following are quasiprojective:

- $\mathbb{P}^n = \mathbb{P}^n \cap \mathbb{P}^n$;
- $\mathbb{A}^n = \mathcal{U}_0 \cap \mathbb{P}^n$;
- Any projective variety $W \subseteq \mathbb{P}^n$, $W = \mathbb{P}^n \cap W$;
- Any closed affine variety $V \subseteq \mathbb{A}^n$, since any affine variety can be viewed as an open subset of its affine cone, $V = \mathcal{U}_0 \cap \overline{V}$.
- Any open set $X \subseteq \mathbb{P}^n$, since $X = X \cap \mathbb{P}^n$;
- Any open set $Y \subseteq \mathbb{A}^n$, since $Y \subseteq \mathbb{P}^n$ is also open.
- $U = \mathbb{A}^1_t \setminus \{0\}$ is a quasiprojective variety. To see this, embed X into $\mathcal{U}_s \subseteq \mathbb{P}^1_{[t:s]}$, i.e. along the line $s = 1$ via $t \mapsto [t : 1]$. Because U is missing the origin, we remove $[0 : 1]$ from $\mathbb{P}^1$, and also, the point at infinity $[1 : 0]$ has no preimage in $\mathbb{A}^1$, so it too is removed. So $U = \mathbb{P}^1_{[t:s]} \setminus \big\{[0 : 1],[1 : 0]\big\}$ is open in $\mathbb{P}^1$, so $U = U \cup \mathbb{P}^1$ is locally closed.

△

The definition of a morphism of quasiprojective varieties is the same as for projective varieties:

Let $X \subseteq \mathbb{P}^n$ and $Y \subseteq \mathbb{P}^m$ be quasiprojective varieties. A map of sets $F : X \to Y$ is a *morphism of quasiprojective varieties* if F is locally a polynomial map at every point of V. That is, for each $p \in V$, there exists an open neighbourhood $U \subseteq X$ of p and homogeneous polynomials $F_0, \dots, F_m \in \mathbb{C}[z_0, \dots, z_n]$ such that

- The F_i do not simultaneously vanish on U;
- The restriction $F|_U : U \to W$ agrees with the map $U \to \mathbb{P}^m$ defined by:

$$[z_0 : \cdots : z_n] \mapsto \big[F_0(z_0, \dots, z_n) : F_1(z_0, \dots, z_n) : \cdots : F_m(z_0, \dots, z_n)\big]$$

Example. Let $X = \mathbb{A}^1_t \setminus \{0\}$ and $Y = \mathbb{V}(xy - 1) \subseteq \mathbb{A}^2_{x,y} \cong \mathbb{V}(xy - z^2) \cap \mathcal{U}_z \subseteq \mathbb{P}^2$. Both X and Y are quasiprojective varieties, and we have a well-defined map

$$\begin{aligned} F : X &\to Y \\ t &\mapsto (t, \tfrac{1}{t}) \end{aligned}$$

We claim that this is a morphism of quasiprojective varieties. To see this, we embed X into $\mathbb{P}^1$ via $t \mapsto [t : 1]$ (as before), and Y into $\mathbb{P}^2$ via $(x,y) \mapsto [x : y : 1]$ (since z does not vanish in $\mathbb{V}(xy - z^2) \cap \mathcal{U}_z$). Then, F agrees everywhere on U with the morphism

$$\begin{aligned} \tilde{F} : \mathbb{P}^1 &\to \mathbb{P}^2 \\ [t : s] &\mapsto [t^2 : s^2 : st] \end{aligned}$$

On $U = \mathbb{P}^1_{[t:s]} \setminus \big\{[0 : 1],[1 : 0]\big\}$, neither t nor s vanish, so setting $t = a/b$, we see

$$U \ni t = [t : 1] \overset{\tilde{F}}{\mapsto} [t^2 : 1 : t] = [t : \tfrac{1}{t} : 1] = (t, \tfrac{1}{t}) \in Y$$

which agrees with F. △

Every morphism of projective varieties is a morphism of quasiprojective varieties, since the definition is effectively identical, but also every morphism of affine varieties is a morphism of quasiprojective varieties.

Having defined quasiprojective varieties, we now redefine the concept of an affine variety.

A quasiprojective variety is *affine* if it is isomorphic to a closed subset of affine space, i.e. to a subvariety of $\mathbb{A}^n$.

Example. The open set $X = \mathbb{A}^1 \backslash \{0\} \subseteq \mathbb{A}^1_t$ is an affine variety, because it is isomorphic as a quasiprojective variety to $Y = \mathbb{V}(xy-1) \subseteq \mathbb{A}^2_{x,y}$: the projection map $G : X \to Y : (x,y) \mapsto x$ is a morphism of quasiprojective varieties and is the inverse of the map $F : U \to V$ defined above. △

The *coordinate ring* $\mathbb{C}[X]$ of an affine quasiprojective variety is the $\mathbb{C}$-algebra $\mathbb{C}[V]$, where $V \subseteq \mathbb{A}^n$ is closed (i.e. is a affine subvariety) and $X \cong V$ as quasiprojective varieties. That is, if $F : X \to V$ is an isomorphism, then the coordinate ring $\mathbb{C}[X]$ is the ring of functions $W \to \mathbb{C}$ that are pullbacks of functions in $\mathbb{C}[V]$.

Example. $\mathbb{C}\big[\mathbb{A}^1 \setminus \{0\}\big] = \frac{\mathbb{C}[x,y]}{\langle xy-1 \rangle} \cong \mathbb{C}[t,t^{-1}]$. △

Similarly, a quasiprojective variety is *projective* if it is isomorphic to a closed subset of projective space, i.e. to a subvariety of $\mathbb{P}^n$. Unlike the case for affine varieties, this redefinition does not enlarge the class of projective varieties.

Theorem 23.8.1. *If X is both affine and projective, then X is isomorphic to a finite set of points.*

Theorem 23.8.2. *If $X \subseteq \mathbb{P}^n$ is a quasiprojective variety and there exists a closed set $Y \subseteq \mathbb{P}^m$ with $X \cong Y$ as quasiprojective varieties, then X is closed in $\mathbb{P}^n$.*

23.8.1 Quasiprojective Varieties are Locally Affine

The Zariski topology for any quasiprojective varieties has a basis of open affine sets.This allows us to think of every quasiprojective variety as "locally affine", in the same way that every manifold is "locally Euclidean". That is, each point in a quasiprojective variety has an open neighbourhood that is an affine subvariety.

First, observe that the complement of any hypersurface in an affine variety is again an affine variety. Specifically, if V is a Zariski-closed subset of $\mathbb{A}^n$ and $f \in \mathbb{C}[V]$, then the open set $U = V \setminus \mathbb{V}(f)$ is an affine variety (though not usually a closed set/affine subvariety of V). To see this, consider the map

$$\begin{aligned} F : U &\to \mathbb{A}^{n+1}_{x_1,\dots,x_n,z} \\ (x_1,\dots,x_n) &\mapsto \left(x_1,\dots,x_n,\frac{1}{f(x_1,\dots,x_n)}\right) \end{aligned}$$

Since f does not vanish on U by definition, this map is well-defined. Moreover, if $x_1,\dots,x_n,z$ denote the coordinates for $\mathbb{A}^{n+1}$, the original defining polynomials $F_1(x_1,\dots,x_n),\dots,F_r(x_1,\dots,x_n)$ for V in $\mathbb{A}^n$ all vanish at the image points of F, as does the polynomial $zf(x_1,\dots,x_n)-1$. So, the image of F is contained in the Zariski-closed subset of $\mathbb{A}^{n+1}$ defined by $W = \mathbb{V}(F_1,\dots,F_r,zf-1)$, and the map $U \to \mathbb{V}(F_1,\dots,F_r,zf-1) \subseteq \mathbb{A}^{n+1}$ is an isomorphism of quasiprojective varieties. So, $V \setminus \mathbb{V}(g) \cong W$ is an affine quasiprojective variety.

Lemma 23.8.3. *The open sets of the form*

$$V \setminus \mathbb{V}(g)$$

where $g \in \mathbb{C}[V]$ is non-zero and non-unit form a basis for the Zariski topology on V.

These sets are called *basic affine open sets* .

Theorem 23.8.4. *There is a basis for the Zariski topology on every quasiprojective variety $V \subseteq \mathbb{P}^n$ consisting of basic affine open sets.*

Corollary 23.8.4.1. *Quasiprojective varieties are also locally affine.*

23.8.2 Regular Functions

Regulan functions are the generalisation of polynomial functions on affine varieties to the case of quasiprojective varieties.

While manifolds locally look like Euclidean space $\mathbb{R}^n$, quasiprojective varieties locally look like affine varieties. The existence of a basis of basic affine open sets means that we can view every variety as a union of affine varieties, and so we can define a regular function locally as a function that restricts on each affine patch to a polynomial function.

Let V be a Zariski-closed subset of $\mathbb{A}^n$. Given $f,g \in \mathbb{C}[V]$, the rational expression $\frac{f}{g}$ is locally well-defined on $V \setminus \mathbb{V}(g)$.

Theorem 23.8.5. *If $W \cong V \setminus \mathbb{V}(g)$, then*

$$\mathbb{C}[W] \cong \mathbb{C}[V][\tfrac{1}{g}] \cong \frac{\mathbb{C}[V][z]}{\langle zg - 1\rangle}$$

On the chart $V \setminus \mathbb{V}(g)$, the function $\frac{1}{g}$ is identified with the polynomial z on $\mathbb{A}^{n+1}$, and the function $\frac{f}{g}$ is identified with the polynomial zf on $\mathbb{A}^{n+1}$. We now extend this definition to affine varieties that are not necessarily closed in $\mathbb{A}^{n+1}$.

Let U be any open subset of a Zariski-closed subset V of affine space. A function $F :\to \mathbb{C}$ is *regular at* $p \in U$ if there exist $f,g \in \mathbb{C}[V]$ such that

- $g(p) \neq 0$;
- there exists an open neighbourhood $W \subseteq U$ of p such that g is non-zero on W, and $F|_W = \frac{f}{g}$.

F is *regular on* U if it is regular at every point $p \in U$.

Example. The slope function

$$\begin{aligned} f : U = \mathbb{A}^2 \setminus \mathbb{V}(x) &\to \mathbb{C} \\ (x,y) &\mapsto \frac{y}{x} \end{aligned}$$

is regular on U. △

We define $\mathcal{O}_V(U)$ to be the set of regular functions $U \to \mathbb{C}$:

$$\mathcal{O}_V(U) = \{F : U \to \mathbb{C} : F \text{ regular}\} \subseteq \mathbb{C}[V]$$

This set is a $\mathbb{C}$-algebra under pointwise addition, multiplication, and scaling of functions.

Example. Let $X = \mathbb{A}^2$ and $Y = \mathbb{A}^2 \setminus \mathbb{V}(x)$. Then, $\frac{1}{x}, \frac{y}{x} \in \mathcal{O}_V(U) = \mathbb{C}[x,y,\frac{1}{x}]$. △

Note that the restriction $\mathbb{C}[V] \to \mathcal{O}_V(U)$ is injective if U is dense in V; in particular, if V is irreducible and U is non-empty.

Theorem 23.8.6. *The inclusion $\mathbb{C}[V] \hookrightarrow \mathcal{O}_V(V)$ is surjective and is hence an isomorphism.*

This is non-obvious, saying that every locally rational function is in fact globally polynomial.

Theorem 23.8.7. *Let $W = V \setminus \mathbb{V}(h)$ be a basic affine open set, and let $U \subseteq W$ be open, not necessarily affine. Then,*

$$\mathcal{O}_V(U) = \mathcal{O}_W(U)$$

We now generalise $\mathcal{O}_V(U)$ from affine V to quasiprojective V. Given a quasiprojective variety V and an open subset $U \subseteq V$,

$$\mathcal{O}_V(U) = \big\{F : U \to \mathbb{C} : \forall p \in U, \text{open neighbourhood } W \subseteq U \text{ of } p \text{ such that } F|_W \in \mathbb{C}[W]\big\}$$

Again, this is naturally a $\mathbb{C}$-algebra.

The definition of a morphism of quasiprojective varieties can also be rephrased locally using regular functions.

Let $X \subseteq \mathbb{P}^n$ and $Y \subseteq \mathbb{P}^m$ be quasiprojective varieties. A map of sets $F : X \to Y$ is a morphism of quasiprojective varieties if for each $p \in V$, there exist open affine neighbourhoods U of p and V of $f(p)$ such that $f(U) \subseteq V$ and $f|_U$ agrees with a map of affine varieties. That is, $f|_U$ is given by a set of regular functions in the coordinates of U.

23.9 The Veronese Embedding

The Veronese embedding is an important example of a morphism of quasiprojective varieties. The Veronese embedding embeds $\mathbb{P}^n$ as a subvariety of a higher dimensional projective space in a non-trivial way.

Consider the set of all homogeneous polynomials of fixed degree d in the polynomial ring $\mathbb{C}[x_0, \dots, x_n]$. This is a finite-dimensional $\mathbb{C}$-vector space, with standard basis given by the monomials of the form

$$\prod_{i=1}^{d} x_i^{d_i}$$

where $\sum_{i=1}^{d} d_i = d$. We define the set of exponent vectors

$$M_{d,n} := \left\{ I = (d_0, \dots, d_n) \in \mathbb{N}^{n+1} : \sum_{i=1}^{n} d_i = d \right\} \cong \{\text{degree } d \text{ monomials in } x_0, \dots, x_n\}$$

with the obvious bijection sending each vector $I = (d_0, \dots, d_n)$ to the monomial $x_0^{d_0} \cdots x_n^{d_n}$. We abbreviate thes monomial to x^I.

Example. If $n = 2$ and $d = 6$, then the vector $I = (0,2,4)$ corresponds to the monomial $x_0^0 x_1^2 x_3^4$. △

These vectors (and hence monomials, by transport along the bijection) are naturally ordered under the lexicographic ordering, where $I = (i_0, \dots, i_n)$ precedes $J = (j_0, \dots, j_n)$ if there exists $k \in \mathbb{N}$ such that $i_k < j_k$, and $i_\ell = j_\ell$ for all $\ell < k$. That is, the first disagreement between I and J in the kth

Example. $x^I := x_0^2 x_1 x_2^3 x_4$ precedes $x^J := x_0^2 x_1 x_2^2, x_3^2$ since

$$I = (2,1,3,0,1) > (2,1,2,2,0) = J$$

△

Theorem 23.9.1. *There are $\binom{n+d}{d}$-many degree d monomials in $n+1$ variables $x_0, \dots, x_n$.*

Proof. Stars and bars. Every monomial can be represented by a string of x_1-many stars, a separating bar, x_2-many stars, etc. of d stars and n separating bars, and there are $\binom{n+d}{d}$ ways to place the d stars amongst the $d+n$ total spaces. ■

The dth Veronese embedding of $\mathbb{P}^n$ is the map $\nu_{d,n}$ defined by the tuple of all monomials of degree d:

$$\nu_{d,n} : \mathbb{P}^n \to \mathbb{P}^{|M_{d,n}|-1} = \mathbb{P}^{\binom{n+d}{d}-1}$$

$$[x_0 : \cdots : x_n] \mapsto [\underbrace{x_0^d : x_0^{d-1}x_1 : \cdots : x_n^d}_{\text{all monomials of degree } d}]$$

This is well-defined, since the polynomials all have the same degree, and cannot simultaneously vanish for any $[x_0 : \cdots : x_n] \in \mathbb{P}^n$, since if $x_i \neq 0$, then $x_i^d \neq 0$.

Example. The 2nd Veronesi embedding in dimension 1 is given by

$$\begin{aligned} \nu_{2,1} : \mathbb{P}^1_{[s:t]} &\to \mathbb{P}^2_{[x:y:z]} \\ [s:t] &\mapsto [s^2 : st : t^2] \end{aligned}$$

and is an isomorphism on to its image $\mathbb{V}(xz - y^2)$.

The 3rd Veronesi embedding in dimension 1 is given by

$$\begin{aligned} \nu_{3,1} : \mathbb{P}^1_{[s:t]} &\to \mathbb{P}^2_{[x:y:z:w]} \\ [s:t] &\mapsto [s^3 : s^2t : st^2 : t^3] \end{aligned}$$

$\triangle$

In general, we may index the coordinates of $\mathbb{P}^{|M_{d,n}|-1}$ by $I \in M_{d,n}$.

Example. The 2nd Veronesi embedding in dimension 2 is given by

$$\begin{aligned} \nu_{2,2} : \mathbb{P}^2_{[x_0:x_1:x_2]} &\to \mathbb{P}^5_{[z_{(2,0,0)}:\cdots:z_{(0,0,2)}]} \\ [x_0 : x_1 : x_2] &\mapsto [\underbrace{x_0^2}_{z_{(2,0,0)}} : \underbrace{x_0x_1}_{z_{(1,1,0)}} : \underbrace{x_0x_2}_{z_{(1,0,1)}} : \underbrace{x_1^2}_{z_{(0,2,0)}} : \underbrace{x_1x_2}_{z_{(0,1,1)}} : \underbrace{x_2^2}_{z_{(0,0,2)}}] \end{aligned}$$

$\triangle$

Theorem 23.9.2. *For all d,n, the Veronesi embedding $\nu_{d,n}$ is an isomorphism from $\mathbb{P}^n$ onto a closed subvariety of $\mathbb{P}^{|M_{d,n}|-1}$.*

Proof. We describe the inverse map.

Let $W \subseteq \mathbb{P}^{\binom{n+d}{d}-1}$ be the image of $\nu_{d,n}$. At each point of W, at least one of the coordinates indexed by the single-variable monomials x_i^d must be non-zero. So,

$$\nu_{d,n}(\mathcal{U}_i) \subseteq \mathcal{U}_{(0,\ldots,\underbrace{1}_{i\text{th position}},\ldots,0)} \subseteq \mathbb{P}^{\binom{n+d}{d}-1}$$

for each i, where $\mathcal{U}_i$ is the subset of $\mathbb{P}^n$ where x_i is non-zero.

Also, for each i,

$$[x_0x_i^{d-1} : x_1x_i^{d-1} : \cdots : x_i^d : \cdots : x_i^{d-1}x_{n-1} : x_i^{d-1}x_n] = [x_0 : \cdots : x_n]$$

so we can define a inverse on each affine chart by:

$$\begin{aligned} \mathcal{U}_{(0,\ldots,\underbrace{1}_{i\text{th position}},\ldots,0)} &\to \mathbb{P}^n \\ z &\mapsto [z_{(1,0,\ldots,d-1,\ldots,0)} : z_{(0,1,\ldots,d-1,\ldots,0)} : \cdots : z_{(0,\ldots,d-1,\ldots,0,1)}] \end{aligned}$$

That is, we send each z to the $(n+1)$-tuple of its coordinates indexed by $x_0x_i^{d-1}, \ldots, x_i^{d-1}x_n$. These maps agree on the overlaps $\mathcal{U}_i \cap \mathcal{U}_j$ and thus glue into a map $W \to \mathbb{P}^n$.

■

23.9.1 Enumerative Problems

A line in $\mathbb{P}^2_{[x:y:z]}$ is a subvariety of the form $\mathbb{V}(Ax + By + Cz)$, where at least one of the coefficients is non-zero. Note that this means that $[A : B : C] \in \mathbb{P}^2$.

So, the space $\mathbb{P}^2_{[A:B:C]}$ is the space of lines in $\mathbb{P}^2_{[x:y:z]}$, called the *dual* of $\mathbb{P}^2_{[x:y:z]}$, also denoted by $\check{\mathbb{P}}^2$.

More generally, a hyperplane in $\mathbb{P}^n_{[x_0:\cdots:x_n]}$ is a linear subvariety $\mathbb{V}(A_0x_0 + \cdots + A_nx_n)$, where at least one of the A_i is non-zero.

23.10 The Segre Map

23.11 The Grassmanian

23.11.1 The Plücker Embedding

23.12 Sheaves

Chapter 24

Algebraic Curves

Chapter 25

Elliptic Curves

Chapter 26

Modular Forms

Chapter 27

Local Fields

Chapter 28

Formal Languages

"An algorithm is a finite answer to an infinite number of questions"

— Stephen Kleene

A *formal language* is a set of words with letters selected from a fixed alphabet, and formed according to a set of rules called a formal grammar. In computational complexity theory, formal languages can encode decision problems, and hence provide a way of comparing the relative strength of various models of computation by checking which languages they are able to parse. In logic, formal languages can be used to represent the syntax of axiomatic and deductive systems, and hence mathematics itself can be reduced to the manipulation of these formal languages.

28.1 Introduction

An *alphabet* is any non-empty set of symbols, often denoted by Σ. A *word* over an alphabet is a finite sequence of letters. Note that the empty string, denoted by ε, is a word.

The *Kleene star* $(-)^*$, is a unary operation on sets of symbols defined as follows. Given a set V, we define $V^0 = \{\varepsilon\}$, where ε is the *empty word* with *length* $|\varepsilon| = 0$, then recursively define

$$V^{i+1} = \{wv : w \in V^i, v \in V\}$$

for each $i > 0$. That is, V^i is the set of strings that can be formed by concatenating i strings in V together. Then, the Kleene star on V is given by

$$V^* = \bigcup_{i \geq 0} V^i$$

That is, V^* is the set of all possible words over V.

Note that if V is countable, then V^* is the countable union of countable sets and is hence countable. Also note that a set of strings has a monoidal structure under concatenation, so the Kleene star of a set V is exactly the free monoid on V.

Then, a *formal language* over an alphabet Σ is a set $L \subseteq \Sigma^*$. Note that there is no requirement that this set be non-empty, so $L = \emptyset \subseteq \Sigma^*$ is a language, called the *empty language.* Also, by definition, we have that the empty word ε is in Σ^* for any alphabet Σ. Note that $L' = \{\varepsilon\}$ is a non-empty language – it contains the empty word.

Given a language $L \subseteq \{0,1\}^*$, we may interpret it as a decision problem by deciding whether a given binary string belongs to L. Conversely, assuming a fixed efficient encoding, we can encode any decision problem as a formal language by taking all strings representing yes-instances of the decision problem to be in our language.

Theorem 28.1.1. *There are functions $f : \mathbb{N} \to \{0,1\}$ that are not computable by any algorithm.*

Proof. Algorithms are finite sequences of finite alphabets of possible instructions, so there are only countably many algorithms possible. Conversely, the set of functions $\mathbb{N} \to \{0,1\}$ has size $2^{\aleph_0} = \mathcal{P}(\mathbb{N}) = \mathfrak{c}$, which is uncountable. ■

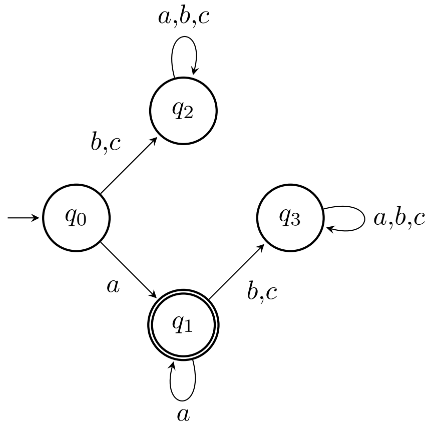
28.2 Regular Languages

28.2.1 Deterministic Finite Automata

A *deterministic finite automata* (DFA) is an abstract machine that either *accepts* or *rejects* a given word by reading through the symbols in the string and deterministically transitioning between different internal states depending on the current symbol and its current state. Formally, a DFA is given by 5-tuple (Q,Σ,q_0,F,δ), consisting of

- a finite set Q of *states*;
- a finite set Σ, the *alphabet*;
- an *initial state* $q_0 \in Q$ in which to start the computation;
- a set $F \subseteq Q$ of *accepting* or *final* states;
- and a transition function $\delta : Q \times \Sigma \to Q$.

We can visually represent a DFA as either a *state diagram*, or a *state transition table*:



	δ	a	b	c
$\to$	q_0	q_1	q_2	q_2
$*$	q_1	q_1	q_3	q_3
	q_2	q_2	q_2	q_2
	q_3	q_3	q_3	q_3

On the left, the transition function is given by the labelled arrows between states; the initial state is marked with a trailing arrow; and any final states are marked by two concentric circles.

On the left, the table on the right simply details the transition function, with the initial state marked with an arrow, and the accepting states marked with an asterisk.

In either case, we *run* the machine on a string by starting at the initial state; consuming the first character from the string; moving to the next state given by the transition function; then iterating this process until the string is empty. If the DFA is in an accepting state when the empty string is reached, then the word is *accepted*, and otherwise *rejected*.

Example. We run the string *abc* on the above DFA using the state diagram. We begin at q_0. The first character is a, so we proceed to q_1 with the remaining string *bc*. The next character is then b, so we move to q_3 with remaining string c. The next character is c, so we remain at q_3, and now the string is empty. q_3 is not an accepting state, so the string *abc* is rejected by this DFA. △

Example. We list the outputs of some more strings:

Input	Output
a	Accept
aa	Accept
aab	Reject
b	Reject
c	Reject
bca	Reject

More concisely, this DFA accepts exactly the strings that consist solely of the character a. △

Let $M = (Q,\Sigma,q_0,F,\delta)$ be a DFA, and let $s = s_1s_2\cdots s_n$ be a string, where $s_i \in \Sigma$ for each i. We define the *run* of M on s as follows:

- The run of M on ε is the state q_0.
- The run of M on the non-empty word s is the sequence of states $(r_i)_{i=0}^n$ given recursively by

$$r_i = \begin{cases} q_0 & i = 0 \\ \delta(r_{i-1},s_i) & i > 0 \end{cases}$$

The run of M on a word s is called an *accepting run* if the last state in the run is an accepting state of M, and we say that a word s is *accepted* or *recognised* by M if the run of M on s is an accepting run. The set of strings that M accepts forms a language over Σ called the language *accepted* or *recognised* by M, denoted by $L(M)$.

$$L(M) \coloneqq \{s \in \Sigma^* : \text{the run of } M \text{ on } s \text{ is an accepting run}\}$$

Note that, for M to accept a language L', it must not only accept only the words in L', but also reject every word in $\Sigma^* \setminus L$.

The transition function $\delta : Q \times \Sigma \to Q$ of a DFA details the change in state upon reading a single symbol. We can expand this function to the *extended transition function* $\hat{\delta}$ that expresses the change in state upon reading an entire *string*.

Formally, we recursively define the extended transition function $\hat{\delta} : Q \times \Sigma^* \to Q$ as follows:

- For every state $q \in Q$, we have $\hat{\delta}(q,\varepsilon) = q$;
- For every state $q \in Q$ and word $s \in \Sigma^*$ with $s = wa$, $w \in \Sigma^*$, $a \in \Sigma$, we have $\hat{\delta}(q,s) = \delta\big(\hat{\delta}(q,w),a\big)$.

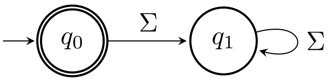
Using the extended transition function we can also write the language recognised by a DFA M as

$$L(M) = \big\{s \in \Sigma^* : \hat{\delta}(q_0,s) \in F\big\}$$

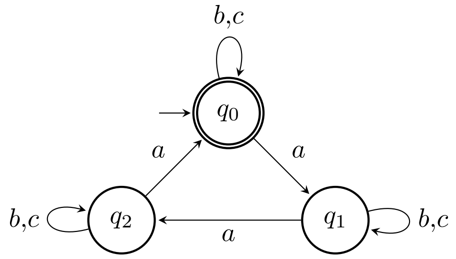
A language L is *regular* if it is accepted by some DFA.

Example.

- The empty language $L = \emptyset$ is regular; it is accepted by any DFA with $F = \emptyset$.
- The language $L = \Sigma^*$ is regular; it is accepted by any DFA with $F = Q$.
- The language $L = \{\varepsilon\}$ is regular; it is accepted by the DFA



- The language $L \subseteq \{a,b,c\}^*$ defined by $L = \{s \in \Sigma^* :$ the number of a's in s is divisible by $3\}$ is regular; it is accepted by the DFA



$\triangle$

28.2.2 Closure Properties of Regular Languages

Because languages are sets, ordinary set operations also apply to languages. Regular languages are *closed* under certain operations, in the sense that the resulting language is also regular.

Regular languages are closed under

- **Complementation**:

  If L is regular, then $\overline{L} = \Sigma^* \setminus L$ is regular; if $L = L(M)$ is accepted by $M = (Q,\Sigma,q_0,F,\delta)$, then $\overline{L} = L(M')$ is accepted by $M' = (Q,\Sigma,q_0,Q \setminus F,\delta)$.

- **Intersection**:

  If L_1 and L_2 are regular, then $L_1 \cap L_2$ is regular. The idea here is to run the DFAs for L_1 and L_2 in parallel by using the Cartesian product of states and applying the transition functions pointwise, and accepting if and only if both original DFAs accept.

  If $L_1 = L(M_1)$ and $L_1 = L(M_1)$ with $M_1 = (Q_1,\Sigma,q_1,F_1,\delta_1)$, $M_2 = (Q_2,\Sigma,q_2,F_2,\delta_2)$, then $L_1 \cap L_2$ is accepted by the DFA
  $$M = \big(Q_1 \times Q_2,\Sigma,(q_1,q_2),F_1 \times F_2,\delta\big)$$
  with $\delta : (Q_1 \times Q_2) \times \Sigma \to Q_1 \times Q_2$ defined pointwise:
  $$\delta\big((p_1,p_2),a\big) = \big(\delta_1(p_1,a),\delta_2(p_2,a)\big)$$

- **Union**:

  If L_1 and L_2 are regular, then $L_1 \cup L_2$ is regular. This follows from De Morgan's laws:
  $$L_1 \cup L_2 = \overline{\overline{L_1} \cap \overline{L_2}}$$
  and the closure properties of complementation and intersection, but we can also give an explicit DFA that recognises this union. As before, the idea is to run the DFAs for L_1 and L_2 in parallel, this time accepting if either of the original DFAs accept.

  If $L_1 = L(M_1)$ and $L_1 = L(M_1)$ with $M_1 = (Q_1,\Sigma,q_1,F_1,\delta_1)$, $M_2 = (Q_2,\Sigma,q_2,F_2,\delta_2)$, then $L_1 \cap L_2$ is accepted by the DFA
  $$M = \big(Q_1 \times Q_2,\Sigma,(q_1,q_2),(F_1 \times Q_2) \cup (Q_1 \times F_2),\delta\big)$$
  where δ is the same as for intersections.

- **Relative difference**:

  If L_1 and L_2 are regular, then $L_1 \setminus L_2$ is regular. This follows from closure under complementation and intersection, as,

  $$L_2 \setminus L_2 = L_1 \cap \overline{L_2}$$

- **Concatenation**:

  If L_1 and L_2 are regular, then $L_1 \cdot L_2 = \{wv : w \in L_1, v \in L_2\}$ is regular. Proof requires more machinery than we currently have.

- **Kleene star**:

  If L is regular, then L^* is regular. Follows from regularity of concatenation and unions:

  $$L^* = \{\varepsilon\} \cup L \cup (L \cdot L) \cup (L \cdot L \cdot L) \cup \cdots$$

Note that L_1 and $L_1 \setminus L_2$ being regular does not imply that L_2 is regular. For instance, if $L_1 = \emptyset$, then $L_1 \setminus L_2 = \emptyset$, regardless of the regularity of L_2.

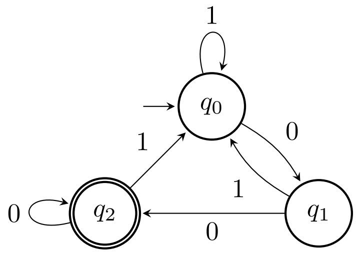
28.2.3 Non-Deterministic Finite Automata

Are regular languages closed under *reversal*? That is, if L is regular, then is the language

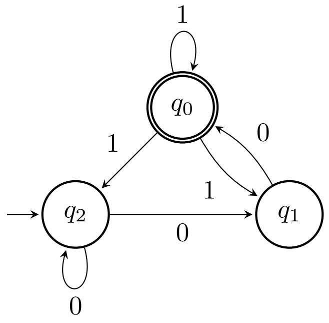
$$L^{\text{rev}} = \{w : w \text{ is the reverse of a string in } L\}$$

regular?

Consider the language $L = \{\text{binary strings ending with } 00\}$, accepted by the DFA



To build a DFA that accepts L^{rev}, we'd might think to reverse all arrows, then swap the start and accepting states:



However, this state diagram now has states with multiple exiting arrows labelled with the same symbol, and some states do not have an exiting arrow for every symbol in the alphabet. Moreover, if we had multiple accepting states, then we would also have multiple initial state in this reverse diagram. So, this state diagram does not describe a DFA. We instead extend the definition of an DFA to a *non-deterministic finite automata* (NFA).

A DFA must have exactly one transition out of a state for each symbol in the alphabet, so each word has a unique run. In contrast, an NFA may have multiple or no transitions out of a state for any given symbol, so an NFA may have multiple choices at each step, and the final state is not *determined* solely by the start state and input word, instead having a branching tree structure.

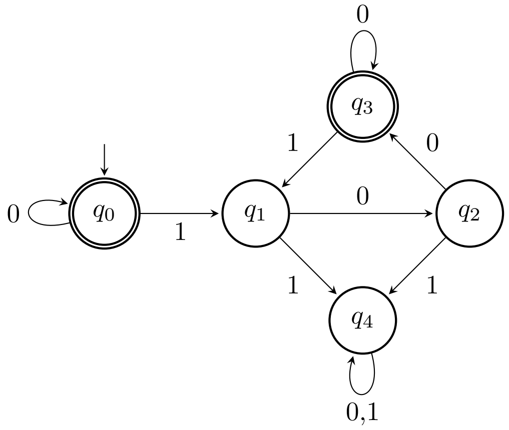
An NFA may also have *ε-transitions* – transitions that do not consume any input. This allows us to deal with multiple initial states by adding a new state to be initial, then adding ε-transitions from this state to all the previous initial states.

Formally, an NFA is a 5-tuple (Q,Σ,q_0,F,δ), consisting of

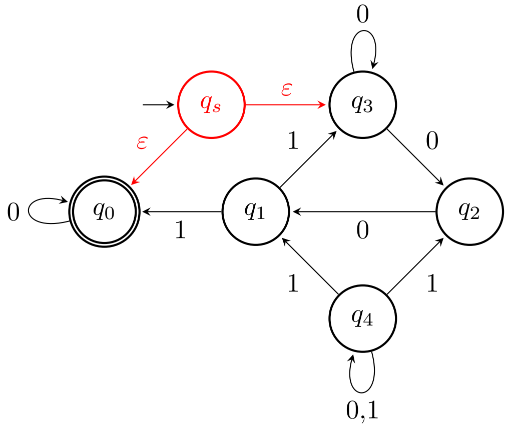
- a finite set Q of states;
- a finite alphabet Σ;
- an initial state $q_0 \in Q$ in which to start the computation;
- a set $F \subseteq Q$ of accepting or final states;
- and a transition function $\delta : Q \times (\Sigma \cup \{\varepsilon\}) \to \mathcal{P}(Q)$.

The first four entries are the same as for DFAs, but because an NFA may have multiple or no transitions out of a state for any given symbol, the transition function instead returns a set of states in $\mathcal{P}(Q)$, and we also include ε in the domain to account for ε-transitions, so the transition function is then a function $\delta : Q \times (\Sigma \cup \{\varepsilon\}) \to \mathcal{P}(Q)$. We also write Σ_ε to denote $\Sigma \cup \{\varepsilon\}$.

Example. Consider the language $L \subseteq \{0,1\}^*$ defined by $L = \{\text{every 1 is followed by 00}\}$, accepted by the DFA



By reversing the arrows and adding a new state equipped with ε-transitions to the two previous accepting states, we obtain a state diagram of an NFA that accepts $L^{\text{rev}} = \{\text{every 1 is preceded by 00}\}$:



We can also represent this NFA as a state transition table:

	δ	0	1	ε
$\rightarrow$	q_s	$\emptyset$	$\emptyset$	$\{q_0,q_3\}$
$*$	q_0	$\{q_0\}$	$\emptyset$	$\emptyset$
	q_1	$\emptyset$	$\{q_0,q_3\}$	$\emptyset$
	q_2	$\{q_1\}$	$\emptyset$	$\emptyset$
	q_3	$\{q_2,q_3\}$	$\emptyset$	$\emptyset$
	q_4	$\{q_4\}$	$\{q_1,q_2,q_4\}$	$\emptyset$

Note that every entry in the table is a set, unlike for a DFA.
We can also see that there is no way to enter state q_4, so we may remove it from the NFA and simplify the state diagram/transition table. △

28.2.4 ε-Closure

What does the extended transition function of an NFA look like? The ordinary transition function already returns sets of states – unlike a DFA, which is *deterministic* and returns a single state – but we also have to deal with ε-transitions at every step in an NFA. For this, we define the ε-closure function,

$\textsc{eclose} : Q \to \mathcal{P}(Q)$.

Informally, given a state q, $\textsc{eclose}(q)$ is the set of states that can be reached from q by following ε-transitions alone (including taking no ε-transitions, so $q \in \textsc{eclose}(q)$). Formally, given a state q, $\textsc{eclose}(q)$ is the minimal set such that

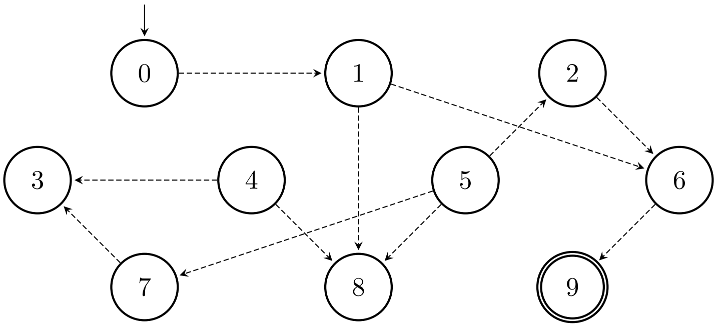
- $q \in \textsc{eclose}(q)$;
- $\forall p,r \in Q, \big(p \in \textsc{eclose}(q) \wedge r \in \delta(p,\varepsilon)\big) \to r \in \textsc{eclose}(q)$.

$\textsc{eclose}$ can be naturally extended to sets of states: given a set $X \subseteq Q$, we define

$$\textsc{eclose}(X) = \bigcup_{x \in X} \textsc{eclose}(x)$$

Note that the nullary union is empty, so $\textsc{eclose}(\emptyset) = \emptyset$.

Example. In the following, dashed arrows represent ε-transitions.



$$\begin{aligned}
\textsc{eclose}(0) &= \{0,1,6,8,9\} & \textsc{eclose}(5) &= \{2,3,5,6,7,8,9\} \\
\textsc{eclose}(1) &= \{1,6,8,9\} & \textsc{eclose}(6) &= \{6,9\} \\
\textsc{eclose}(2) &= \{2,6,9\} & \textsc{eclose}(7) &= \{3,7\} \\
\textsc{eclose}(3) &= \{3\} & \textsc{eclose}(8) &= \{8\} \\
\textsc{eclose}(4) &= \{3,4,8\} & \textsc{eclose}(9) &= \{9\}
\end{aligned}$$

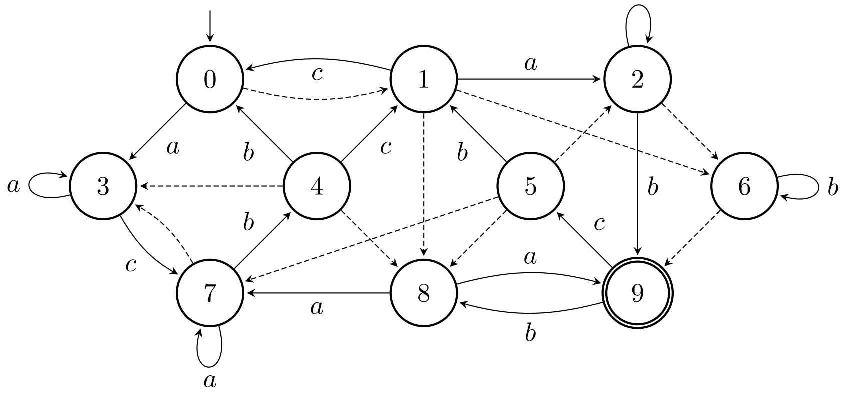
△

The extended transition function $\hat{\delta}$ for an NFA $N = (Q,\Sigma,q_0,F,\delta)$ is then a function $\delta : Q \times \Sigma^* \to \mathcal{P}(Q)$ defined as follows:

- For every state $q \in Q$, we have $\hat{\delta}(q,\varepsilon) = \textsc{eclose}(q)$;
- For every state $q \in Q$ and word $s \in \Sigma^*$ with $s = wa$, $w \in \Sigma^*$, $a \in \Sigma$, we have

$$\hat{\delta}(q,s) = \textsc{eclose}\left(\bigcup_{p \in \hat{\delta}(q,w)} \delta(p,a)\right)$$

Example. The ε-transitions (and hence $\textsc{eclose}$ sets on singletons) are the same as in the previous diagram.



△

$$
\begin{aligned}
\hat{\delta}(0,a) &= \textsc{eclose} \bigcup_{p\in\hat{\delta}(0,\varepsilon)} \delta(p,a)\\
&= \textsc{eclose} \bigcup_{p\in\textsc{eclose}(0)} \delta(p,a)\\
&= \textsc{eclose}\big(\delta(0,a)\cup\delta(1,a)\cup\delta(6,a)\cup\delta(8,a)\cup\delta(9,a)\big)\\
&= \textsc{eclose}\big(\{3\}\cup\{2\}\cup\emptyset\cup\{7,9\}\cup\emptyset\big)\\
&= \textsc{eclose}\big(\{2,3,7,9\}\big)\\
&= \bigcup_{z\in\{2,3,7,9\}} \textsc{eclose}(x)\\
&= \{2,6,8,9\}\cup\{3\}\cup\{3,7\}\cup\{9\}\\
&= \{2,3,6,7,8,9\}
\end{aligned}
$$

$$
\begin{aligned}
\hat{\delta}(0,aa) &= \textsc{eclose} \bigcup_{p\in\hat{\delta}(0,a)} \delta(p,a)\\
&= \textsc{eclose} \bigcup_{p\in\{2,3,6,7,8,9\}} \delta(p,a)\\
&= \textsc{eclose}\big(\delta(2,a)\cup\delta(3,a)\cup\delta(6,a)\cup\delta(7,a)\cup\delta(8,a)\cup\delta(9,a)\big)\\
&= \textsc{eclose}\big(\emptyset\cup\{3\}\cup\emptyset\cup\{7\}\cup\{7,9\}\cup\emptyset\big)\\
&= \textsc{eclose}\big(\{3,7,9\}\big)\\
&= \bigcup_{z\in\{3,7,9\}} \textsc{eclose}(x)\\
&= \{3\}\cup\{3,7\}\cup\{9\}\\
&= \{3,7,9\}
\end{aligned}
$$

$$
\begin{aligned}
\hat{\delta}(0,b) &= \{6,8,9\} & \hat{\delta}(1,a) &= \{2,3,6,7,8,9\}\\
\hat{\delta}(0,c) &= \{0,1,6,8,9\} & \hat{\delta}(2,b) &= \{2,6,9\}\\
\hat{\delta}(8,a) &= \{3,7,9\} & \hat{\delta}(3,c) &= \emptyset\\
\hat{\delta}(9,c) &= \{2,3,5,6,7,8,9\} & \hat{\delta}(4,a) &= \{1,3,6,7,8,9\}
\end{aligned}
$$

28.2.5 Languages Recognised by NFA

Previously, we defined the language $L(M)$ recognised by a DFA M to be the set of words accepted by M. That is,

$$\begin{aligned} L(M) &\coloneqq \{s \in \Sigma^* : \text{the run of } M \text{ on } s \text{ is an accepting run}\} \\ &= \left\{s \in \Sigma^* : \hat{\delta}(q_0,s) \in F\right\} \end{aligned}$$

However, unlike a DFA, which is *deterministic* and always returns the same output, the run of an NFA on the same word may be different across several computations.

Let $N = (Q,\Sigma,q_0,F,\delta)$ be an NFA, and let s be a string. *A* run of N on s is a sequence of states $(r_i)_{i=1}^{n}$ such that

- $r_0 = q_0$;
- There exists a decomposition $s = s_1 s_2 \cdots s_n$, with $s_i \in \Sigma \cup \{\varepsilon\}$ for each i, such that for each $i > 0$, $r_i \in \delta(r_{i-1},s_i)$.

Then, an NFA N *accepts* or *recognises* a word s if there exists *some* accepting run.

Let $N = (Q,\Sigma,q_0,F,\delta)$ be an NFA. Then, the language $L(N)$ accepted or recognised by N is defined by

$$L(N) \coloneqq \{s \in \Sigma^* : \textit{some} \text{ run of } N \text{ on } s \text{ is an accepting run}\}$$

which we can again write in terms of the extended transition function:

$$= \left\{s \in \Sigma^* : \hat{\delta}(q_0,s) \cap F \neq \emptyset\right\}$$

Because the extended transition function returns the set of possible states after reading a word, we just check that it has non-empty intersection with the set of accepting states.

28.2.6 The Subset Construction

Are NFAs more powerful than DFAs?

Firstly, what do we even mean by “more powerful”? Intuitively, a computer is “more powerful” than a simple pocket calculator, but how do we formalise this notion? We might notice that a computer can have a calculator application within it – so a computer can do every task a calculator can. This shows that a computer is at least as powerful as a calculator. Importantly, to make this comparison strict, we note that there are tasks that a computer can do that a calculator cannot.

Given two computational models A and B, we say that A is *more powerful* or *expressive* than B if the class of languages recognised by A is a strict superset of the class of languages accepted by B. Note that it may be the case that two distinct computational models are incomparable under this relation if the class of languages they accept are not supersets of each other in either direction.

Clearly, every DFA is an NFA, as an NFA is a relaxation of the requirements of a DFA, so NFAs are at least as powerful as DFAs. However, are they strictly more powerful? It turns out that, for any NFA, we may *determinise* it and construct an equivalent DFA that recognises precisely the same language via the *subset* or *powerset construction.*

When a DFA is run on a word, we just keep track of a single state; that is, the state q_i that is reached upon reading a prefix of the input string, that can then be overwritten by the state that is reached upon reading the next symbol s.

In contrast, when running an NFA, we need to keep track of the set of all states that could be reached after seeing the same prefix, according to the non-deterministic choices made by the automaton. If, however, after a certain prefix has been read, a set S of states can be reached, then the set of symbols reachable upon reading the next symbol s is a deterministic function of S and s. That is, while the states reached at each step in an individual run is non-deterministic, the set of states reachable at each step over all possible runs is fully deterministic, and as such, traversing sets of reachable states in this way describes the action of a DFA. This is the strategy of our construction.

Let $N = (Q,\Sigma,q_0,F,\delta)$ be the NFA to be determinised. Then, the DFA $M = (Q',\Sigma,q_0',F',\delta')$ defined by

- $Q' = \mathcal{P}(Q)$;
- $q_0' = \text{ECLOSE}(q_0)$;
- $F' = \{X \subseteq Q : X \cap F \neq \emptyset\}$;
- $$\begin{aligned}\delta'(X,a) &= \bigcup_{x\in X} \text{ECLOSE}\big(\delta(x,a)\big)\\ &= \Big\{z : \exists x \in X : z \in \text{ECLOSE}\big(\delta(x,a)\big)\Big\}\end{aligned}$$

accepts the same language.

Theorem 28.2.1. *For all $s \in \Sigma^*$, $\hat{\delta}(q_0,s) = \hat{\delta}'(q_0',s)$.*

Proof. We induct on $|s|$. If $|s| = 0$, then $\hat{\delta}(q_0,s) = \hat{\delta}(q_0,\varepsilon) = \text{ECLOSE}(q_0) = q_0' = \delta'(q_0',\varepsilon) = \hat{\delta}'(q_0',s)$. Otherwise, suppose $s = wa$ with $w \in \Sigma^*$, $a \in \Sigma$. Then,

$$\begin{aligned}\hat{\delta}(q_0,s) &= \bigcup_{p\in\hat{\delta}(q_0,w)} \text{ECLOSE}\big(\delta(p,a)\big)\\ &= \bigcup_{p\in\hat{\delta}'(q_0',w)} \text{ECLOSE}\big(\delta(p,a)\big)\\ &= \delta'\big(\hat{\delta}(q_0',w),a\big)\\ &= \hat{\delta}(q_0',s)\end{aligned}$$

■

28.2.7 Regular Expressions

A *regular expression*, *regex*, or a *pattern*, over an alphabet is a construction that specifies or *matches* a language over that alphabet. Regular expressions are defined recursively as follows:

Given an alphabet Σ, the following constants are *basic* regular expressions:

- (Empty set) $\emptyset$ is a valid regular expression, matching the empty language – $L(\emptyset) = \emptyset$;
- (Empty string) ε is a valid regular expression, matching the language containing the empty string – $L(\varepsilon) = \{\varepsilon\}$;
- (Literal character) $a \in \Sigma$ is a valid regular expression, matching the language containing only the character – $L(a) = \{a\}$;

and given two regular expressions R and S, we have:

- (Concatenation) $R \cdot S$, or RS, is a regular expression, matching the set of strings that can be obtained by concatenating a string accepted by R with a string accepted by S – $L(R \cdot S) = L(R) \cdot L(S)$;
- (Union) $R + S$, or $R|S$, is a regular expression, matching the union of the sets matched by R and S – $L(R + S) = L(R) \cup L(S)$;

- (Kleene Star) R^* is a regular expression, matching the smallest superset of the set matched by R that contains ε and is closed under concatenation – $L(R^*) = L(R)^*$;

In decreasing order, these operations have precedence *, $\cdot$, $+$.

Example. Let $\Sigma = \{a,b\}$ and $R = (a+b)^*$ be a regular expression over Σ. Intuitively, $(a+b)$ matches "a" or "b", so $L(R) = \Sigma^*$, but we can also unfold the definition algebraically:

$$\begin{aligned} L(R) &= L\big((a+b)^*\big) \\ &= L\big((a+b)\big)^* \\ &= \big(L(a) \cup L(b)\big)^* \\ &= \big(\{a\} \cup \{b\}\big)^* \\ &= \Sigma^* \end{aligned}$$

△

Example. If $\Sigma = \{a,b\}$, and $R = (a+b)^*(a+bb)$, then

$$\begin{aligned} L(R) &= L\big((a+b)^*(a+bb)\big) \\ &= L\big((a+b)^*\big)L\big((a+bb)\big) \\ &= \Sigma^* L\big((a+bb)\big) \\ &= \{\text{all strings over } \{a,b\} \text{ that end with } a \text{ or } bb\} \end{aligned}$$

△

Example. If $\Sigma = \{a,b\}$, and $R = (aa)^*(bb)^*b$, then

$$\begin{aligned} L(R) &= L\big((aa)^*(bb)^*b\big) \\ &= L\big((aa)\big)^* L\big((bb)\big)^* L\big(b\big) \\ &= \{\text{all strings over } \{a,b\} \text{ with an even number of } a\text{'s followed by an odd number of } b\text{'s}\} \end{aligned}$$

△

28.2.8 Generalised Non-Deterministic Finite Automata

Using regular expressions, we can define a *generalised non-deterministic finite automaton* (GNFA). A GNFA is a variation of a NFA where each transition may be any regular expression, and there may only be one transition between any two states, unlike a DFA or an NFA, which may have multiple such transitions. Furthermore, A GNFA must have exactly one initial state and one accepting state, and these states must be distinct.

A GNFA is a 5-tuple $(Q,\Sigma,q_{\text{start}},q_{\text{accept}},\delta)$, consisting of

- a finite set Q of states;
- a finite alphabet Σ;
- the start state, $q_{\text{start}} \in Q$;
- the accept state, $q_{\text{accept}} \in Q$;
- and a transition function $\delta : \big(Q \setminus \{q_{\text{accept}}\}\big) \times \big(Q \setminus \{q_{\text{start}}\}\big) \to \mathcal{R}$, where $\mathcal{R}$ is the set of all regular expressions over Σ.

28.2.9 Languages Recognised by Regular Expressions

As suggested by the name, regular expressions recognise precisely the class of regular languages, so NFAs, DFAs, and regular expressions are equally as expressive.

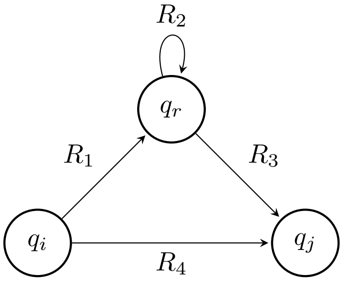
Theorem 28.2.2. *A language is regular if and only if it is described by a regular expression.*

Proof. Given a regular language $L = L(M)$ accepted by a DFA, we may convert this DFA into a GNFA as follows:

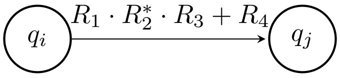
1. Add a new start state with an ε-transition to the previous start state.
2. Add a new accepting state with ε-transitions from the previous accepting states to this state.
3. Any transitions with multiple labels may be replaced with a transition labelled with the union of the previous labels.
4. For any ordered pair of states that do not end at the start state nor begin at the accept state and are disconnected, add a new transition between them labelled with $\emptyset$.

This GNFA may then be converted into a regular expression as follows:

1. If there are only two states, then we are done, as these must be the unique start and accepting states, and transition connecting them is a regular expression.
2. Otherwise, select some state $q_r \in Q \setminus \{q_{\text{start}},q_{\text{end}}\}$. Then, for all $(q_i,q_j) \in \big(Q \setminus \{q_{\text{start}},q_r\}\big) \times \big(Q \setminus \{q_{\text{end}},q_r\}\big)$, we may replace the transitions



by the single edge

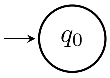


Once this has been done, we may remove q_r from the diagram, then pick a new state to be q_r, until the diagram has only the initial and accepting state remaining.

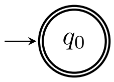
For the reverse implication, suppose we have a regular expression R that accepts $L(R)$. Then, we can construct a NFA N such that $L(N) = L(R)$.

We give an NFA for each of the basic regular expressions:

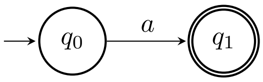
1. If $R = \emptyset$, then $L(R) = \emptyset$ is recognised by the NFA $N = \big(\{q_0\},\Sigma,q_0,\emptyset,\delta\big)$,



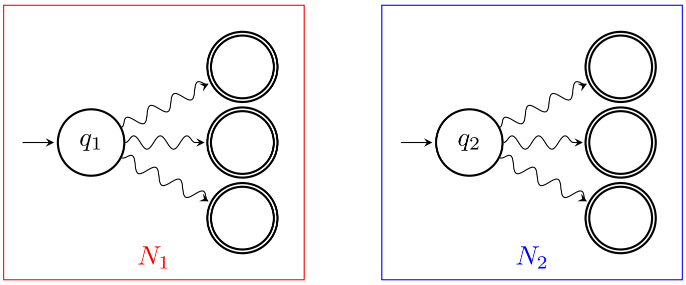
2. If $R = \varepsilon$, then $L(R) = \{\varepsilon\}$ is recognised by the NFA $N = \big(\{q_0\},\Sigma,q_0,\{q_0\},\delta\big)$,



3. If $R = a \in \Sigma$, then $L(R) = \{a\}$ is recognised by the NFA $N = \big(\{q_0,q_1\},\Sigma,q_0,\{q_1\},\delta\big)$,

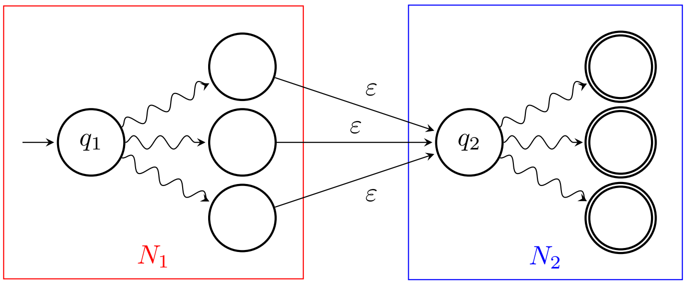


Then, given regular expressions R_1 and R_2 with NFAs $N_1 = (Q_1,\Sigma,q_1,F_1,\delta_1)$ and $N_2 = (Q_2,\Sigma,q_2,F_2,\delta_2)$

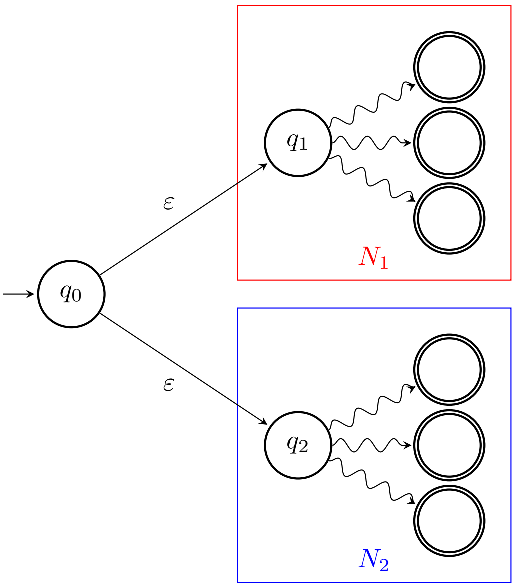


accepting $L(R_1)$ and $L(R_2)$, respectively,

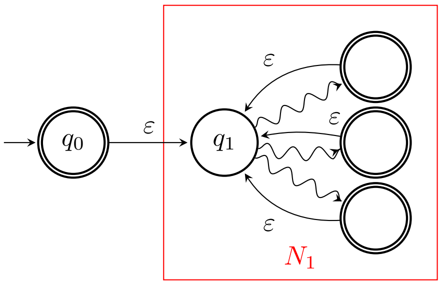
1. The language recognised by the concatenation $R_1 \cdot R_2$ is recognised by the NFA $N = \big(Q_1 \cup Q_2,\Sigma,q_1,F_2 \setminus F_1,\delta\big)$ formed by making the accepting states of R_1 no longer accepting, then attaching ε-transitions from the old accepting states to the initial state of R_2:



2. The language recognised by the union $R_1 + R_2$ is recognised by the NFA $N = \big(Q_1 \cup Q_2 \cup \{q_0\},\Sigma,q_1,F_2 \cup F_1,\delta\big)$ formed by adding a new starting state q_0 with ε-transitions to the previous start states:



3. The language recognised by the Kleene star R_1^* is recognised by the NFA $N = \left(Q_1 \cup \{q_0\}, \Sigma, q_1, F_1, \delta\right)$ formed by adding a new starting state q_0 that is also accepting – in order to accept the empty string – with an ε-transition to the previous start state, then adding ε-transitions from the previous accepting states to the start state – to allow for arbitrary concatenations of the R_1 NFA:



By induction, this construction extends to any regular expression. ■

28.3 Non-Regular Languages

28.3.1 The Myhill-Nerode Theorem

Because the change in state of a DFA is determined entirely by the current state and the next character, a DFA is effectively *memoryless*. That is, if two different strings converge to the same state, then the DFA will respond in precisely the same way to any further characters appended to them, so the two initial strings are, with respect to this DFA, identical. This motivates our next definition.

Two strings $x,y \in \Sigma^*$ are *distinguishable* by a language L if there exists a string $z \in \Sigma^*$ such that $x \cdot z \in L$ and $y \cdot z \notin L$, or vice versa, and we call z the *certificate* or *witness* of the distinguishability of x and y.

If two strings $x,y \in \Sigma^*$ are not distinguishable by L, then we say they are *indistinguishable*, and we write $x \equiv_L y$ to denote this relation. Note that this relation is on the *language* L, and is independent from any specific implementation in a particular DFA.

Furthermore, this relation forms an equivalence relation on L (i.e., it is transitive, reflexive, and symmetric), and we call the number of equivalence classes of L under $\equiv_L$ the *index* of $\equiv_L$.

Theorem (Myhill-Nerode)**.** *A language L is regular if and only if $\equiv_L$ has finite index.*

So, to prove a language is non-regular, we can find an infinite set of strings and show that they are pairwise distinguishable by L, and hence $\equiv_L$ has infinite index.

Example. Let $L = \{0^n1^n : n \in \mathbb{N}\}$ be a language over $\Sigma = \{0,1\}$. Then, the set $\left\{0^i \in \Sigma^* : i \in \mathbb{N}\right\}$ is infinite, and two strings 0^i and 0^j with $i \neq j$ are distinguishable with 1^i as witness. It follows that $\equiv_L$ has infinite index, and hence L is non-regular by Myhill-Nerode. △

Example. Let $L = \{0^p : p \text{ prime}\}$ be a language over $\Sigma = \{0,1\}$.

Let $i \neq j$ and p be prime, and consider the sequence of integers

$$\begin{gathered} p + 0(j-i) \\ p + 1(j-i) \\ p + 2(j-i) \\ \vdots \\ p + p(j-i) \end{gathered}$$

Note that the first integer is $p+0(j-i) = p$, which is prime, and the final integer is $p+p(j-i) = p(1+j-i)$, which is composite. Let $1 \leq k \leq p$ be the least integer for which $p + k(j-i)$ is composite. Then, $1^{p+(k-1)(j-i)-i}$ is a certificate for the distinguishability of 1^i and 1^j:

$$\begin{aligned} 1^i \cdot 1^{p+(k-1)(j-i)-i} &= 1^{p+(k-1)(j-i)} && \in L \\ 1^j \cdot 1^{p+(k-1)(j-i)-i} &= 1^{p+k(j-i)} && \notin L \end{aligned}$$

So, every pair of elements of the infinite set $\left\{1^i \in \Sigma^* : i \in \mathbb{N}\right\}$ are distinguishable, so $\equiv_L$ has infinite index, and hence L is non-regular by Myhill-Nerode. △

Another way to show pairwise distinguishability is to order the infinite set of strings, $(s_i)_{i=1}^{\infty}$, then show that for all i, s^i is distinguishable from s^j for all $j > i$.

Example. Let $n_i(s)$ denote the number of occurrences of the character i in a string s, and let $L = \{s \in \Sigma^* : n_a(s) < n_b(s)\}$ be a language over $\Sigma = \{a,b\}$.

The set $\left\{a^i \in \Sigma^* : i \in \mathbb{N}\right\}$ is infinite, and two strings a^i and a^j with $i > j$ are distinguishable with b^{i+1} as witness, so L is non-regular. △

28.3.2 The Pumping Lemma for Regular Languages

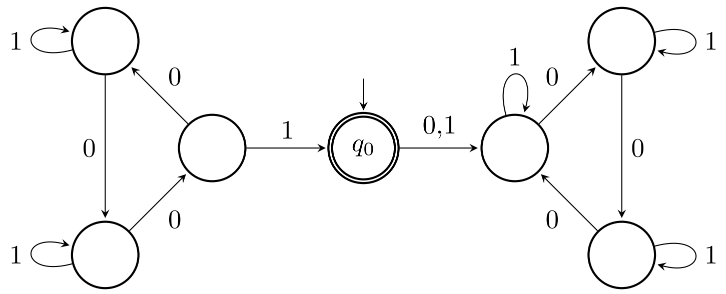
Let M be a DFA. If there is a cycle in the state diagram of M traversed by a string c, then we may traverse that cycle arbitrarily many times, and return to the same state. Again, because DFAs are memoryless, traversing the cycle once is the same as traversing it 2 times, or 3 times, or n times. If that cycle is reachable from the initial state, and can also reach an accepting state, this means that, given a word s accepted by M whose run intersects this cycle, we may add as many copies of c in the middle of s as we want, and the word will still be recognised by M.

That is, if there is a cycle reachable from the initial state that can also reach an accepting state in the state diagram of the DFA M, then the language $L(M)$ is infinite.

In fact, this is a complete characterisation of the DFAs that accept an infinite language: because DFAs must have finitely many states, this is the only way an infinite language can arise.

Theorem 28.3.1. *A language $L = L(M)$ is infinite if and only if there is a cycle reachable from the initial state that can also reach an accepting state in the state diagram of the DFA M.*

Note that it must be the same cycle that is reachable from the initial state and can reach an accepting state. For instance, the DFA M with state diagram



has both a cycle that is reachable from the start state, and a cycle that can reach an accepting state, but these cycles do not coincide, and $L(M) = \{\varepsilon\}$ is finite.

Lemma (Pumping Lemma)**.** *Let L be a regular language. Then, there exists an integer $p \geq 1$ (the pumping length), such that for every string $s \in L$ with length $|s| \geq p$, there exists a decomposition $s = x \cdot y \cdot z$ such that*

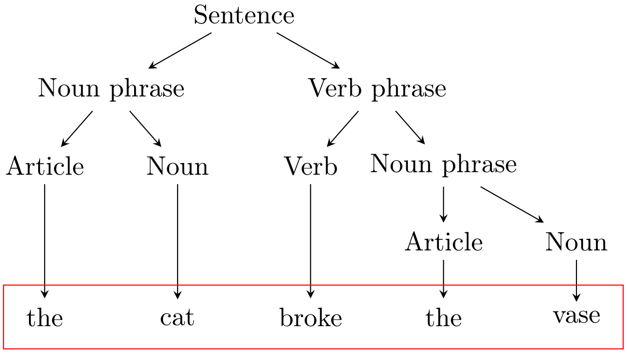
- $|y| \geq 1$*;*
- $|xy| \leq p$*;*
- *for all $n \geq 0$, $x \cdot y^n \cdot z \in L$.*

Example. Let $L = \{0^n 1^n : n \in \mathbb{N}\}$ be a language over $\Sigma = \{0,1\}$. Suppose there exists $p \geq 1$ such that the string $s = 0^p 1^p$ with length $|s| = 2p \geq p$ has a decomposition $s = x \cdot y \cdot z$ satisfying $|x \cdot y| \leq p$ and $|y| \geq 1$. From the former condition, y consists of only instances of 0; and from the latter, y contains at least one instance of 0. Pumping y to obtain $x \cdot y^2 \cdot z$ adds more 0s to the string without adding any 1s, so $x \cdot y^2 \cdot z \notin L$, contradicting the pumping lemma, and hence L is non-regular. △

Note that the pumping lemma is not biconditional; there exist non-regular languages that satisfy the pumping lemma, so the pumping lemma cannot be used to show that a language is regular.

28.4 Grammars

Ordinary languages not only contain words, but also have particular rules, called a *grammar*, that dictate how they can fit together. For instance, we could have a grammar fragment that says that sentences may be composed of a noun phrase and a verb phrase, which can each then be further decomposed:



so we have *derived* this sentence from the given grammar.

A *grammar* G is a 4-tuple (V,Σ,R,S), consisting of

- A finite set V of *variables* or *non-terminal* symbols;
- a finite set Σ, the alphabet, of *terminal* symbols;

- a finite set R of *substitution rules* or *productions*, where a substitution rule is a string of the form
$$\alpha \to \beta$$
where $\alpha,\beta \in (V \cup \Sigma)^*$, and $\alpha \neq \varepsilon$.
- and an *initial variable* $S \in V$.

Example. In the above tree, "noun phrase" is a variable, while "the" is a terminal symbol, and "Article → the" is a substitution rule. △

If there are multiple substitution rules with the same term on the left, i.e. $\alpha \to \beta$ and $\alpha \to \gamma$, then we may abbreviate this by writing $\alpha \to \beta \mid \gamma$.

To generate a string from a given grammar G, we start with the initial variable, then, given a production $\alpha \to \beta$, replace an instance of α with β, and repeat, until there are only terminal symbols remaining.

The sequence of substitutions to generate a string from a grammar is then called a *derivation.*

Example. Let $G = \big(\{S,T\},\{0,1\},R,S\big)$ be a grammar, where
$$R = \begin{Bmatrix} S \to TT, \\ T \to 0T1 \mid \varepsilon \end{Bmatrix}$$
Some derivations are as follows:
$$\begin{aligned}
S &\Longrightarrow TT \Longrightarrow 0T1T \Longrightarrow 01T && \Longrightarrow 010T1 \Longrightarrow 0101 \\
S &\Longrightarrow TT \Longrightarrow 0T1T \Longrightarrow 00T11T && \Longrightarrow 0011T \Longrightarrow 0011 \\
S &\Longrightarrow TT \Longrightarrow T\varepsilon \quad \Longrightarrow 0T1 && \Longrightarrow 01 \\
S &\Longrightarrow TT \Longrightarrow T\varepsilon \quad \Longrightarrow \varepsilon
\end{aligned}$$

△

A derivation is a *left-most derivation* if at each step, a production is applied to the left-most variable in the expression; *right-most derivations* are defined similarly.

Example. The first derivation in the previous example is a left-most derivation, and the last derivation is a right-most derivation. △

For any two strings $\alpha,\beta \in (V \cup \Sigma)^*$, we say that

- α *directly yields* β and write $\alpha \Longrightarrow \beta$ if α may be rewritten as β by applying a single production rule once;
- α *yields* β, or β is *derived from* α and write $\alpha \overset{*}{\Longrightarrow} \beta$ if α may be rewritten as β by applying a finite sequence of productions.

Given a grammar G, we then define the language $L(G)$ to be the set of all strings generated by G:
$$\begin{aligned}
L(G) &:= \{s \in \Sigma^* : s \text{ is derivable from } S \text{ using production rules in } G\} \\
&= \{s \in \Sigma^* : S \overset{*}{\Longrightarrow} s\}
\end{aligned}$$

Example. Let $G = (\{S\},\{0,1\},R,S)$ be a CFG, where
$$R = \{S \to S\}$$
Then, we have the (unique) production
$$S \Rightarrow S \Rightarrow S \Rightarrow S \Rightarrow \cdots$$
so G does not generate any strings, and hence $L(G) = \emptyset$. △

Example. Let $G = \big(\{S\},\{0,1\},R,S\big)$ be a grammar, where

$$R = \begin{Bmatrix} S \to 0S1, \\ S \to \varepsilon \end{Bmatrix}$$

or equivalently,

$$R = \{S \to 0S1 \mid \varepsilon\}$$

Then, we may generate the strings

$$\begin{aligned}
&S \Rightarrow 0S1 \Rightarrow 01 \\
&S \Rightarrow 0S1 \Rightarrow 00S11 \Rightarrow 0011 \\
&S \Rightarrow 0S1 \Rightarrow 00S11 \Rightarrow 000S111 \Rightarrow 000111 \\
&\quad\vdots \\
&S \Rightarrow 0S1 \Rightarrow \cdots \Rightarrow 0^nS1^n \Rightarrow 0^n1^n
\end{aligned}$$

so $L(G) = \{0^n1^n : n \in \mathbb{N}\}$. △

Evidently, grammars can generate non-regular languages.

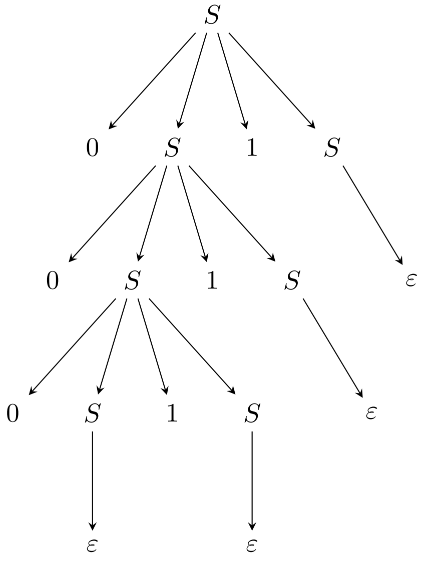
28.4.1 Parse Trees

We can represent a derivation with a *parse tree*, like the one at the beginning of this section.

Example. Let $G = \big(\{S\},\{0,1\},R,S\big)$ be a grammar, where

$$R = \{S \to 0S1S \mid \varepsilon\}$$

Then, the parse tree for one possible left-most derivation is:



Then, reading the terminals of the tree with an inorder depth first search, we obtain the string

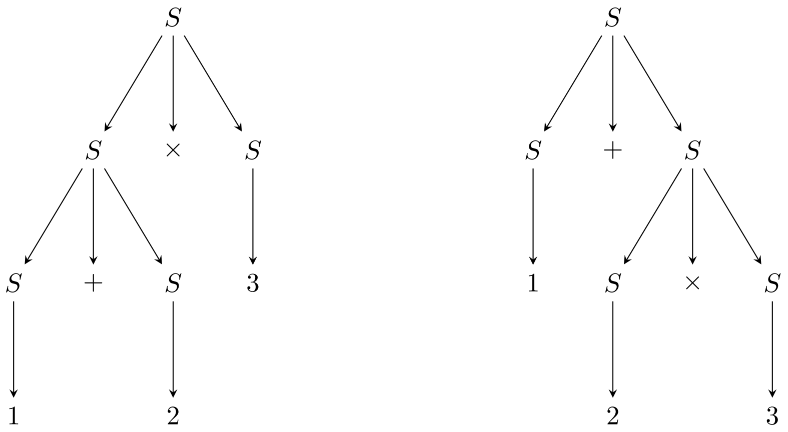
$$000\varepsilon1\varepsilon1\varepsilon1\varepsilon = 000111$$

also called the *yield* of the tree. △

Let $G = \big(\{S\},\{+, \times, 0,1,\dots,9\},R,S\big)$ be a CFG, where

$$R = \begin{Bmatrix} S \to S + S, \\ S \to S \times S, \\ S \to (S) \\ S \to 0 \mid 1 \mid \cdots \mid 9 \end{Bmatrix}$$

Consider the following pair of parse trees:



These trees both yield the string $1 + 2 \times 3$, so there isn't a unique parse tree for this string.

A grammar G is *ambiguous* if it can generate the same string with multiple parse trees, or equivalently, if the same string can be derived from two left-most derivations.

Some ambiguous grammars G may be rewritten as an equivalent unambiguous grammar H, with $L(G) = L(H)$. However, not all grammars admit an unambiguous equivalent. Such grammars are called *inherently ambiguous* grammars.

The problem of determining whether a grammar is ambiguous or not is undecidable.

28.4.2 Right/Left-Linear Grammars

A *linear grammar* is a grammar that has at most one variable in the right side of each substitution rule.

Example. The grammar $G = \big(V,\{0,1\},R,S\big)$ with rules

$$R = \begin{Bmatrix} S \to 0S1, \\ S \to \varepsilon \end{Bmatrix}$$

is linear, and generates the language $L(G) = \{0^n1^n : n \in \mathbb{N}\}$. △

As demonstrated by this example, linear grammars may accept some non-regular languages. However, we may add a further restriction:

A *right-linear grammar* is a grammar $G = (V,\Sigma,R,S)$ where each rule is of the following form

- $A \to xB$;
- $A \to x$;

where $A,B \in V$ are variables and $x \in \Sigma^*$ is a string of terminals. A *left-linear grammar* is defined similarly, with the first production rule replaced by $A \to Bx$.

Any derivation of a word from a strict right-linear grammar is of the form

$$S \Longrightarrow a_1A_1 \Longrightarrow a_1a_2A_2 \Longrightarrow \cdots \Longrightarrow a_1a_2\cdots a_n$$

where $(A_i)_i \subseteq V$ and $(a_i)_i \subseteq \Sigma^*$; that is, strings grow towards the right as a derivation progresses.

It turns out that we can strengthen this restriction even further (and it will be convenient for us to do so) without changing the class of languages the grammar accepts:

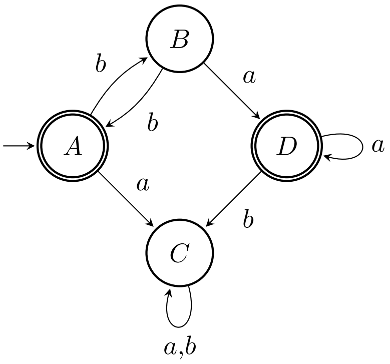
A *strictly right-linear grammar* is a grammar $G = (V,\Sigma,R,S)$ where each rule is of the following form

- $A \to xB$, where $A \in V$, $x \in \Sigma \cup \{\varepsilon\}$;
- $A \to x$.

Strings still grow to the right during a derivation, but the productions now only add a single symbol at a time.

Strightly (right/left)-linear grammars cannot accept all context-free languages. In fact, they accept precisely the regular languages.

Example. Consider the DFA



We construct the associated strightly right-linear grammar. We represent states as variables, with the starting variable representing the starting state, and the alphabet should be the same, so the grammar will be of the form $G = \big(\{A,B,C,D\},\{0,1\},R,A\big)$.

Then, productions should match up with the *outgoing* transitions, and whenever we have an accepting state, we allow a production of the empty string from the variable representing that state.

For instance, at the initial state A, we have transitions $\delta(A,0) = C$ and $\delta(A,1) = B$, so we have productions

$$A \to aC$$
$$A \to bB$$

A is also an accepting state, so we also have

$$A \to \varepsilon$$

For the other states, we have

$$B \to aD$$
$$B \to bA$$
$$C \to aC$$
$$C \to bC$$

$$
\begin{aligned}
&D \to aD\\
&D \to bC\\
&D \to \varepsilon
\end{aligned}
$$

Note that all the production rules are of the required form for this grammar to be right-linear.

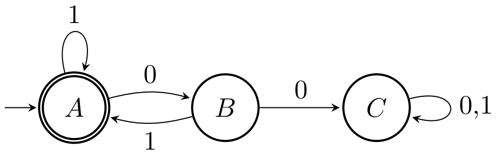
The only way to remove a variable is by replacing it with ε, but by construction, this happens only when the current state is an accepting state. Thus, this grammar generates precisely the language that the DFA recognises. △

We can also go the other way and construct a DFA given any strictly right-linear grammar.

Example. Consider the strictly right-linear grammar $G = \big(\{A,B,C\},\{0,1\},R,A\big)$, where

$$
R = \begin{Bmatrix} A \to 0B \mid 1A \mid \varepsilon, \\ B \to 0C \mid 1A, \\ C \to 0C \mid 1C \end{Bmatrix}
$$

The process is largely the same as the previous in reverse: we introduce a new state for each variable, add transitions $\delta(A,b) = C$ for each production $A \to bC$, and mark any variables with productions $A \to \varepsilon$ as accepting states:



△

Together, these constructions show more generally that:

Theorem 28.4.1. *A language L is accepted by a strict right-linear grammar if and only if L is regular.*

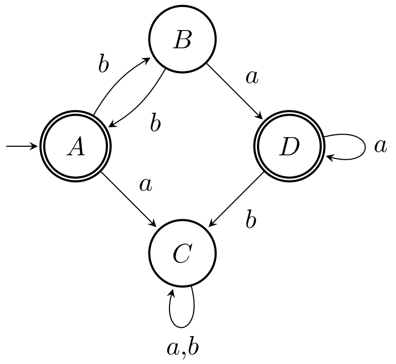
Proof. Let $L = L(M)$ for a DFA $M = (Q,\Sigma,q_0,F,\delta)$. Then, L is recognised by the right-linear grammar $G = (Q,\Sigma,R,q_0)$, where

$$
R = \begin{Bmatrix} q \to ap & \forall q,a \in Q : \delta(q,a) = p \\ q \to \varepsilon & \forall q \in F \end{Bmatrix}
$$

The reverse construction is analogous: given a strictly right-linear grammar $G = (V,\Sigma,R,S)$, introduce a state for each variable $A \in V$; for each rule $A \to bC$ with $A,C \in V$, $b \in \Sigma$, define $\delta(A,b) = C$; and for each rule $A \to \varepsilon$, make the state A an accepting state. ■

We can also construct the strictly left-linear grammar with a similar modified method:

Example. Consider the same DFA as in the previous example:



To construct the associated strictly left-linear grammar, we proceed in much the same way, representing states as variables, but this time, we start at the *final* state, and represent *incoming* transitions with production rules, and finally add a production for the empty string only at the starting state.

The first problem is that there are multiple final states, but a grammar only allows one starting variable. We resolve this analogously to adding ε-transitions in an NFA: add a new state q^* and add productions of the form $q^* \to q$ for every final state q.

So, in this example, we have

$$\begin{aligned} q^* &\to A \\ q^* &\to D \end{aligned}$$

(Note that this is of the required form, since $A = A\varepsilon$.)

Then, we add production rules corresponding to incoming transitions:

$$\begin{aligned} A &\to \varepsilon \mid Bb \\ B &\to Ab \\ C &\to Aa \mid Ca \mid Cb \mid Db \\ D &\to Ba \mid Da \end{aligned}$$

This defines the required grammar $G = \big(\{q^*,A,B,C,D\},\{a,b\},R,q^*\big)$. △

The reverse construction is again similar. Thus, we have:

Theorem 28.4.2. *A language L is accepted by a strictly left-linear grammar if and only if L is regular.*

Proof. Let $L = L(M)$ for a DFA $M = (Q,\Sigma,q_0,F,\delta)$. Then, L is recognised by the strictly left-linear grammar $G = (Q \cup \{q^*\},\Sigma,R,q^*)$, where

$$R = \begin{Bmatrix} q_0 \to \varepsilon & \\ p \to qa & \forall q,a \in Q : \delta(q,a) = p \\ q^* \to q & \forall q \in F \end{Bmatrix}$$

Conversely, given a strictly left-linear grammar $G = (V,\Sigma,R,S)$, introduce a state for each variable $A \in V \setminus S$; for each rule $A \to Bc$ with $A,B \in V$ and $c \in \Sigma$, define $\delta(B,c) = A$; and for each rule $S \to A$, make the state A an accepting state. ■

28.4.3 Chomsky Hierarchy of Grammars

As we have seen previously, grammars can generate a wider class of languages than just regular languages. We can precisely classify when this happens in terms of constraints on what kind of productions are allowed in a grammar.

Let $G = (V,\Sigma,R,S)$ be a grammar, $A,B \in V$ be variables, $\alpha,\beta,\gamma,\delta \in (V \cup \Sigma)^*$ be strings of arbitrary symbols, and $x \in \Sigma^*$ be a string of terminals.

Grammar	Languages	Recognising Automata	Constraints
Type-3	Regular/(Right/Left)-Linear	Finite automata	$\left.\begin{cases} A \to x \\ A \to xB \end{cases}\right\}$ (right regular) or $\left.\begin{cases} A \to x \\ A \to Bx \end{cases}\right\}$ (left regular)
Type-2	Context-free	Non-deterministic pushdown automata	$A \to \alpha$
Type-1	Context-sensitive	Linear-bounded non-deterministic Turing machine	$\alpha A \beta \to \alpha \gamma \beta$
Type-0	Recursively enumerable	Turing machine	$\alpha \to \beta$ $(\alpha \neq \varepsilon)$

Each type is a proper subset of the next, so there are recursively enumerable languages that are not context-sensitive, context-sensitive languages that are not context-free, and context-free languages that are not regular.

We call a type-2 grammar a *context-free grammar* (CFG), and the language generated by a CFG is called a *context-free language* (CFL).

28.5 Context-Free Languages

28.5.1 Pushdown Automata

A (*non-deterministic*) *pushdown automaton* (PDA) is a 6-tuple $P = (Q,\Sigma,\Gamma,\delta,q_0,F)$ consisting of

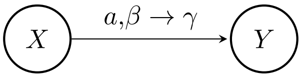
- a finite set Q of states;
- a finite set Σ, the input alphabet;
- a finite set Γ, the *stack symbol* alphabet;
- a transition function $\delta : Q \times \Sigma_\varepsilon \times \Gamma_\varepsilon \to \mathcal{P}(Q \times \Gamma_\varepsilon)$;
- an initial state $q_0 \in Q$;
- a set $F \subseteq Q$ of accepting states.

A PDA is effectively an NFA equipped with some limited memory in the form of a stack, to which we can push symbols from Γ and pop with the transition function. We will also usually assume that $\Gamma \supseteq \Sigma$ so we can store any read symbols on the stack.

As we will see, this additional memory allows us to recognise a strictly larger class of languages than NFA/DFA, such as the non-regular palindrome languages $L = \{ww^{\text{rev}} : w \in \Sigma^*\}$.

Also note that we are working with non-deterministic PDA. We will not discuss them in detail here, but unlike NFAs and DFAs which are equivalent, *deterministic* pushdown automata are provably less expressive than their non-deterministic counterparts.

As usual, we may represent a PDA as a state transition diagram; but now, the transitions are controlled not only by the currently read symbol, but also by the state of the stack. The labels in a state diagram are given in the form $a,\beta \to \gamma$, where $a \in \Sigma_\varepsilon$ and $\beta,\gamma \in \Gamma_\varepsilon$, where the arrow indicates the stack operation of popping β then pushing γ.



That is, when we are in the state S, this transition may only be traversed if the current read symbol is a and the top element of the stack is β.

In terms of the transition function, the label $a,\beta \to \gamma$ from a state X to a state Y represents the element $(Y,\gamma) \in \delta(X,a,\beta)$.

Any or all of a, β, and γ may be the empty string: if $a = \varepsilon$, then the transition consists only of the stack operation $\beta \to \gamma$, and it may be traversed without reading any symbols from the input string; if $\beta = \varepsilon$, then the stack operation just pushes γ to the stack, as popping the empty string effectively does not change the state of the stack, since we may assume infinitely many empty strings are on top of the stack; and if $\gamma = \varepsilon$, the stack operation just pops β from the stack, as pushing the empty string again does not change the state of the stack.

We also write $a,\beta \to \gamma_1 \dots \gamma_n$ to denote pushing multiple symbols onto the stack (note that γ_n is pushed first, and γ_1 last, in this notation). This can be converted into an ordinary transition via the provision of some new states such that the intermediary transitions only push one of the γ_i at a time.

Because the stack may always be regarded as having infinitely many empty strings on top, it is difficult to determine whether the stack is empty or not, so for convenience, we also often include the string \$ in Γ which we immediately push on to the stack at the beginning of a computation using the transition $\varepsilon,\varepsilon \to \$$. From then on, we can use the \$ symbol to detect when the stack is intended to be "empty", and we can use the transition $\varepsilon,\$ \to \varepsilon$ to fully empty the stack.

28.5.2 Languages Recognised by PDA

Given a PDA $P = (Q,\Sigma,\Gamma,\delta,q_0,F)$, we say that P *accepts* or recognises a string $s_1 \cdots s_k = s \in \Sigma^*$ if there exists a sequence $(r_i)_{i=1}^k \subseteq Q$ of states and a sequence $(\sigma_i)_{i=1}^k \subseteq \Gamma$ of stack symbols such that

- $r_0 = q_0$;
- $r_k \in F$;
- For all i, $(r_i,\beta) \in \delta(x_{i-1},s_i,\alpha)$ where $r_{i-1} = \alpha \cdot t$ and $r_i = \beta \cdot t$ for some $\alpha,\beta \in \Gamma_\varepsilon$, $t \in \Gamma^*$.

Example. Consider the CFG $G = \big(\{S\},\{0,1\},R,S\big)$ where

$$R = \begin{Bmatrix} S \to 0S1 \\ S \to \varepsilon \end{Bmatrix}$$

which generates the language

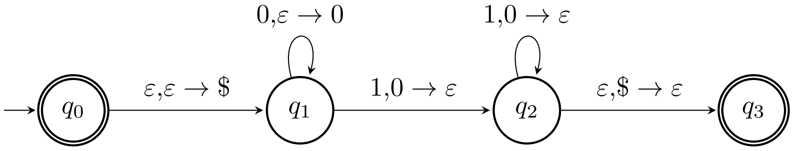
$$L(G) = \{0^n1^n : n \geq 0\}$$

We construct a PDA that recognises this language as follows.

First, push the empty stack symbol \$ to the stack with $\varepsilon,\varepsilon \to \$$ from the initial state q_0 to a state q_1. Then, whenever we read a 0 from the string, push it on to the stack with $0,\varepsilon \to 0$, and we can repeat this arbitrarily many times, so this transition is a loop on q_1.

When we read a 1 from the string for the first time, we pop a 0 from the stack with $1,0 \to \varepsilon$ and move to a new state q_2, as we we will not allow any more 0s to be read from the input string. In this new state, we can then read 1s from the string and pop 0s from the stack.

If at any point, the stack is empty (i.e. we can read the \$ symbol), we can move to a final accepting state q_3.



△

We claim that PDA recognise precisely the class of context-free languages. To show this, we will show that every CFG can be converted into a PDA and vice versa.

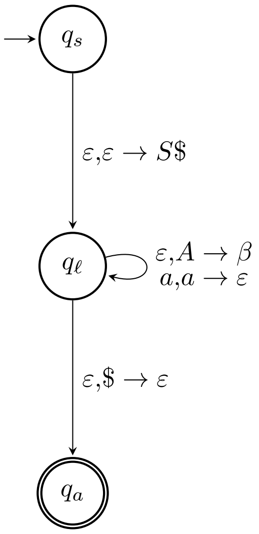
Lemma 28.5.1. *If a language L is context-free, then there exists a PDA that recognises L.*

Proof. Since L is context-free, there is a CFG $G = (V,\Sigma,R,S)$ such that $L(G) = L$.

We can decide whether a given string T is derivable from some fixed grammar using the following algorithm:

1. Push $S\$$ on to the stack.
2. While the symbol on the top of the stack is not \$:
   - If the top of the stack is a variable A, pop A and non-deterministically select a grammar rule for A from R, and push the production to the stack.
   - Otherwise, the top of the stack is a terminal a. Read the next input symbol in T and check if it equals a. If so, pop a from the stack and continue. Otherwise, reject T.
3. Once the top of the stack is \$, accept.

We implement this algorithm as a PDA as follows:


q_s
$\varepsilon,\varepsilon \to S\$$
q_ℓ
$\varepsilon,A \to \beta$
$a,a \to \varepsilon$
$\varepsilon,\$ \to \varepsilon$
q_a

where the loop in the middle has a transition rule of the form $\varepsilon,A \to \beta$ for every terminal with production $A \to \beta$ in R; and a transition of the form $a,a \to \varepsilon$ for every terminal $a \in A$.

Note that we are implicitly using more than just 3 states when pushing multiple symbols to the stack. The vertical transitions represent steps 1 and 3 of the algorithm above, and the self-loops on q_ℓ simulating the grammar represents step 2.

The first kind of loops of the form $\varepsilon,A \to w$ correspond to a production being applied to the left-most variable A in a derivation. The second kind of loop $a,a \to \varepsilon$ correspond to matching the input T with the currently generated string. It is safe to do this incrementally, since a terminal is never replaced in a derivation, so if the first symbol is a terminal at any point, it will always be the same terminal at any point onwards in the derivation. ■

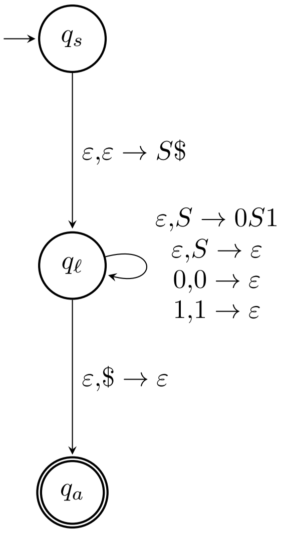
Example. Consider the CFG $G = \big(\{S\},\{0,1\},R,S\big)$ where

$$R = \begin{Bmatrix} S \to 0S1 \\ S \to \varepsilon \end{Bmatrix}$$

which generates the language

$$L(G) = \{0^n1^n : n \geq 0\}$$

The corresponding PDA is given by


q_s
$\varepsilon,\varepsilon \to S\$$
$\varepsilon,S \to 0S1$
$\varepsilon,S \to \varepsilon$
$0,0 \to \varepsilon$
$1,1 \to \varepsilon$
q_ℓ
$\varepsilon,\$ \to \varepsilon$
q_a

△

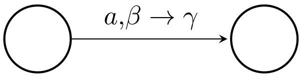
Before we prove the converse, we define a convenient normal form of PDA.

A *normalised PDA* is a PDA such that

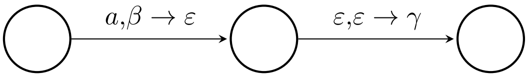
- N has a single accept state;
- N empties its stack before accepting;
- Each transition does exactly one of the following;
  - Push a symbol onto the stack;
  - Pop a symbol from the stack.

Every PDA can be converted into a normalised PDA:

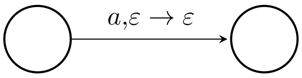
- Add a new accept state q_{accept} and add transitions of the form $\varepsilon,\$ \to \varepsilon$ from the old accept states to q_{accept};
- For every old accepting state and every stack symbol $\gamma \in \Gamma$, add a loop of the form $\varepsilon,\gamma \to \varepsilon$ to empty the stack;
- For every transition with both stack instructions,



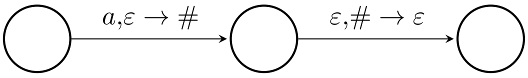
add a new intermediary state with transitions


$a,\beta \to \varepsilon$
$\varepsilon,\varepsilon \to \gamma$

For transitions with no stack instructions,



instead push and pop a new symbol # on the new transitions:



So, we may assume without loss of generality that all of our PDA are in normalised form.

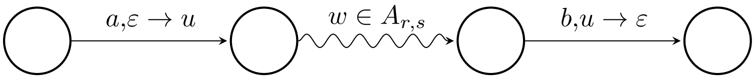
Lemma 28.5.2. *If M is a PDA, then $L(M)$ is a context-free language.*

Proof. By the above, there exists a normalised PDA $N = (Q,\Sigma,\Gamma,\delta,q_0,\{q_a\})$ such that $L(N) = L(M)$. We construct a CFG $G = (V,\Sigma,R,S)$ as follows.

For each pair of states $p,q \in Q$, add a variable $A_{p,q}$ to V. If we can guarantee that $A_{p,q}$ generates precisely the set of strings that take us from p (starting with an empty stack) to q (ending with an empty stack), then we are done, as $S \coloneqq A_{q_0,q_a}$ would generate precisely the language accepted by N.

To achieve this, we define three types of rules.

1. For each state $p \in Q$, add the rule
$$A_{p,p} \to \varepsilon$$
since not reading any characters is a valid path from p to p with empty stacks.
2. For all states $p,q,r \in Q$, add the rule
$$A_{p,q} \to A_{p,r}A_{r,q}$$
since travelling from p to r with empty stacks, then r to q with empty stacks, is a valid path from p to q with empty stacks;
3. For all states $p,q,r,s \in Q$ and stack symbol $u \in \Gamma$ and (possibly empty) letters $a,b \in \Sigma_\varepsilon$, if $(r,u) \in \delta(p,a,\varepsilon)$ and $(q,\varepsilon) \in \delta(s,b,u)$, add the rule
$$A_{p,q} \to aA_{rs}b$$
since if r is reachable from p by pushing u to the stack, and q is reachable from s by popping u, then concatenating with any string w given by $A_{r,s}$ gives a path from p with empty stack to q with empty stack, since by construction, w leaves the stack unchanged, so the u is still available to be popped during the final transition.



■

The previous two results imply:

Theorem 28.5.3. *A language L is context-free if and only if there is a PDA recognising L.*

28.5.3 Chomsky Normal Form

A CFG $G = \{V,\Sigma,R,S\}$ is in *Chomsky Normal Form* (CNF) if every substitution rule in R is one of the following form:

- $S \to \varepsilon$;
- $A \to s$, where $A \in V$, $s \in \Sigma$;
- $A \to PQ$, where $A \in V$, $P,Q \in V \setminus \{S\}$.

Note that every parse tree generated by a CFG in CNF is a binary tree: every production either splits a variable into two variables, or replaces it with a terminal, creating a leaf node.

This also implies that every string of length n can be derived in exactly $2n-1$ production steps, since it requires exactly $n-1$ steps to produce n variables from S, and n more steps to replace the variables with terminals.

Theorem 28.5.4. *Every CFL is generated by a CFG in CNF.*

Proof. Given a CFG $G = \{V,\Sigma,R,S\}$, we convert G to CNF via the following procedure:

- **Eliminate the start symbol from right sides**:

  Add a new initial variable S_0 and a substitution rule $S_0 \to S$ to the old initial variable. This does not change the grammar's produced language, and the new initial variable S_0 will not occur on any rule's right side.

- **Eliminate rules with non-solitary terminals**

  To eliminate a rule where the right side is some combination of terminals $(a_i)_i$ and variables $(X_j)_j \neq \emptyset$,
  $$A \to (a_n \mid X_n)_n$$
  introduce for each such terminal a_i a new variable N_{a_i} and a new rule $N_{a_i} \to a_i$, then replace the rules
  $$A \to (a_n \mid X_n)_n$$
  with
  $$A \to (N_{a_n} \mid X_n)_n$$
  That is, replace any instance of a_i with N_{a_i}.

- **Eliminate rules with more than 2 variables**

  Replace each rule where the right side is some combination of $k \geq 3$ variables
  $$A \to X_1 \cdots X_k$$
  by introducing new variables $(A_i)_{i=1}^{n-2}$
  $$\begin{aligned} A &\to X_1A_1 \\ A_1 &\to X_2A_2 \\ A_2 &\to X_3A_3 \\ &\vdots \\ A_{n-3} &\to X_{n-2}A_{n-2} \\ A_{n-2} &\to X_{n-1}X_n \end{aligned}$$
  (Similar to currying.)

- **Eliminate ε-rules**

  Remove any rules of the form $A \to \varepsilon$ for $A \neq S_0$. Then, for each rule containing n occurrences of A, add 2^n new copies of that rule with each possible combination of A replaced by ε (i.e. removed). Similar to inline expansion or β-reduction.

  For instance, if we had rules $A \to \varepsilon$ and $R \to aAbAcAd$, then we would remove this two rules and add in

  $$\begin{aligned}
  &R \to aAbAcAd\\
  &R \to aAbAcd\\
  &R \to aAbcAd\\
  &R \to aAbcd\\
  &R \to abAcAd\\
  &R \to abAcd\\
  &R \to abcAd\\
  &R \to abcd
  \end{aligned}$$

  Directly remove any ε-rules introduced by this step.

- **Eliminate unit rules**

  A *unit rule* is a rule of the form
  $$A \to B$$
  for some variables A,$B \in V$.

  To remove a unit rule $A \to B$, for each rule

  $$B \to \Lambda_1 \cdots \Lambda_n$$

  where $(\Lambda_i)_{i=1}^n \subseteq V \cup \Sigma$ is some combination of variables and terminals, add a new rule

  $$A \to \Lambda_1 \cdots \Lambda_n$$

  unless this is a unit rule already removed.

■

Example. Consider a grammar with production rules

$$\begin{aligned}
S &\to ASB\\
A &\to aAS \mid a \mid \varepsilon\\
B &\to SbS \mid A \mid bb
\end{aligned}$$

We will convert this into CNF.

Start by eliminating the start variable S:

$$\begin{aligned}
S_0 &\to S\\
S &\to ASB\\
A &\to aAS \mid a \mid \varepsilon\\
B &\to SbS \mid A \mid bb
\end{aligned}$$

Then, we eliminate the terminals in $A \to aAS$, $B \to SbS$ and $B \to bb$:

$$\begin{aligned}
S_0 &\to S \\
S &\to ASB \\
A &\to N_aAS \mid a \mid \varepsilon \\
B &\to SN_bS \mid A \mid N_bN_b \\
N_a &\to a \\
N_b &\to b
\end{aligned}$$

Now, we eliminate the rules with 3 or more variables, $S \to ASB$, $A \to N_aAS$, and $B \to SN_bS$:

$$\begin{aligned}
S_0 &\to S \\
S &\to AV_1 \\
A &\to N_aV_2 \mid a \mid \varepsilon \\
B &\to SV_3 \mid A \mid N_bN_b \\
N_a &\to a \\
N_b &\to b \\
V_1 &\to SB \\
V_2 &\to AS \\
V_3 &\to N_bS
\end{aligned}$$

Now, we eliminate the ε-rule $A \to \varepsilon$

$$\begin{aligned}
S_0 &\to S \\
S &\to AV_1 \mid V_1 \\
A &\to N_aV_2 \mid a \\
B &\to SV_3 \mid A \mid N_bN_b \mid \cancel{\varepsilon} \\
N_a &\to a \\
N_b &\to b \\
V_1 &\to SB \\
V_2 &\to AS \mid S \\
V_3 &\to N_bS
\end{aligned}$$

Now, we eliminate the unit rules $S \to V_1$, $B \to A$, and $V_2 \to S$:

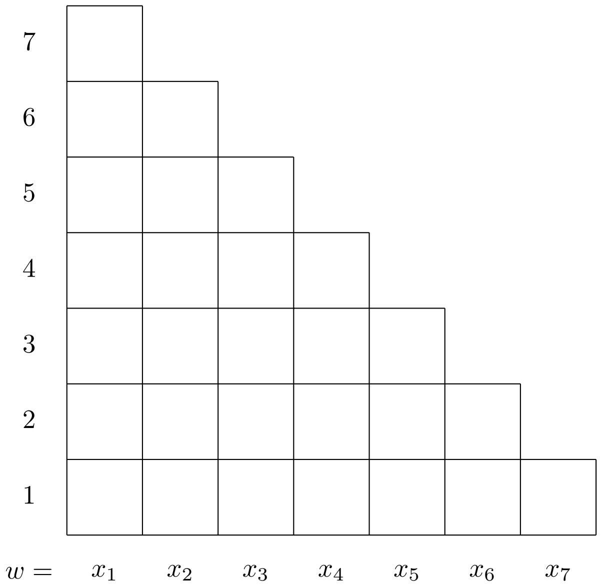
$$\begin{aligned}
S_0 &\to S \\
S &\to AV_1 \mid SB \\
A &\to N_aV_2 \mid a \\
B &\to SV_3 \mid N_bN_b \mid N_aV_2 \mid a \\
N_a &\to a \\
N_b &\to b \\
V_1 &\to SB \\
V_2 &\to AS \mid AB_1 \mid SB \\
V_3 &\to N_bS
\end{aligned}$$

△

28.5.4 Cocke–Younger–Kasami (CYK) Parsing

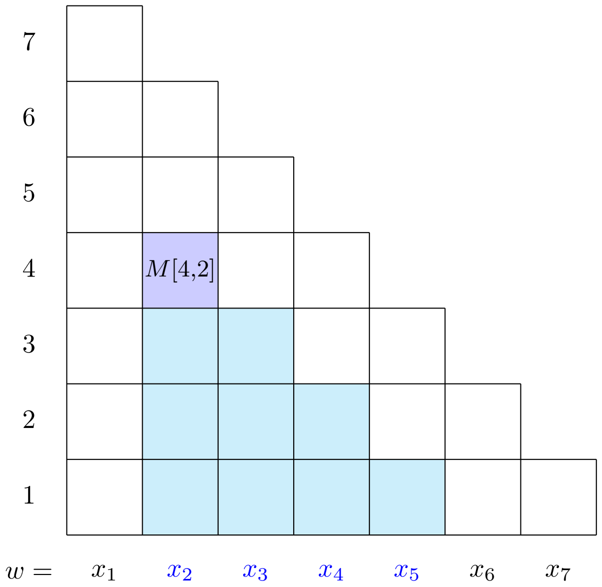
The CYK algorithm is a $\Theta(n^3 \cdot |G|)$ time dynamic programming algorithm for the bottom-up parsing of a CFG in CNF.

We build up a lower triangular table with width and height equal to the length of the word w we are parsing. For instance, if $w = x_1x_2x_3x_4x_5x_6x_7$, then the table is initialised as:



Each entry $M[i{,}j]$ will contain the set of variables that can generate the substring $x_jx_{j+1}\cdots x_{i+j-1}$ of w. Note that the row number corresponds to the length of the substring.

For instance, $M[4{,}2]$ contains the set of variables that can generate the substring $x_2x_3x_4x_5$. Visually, this is the substring at the base of the triangular "cone" under the entry:



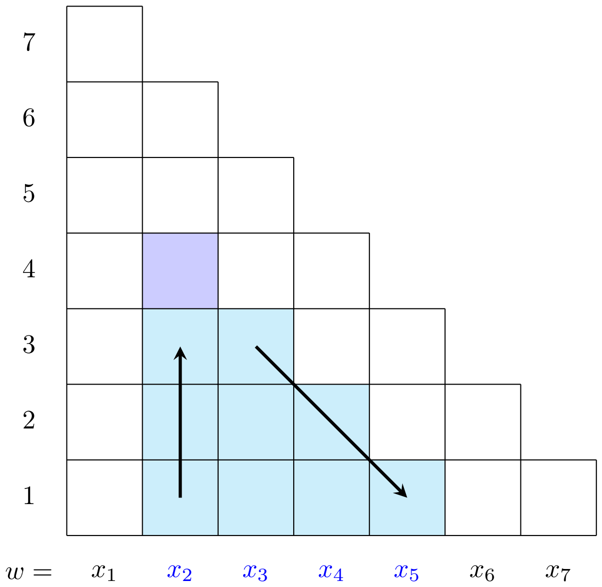
We start from the bottom row, and recursively fill in the table by considering how each substring can be constructed by concatenating previously constructed strings. For instance, the string $x_2x_3x_4x_5$ represented by the $M[4{,}2]$ cell can be produced by concatenating

- x_2 with $x_3x_4x_5$;
- x_2x_3 with x_4x_5;
- $x_2x_3x_4$ with x_5.

In the first case, the set of variables that generate x_2 is given by $M[1{,}2]$, and the set of variables that generate $x_3x_4x_5$ is given by $M[3{,}3]$. We would then take each possible concatenation of a variable in the first set with a variable in the second set, check if any of these concatenations can be generated by the grammar, and place them in $M[4{,}2]$ if they are.

The second case, we would check the concatenation of variables in $M[2{,}2]$ and $M[4{,}2]$, and in the third case, we would check the concatenation of variables in $M[3{,}2]$ and $M[5{,}1]$.

Note that since the left substring is growing by one character in each case, the variables are given by the cells in the column below the current cell, starting at the bottom, as the cones of these cells cover one more character each level up. Similarly, the second substring is given by the cells along the right descending diagonal of the current cell:



Once the table has been filled, the unique cell in the top row will contain the starting variable S if and only if the given word w can be derived from the grammar.

Example. Consider the CFG $G = \big(\{S,A,B,C\},\{a,b\},R,S\big)$ where

$$R = \left\{\begin{array}{l} S \to AB \mid BC, \\ A \to BA \mid a, \\ B \to CC \mid b, \\ C \to AB \mid a \end{array}\right\}$$

Note that G is in CNF, as required.

We will parse the string $w = baaba$. The table is initialised as:

5					
4					
3					
2					
1					
	b	a	a	b	a

The first row is simple to fill in, as the substrings consist of single terminals:

5					
4					
3					
2					
1	B	A,C	A,C	B	A,C
	b	a	a	b	a

Now, consider the first cell on the second row, $M[2,1]$. This corresponds to the substring ba.

This substring can be obtained by concatenating any variable that produce b followed by any variable that produces a, which we have already computed in the cells contained in the cone below this cell as $\{B\}$ and $\{A,C\}$.

We have the possible concatenations BA and BC. BA is produced by A, and BC is produced by S, so $M[2,1]$ contains S,A.

We continue similarly for the rest of the row. For the second cell, we have concatenations AA, AC, CA, and CC. CC can be produced by B, and the others cannot be produced, so this cell contains B. The third cell has concatenations AB and CB, which are produced by C and S, respectively. The fourth cell has concatenations BA and BC, which are produced by A and S, respectively.

5					
4					
3					
2	S,A	B	S,C	S,A	
1	B	A,C	A,C	B	A,C
	b	a	a	b	a

Moving onto the third row, the first cell $M[3,1]$ corresponds to the substring baa.

This can be produced by concatenating b with aa, or ba with a. From the previously computed cells in the cone below $M[3,1]$, we already know how to produce these substrings:

- We can produce b with B ($M[1,1]$) and aa with B ($M[2,2]$), so the concatenations are BB;
- We can produce ba with S,A ($M[2,1]$) and a with A,C ($M[1,3]$), so the concatenations are SA, SC, AA, and AC.

No productions generate any of these, so $M[3,1]$ is empty.

The next cell corresponds to the substring aab, which can be produced by concatenating a with ab, or aa with b.

- We can produce a with A,C and ab with S,C, so the concatenations are AS, AC, CS, and CC;
- We can produce aa with B and b with B, so the concatenations are BB.

Only CC can be produced (by B), so $M[3,2]$ contains B.

The next cell corresponds to the substring aba, which can be produced by concatenating a with ba, or ab with a.

- We can produce a with A,C and ba with S,A, so the concatenations are AS, AA, CS, and CA;
- We can produce ab with S,C and a with A,C, so the concatenations are SA, SC, CA, and CC.

Only CC can be produced (by B), so $M[3,3]$ contains B.

5					
4					
3	$\varnothing$	B	B		
2	S,A	B	S,C	S,A	
1	B	A,C	A,C	B	A,C
	b	a	a	b	a

On the fourth row, the first cell represents the substring $baab$, which can be constructed as:

- $b + aab$, produced by B and B, so the concatenations are BB;
- $ba + ab$, produced by S,A and S,C, so the concatenations are SS, SC, AS, AC;
- $baa + b$, produced by $\varnothing$ and B, so there are no concatenations.

(The pattern in the table should now be more apparent: we use the entries on each "leg" of the cone, starting from the bottom of one, and the top of the other.)

None of these concatenations are producible, so $M[4{,}1]$ is empty.

The second cell represents the substring $aaba$, which can be constructed as

- $a + aba$, produced by A,C and B, so the concatenations are AB, CB;
- $aa + ba$, produced by B and S,A, so the concatenations are BS, BA;
- $aab + a$, produced by B and A,C, so the concatenations are BA, BC.

AB can be produced by S and C, BA by A, and BC by S, so $M[4{,}2]$ contains S, A, and C:

5					
4	$\varnothing$	S,A,C			
3	$\varnothing$	B	B		
2	S,A	B	S,C	S,A	
1	B	A,C	A,C	B	A,C
	b	a	a	b	a

Finally, the unique cell at the top represents the entire string, which can be constructed as

- $b + aaba$, produced by B and S,A,C, giving BS, BA, BC;
- $ba + aba$, produced by S,A and B, giving SB, AB;
- $baa + ba$, produced by $\emptyset$ and S,A, so no concatenations;
- $baab + a$, produced by $\emptyset$ and A,C, so no concatenations.

BA can be produced by A, BC by S, and AB by S and C, so $M[5{,}1]$ contains S, A, and C:

5	S,A,C				
4	$\varnothing$	S,A,C			
3	$\varnothing$	B	B		
2	S,A	B	S,C	S,A	
1	B	A,C	A,C	B	A,C
	b	a	a	b	a

S is in the cell on the top row, so $w = baaba$ can be generated by this grammar. △

We can also recover the parse tree from the diagram by connecting non-empty entries along the legs of the cones.

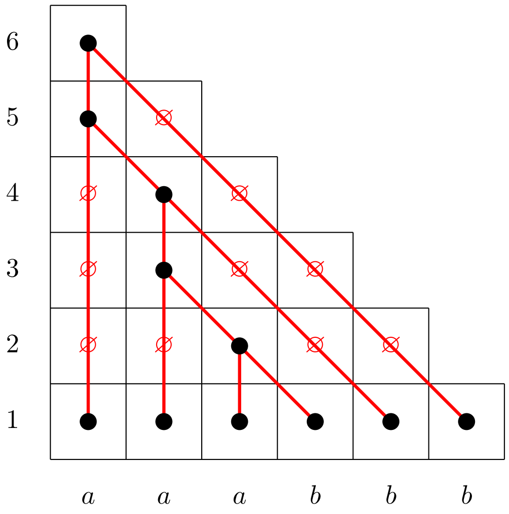
Example. Consider the CFG $G = \big(\{S,T,X,A,B\},\{a,b\},R,S\big)$ where

$$R = \left\{ \begin{array}{l} S \to AB \mid XB \mid \varepsilon, \\ T \to AB \mid XB, \\ X \to AT, \\ A \to a, \\ B \to b \end{array} \right\}$$

with the following CYM table parsing the word $aaabbb$:

6	S					
5	X	$\varnothing$				
4	$\varnothing$	T	$\varnothing$			
3	$\varnothing$	X	$\varnothing$	$\varnothing$		
2	$\varnothing$	$\varnothing$	T	$\varnothing$	$\varnothing$	
1	A	A	A	B	B	B
	a	a	a	b	b	b

Then, the parse tree is given by:



△

28.6 Non-Context-Free Languages

28.6.1 The Pumping Lemma for Context-Free Languages

Recall the pumping lemma for regular languages:

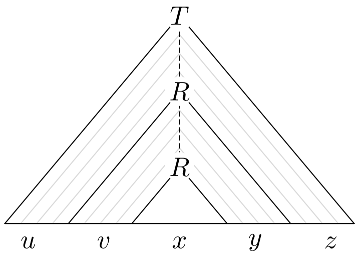
Lemma (Pumping Lemma)**.** *Let L be a regular language. Then, there exists an integer $p \geq 1$ (the pumping length), such that for every string $s \in L$ with length $|s| \geq p$, there exists a decomposition $s = x \cdot y \cdot z$ such that*

- $|y| \geq 1$*;*
- $|xy| \leq p$*;*
- *for all* $n \geq 0$*,* $x \cdot y^n \cdot z \in L$*.*

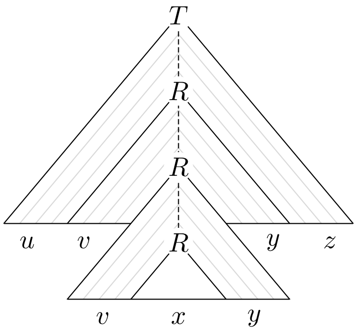
The intuition is that, if an input string is too long, then it must loop somewhere inside the DFA.

A similar result holds for context-free languages, in that if a derived string is too long, it must repeat a variable somewhere in its parse tree. Similarly to before we can generate an infinite set of strings in the language by pumping this repeated variable:

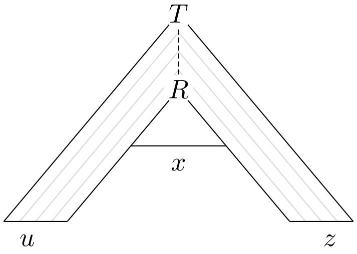
Suppose we repeat a variable R, and that the derivation between the first and second R yields a string that starts with a string v and ends with a string y.



Then, if we repeat R again, the derivation will repeat v and y:



We can also remove the repeated R:



Lemma 28.6.1 (Pumping Lemma)**.** *Let L be a context free language. Then, there exists an integer $p \geq 1$ (the pumping length) such that for every string $s \in L$ with length $|s| \geq p$, there exists a decomposition $s = u \cdot v \cdot x \cdot y \cdot z$ such that*

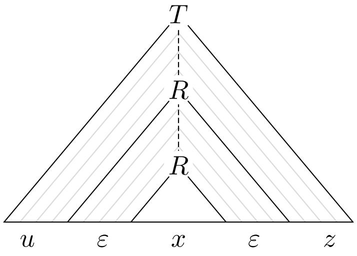
- $|vy| \geq 1$*;*
- $|vxy| \leq m$*;*
- *for all $n \geq 0$, $u \cdot v^n \cdot x \cdot y^n \cdot z \in L$.*

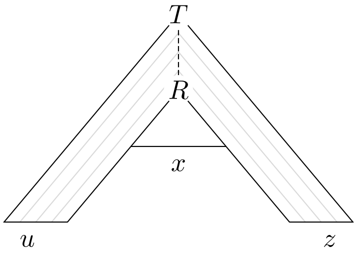
Proof. Let $G = (V,\Sigma,R,S)$ be a CFG, and suppose that the length of the longest string in the right side of a production rule in R is b. Then, any node in a parse tree yielding a string in $L(G)$ has at most b children. If the height of the tree is h, then it has at most b^h leaves, as each layer can only have b times as many nodes as the previous layer.

How long must a string be such that a variable is repeated on some root to leaf path of any parse tree of this string?

We claim that $p \coloneqq b^{|V|+1}$ is sufficient. Indeed, the parse tree of such a word must have $b^{|V|+1}$ leaves, so the tree must have height at least $|V| + 1$. But, there are only $|V|$ variables, so one must repeat.

Take the smallest parse tree for the string $w = uvxyz$. If $|vy| \ngeq 1$, then $v = y = \varepsilon$, and the section of the parse tree between the two Rs don't contribute anything, so $w = uxz$, and we have an even smaller parse tree for w, contradicting the minimality of the first tree.





If $|vxy| > p$, then the subtree rooted at the first R is itself long enough to have a repeated variable on a root to leaf path. So, we can split x into $x = u'x'y'$, so $w = (uv)v'x'y'(yz)$, which is still of the required form, but the middle substring has now decreased in length. We can then repeat this process until the middle substring has length at most p. ■

Example. Consider the language

$$L = \{a^n b^n c^n : n \geq 0\}$$

Suppose that L is context-free, and let p be the pumping length given by the pumping lemma. Take the string $w = a^p b^p c^p \in L$. Clearly, $|w| \geq m$.

Take an arbitrary decomposition $w = uvxyz$ with $|vxy| \leq p$ and $|vy| \geq 1$.

Since $|vy| \geq 1$, v and y contain at least 1 symbol from $\{a,b,c\}$. Because the middle of w contains p many bs and $|vxy| \leq p$, vxy contains at most 2 of the symbols. So, if we pump vxy 0 times, uv^0xy^0z will lose some number of only 2 of the characters. Hence $uv^0xy^0z \notin L$, and L is not context-free. △

Example. Consider the language

$$L = \left\{xx : x \in \{0,1\}^*\right\}$$

Suppose that L is context-free, and let p be the pumping length given by the pumping lemma. Take the string $w = 0^{p+1}1^{p+1}0^{p+1}1^{p+1} \in L$. Clearly, $|w| \geq p$.

Take an arbitrary decomposition $w = uvxyz$ with $|vxy| \leq p$ and $|vy| \geq 1$.

If we pump w 0 times, we obtain $uv^0xy^0z = 0^\alpha 1^\beta 0^\gamma 1^\delta$. Because $|vxy| \leq p$, vxy can intersect at most two adjacent sections of continguous 1s or 0s in w. If α and β or γ and δ are changed, then only one half of $w = \omega\omega$ has changed, so $uv^0xy^0z \notin L$. If β and γ are changed, then we have changed the number of 1s in the first ω, and/or the number of 0s in the second. In any case, $uv^0xy^0z \notin L$, so L is not context-free. △

28.6.2 Finiteness of Context-Free Languages

If L is a context-free language that has a string longer than the pumping length p, then the pumping lemma allows us to pump this string to generate infinitely many strings.

The converse also holds: if L contains no strings of length longer than p, then L is finite. In particular, it is also regular.

We also have that if L contains even one string of length longer than p, then we also have a bound on the maximal length of a shortest string as follows. Given any arbitrary string w, the pumping lemma gives a decomposition $w = uvxyz$ with $|vxy| \leq p$ and $|vy| \geq 1$. In particular, $|vy|$ is at most p when $x = \varepsilon$. So, pumping w down reduces its length by p. Repeating this process will eventually return a string with length between p and $2p - 1$.

28.6.3 Closure Properties of Context-Free Languages

Recall (§28.2.2) that regular languages are closed under

- Intersection;
- Union;
- Concatenation;
- Complementation;
- Kleene star.

The proof for intersections was to construct a DFA that simulated the two original DFAs in parallel, and accepting if both original DFAs accepted.

However, simulating a pair of PDAs using one PDA would require us to simulate two stacks with just one stack. This is provably impossible, and context-free languages are not closed under intersection.

Example. Consider the following CFLs:

$$L_1 = \{a^i b^i c^j : i,j, \geq 0\}$$
$$L_1 = \{a^i b^j c^j : i,j, \geq 0\}$$

Then, $L_1 \cap L_2 = \{a^n b^n c^n : n > 0\}$ is non-context-free. △

However, the intersection of a context-free language and a regular language *is* context-free, since we then only have one stack to simulate (i.e. just use the stack).

In contrast, for unions, we can use non-determinism to our advantage. Unlike for intersections, we only need to simulate one stack at a time in a union – not both simultaneously. We add a new start state for the PDA and an ε-transition to the start states of the previous PDAs. Then, this new PDA will accept if either of the previous PDAs accept.

Other similar construction works for grammars:

- **Union**:

  If $L_1 = L(G_1)$ and $L_2 = L(G_2)$ with $G_1 = (V_1,\Sigma_1,R_1,S_1)$ and $G_2 = (V_2,\Sigma_1,R_2,S_2)$, then $L_1 \cup L_2$ is generated by the grammar

  $$G = \big(V_1 \sqcup V_2, \quad \Sigma_1 \cup \Sigma_2, \quad R_1 \cup R_2 \cup \{S \to S_1 \mid S_2\}, \quad S\big)$$

- **Concatenation**:

  If $L_1 = L(G_1)$ and $L_2 = L(G_2)$ with $G_1 = (V_1,\Sigma_1,R_1,S_1)$ and $G_2 = (V_2,\Sigma_1,R_2,S_2)$, then $L_1 \cdot L_2$ is generated by the grammar

  $$G = \big(V_1 \sqcup V_2, \quad \Sigma_1 \cup \Sigma_2, \quad R_1 \cup R_2 \cup \{S \to S_1 S_2\}, \quad S\big)$$

- **Kleene star**:

  If $L = L(G_1)$ with $G_1 = (V_1,\Sigma_1,R_1,S_1)$, then L^* is generated by the grammar

  $$G = \big(V_1, \quad \Sigma_1, \quad R \cup \{S \to \varepsilon \mid S_1 S\}, \quad S\big)$$

Context-free languages are also not closed under complementation: using De Morgan's laws, we can write an intersection as:

$$L_1 \cap L_2 = \overline{\overline{L_1} \cup \overline{L_2}}$$

Since context-free languages are closed under union, if they were also closed under complementation, they would be closed under intersection. But this cannot be the case.

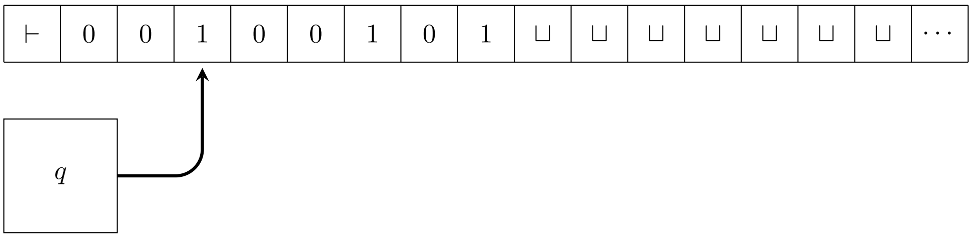
28.7 Recursively Enumerable Languages

A *Turing machine* is a 7-tuple $(Q,\Sigma,\Gamma,\delta,q_0,q_{\text{accept}},q_{\text{reject}})$, consisting of

- a finite set Q of states;
- a finite set Σ, the input alphabet;

- a finite set Γ, the *tape alphabet* not containing the *blank symbol* $\sqcup$;
- a transition function $\delta : Q \times \Gamma \to Q \times \Gamma \times \{L,R\}$;
- an initial state $q_0 \in Q$;
- an accepting state $q_{\text{accept}} \in Q$;
- an rejecting state $q_{\text{reject}} \in Q$, where $q_{\text{reject}} \neq q_{\text{accept}}$.

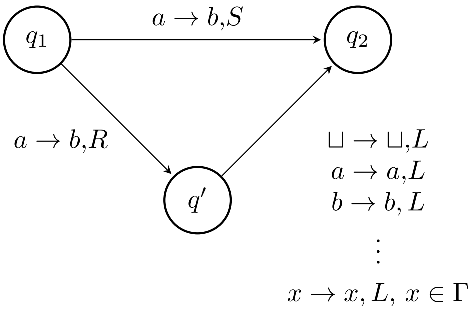
Instead of a stack, a Turing machine is equipped with a *tape* of memory – an array of memory that extends infinitely in one direction – and a *read/write head* pointing to one of the cells on the tape. The $\{L,R\}$ in the transition function indicates which way the read/write head should move after each instruction.



We mark the first cell with the reserved symbol $\vdash$, and unless otherwise specified, the other infinitely many cells with the reserved blank symbol $\sqcup$.

We can, as usual, represent a Turing machine as a state transition diagram. The labels in a state transition diagram are given in the form $a \to b,m$, where $m \in \{L,R\}$ is a movement instruction, and $a,b \in \Gamma$ are tape symbols. This time, the arrow indicates reading an a at the current tape position, writing a b, then moving the read/write head the specified direction.

Note that by this definition, the read/write head *must* move after every instruction. However, we can perform a memory operation $a \to b$ without any net movement, which we denote by $a \to b, S$, by performing the memory operation, moving right, reading/writing the same symbol back into this cell, then moving back left:



A *configuration* of a Turing machine $M = (Q,\Sigma,\Gamma,\delta,q_0,q_{\text{accept}},q_{\text{reject}})$ is a triple (u,q,v), where

- $q \in Q$ is the current state of the machine;
- The string on the tape is $u \circ v$, and the read/write head is on the first symbol of v.

Example. The configuration displayed on the tape above is

$$\big(\{0,0\},q,\{1,0,0,1,0,1\}\big)$$

△

Example.

- $(\vdash, q_0, w)$ is the *start* configuration;
- $(u, q_{\text{accept}}, v)$ is the *accepting* configuration;
- $(u, q_{\text{reject}}, v)$ is the *rejecting* configuration;

△

We say that a configuration $X = (u,q,v)$ *yields* another configuration $Y = (u',p,v')$ if the Turing machine can go from X to Y with a single state transition.

The *run* of a Turing machine $M = (Q,\Sigma,\Gamma,\delta,q_0,q_{\text{accept}},q_{\text{reject}})$ on a word w is a sequence of configurations $C_1,\dots,C_r$ such that $C_1 = (\vdash, q_0, w)$ is the start configuration (i.e., the word w is written on the tape), and such that each C_i yields C_{i+1}.

The run is *accepting* if C_r is the accepting configuration, and *rejecting* if C_r is the rejecting configuration. Note that, unlike for the weaker automata we have seen, we have a new possible end state, because the run of a Turing machine on any given word may not necessarily be finite. If the run is infinite, then we say that the Turing machine does not *halt*.

We define the language $L(M)$ to be the set of words w such that M *accepts*:

$$L(M) \coloneqq \{s \in \Sigma : \text{the run of } M \text{ on } w \text{ is accepting}\}$$

A language L is *Turing-recognisable* or *recursively enumerable* (RE) if there exists a Turing machine M which accepts precisely the strings in L. We do not require that M explicitly rejects strings outside of L (since in this case, it may not halt), only that it does not accept any strings outside of L.

In contrast, a language L is *Turing-decidable* or *recursive* if there exists a Turing machine M which accepts precisely the strings in L, and always halts. That is, it must also explicitly reject strings outside of L.

If a Turing machine always halts, then we call it a *decider* or a *total Turing machine.*

Turing-decidable languages are precisely those for which an algorithm to determine membership exists.

Example. Is there a decider D, which, when given a Turing machine M, decides whether M:

(*i*) has more than 847 states?

(*ii*) takes more than 847 steps on the input ε?

(*iii*) takes more than 847 steps on *some* input?

(*iv*) takes more than 847 steps on *all* inputs?

(*v*) ever moves the read/write head more than 847 tape cells away from the endmarker on input ε?

(*vi*) accepts ε?

(*vii*) accepts any string at all?

(*viii*) accepts every string?

(*i*) Yes, just count the states until we reach 847. If we reach 847, accept M. Otherwise, reject M. As we have bounded the maximum count, this procedure will always halt, so this describes a decider.

(*ii*) Yes, just simulate the run of M on ε for 847 steps, or until M halts, whichever happens earlier. If M halts before 847 steps, reject M. Otherwise, accept M.

As we have bounded the number of steps, this will always halt, so this describes a decider.

(*iii*) Yes. Simulate the run of M on all possible inputs of size at most 847 for 847 steps, or until M halts, whichever happens earlier.

There are $(\Sigma + 1)^{847}$ many such inputs, which is finite, so the program will eventually halt.

It is also sufficient to only check these inputs, since if M runs for 847 steps on a word p of length greater than 847, then M also runs for at least 847 steps when the input is only the first 847 symbols of p, since M can only possibly access the first 847 symbols of p in 847 steps.

(*iv*) Yes. The program for the previous case also decides this problem.

(*v*) Yes. If the read/write head of M never goes past the 847th cell, then there are only finitely many possible configurations that M could be in, since a Turing machine only has finitely many states and tape symbols. Namely, there are

$$t \coloneqq 847 \cdot |Q| \cdot |\Gamma|^{847}$$

possible configurations. So, we simulate the run of M on ε for $t + 1$ steps, or until it halts, whichever happens earlier. If M ever moves the read/write head past the 847th cell, stop and accept M. Otherwise, reject M.

(*vi*) No.

(*vii*) No.

(*viii*) No.

We will prove these last three cases later as a special case of the *Membership Problem.* △

28.7.1 Modifications of Turing Machines

There are many useful ways in which we might modify a Turing machine. However, this model of computation is very robust, and many of these modifications end up being equivalent to an ordinary Turing machine:

Example. What if we equipped a Turing machine with three tapes instead of one?

It turns out that we can simulate a Turing machine M_3 with three tapes with a Turing machine M using just one.

We make the tape alphabet of M a tuple to store three symbols in each cell. Then, to mark the position of the read/write head, for each symbol a in the tape alphabets of M_3, add a marked copy $\hat{a}$ to M. Then, for each instruction moving a read/write head of M_3, we instead replace the marked symbols. △

Using this, we can indeed simulate a Turing machine M using another Turing machine, as we have been claiming in the previous example, by implementing two tapes and writing the source code of M on the first tape and the input to be run on the second.

Example. What if the tape were infinite in both directions?

We can "fold" the tape somewhere, forming two tapes that are infinite in one direction only. As seen by the previous example, we can simulate this on a Turing machine. △

An *enumeration machine* or *enumerator* is a modification of a Turing machine equipped with two tapes,

- an ordinary read/write tape, assumed to always start blank;
- an write-only *output tape*;

and a special *enumeration* state. When the machine enters the enumeration state, we say that it has *enumerated* whatever word is on the output tape. Then, the machine erases the output tape and sends the write-only head to the beginning of the output tape, before continuing.

Given an enumerator E, we define the language $L(E)$ of E to be the set of words enumerated by E.

$$L(E) \coloneqq \{s : E \text{ enumerates } s\}$$

Theorem 28.7.1. *A language L is Turing-recognisable if and only if there exists an enumerator E such that $L = L(E)$.*

Proof. Suppose $L = L(E)$. From E, construct a Turing machine M that operates on an input w as follows:

- Run E on w;
- Each time E enumerates a word u, compare it to w;
- If $w = u$, accept. Otherwise, continue.

If L accepts w, then it will eventually be enumerated by E. If L does not accept E, then M never halts. So, $L(M) = L(E)$.

Now, suppose that L is Turing-recognisable, and let M recognise L. We construct an enumerator E that operates on an input w as follows.

- Order all possible strings (i.e. via lexicographical order) as $s_1,s_2,\dots$
- For each $i = 1,2,3,\dots$, repeat the following:
  - Run M for i steps on $s_1,\dots,s_i$.
  - If any computations accept, print out the corresponding s_j.

■

A *Universal Turing Machine* (UTM) U takes an encoding of a Turing machine M and a word w, written as $\mathrm{Enc}(M)\#w$ or just $\langle M,w\rangle$, and simulates the run of M on w.

Theorem 28.7.2. *A language L is Turing-recognisable if and only if there exists an enumerator E such that $L = L(E)$.*

Proof. Suppose $L = L(E)$. From E, construct a Turing machine M that operates on an input w as follows:

- Run E on w;
- Each time E enumerates a word u, compare it to w;
- If $w = u$, accept. Otherwise, continue.

If L accepts w, then it will eventually be enumerated by E. If L does not accept E, then M never halts. So, $L(M) = L(E)$.

Now, suppose that L is Turing-recognisable, and let M recognise L. We construct an enumerator E that operates on an input w as follows.

- Order all possible strings (i.e. via lexicographical order) as $s_1,s_2,\dots$
- For each $i = 1,2,3,\dots$, repeat the following:
  - Run M for i steps on $s_1,\dots,s_i$.
  - If any computations accept, print out the corresponding s_j.

■

28.7.2 Undecidability

28.7.2.1 The Halting Problem

Consider the language

$$\mathbf{HP} \coloneqq \big\{\langle M,x\rangle : M \text{ halts on } x\big\}$$

This language is Turing-recognisable by a universal Turing machine as we can simply simulate the run of M on x.

If M halts on x, then U accepts $\langle M,x\rangle$. Otherwise, U also fails to halt, which is allowed, as we only prohibit explicit acceptance in Turing-recognisability. So, **HP** is Turing-recognisable.

However, is it Turing-*decidable*?

That is, given an instance $\langle M,x\rangle$ of the halting problem, can we decide with a definite yes-or-no answer whether or not M will halt on any given string x?

Theorem 28.7.3. **HP** *is undecidable.*

Proof. Observe that any Turing machine M consists of only a finite amount of data. As such, this information can be encoded in binary. With such an encoding, we can inversely interpret every binary string b as a Turing machine M_b.

Suppose there exists a Turing machine U that decides the halting problem instance $\langle M_b,x\rangle$, given an encoding b of a Turing machine and an input x. Construct a table of its outputs on all possible strings:

	ε	0	1	00	01	10	11	000	001	010	$\cdots$
M_ε	H	L	L	H	L	H	H	H	L	L	$\cdots$
M_0	L	H	H	H	L	L	L	L	H	H	$\cdots$
M_1	L	H	L	L	H	L	L	H	H	H	$\cdots$
M_{01}	H	H	H	H	L	H	H	L	L	L	$\cdots$
M_{10}	H	L	H	L	L	L	L	L	H	L	$\cdots$
M_{11}	H	H	H	L	H	H	L	L	L	L	$\cdots$
M_{000}	L	L	L	H	H	H	H	L	H	H	$\cdots$
M_{001}	L	H	H	L	L	H	H	H	H	L	$\cdots$
M_{010}	H	L	L	H	L	L	H	H	L	L	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

We construct a Turing machine K as follows.

- Given an input string b, construct M_b.
- Simulate U on $\langle M_b,b\rangle$.
- If U halts, go into an infinite loop.
- If U accepts, halt.

That is, K halts on b if and only if M_b does not half on b, and vice versa, so its output is the reverse of the diagonal entries on the above table.

So, by construction, K disagrees with every Turing machine M_b on at least the input b, so K cannot be on the table, contradicting that this table contains every Turing machine. ■

28.7.2.2 The Membership Problem

Consider the language

$$\mathbf{MP} \coloneqq \big\{ \langle M,x \rangle : x \in L(M) \big\}$$

Again, this language is Turing-recognisable by a universal Turing machine as we can simply simulate the run of M on x.

However, the *Membership Problem* is again undecidable:

Theorem 28.7.4. **MP** *is undecidable.*

Proof. Suppose U decides **MP**. From U, we will construct a Turing machine K that decides **HP** as follows:

- Given an input $\langle M,x \rangle$, K constructs a new Turing machine M' as follows:
  - Add a new accept state.
  - Redirect all incoming transitions to the old accept and reject states to the new accept state.
- Run U on M'.
- If U accepts M', accept M. Otherwise, reject M.

We have that K accepts M if and only if U accepts M' if and only if M accepts or rejects x. That is, if and only if M halts on x. So, K decides **HP**, which is undecidable. ■

28.7.3 Computability and Reductions

A function σ is *computable* if there is a decider such that, when run on input x, halts with $\sigma(x)$ on the output tape.

Given subsets $A \subseteq \Sigma^*$ and $B \subseteq \Delta^*$, a function $\sigma : \Sigma^* \to \Delta^*$ is a *mapping reduction* if

- For all $x \in \Sigma$, $x \in A$ if and only if $\sigma(x) \in B$;
- σ is computable;

and we write $A \leq_m B$ if such a reduction exists.

Theorem 28.7.5.

- *If $A \leq_m B$ and B is decidable, then A is decidable;*
- *If $A \leq_m B$ and A is undecidable, then B is undecidable.*

Proof. Let M decide B and let σ be a reduction from A to B. Then, we define a decider N for A as follows:

- Given an input w, compute $\sigma(w)$.
- Simulate the run of M on $\sigma(w)$, and return whatever M returns.

Because σ is a rediction from A to B, if $w \in A$, then $\sigma(w) \in B$. So, M accepts $\sigma(w)$ whenever $w \in A$, so N decides A.

The second claim is the contrapositive. ■

Here are some more languages:

$$\varepsilon\text{-}\textsc{Acceptance} \coloneqq \big\{ \langle M \rangle : \varepsilon \in L(M) \big\}$$
$$\exists\text{-}\textsc{Acceptance} \coloneqq \big\{ \langle M \rangle : L(M) \neq \emptyset \big\}$$

$$\forall\text{-}\textsc{Acceptance} := \big\{\langle M\rangle : L(M) = \Sigma^*\big\}$$

Consider the following construction:

Let $\langle M{,}x\rangle$ be an instance of **HP**. Define the Turing machine M'_x as follows:

- Given an input y, ignore y and simulate the run of M on input x.
- If M halts, accept y. Otherwise, reject y.

By construction, M'_x accepts x if and only if M halts on x. So, $\langle M{,}x\rangle \in$ **HP** if and only if $\langle M'_x{,}x\rangle \in$ **MP**.

In fact, $\langle M{,}x\rangle \in$ **HP** if and only if:

(i) $\langle M'_x\rangle \in \varepsilon$-Acceptance;

(ii) $\langle M'_x\rangle \in \exists$-Acceptance;

(iii) $\langle M'_x\rangle \in \forall$-Acceptance.

Because M'_x ignores its input, it accepts any and all inputs if and only if M accepts x. That is, if and only if $\langle M{,}x\rangle \in$ **MP**. So none of these languages are decidable.

Example. Is the following language is decidable?

$$L := \big\{\langle M_1{,}M_2\rangle : M_1 \text{ accepts a word that } M_2 \text{ does not.}\big\}$$

No, because we can decide ε-Acceptance given a decider M for L. We construct a decider U as follows:

- Given an input M_b, construct a Turing machine M' that accepts the string x if and only if $x \neq \varepsilon$ and M_b accepts x.
- Simulate the run of M on $\langle M_b{,}M'\rangle$.
- Accept M_b if M accepts. Otherwise, reject M_b.

By construction, M_b and M' accept the same non-empty words, but M' cannot accept the empty word. Thus, M_b accepts a word that M' does not if and only if M_b accepts ε. So, U decides ε-Acceptance. △

Theorem 28.7.6.

- *If $A \leq_m B$ and B is Turing-recognisable, then A is Turing-recognisable;*
- *If $A \leq_m B$ and A is not Turing-recognisable, then B is not Turing-recognisable.*

Proof. Identical to the previous, with recognisers replacing deciders. ■

Theorem 28.7.7. *A language L is decidable if and only if L and $\overline{L}$ are both Turing-recognisable.*

Proof. If L is decidable, then a decider for L also functions as a recogniser for L. For $\overline{L}$, use the same decider and complement the answer.

Conversely, if L and $\overline{L}$ are both Turing-recognisable with recognisers P and Q, then define the decider M as follows:

- Given an input x, simulate P and Q simultaneously with input x.
- If P accepts, halt and accept. If Q accepts, halt and reject.

■

28.7.4 Closure Properties of Turing-Recognisable and Turing-Decidable Languages

- **Complementation**:

  Decidable languages are closed under complementation, as we can simply invert the return value of a decider.

  However, Turing-recognisable languages are *not* closed under complementation.

- **Union**:

  Decidable and Turing-recognisable languages are both closed under union.

- **Intersection**:

  Decidable and Turing-recognisable languages are both closed under intersection.

- **Kleene star**:

  Decidable and Turing-recognisable languages are both closed under Kleene star.

- **Concatenation**:

  Decidable and Turing-recognisable languages are both closed under concatenation.

28.7.5 Pairwise Intersection Closures Properties

	Regular	CFL	Decidable	RE
Regular	Regular			
CFL	CFL	Decidable		
Decidable	Decidable	Decidable	Decidable	
RE	RE	RE	RE	RE

Chapter 29

Boolean Functions

29.1 Introduction

Let $\mathcal{B} := \{0,1\}$, and let n be a positive integer. The points of the set $\mathcal{B}^n$ are called *binary vectors* or *Boolean points.* To simplify notation, we will write the elements of $\mathcal{B}^n$ without commas or parentheses, e.g.

$$\mathcal{B}^2 = \{00,01,10,11\}$$

A *Boolean function of n variables* is a function $f : \mathcal{B}^n \to \mathcal{B}$. A point $X = (x_1, \ldots, n_n) \in \mathcal{B}^n$ is a *true point* of f if $f(X) = 1$, and is a *false point* if $f(X) = 0$. The set of true points of f is denoted by $T(f)$, and the set of false points by $F(f)$.

The most elementary way to define a Boolean function is via its *truth table*, i.e. a list of all of the points of $\mathcal{B}^n$, along with the value of the function at each point.

Example. A Boolean function f of three variables, x, y, and z, defined by its truth table:

x,y,z	$f(x,y,z)$
000	1
001	1
010	1
011	1
100	0
101	0
110	0
111	1

This function has five true points

$$T(f) = \{000,001,010,011,111\}$$

and three false points

$$F(f) = \{100,101,110\}$$

△

In a truth table, the Boolean points are normally listed in lexicographic order, in which case, we only need the output values, which can be represented as a *vector of values.* For instance, the function above as vector of values 11110001.

Theorem 29.1.1. *The number of Boolean functions of n variables is 2^{2^n}.*

Proof. There are 2^n Boolean points, and for each one, a Boolean function can take one of two possible values. ■

29.1.1 Boolean Functions of One or Two Variables

There are four Boolean functions of one variable.

x	$g_1(x) \equiv \mathbf{0}_1$	$g_2(x) \equiv \mathbf{1}_1$	$g_3(x) \equiv x$	$g_4(x) \equiv \overline{x}$
0	0	1	0	1
1	0	1	1	0

- $\mathbf{0}_n$ is the *constant zero* function of n variables that takes the value 0 on all points of $\mathcal{B}_n$;
- $\mathbf{1}_n$ is the *constant one* function of n variables that takes the value 1 on all points of $\mathcal{B}_n$;
- $\overline{x}$ is the *negation*, *complementation*, or *Boolean NOT* of x.

There are sixteen Boolean functions of two variables:

x,y	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}
00	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
01	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
10	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
11	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

Many of these have special names and notations:

- $f_2(x,y) = x \wedge y = x\&y = xy$ is *conjunction* or *Boolean AND*;
- $f_8(x,y) = x \vee y = x\&y = x + y$ is *disjunction* or *Boolean OR*;
- $f_7(x,y) = x \oplus y$ is addition modulo 2 or *Boolean XOR*;
- $f_9(x,y) = x \downarrow y$ is the *Peirce arrow* or *Boolean NOR*;
- $f_{10}(x,y) = x \sim y$ is *equivalence*;
- $f_{14}(x,y) = x \to y$ is *implication*;
- $f_{15}(x,y) = x \uparrow y$ is the *Sheffer stroke* or *Boolean NAND*;

29.1.2 An Aside on Set Systems, Hypergraphs, and Graphs

A *set system* is a pair $(V,\mathcal{E})$ consisting of a finite set V called the *ground set* or *universe*, and a collection of subsets $\mathcal{E} \subseteq \mathcal{P}(V)$.

If $V = \{v_1,\dots,v_n\}$, then any subset $A \subseteq V$ can be described by its characteristic vector e_A. That is, a binary vector $(a_1,\dots,a_n) \in \mathcal{B}^n$ such that $a_i = 1$ if and only if $v_i \in A$.

Every set system over a totally ordered universe of n elements uniquely corresponds to a Boolean function f of n variables by mapping a Boolean point to true under f if and only if it is the characteristic vector of an subset in $\mathcal{E}$. This correspondence also establishes a relationship between set operations and Boolean functions. For instance, the union of two sets corresponds to disjunction in that $C = A \cup B$ if and only if $e_C = e_A \vee e_B$, where the disjunction is taken pointwise over the vector. Similarly, intersection correspond to conjunction, symmetric difference to addition modulo 2, relative difference to the function f_3 in the Table 29.1, and complementation to negation.

Set systems can also be interpreted as *hypergraphs*, with the ground set V containing *vertices* and $\mathcal{E}$ containing *hyperedges*. In particular, a hypergraph in which every hyperedge consists of two vertices is a *graph*, in which case the hyperedges are called *edges*. A *directed graph* is a graph in which every edge is an ordered pair of vertices, in which case the edges are called *arcs*.

29.1.3 Basic Identities

Theorem 29.1.2. *For all x,y,$z \in \mathcal{B}$,*

(*i*) $x \vee 1 = 1$ *and* $x \wedge 0 = 0$*;*

(*ii*) $x \vee 0 = x$ *and* $x \wedge 1 = x$*;*

(*iii*) $x \vee y = y \vee x$ *and* $x \wedge y = y \wedge x$ *(commutativity);*

(*iv*) $(x \vee y) \vee z = x \vee (y \vee z)$ *and* $(x \wedge y) \wedge z = x \wedge (y \wedge z)$ *(associativity);*

(*v*) $x \vee x = x$ *and* $x \wedge x = x$ *(idempotency);*

(*vi*) $x \vee (x \wedge y) = x$ *and* $x \wedge (x \vee y) = x$ *(absorption);*

(*vii*) $x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)$ *and* $x(y \vee z) = (x \wedge y) \vee (x \wedge z)$ *(distribution);*

(*viii*) $x \vee \overline{x} of 1$ *and* $x \wedge \overline{x} = 0$*;*

(*ix*) $\overline{\overline{x}} = x$ *(involution);*

(*x*) $\overline{x \vee y} = \overline{x} \wedge \overline{y}$ *and* $\overline{x \wedge y} = \overline{x} \vee \overline{x}$ *(De Morgan's laws)*

(*xi*) $x \vee (\overline{x} \wedge y) = x \vee y$ *and* $x \wedge (\overline{x} \vee y) = xy$ *(Boolean absorption);*

Proof. All easily verified through truth tables. ■

29.1.4 Boolean Expressions

Given a finite collection of Boolean variables $x_1, \dots ,x_n$, a *Boolean expression* or *Boolean formula* in the variables $x_1, \dots ,x_n$ is defined inductively as follows:

- The constants 0 and 1, and the variables $x_1, \dots ,x_n$ are Boolean expressions in $x_1, \dots ,x_n$;
- If ϕ and ψ are Boolean expressions in $x_1, \dots ,x_n$, then $(\phi \vee \psi)$, $(\phi \wedge \psi)$, and $\overline{\phi}$ are Boolean expressions in $x_1, \dots ,x_n$.

Given a Boolean expression ϕ, we call $\vee$, $\wedge$, and the negation operator $\overline{\bullet}$ the *operators* of the expression. A *literal* is either a variable or its complement.

We will adopt the convention that the operators have precedence, in decreasing order: complementation, conjunction, then disjunction. Along with the associativity properties, these precedence assumptions allow us to simplify Boolean expressions by removing extraneous parentheses. For instance, $(((\overline{x} \vee y)(y \vee \overline{z})) \vee ((xy)z))$ simplifies to $(\overline{x} \vee y)(y \vee \overline{z}) \vee xyz$.

Every Boolean expression ψ represents a unique boolean function f_ψ in the obvious way. Two Boolean expresions ϕ and ψ are *equivalent* if they represent the same Boolean function, and we write $\phi = \psi$ to denote this relation.

Note the basic identities in the previous theorem preserve equivalence of Boolean expressions.

29.1.5 Duality

Given a Boolean function f, we define the *dual* f^d of f to be the Boolean function

$$f^d(x_1, \dots ,x_n) \coloneqq \overline{f(\overline{x}_1, \dots ,\overline{x}_n)}$$

If $X = (x_1, \dots ,x_n)$, we abbreviate the vector of complemented variables by $\overline{X} = (\overline{x}_1, \dots ,\overline{x}_n)$, so $f^d(X) = \overline{f(\overline{X})}$.

The vector of values of the dual f^d is given by vertically reflecting the vector of values of f, then complementing pointwise.

Example.

x,y	$f(x,y)$	$f^d(x,y)$
00	0	0
01	1	0
10	1	0
11	1	1

This shows that the dual of disjunction is conjunction, and vice versa. △

Theorem 29.1.3. *If f and g are Boolean functions, then,*

(i) $(f^d)^d = f$*;*

(ii) $\overline{f}^d = \overline{f^d}$*;*

(iii) $(f \vee g)^d = f^d \wedge g^d$*;*

(iv) $(f \wedge g)^d = f^d \vee g^d$*.*

Proof.

(i) Since complementation is involutive,

$$\begin{aligned}(f^d)^d(X) &= \overline{f^d(\overline{X})} \\ &= \overline{\overline{f(\overline{\overline{X}})}} \\ &= f(X)\end{aligned}$$

(ii) Similarly,

$$\begin{aligned}\overline{f}^d(X) &= \overline{\overline{f(\overline{X})}} \\ &= \overline{f^d(X)}\end{aligned}$$

(iii) By De Morgan's laws,

$$\begin{aligned}(f \vee g)^d &= \overline{(f \vee g)(\overline{X})} \\ &= \overline{f(\overline{X}) \vee g(\overline{X})} \\ &= \overline{f(\overline{X})} \wedge \overline{g(\overline{X})} \\ &= f^d(X) \wedge g^d(X)\end{aligned}$$

(iv) Dually,

$$\begin{aligned}(f \wedge g)^d &= \overline{(f \wedge g)(\overline{X})} \\ &= \overline{f(\overline{X}) \wedge g(\overline{X})} \\ &= \overline{f(\overline{X})} \vee \overline{g(\overline{X})} \\ &= f^d(X) \vee g^d(X)\end{aligned}$$

■

Given a Boolean expression ϕ, we define the *dual* ϕ^d of ϕ to be the Boolean expression obtained from ϕ by interchanging the operators $\vee$ and $\wedge$, and the constants 0 and 1.

Theorem 29.1.4. *If the Boolean expression ϕ represents the Boolean function f, then ϕ^d represents f^d.*

Proof. We proceed by structural induction on ϕ. If ϕ is a constant or literal, then this is clear by involution. Otherwise, suppose $\phi = \phi_1 \vee \phi_2$ for some expressions ϕ_1 and ϕ_2. By the induction hypothesis, the expressions ϕ_1^d and ϕ_2^d represent the duals $f_{\phi_1}^d$ and $f_{\phi_2}^d$ respectively, so $\phi^d = \phi_1^d \wedge \phi_2^d$ represents the dual $f_{\phi_1}^d \wedge f_{\phi_2}^d$ of $f_{\phi_1} \vee f_{\phi_2} = f_\phi$. The proof for $\phi = \phi_1 \wedge \phi_2$ and $\phi = \overline{\psi}$ are similar. ■

29.1.6 Normal Forms

An *elementary conjunction* or a *term* is a conjunction of literals, and an *elementary disjunction* or a *clause* is a disjunction of literals. That is, an elementary conjunction is an expression of the form

$$C = \bigwedge_{i \in A} x_i \wedge \bigwedge_{j \in B} \overline{x}_j, \qquad A \cap B = \varnothing$$

and an elementary disjunction is an expression of the form

$$D = \bigvee_{i \in A} x_i \vee \bigvee_{j \in B} \overline{x}_j, \qquad A \cap B = \varnothing$$

That is, a variable can appear only once, and only either complemented or not.

A *disjunctive normal form (DNF)* is a disjunction of terms, and a *conjunctive normal form (CNF)* is a conjunction of clauses. That is, a DNF is a expression of the form

$$\bigvee_{k=1}^{m} C_k = \bigvee_{k=1}^{m} \left(\bigwedge_{i \in A_k} x_i \wedge \bigwedge_{j \in B_k} \overline{x}_j \right)$$

where each C_k is a term, and a CNF is an expression of the form

$$\bigwedge_{k=1}^{m} D_k = \bigwedge_{k=1}^{m} \left(\bigvee_{i \in A_k} x_i \vee \bigvee_{j \in B_k} \overline{x}_j \right)$$

where each D_k is a clause.

Theorem 29.1.5. *Every Boolean function admits a DNF and a CNF representation.*

Proof. Let f be a Boolean function and let $T = T(f)$ be the set its of true points. Consider the DNF

$$\phi_f(x_1, \ldots, x_n) = \bigvee_{Y \in T} \left(\bigwedge_{i : y_i = 1} x_i \wedge \bigwedge_{j : y_j = 0} \overline{x}_j \right) \tag{1}$$

Then, a point X^* is a true point of the function F represented by ϕ_f if and only if there exists a true point $Y = (y_1, \ldots, y_n) \in T$ of f such that

$$\bigwedge_{i : y_i = 1} x_i^* \wedge \bigwedge_{j : y_j = 0} \overline{x}_j^* = 1$$

But this just means that $x_i^* = 1$ whenever $y_i = 1$ and $x_j^* = 0$ whenever $y_i = 0$. That is, $X^* = Y$. Hence X^* is a true point of F if and only if it is a true point of f, and hence $f = F$ is represented by ϕ_f.

Similar reasoning establishes that f is also represented by the CNF

$$\phi_f(x_1, \ldots, x_n) = \bigwedge_{Y \in F} \left(\bigvee_{i : y_i = 0} x_i \vee \bigvee_{j : y_j = 1} \overline{x}_j \right) \tag{2}$$

where $F = F(f)$ is the set of false points of f. Alternatively, this expression can be obtained as the dual of (1). ■

The expressions in (1) and (2) are of a special form. In particular, every term in (1) contains n literals. Such a term is called a *minterm*, and the whole expression (1) representing f is called the *minterm expression* or *canonical DNF* of f. Similarly, every clause in (2) contains n literal and is called a *maxterm*, and the whole expression (2) is called the *maxterm expression* or *canonical CNF* of f.

The proof above gives an easy way to construct the canonical DNF/CNF of a function from its truth table. For the canonical DNF:

- Identify all rows where the function output is 1;
- For each row, write a minterm by including each variable with polarity corresponding to its appearance in the row;
- Take the disjunction of all the minterms.

That is, each minterm corresponds to a unique input combination, so we take the ones where the function is true, then allow any of the combinations by taking a disjunction.

Similarly, for the canonical CNF:

- Identify all rows where the function output is 0;
- For each row, write a maxterm by including each variable with polarity opposite to its appearance in the row;
- Take the conjunction of all the maxterms.

Dually, each maxterm corresponds to the complement of a unique input combination, i.e. evaluates to false for only one input, so we take the ones where the function is false, then mask out the false rows by taking a conjunction.

Example. Consider the following function:

x,y,z	$f(x,y,z)$
000	1
001	0
010	1
011	1
100	0
101	0
110	1
111	1

This function is represented by the DNF

$$\overline{x}\,\overline{y}\,\overline{z} \vee \overline{x}y\overline{z} \vee \overline{x}yz \vee xy\overline{z} \vee xyz$$

and the CNF

$$(x \vee y \vee \overline{z})(\overline{x} \vee y \vee z)(\overline{x} \vee y \vee \overline{z})$$

△

A canonical DNF and CNF is unique up to permutation of its terms and literals. However, a function generally admits many other non-canonical DNF and CNF representations.

Example. The function from the previous example can also be represented by the DNF

$$\overline{x}\,\overline{z} \vee y$$

△

The *degree* of a term $C = \bigwedge_{i\in A} x_i \bigwedge_{j\in B} \overline{x}_j$ is the number of literals appearing in C. That is, $|A| + |B|$.

More generally, the *degree* of a DNF $\phi = \bigvee_{k=1}^{m} C_k$ is the maximum degree of the C_k. A DNF is called *linear* if its degree is at most 1, *quadratic* if at most 2, *cubic* if at most 3, etc. We write $|\phi|$ for the number of literals in ϕ, also called its *length*; and $\|\phi\|$ for the number of terms in ϕ.

Example. The DNF $\phi = \overline{x}\,\overline{z} \vee y$ has $|\phi| = 3$ literals and $\|\phi\| = 2$ terms, and has term degrees 2 and 1, and is thus of degree 2, or is quadratic. △

The following example shows that a shortest DNF may have a length exponential in the number of variables.

Example. The function represented by the CNF

$$\psi(x_1,\dots,x_{2n}) = (x_1 \vee x_2)(x_3 \vee x_4)\cdots(x_{2n-1} \vee x_{2n})$$

has a unique shortest DNF consisting of 2^n terms, each containing exactly one literal from each clause of ψ. △

29.1.7 Orthogonal DNFs

A DNF

$$\phi = \bigvee_{k=1}^{m} C_k = \bigvee_{k=1}^{m} \left(\bigwedge_{i\in A_k} x_i \wedge \bigwedge_{j\in B_k} \overline{x}_j \right), \qquad A_k \cap B_k = \varnothing$$

is *orthogonal* or is a *sum of disjoint products* if $(A_k \cap B_\ell) \cup (A_\ell \cap B_k) \neq \varnothing$ for all $k \neq \ell$.

That is, a DNF is orthogonal if every pair of terms is “conflicting” in at least one variable; there must exist a variable that appears complemented in one term and uncomplemented in the other. Alternatively, a DNF is orthogonal if and only if the product (conjunction) of every pair of its terms is 0 at every Boolean point (since conflicting variables would yield 0 in the product).

Example. The DNF $\phi = \overline{x}_1\overline{x}_2x_4 \vee \overline{x}_1x_3x_4$ is not orthogonal since x_1 is negative in both terms, x_2 and x_3 do not appear in both terms, and x_4 is positive in both terms, so none of the variables are in conflict. Alternatively, both terms (and hence their product) are equal to 1 at 0011. △

Note that orthogonality is not preserved under equivalence.

Example. The DNF ϕ is equivalent to the DNF $\psi = \overline{x}_1\overline{x}_2x_4 \vee \overline{x}_hx_2x_3x_4$, which is orthogonal since its only pair of terms are conflicting at the variable x_2. △

Note that the minterm expression constructed in the proof of Theorem 29.1.5 is orthogonal, so we can specialise the result further to:

Theorem 29.1.6. *Every Boolean function can be represented by an orthogonal DNF.*

One of the main motivations for our interest in orthogonal DNFs is that the number of true points $\omega(f) := \big|T(f)\big|$ of a function f expressed in this form can be efficiently computed.

Theorem 29.1.7. *If a Boolean function f on $\mathcal{B}^n$ is represented by an orthogonal DNF, then the number of its true points is*

$$\omega(f) = \sum_{k=1}^{m} 2^{n-|A_k|-|B_k|}$$

Proof. The DNF takes the value 1 precisely when one of its terms takes the value 1. ■

The *Chow parameters* of a Boolean function f on $\mathcal{B}^n$ are the $n+1$ integers $(\omega_1,\dots,\omega_n,\omega)$ where $\omega = \omega(f)$ is the number of true points of f and ω_i is the number of true points $X^* = (x_1^*,\dots,x_n^*)$ of f with $x_i^* = 1$.

Example. The function f represented by the orthogonal DNF $\psi = \overline{x}_1\overline{x}_2\overline{x}_4 \vee \overline{x}_1x_2x_3x_4$ is true at 0001, 0011, and 0111. None of these have $x_1 = 1$; 1 has $x_2 = 1$; 2 have $x_3 = 1$; 3 have $x_4 = 1$; and there are 3 total true points; so its Chow parameters are $(\omega_1,\omega_2)(0,1,2,3,3)$. △

The same reasoning as in the proof of the previous theorem also shows that the Chow parameters of a function represented in orthogonal form can be efficiently computed: for ω, this is precisely the statement of the theorem; for ω_i, this follows from the fact that the DNF obtained by fixing $x_i = 1$ in an orthogonal DNF is still orthogonal.

29.1.8 Implicants

Given two Boolean functions $f,g : \mathcal{B}^n \to \mathcal{B}$, we say that f *implies* g, that f is a *minorant* of g, or that g is a *majorant* of f, and write $f \leq g$, if $f(X) = 1$ implies $g(X) = 1$ for all $X \in \mathcal{B}^n$. That is, f implies g pointwise. Equivalently, if we identify each function f with its set of true points, then this relation is precisely the subset containment relation.

This definition extends to Boolean expressions in the obvious way. We will often identify Boolean functions with representing Boolean expressions, and write, for instance, $\psi \leq f$ for $\psi \leq \phi_f$.

Theorem 29.1.8. *For all Boolean functions $f,g : \mathcal{B}^n \to \mathcal{B}$, the following are equivalent:*

(*i*) $f \leq g$;

(*ii*) $f \vee g = g$;

(*iii*) $\overline{f} \vee g = \mathbf{1}_n$;

(*iv*) $f \wedge g = f$;

(*v*) $f \wedge \overline{g} = \mathbf{0}_n$.

Proof. It suffices to note that each of these statements fail precisely when there exists $X \in \mathcal{B}^n$ such that $f(X) = 1$ but $g(X) = 0$. ■

Theorem 29.1.9. *For all Boolean functions $f,g,h : \mathcal{B}^n \to \mathcal{B}$,*

(*i*) $\mathbf{0}_n \leq f \leq \mathbf{1}_n$;

(*ii*) $f \wedge g \leq f \leq f \vee g$;

(*iii*) $f = g$ *if and only if* $f \leq g$ *and* $g \leq f$;

(*iv*) ($f \leq h$ *and* $g \leq h$) *if and only if* $f \vee g \leq h$;

(*v*) ($f \leq g$ *and* $f \leq h$) *if and only if* $f \leq g \wedge h$;

(*vi*) *if* $f \leq g$, *then* $f \wedge h \leq g \wedge h$;

(*vii*) *if* $f \leq g$, *then* $f \vee h \leq g \vee h$;

For two Boolean function represented by arbitrary Boolean expressions, it can be non-trivial to verify whether or not f implies g. However, for elementary conjunctions, implication is easy to verify. An elementary conjunction implies another elementary conjunction if and only if the latter results from the former by deleting literals, i.e. by removing constraints.

Theorem 29.1.10. *The elementary conjunction $C_{AB} = \bigwedge_{i\in A} x_i \wedge \bigwedge_{j\in B} \overline{x}_j$ implies the elementary conjunction $C_{XY} = \bigwedge_{i\in F} x_i \wedge \bigwedge_{j\in G} \overline{x}_j$ if and only if $F \subseteq A$ and $G \subseteq B$.*

Proof. Suppose that $F \subseteq A$ and $G \subseteq B$, and let $X \in \mathcal{B}^n$. If $C_{AB}(X) = 1$, then $x_i = 1$ for all $i \in A$ and $x_j = 0$ for all $j \in B$. So in particular, $x_i = 1$ for all $i \in F$ and $x_j = 0$ for all $j \in G$, so $C_{FG} = 1$ and hence $C_{AB} \leq C_{FG}$.

Conversely, suppose $C_{AB} \leq C_{FG}$, and suppose for a contradiction that $F \not\subseteq A$, so there exists $k \in F \setminus A$. Fix $x_i = 1$ for all $i \in A$ and $x_j = 0$ for $j \notin A$, and let $X = (x_1, \ldots, x_n)$. Then, $C_{AB}(X) = 1$, but $C_{FG}(X) = 0$ since $x_k = 0$ and $k \in F$. ■

Let f be a Boolean function and C be an elementary conjunction. Then, C is an *implicant* of f if $C \leq f$.

Theorem 29.1.11. *If ϕ is a DNF representation of f, then every term of ϕ is an implicant of f. Moreover, if an elementary conjunction C is an implicant of f, then the DNF $\phi \vee C$ also represents f.*

Proof. Firstly, note that whenever any term of ϕ takes the value 1, then the entire disjunction ϕ, and hence f, takes the value 1. Then, $\phi \vee C \leq f$ since ϕ and C both imply f, and $f \leq \phi \leq \phi \vee C$, so f is represented by $\phi \vee C$. ■

Example. Let $f = xy \vee x\overline{y}z$. Then, the terms xy and $x\overline{y}z$ are implicants of f. The term xz is also an implicant of f, so $xy \vee x\overline{y}z \vee xz$ also represents f. △

Let f be a Boolean function and C_1 and C_2 be implicants of f. Then, C_1 *absorbs* C_2 if $C_1 \vee C_2 = C_1$, or equivalently, if $C_2 \leq C_1$.

Let f be a Boolean function and C be an implicant of f. Then, C is a *prime implicant* of f if C is not absorbed by any other implicant of f. That is, a prime implicant is a maximal conjunction implying f.

Theorem 29.1.12. *Every Boolean function can be represented by the disjunction of all its prime implicants.*

Proof. Let f be a Boolean function on $\mathcal{B}^n$ with prime implicants $P_1, \ldots, P_m$. Consider any DNF representation of f, say $\phi = \bigvee_{k=1}^{r} C_k$. Then, the DNF

$$\psi = \bigvee_{k=1}^{r} C_k \vee \bigvee_{k=1}^{m} P_j$$

Now, every term C_k is an implicant of f and is hence absorbed by some prime implicant P_j. So $C_k \vee P_j = P_j$, and $\psi = \bigvee_{j=1}^{m} P_j$ represents f. ■

The DNF of all prime implicants of a Boolean function is called the *complete DNF* or *Blake canonical form* of the function.

Example. Consider again the function $f = xy \vee x\overline{y}z$. Its prime implicants are xy and xz, so $f = xy \vee xz$ is its complete DNF. △

An interesting corollary of this theorem is that every Boolean function is uniquely identified by the list of its prime implicants. Equivalently, two Boolean functions are equal if and only if they have the same complete DNF.

Let $\phi = \bigvee_{k \in \Omega} C_k$ be a DNF representation of a Boolean function f on $\mathcal{B}^n$. We say that ϕ is a *prime DNF* of f if each term C_k is a prime implicant of f. We say that ϕ is an *irredundant DNF* of f if there is no $j \in \Omega$ such that $\psi = \bigvee_{k \in \Omega \setminus \{j\}} C_k$ represents f.

The notion of (prime) implicants naturally have a dual notion for disjunctions. Let f be a Boolean function and D be an elementary disjunction. Then, D is an *implicate* of f if $f \leq D$, and is furthermore a *prime implicate* if it is not implied by any other implicate of f. That is, a prime implicate is a minimal disjunction implied by f.

Theorem 29.1.13. *Every Boolean function can be represented by the conjunction of all its prime implicates.*

Example. The function $g = x\overline{y} \vee \overline{x}y \vee x\overline{z}$ has four implicates, namely $(x \vee y)$, $(x \vee y \vee z)$, $(x \vee y \vee \overline{z})$, and $(\overline{x} \vee \overline{y} \vee \overline{z})$. Only the first and last implicate in this list are prime, so $g = (x \vee y)(\overline{x} \vee \overline{y} \vee \overline{z})$. △

29.1.9 Generation of All Prime Implicates from a DNF Representation

If xC and $\overline{x}D$ are two elementary conjunctions such that CD is not identically 0, then we say that CD is the *consensus* of xC and $\overline{x}D$, and that CD is *derived from xC and $\overline{x}D$ by consensus on x.*

Given an arbitrary DNF ϕ, the *consensus procedure* generates the complete DNF equivalent to ϕ by repeatedly applying the operations of absorption ($x \vee xy = x$) and consensus:

- If there exist two terms C and D of ϕ such that C absorbs D, remove D from ϕ.
- If there exist two terms x_iC and $\overline{x}_iD$ of ϕ such that x_iC and $\overline{x}_iD$ have a consensus CD that is not absorbed by another term of ϕ, then add CD to ϕ.

The procedure halts when:

- the absorption operation cannot be applied, and;
- either the consensus operation cannot be applied, or all the terms that can be produces by consensus are absorbed by other terms of ϕ.

A DNF is *closed under absorption* if it satisfies the first stopping condition, and is *closed under consensus* if it satisfies the second.

The consensus procedure always terminates and produces a DNF closed under consensus and absorption in a finite number of steps. Indeed, the number of terms in the given variables is finite, and once a term is removed by absorption, it will never be added by consensus.

Example. Consider the DNF

$$\phi(x_1,x_2,x_3,x_4) = x_1\overline{x}_2x_3 \vee \overline{x}_1\overline{x}_2x_4 \vee x_2x_3x_4$$

Absorption is not possible. We can apply consensus on x_1 on the first two terms to obtain the term $\overline{x}_2x_3\overline{x}_2x_4 = \overline{x}_2x_3x_4$ not absorbed by any term of ϕ, to obtain the DNF

$$\phi'(x_1,x_2,x_3,x_4) = x_1\overline{x}_2x_3 \vee \overline{x}_1\overline{x}_2x_4 \vee x_2x_3x_4 \vee \overline{x}_2x_3x_4$$

Again, absorption is not possible. We can apply consensus on x_2 on the last two terms to obtain the term x_3x_4 not absorbed by any existing terms:

$$\phi''(x_1,x_2,x_3,x_4) = x_1\overline{x}_2x_3 \vee \overline{x}_1\overline{x}_2x_4 \vee x_2x_3x_4 \vee \overline{x}_2x_3x_4 \vee x_3x_4$$

The new term absorbs the previous two:

$$\phi''(x_1,x_2,x_3,x_4) = x_1\overline{x}_2x_3 \vee \overline{x}_1\overline{x}_2x_4 \vee x_3x_4$$

Now, neither consensus and absorption can be applied, so the procedure stops, and these three terms are the prime implicants of ϕ. △

We have already observed that the operations of absorption and consensus transform DNFs,but do not change the Boolean functions that they represent. This is implied by the two lemmata below, which easily follow from the basic Boolean identities.

Lemma 29.1.14. *For any two elementary conjunctions C and CD,*

$$C \vee CD = C$$

Lemma 29.1.15. *For any two elementary conjunctions xC and $\overline{x}D$,*

$$xC \vee \overline{x}D = xC \vee \overline{x}D \vee CD$$

The importance of the consensus procedure is due to the fact that it produces the complete DNF. To prove this, we need one more lemma.

Lemma 29.1.16. *Given any DNF ϕ of a Boolean function f, if C is an implicant of f that involves all variables present in ϕ, then C is absorbed by a term of ϕ.*

Proof. If C contains all the variables of ϕ, then the valuation of that makes $C = 1$ assigns values to all the variables in ϕ. Since C is an implicant of f, this assignment also makes $\phi = 1$, and hence at least one term of ϕ is 1. This term absorbs C. ■

Theorem 29.1.17. *Given any DNF ϕ of a Boolean function f, the consensus procedure applied to ϕ yields the complete DNF of f.*

Proof. Suppose otherwise that there exists a Boolean function f and a DNF ϕ of f such that the consensus procedure produces a DNF ψ that does not contain a prime implicant C_0 of f. Then, by Theorem 29.1.20, C_0 only involves variables present in ψ. Consider the set $\mathcal{S}$ of elementary conjunctions C satisfying the following conditions:

- C only contains variables present in ψ;
- $C \leq C_0$ (and therefore C is an implicant of f);
- C is not absorbed by any term in ψ.

The set $\mathcal{S}$ is non-empty since C_0 satisfies all three conditions, so let C^m be a term of maximum degree in $\mathcal{S}$. Since C^m is not absorbed by any term in ψ, C^m cannot involve all the variables present in ψ by the previous lemma. Let x be a variable present in ψ not present in C^m. The degree of the elementary conjuctions xC^m and $\overline{x}C^x$ exceeds that of C^m, and since the degree of C^tm is maximum, xC^m and $\overline{x}C^m$ do not belong to $\mathcal{S}$ and therefore cannot satisfy all three conditions. Since they clearly must satisfy the first two conditions, they must violate the third, so there must exist terms C' and C'' in ψ such that $xC^m \leq C'$ and $\overline{x}C^m \leq C''$. Since C^m is not absorbed by either C' nor C'', it follows that $C' = xD'$ and $C'' = \overline{x}D''$, where D' and D'' are elementary conjunctions that absorb C^m. This implies that D' and D'' do not conflict in any variable. Therefore, the consensus of C' and C'' exists, namely $D'D''$, and this term absorbs C^m. Since the consensus procedure stops on the DNF ψ, there must exist a term C''' in ψ that absorbs $D'D''$. Then, C''' must also absorb C^m, contradicting the assumption that $C^m \in \mathcal{S}$. ■

29.1.10 Restrictions of Functions, Essential Variables

Let f be a Boolean function on $\mathcal{B}^n$ and let $k \in [n]$. We define the *restrictions* $f|_{x_k=1}$ and $f|_{x_k=0}$ to be the Boolean functions on $\mathcal{B}^{n-1}$ defined by:

$$\begin{aligned}
f|_{x_k=1}(x_1,\ldots,\widehat{x_k},\ldots,x_n) &= f(x_1,\ldots,x_{k-1},1,x_{k+1},\ldots,x_n)\\
f|_{x_k=0}(x_1,\ldots,\widehat{x_k},\ldots,x_n) &= f(x_1,\ldots,x_{k-1},0,x_{k+1},\ldots,x_n)
\end{aligned}$$

That is, the functions obtained from f by fixing its kth argument to 1 and 0 respectively.

Example. Consider the function $f(x,y,z) = (xz \vee y)(x \vee \overline{z}) \vee \overline{x}\,\overline{y}$. Then,

$$\begin{aligned}
f|_{x=1}(y,z) &= (1z \vee y)(1 \vee \overline{z}) \vee \overline{1}\overline{y}\\
&= z \vee y
\end{aligned}$$

$$\begin{aligned}
f|_{x=0}(y,z) &= (0z \vee y)(0 \vee \overline{z}) \vee \overline{0}\overline{y}\\
&= y\overline{z} \vee \overline{y}\\
&= \overline{y} \vee \overline{z}
\end{aligned}$$

$\triangle$

Theorem 29.1.18. *Let f be a Boolean function on $\mathcal{B}^n$, let ψ be a representation of f and let $k \in [n]$. Then, the expression obtained by substituting the constant* 1 *(respectively,* 0*) for every occurrence of x_k in ψ represents $f|_{x_k=1}$ (respectively, $f|_{x_k=0}$).*

Theorem 29.1.19. *Let f be a Boolean function on $\mathcal{B}^n$ and let $k \in [n]$. Then,*

$$f(x_1,\dots,x_n) = x_k f|_{x_k=1} \vee \overline{x}_k f|_{x_k=0}$$

Proof. Clear by case analysis on the value of x_k. ■

The right side of this identity is called the *Shannon expansion* of f with respect to x_k. By applying the Shannon expansion to a function and its successive restrictions until these restrictions become constants or literals, we obtain an orthogonal DNF of the function, which can easily be proved by induction. However, not every orthogonal DNF can be obtained in this way, since the Shannon expansion necessarily produces a DNF in which one of the variables appears in all of the terms.

Example. Consider the function $f = (xz \vee y)(x \vee \overline{z}) \vee \overline{x}\,\overline{y}$ from the previous example. The Shannon expansion of $f|_{x=1}$ with respect to y is

$$\begin{aligned} f|_{x=1} &= yf|_{x=1,y=1} \vee \overline{y}f|_{x=1,y=0} \\ &= y(z \vee 1) \vee \overline{y}(z \vee 0) \end{aligned}$$

Note that $z \vee 1 = 1$ is a constant and $z \vee 0 = z$ is a literal, so we stop here for the expansion of $f|_{y=1}$. The Shannon expansion of $f|_{x=0}$ with respect to y is

$$\begin{aligned} f|_{x=0} &= yf|_{x=0,y=1} \vee \overline{y}f|_{x=0,y=0} \\ &= y(\overline{1} \vee \overline{z}) \vee \overline{y}(\overline{0} \vee \overline{z}) \end{aligned}$$

Here, $\overline{1} \vee \overline{z} = \overline{z}$ is a literal and $\overline{0} \vee \overline{z} = 1$ is a constant, so we stop here for the expansion of $f|_{x=0}$. So, an orthogonal DNF of f is given by:

$$\begin{aligned} f(x,y,z) &= xf|_{x=1} \vee \overline{x}f|_{x=0} \\ &= x(yf|_{x=1,y=1} \vee \overline{y}f|_{x=1,y=0}) \vee \overline{x}(yf|_{x=0,y=1} \vee \overline{y}f|_{x=0,y=0}) \\ &= x(y1 \vee \overline{y}z) \vee \overline{x}(y\overline{z} \vee \overline{y}1) \\ &= xy \vee x\overline{y}z \vee \overline{x}y\overline{z} \vee \overline{x}\,\overline{y} \end{aligned}$$

Another orthogonal DNF of f is

$$xy \vee \overline{x}\,\overline{z} \vee \overline{y}z$$

However, this DNF cannot be obtained from successive Shannon expansions since there is no variable common to all three terms. △

Let f be a Boolean function on $\mathcal{B}^n$ and let $k \in [n]$. The variable x_k is *inessential* for f, or that f does not *depend* on x_k if $f|_{x_k=1}(X) = f|_{x_k=0}(X)$ for all $X \in \mathcal{B}^{n-1}$. That is, the value of f is the same, regardless of the value of x_k.

Theorem 29.1.20. *Let f be a Boolean function on $\mathcal{B}^n$ and let $k \in [n]$. Then, the following are equivalent:*

(*i*) *The variable x_k is inessential for f;*

(*ii*) *The variable x_k does not appear in any prime implicant of f;*

(*iii*) *f has a DNF representation in which the variable x_k does not appear.*

Proof. (*ii*) → (*iii*) since any function can be represented by a DNF of its prime implicants (Theorem 29.1.12), and (*iii*) → (*i*) by Theorem 29.1.18.

Now, suppose the variable x_k is inessential for f, and consider an implicant $C_{AB} = \bigwedge_{i\in A} x_i \bigwedge_{j\in B} \overline{x}_j$ of f. Suppose that $k \in A$ (the argument being symmetric for $k \in B$), and consider the conjunction C obtained by deleting x_k from C_{AB}:

$$C = \bigwedge_{i\in A\setminus\{k\}} x_i \bigwedge_{j\in B} \overline{x}_j$$

We claim that C is an implicant of f, and therefore any prime implicant need not involve x_k. Let $X = (x_1, \dots ,x_n) \in \mathcal{B}^n$ such that $C(X) = 1$. Since neither C nor f depend on x_k, we may suppose that $x_k = 1$ in X. Then, $C(X) = X_{AB}(X) = 1$, and hence $f(X) = 1$. ■

It should be clear that any particular representation of a Boolean function may involve a variable that the function does not depend on.

Example. The DNF $\phi(x_1,x_2,x_3,x_4) = x_1x_2 \vee x_1\overline{x}_2 \vee \overline{x}_1x_2 \vee \overline{x}_1\overline{x}_2$ represent the constant function $\mathbf{1}_4$. In particular, ϕ does not depend on any of its variables. △

Finally, let us mention an interesting connection between essential variables and Chow parameters.

Theorem 29.1.21. *Let f be a Boolean function on $\mathcal{B}^n$, let $(\omega_1, \dots ,\omega_n,\omega)$ be its vector of Chow parameters, and let $k \in [n]$. Then, if the variable x_k is inessential for f, then $\omega = 2\omega_k$.*

Proof. The sets $A = \{X \in T(f) : x_k = 1\}$ and $B = \{X \in T(f) : x_k = 0\}$ of true points with $x_k = 1$ and $x_k = 0$, respectively, partition $T(f)$, with $|A| = \omega_k$ and $|B| = \omega - \omega_k$. If x_k is inessential, then $|A| = |B|$, so $\omega = 2\omega_k$. ■

The converse fails in general, however. For instance, the function $f(x_1,x_2) = x_1\overline{x}_2 \vee \overline{x}_1x_2$ has Chow parameters (1,1,2), and both variables x_1,x_2 are essential.

29.1.11 Monotone Boolean Functions

Let f be a Boolean function on $\mathcal{B}^n$ and let $k \in [n]$. We say that f is *positive* in the variable x_k if $f|_{x_k=0} \leq f|_{x_k=1}$. Dually, f is *negative* in x_k if $f|_{x_k=1} \leq f|_{x_k=0}$. More generally, f is *monotone* in x_k if it is positive or negative in x_k.

To check that a variable is positive or negative, we can just check that flipping that variable from 0 to 1 does not decrease the value of the function.

Example. Consider the function:

x,y,z	$f(x,y,z)$
000	0
001	0
010	0
011	1
100	0
101	1
110	1
111	1

Notation: input $\overset{\text{output}\to\text{output}}{\to}$ input.

- x is positive, since we have $000 \overset{0\to0}{\to} 100$, $001 \overset{0\to1}{\to} 101$, $010 \overset{0\to1}{\to} 110$, and $011 \overset{0\to1}{\to} 111$ all non-decreasing;
- y is positive, as we have $000 \overset{0\to0}{\to} 010$, $001 \overset{0\to1}{\to} 011$, $100 \overset{0\to1}{\to} 110$, and $101 \overset{1\to1}{\to} 111$ all non-decreasing;
- z is positive, as we have $000 \overset{0\to0}{\to} 001$, $010 \overset{0\to0}{\to} 011$, $100 \overset{0\to1}{\to} 101$, and $110 \overset{1\to1}{\to} 111$ all non-decreasing.

△

The function f is furthermore *positive* (respectively, *negative*) if it is positive (respectively, negative) in every variable, and *monotone* if it is positive or negative in each variable (not necessarily all positive or all negative).

Example. The function above is positive, since it is positive in all of its variables. △

Theorem 29.1.22. *Let f be the Boolean function on $\mathcal{B}^n$, and let g be the function defined by*

$$g(x_1,x_2,\dots,x_n) = f(\overline{x}_1,x_2,\dots,x_n)$$

Then, g is negative in x_1 if and only if f is positive in x_1.

Proof. If f is positive in x_1, then changing x_1 from 0 to 1 does not decrease the value of f. But changing x_1 from 0 to 1 in f is same as changing x_1 from 1 to 0 in g, and g shares the same output values as f when x_1 is reversed, so changing x_1 from 1 to 0 does not decrease the value of g, i.e. g is negative in x_1. ■

Theorem 29.1.23. *A Boolean function f on $\mathcal{B}^n$ is positive if and only if $f(X) \leq f(Y)$ for all $X,Y \in \mathcal{B}^n$ such that $X \leq Y$.*

Proof. Suppose that f is positive, so $f|_{x_k=0} \leq f|_{x_k=1}$ for all $k \in [n]$, and let $X,Y \in \mathcal{B}^n$ with $X \leq Y$. Consider the sequence of Boolean points

$$X = Z^1 \leq Z^2 \leq \cdots \leq Z^k = Y$$

where Z^{i+1} is obtained from Z^i by flipping the first bit of Z^i that disagrees with Y, say in the k_ith position. Note that since $X \leq Y$, all such flip change a 0 to a 1. Then, since f is positive in each variable and in particular in x_{k_i}, we have

$$f(Z^i) = f|_{x_{k_i}=0}(Z^{i+1}) \leq f|_{x_{k_i}=1}(Z^{i+1}) = f(Z^{i+1})$$

so by induction,

$$f(X) = f(Z^1) \leq f(Z^2) \leq \cdots \leq f(Z^k) = f(Y)$$

Conversely, suppose $f(X) \leq f(Y)$ whenever $X \leq Y$. Let $k \in [n]$, and let $\tilde{X} = (x_1,\dots,\widehat{x_k},\dots,x_n) \in \mathcal{B}^{n-1}$ be any assignment of values to all variables apart from x_k. Then, the Boolean points $X,Y \in \mathcal{B}^n$ obtained from $\tilde{X}$ by assigning the value $x_k = 0$ and $x_k = 1$, respectively,satisfy $X \leq Y$, and thus,

$$f|_{x_k=0}(\tilde{X}) = f(X) \leq f(Y) = f|_{x_k=1}(\tilde{X})$$

so $f|_{x_k=0} \leq f|_{x_k=1}$. ■

Let $\psi(x_1,\dots,x_n)$ be a DNF and let $k \in [n]$. Then,

- ψ is *positive* (respectively, *negative*) in the variable x_k if the complemented literal $\overline{x}_k$ (respectively, uncomplemented literal x_k) does not appear in ψ;
- ψ is *monotone* in x_k if ψ is either positive or negative in x_k;
- ψ is *positive* (respectively, *negative*) if it is positive (respectively, negative) in all of its variables;
- ψ is *monotone* if it is positive or negative in each of its variables.

Example.

- Every elementary conjunction is monotone since a variable appears in it at most once.

- The DNF $\phi(x,y,z) = xy \vee x\overline{y}\,\overline{z} \vee xz$ is positive in x, and is neither positive nor negative (i.e. is not monotonic) in x and y.
- The DNF $\psi(x,y,z) = xy \vee x\overline{z} \vee y\overline{z}$ is positive in x, positive in y, and negative in z and is thus monotone, but neither negative nor positive. △

Every positive DNF represents a positive function: since all literals are positive, increasing any input variable from 0 to 1 cannot cause a term to decrease.

However, the converse is not true in general: a non-positive, or even non-monotone DNF may represent a positive function. For instance, $\phi(x,y,z) = xy \vee x\overline{y}\,\overline{z} \vee xz$ represents the positive function $f(x,y,z) = x$.

The next theorem characterising positive variables closely mirrors a previous result characterising inessential variables, Theorem 29.1.20, in the sense that f being positive in x_k means that the negative literal $\overline{x}_k$ is "inessential" in f.

Theorem 29.1.24. *Let f be a Boolean function on $\mathcal{B}^n$ and let $k \in [n]$. Then, the following are equivalent:*

(i) f is positive in x_k;

(ii) The literal $\overline{x}_k$ does not appear in any prime implicant of f;

(iii) f has a DNF representation in which the literal $\overline{x}_k$ does not appear.

Proof. As for Theorem 29.1.20, *(ii)* → *(iii)* since any function can be represented by a DNF of its prime implicants (Theorem 29.1.12), and *(iii)* → *(i)* by Theorem 29.1.18, since if $\phi = \bigvee_{j=1}^{m} C_j$ is any DNF representing f where $\overline{x}_k$ does not appear, then the substitution $x_k = 1$ has no effect on any terms involving x_k, while the substitution $x_k = 0$ deletes any such terms, and hence $f|_{x_k=0} \leq f|_{x_k=1}$.

Now, suppose f is positive in x_k and consider an prime implicant $C_{AB} = \bigwedge_{i\in A} x_i \bigwedge_{j\in B} \overline{x}_j$ of f. If $k \notin B$, then we are done, so otherwise suppose $k \in B$ and consider the conjunction C obtained by deleting $\overline{x}_k$ from C_{AB}:

$$C = \bigwedge_{i\in A} x_i \bigwedge_{j\in B\setminus\{k\}} \overline{x}_j$$

Since C_{AB} is prime, C is not an implicant of f, so there exists a point $X = (x_1, \ldots, x_n) \in \mathcal{B}^n$ such that $C(X) = 1$ but $f(X) = 0$. Since C_{AB} is an implicant of f, we must have $C_{AB}(X) = 0$ and hence $x_k = 1$. Now consider the point $Y \in \mathcal{B}^n$ equal to X in every component apart from the kth, where $y_k = 0$. Then, $C_{AB}(Y) = 1$, and hence $f(Y) = 1$, contradicting that f is positive in x_k. ■

Corollary 29.1.24.1. *A Boolean function is positive if and only if it can be represented by an expression with no complemented variables.*

Theorem 29.1.25. *Let ϕ and ψ be DNFs and suppose that ψ is positive. Then, ϕ implies ψ if and only if each term of ϕ is absorbed by some term of ψ.*

Proof. Suppose without loss of generality that ϕ and ψ are expressions in the same n variables. The forward direction follows immediately from Theorem 29.1.10. Conversely, suppose ϕ implies ψ and consider some term of ϕ, say,

$$C_k = \bigwedge_{i\in I} x_i \bigwedge_{j\in B} \overline{x}_j$$

Let e_A be the characteristic vector of A, so $C_k(e_A) = \phi(e_A) = 1$. Since ϕ implies ψ, we have $\phi(e_A) = 1$, and thus some term $C_j = \bigwedge_{i\in F} x_i$ of ψ satisfies $C_j(e_A)$, and thus $F \subseteq A$, so C_j absorbs C_k as required. ■

Theorem 29.1.26. *The complete DNF of a positive Boolean function f is positive and irredundant, and is furthermore the unique prime DNF of f.*

Proof. Let f be a positive Boolean function with prime implicants $P_1,\ldots,P_n$, and let $\psi = \bigvee_{i=1}^{m} P_i$ be the complete DNF of f. By Theorem 29.1.24, ψ is positive. Now, let $\phi = \bigvee_{k=1}^{r} P_k$ be any prime expression of f, where $r \in [m]$. Since $f = \phi = \psi$, we have in particular that ψ implies ϕ, so by the previous theorem, each term of ψ is absorbed by some term of ϕ. In particular, if $m > r$, then P_m must be absorbed by some other prime implicant P_k for $k \leq r$, contradicting the primality of P_m. So, $r = m$, and hence $\psi = \phi$ is irredundant and unique. ■

This theorem shows that the complete DNF provides a "canonical" shortest DNF representation of a positive Boolean function. Since a shortest DNF representation is necessarily prime and irredundant, no other DNF representation of a positive Boolean function can be as short as its complete DNF.

Theorem 29.1.27. *Let $\phi = \bigvee_{k=1}^{m} \left(\bigwedge_{i \in A_k} x_i \bigwedge_{j \in B_k} \overline{x}_j \right)$ be a DNF representation of a positive Boolean function f. Then, $\psi = \bigvee_{k=1}^{m} \bigwedge_{i \in A_k} x_i$ is a positive DNF representation of f, and the prime implicants of f are the terms of ψ which are not absorbed by other terms of ψ.*

Proof. Since every term of ϕ is absorbed by some term in ψ, i.e. $\bigwedge_{i \in A_k} x_i \bigwedge_{j \in B_k} \overline{x}_j$ by $\bigwedge_{i \in A_k} x_i$ and ψ is positive, $\phi = f$ implies ψ. For the reverse inequality, consider any point $X = (x_1,\ldots,x_n) \in \mathcal{B}^n$ such that $\psi(X) = 1$. Then, there is a term of ψ that takest the value 1 at X, or equivalently, one of the terms defined by A_k has $x_i = 1$ for all $i \in A_k$. Let e_{A_k} be the characteristic vector of this A_k, so $\phi(e_{A_k}) = f(e_{A_k}) = 1$. Moreover, $e_{A_k} \leq X$, and therefore, by the positivity of f, $f(X) = 1$, and hence $\psi \leq f$. So $\psi = f$.

For the second part of the statement, consider the complete DNF ψ^* of f. Since ϕ is positive and $\psi*$ implies ψ, every term of ψ^* is absorbed by some term of ψ. However, the terms of ψ are implicants of f, while the terms of ϕ^* are prime implicants of f, so all prime implicants of f must appear amongst the terms of ψ. ■

Example. As mentioned previously, the DNF $\phi(x,y,z) = xy \vee x\overline{y}\,\overline{z} \vee xz$ represents the positive function $f(x,y,z) = x$. By deleting all the negative literals from ϕ, we obtain the DNF $\psi = xy \vee x \vee xz$. The terms xy and xz are absorbed by x, so they are not prime. The remaining term x is thus the only prime implicant of f. △

Let f be a Boolean function on $\mathcal{B}^n$ and let $X \in T(f)$ be a true point of f. Then, X is a *minimial true point* of f if there is no distinct true point $Y \neq X$ such that $Y \leq X$ (pointwise). Dually, $X \in F(f)$ is a *maximal false point* of f if there is no distinct false point $Y \neq X$ such that $X \leq Y$.

We denote by $\min T(f)$ the set of minimal true points of f, and $\max F(f)$ the set of maximal false points of f.

Theorem 29.1.28. *Let f be a positive Boolean function on $\mathcal{B}^n$ and let $Y \in \mathcal{B}^n$. Then,*

- *(i) Y is a true point of f if and only if there exists a minimal true point X of f such that $X \leq Y$;*
- *(ii) Y is a false point of f if and only if there exists a maximal false point X of f such that $Y \leq X$.*

Proof. The forward implications are trivial in both cases and are independent of the positivity assumption. The reverse implications are straightforward corollaries of the characterisation of positivity as $f(X) \leq f(Y)$ whenever $X \leq Y$. ■

Theorem 29.1.29. *Let f be a positive Boolean function on $\mathcal{B}^n$, let $C_A = \bigwedge_{i \in A} x_i$ be an elementary conjunction, and let e_A be the characteristic vector of A. Then,*

- *(i) C_A is an implicant of f if and only if e_A is a true point of f;*
- *(ii) C_A is a prime implicant of f if and only if e_A is a minimal true point of f.*

Proof.

(i) The forward implication is again trivial and independent of the positivity assumption. Conversely, if e_A is a true point of f, then $\bigwedge_{i\in a} x_i \bigwedge_{j\notin A} \overline{x}_j$ is an implicant of f. Then, by the positivity of f, C_A is also an implicant of f by the same reasoning as in the proof of Theorem 29.1.27.

(ii) Let C_B be an elementary conjunction, and let e_B be the characteristic vector of B. Note that $C_A \leq C_B$ if and only if $B \subseteq A$, or equivalently, $e_B \leq e_A$. Together with (i), this implies that C_A is a prime implicant of f if and only if e_A is a minimal true point of f.

■

Example. Consider the positive function $f(x,y,z,w) = xy \vee xzw \vee yz$. Each term is a implicant, so the indicator vectors 1100, 1011, and 0110 are true points of f.

Note that the positivity requirement in this theorem is essential:

(i) The function $g(x,y) = x\overline{y}$ has true point 10, but x is not an implicant of g;

(ii) The function $h(x,y,z) = xy \vee \overline{x}\,\overline{z}$ has prime implicants xy and $\overline{x}\,\overline{z}$ with the corresponding true points 110 and 101, but 000 is the unique minimal true point of g.

△

A dual correspondence also holds between maximal false points and prime implicates of a positive function:

Theorem 29.1.30. *Let f be a positive Boolean function on $\mathcal{B}^n$, let $D_A = \bigvee_{i\in A} x_i$ be an elementary disjunction, and let $e_{[n]\setminus A}$ be the characteristic vector of $[n] \setminus A$. Then,*

(i) *D_A is an implicate of f if and only if $e_{N\setminus A}$ is a false point of f;*

(ii) *D_A is a prime implicate of f if and only if $e_{N\setminus A}$ is a maximal false point of f.*

Proof. The structure of the previous proof also suffices for this result with minor modifications. Alternatively, De Morgan's laws and simple duality arguments can be applied to the previous theorem statement. ■

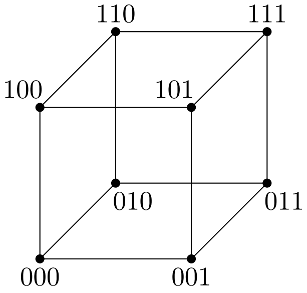
Example. Again, consider the positive function $f(x,y,z,w) = xy \vee xzw \vee yz$. Its prime implicates are $x \vee y$, $x \vee z$, $y \vee z$, and $y \vee w$, so the complementary indicator vectors 0011, 0101, 1001, and 1010 are maximal false points of f. △

29.1.12 Other Representations of Boolean Functions

Boolean functions can be represented in many other ways than just truth tables and Boolean expressions. In this section, we briefly outline some other representations.

29.1.12.1 Geometric Interpretation

The *hypercube* Q_n is the graph with vertex set $\mathcal{B}^n$, where two vertices are adjacent if and only if they differ in one coordinate.



The hypercube Q_3.

A subset $S \subseteq \mathcal{B}^n$ is a *subcube* of Q_n if $|S| = 2^k$ for some $k \leq n$, and there are $n-k$ coordinates in which all the vectors of S coincide. That is, a subcube (of dimension k) is obtained from Q_n by fixing $n-k$ coordinates to 0 or 1.

Lemma 29.1.31. *Let $C = \bigwedge_{i\in A} x_i \bigwedge_{j\in B} \overline{x}_j$ be an elementary conjunction of length $k = |A \cup B| \leq n$. Then, the set of true points of C consists of 2^{n-k} points and defines a subcube of Q_n of codimension k.*

Proof. Let $F = A \cup B$ be the set of (indices of) variables present in C. Then, changing any of the $n-k$ variables not in F does not affect the value of C, so there are 2^{n-k} true points of C. Moreover, these true points all agree in the $n-k$ coordinates fixed by F and hence describe a subcube of codimension k. ■

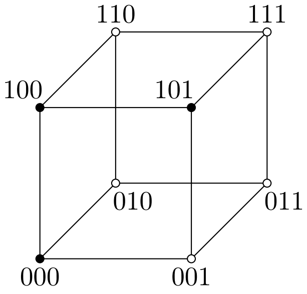
Example. The elementary conjunction $\overline{x}_1 x_3$ on $\mathcal{B}^3$ has true points 001 and 011, which has first and last coordinates fixed and hence describes a codimension-2 subcube. △

Corollary 29.1.31.1. *There is a bijection between elementary conjunctions and subcubes of Q_n.*

Proof. The correspondence described in the previous theorem is a bijection. ■

We can represent a Boolean function f on $\mathcal{B}^n$ by colouring the vertices corresponding to true points white, and vertices corresponding to false points black.

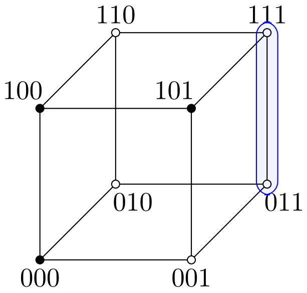
Example. The function f on $\mathcal{B}^3$ with true points $T(f) = \{001{,}010{,}011{,}110{,}111\}$ is represented by:



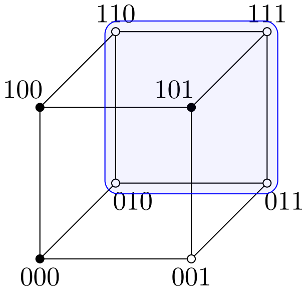
△

In view of the previous discussion, an implicant of f then corresponds to a subcube of Q_n that does not contain any false points, and is furthermore prime if it is maximal with this property, i.e. is not contained inside a larger subcube without false points.

Example. The expression x_2x_3 (i.e. the vertices of the form _11) is an implicant of the function f above, since it defines the subcube:



However, it is not prime, since it is contained in the subcube



△

Let $\phi = \bigvee_{k=1}^{m} C_k$ be a DNF representing f. The set of true points of f then coincides with the union of the sets of true points of the terms C_k. So, a DNF representing f can be viewed as a collection of subcubes of Q_n that cover precisely the true points of f. In particular, an orthogonal DNF is one for which the subcubes in the collection are all disjoint.

29.1.12.2 Representations of Boolean Functions over GF(2)

The *exclusive-or function*, *Boolean XOR*, or *parity function* is the Boolean function $\oplus : \mathcal{B}^n \to \mathcal{B}$ defined by

$$\oplus(x_1,x_2) \coloneqq x_1\overline{x}_2 \vee \overline{x}_1 x_2$$

We write this function in infix notation as $x_1 \oplus x_2$. When viewed as a binary operation, $\oplus$ is commutative and associative, and the iteration

$$f(x_1,\ldots,x_n) = x_1 \oplus \cdots \oplus x_n$$

takes the value 1 precisely when the number of 1s in $(x_1,\ldots,x_n)$ is odd. Moreover, $\oplus$ defines addition modulo 2 in the Galois field $\mathrm{GL}(2) = \big(\{0,1\}, \oplus, \wedge\big) \cong \mathbb{Z}/2$.

Theorem 29.1.32. *For every Boolean function f on $\mathcal{B}^n$, there exists a unique mapping $c : \mathcal{P}\big([n]\big) \to \{0,1\}$ such that*

$$f(x_1,\ldots,x_n) = \bigoplus_{A \in \mathcal{P}\left([n]\right)} c(A) \prod_{i \in A} x_i$$

Proof. We provide a constructive proof from first principles. To establish the existence of this representation, we induct on the dimension n. For $n = 1$, such a representation exists since $x = x$ and $\overline{x} = x \oplus 1$. Then, for $n > 1$, existence of the representation follows from the trivial identity,

$$f = f|_{x_n=0} \oplus x_n f|_{x_n=0} \oplus x_n f|_{x_n=1}$$

Indeed, by the inductive hypothesis, $f|_{x_n=1}$ and $f|_{x_n=1}$ have representations of the required form, and hence, after removing any duplicated terms with the identity $x \oplus x = 0$, f also has a representation of this form.

For uniqueness, it suffices to observe that there are exactly 2^{2^n} expressions of this form, and this is also the number of Boolean functions on $\mathcal{B}^n$. ■

These representations of Boolean functions over GL(2) are sometimes called *Zhegalkin polynomials*, *Reed-Muller expansions*, or *algebraic normal forms* (ANF).

To compute the Zhegalkin polynomials of a function, we proceed essentially by comparing coefficients.

Example. We represent the function $f = (x_1 \vee x_2)(x_2 \vee \overline{x}_3)$ as a Zhegalkin polynomial.

Let us write $c_{123} = c(\{x_1,x_2,x_3\})$, etc. for the coefficient mappings, so the general form of a Zhegalkin polynomial on three variables is:

$$(x_1 \vee x_2)(x_2 \vee \bar{x}_3) = c_{123}x_1x_2x_3 \oplus c_{12}x_1x_2 \oplus c_{13}x_1x_3 \oplus c_{23}x_2x_3 \oplus c_1x_1 \oplus c_2x_2 \oplus c_3x_3 \oplus c_0$$

Now, compare coefficients by evaluating each side at each Boolean point in $\mathcal{B}^3$:

(*i*) $(x_1,x_2,x_3) = (0{,}0{,}0)$

$$0 = c_0$$

(*ii*) $(x_1,x_2,x_3) = (0{,}0{,}1)$

$$0 = c_3$$

(*iii*) $(x_1,x_2,x_3) = (0{,}1{,}0)$

$$1 = c_2$$

(*iv*) $(x_1,x_2,x_3) = (0{,}1{,}1)$

$$\begin{aligned} 1 &= c_{23} \oplus c_2 \oplus c_3 \\ 1 &= c_{23} \oplus 1 \oplus 0 \\ 0 &= c_{23} \end{aligned}$$

(*v*) $(x_1,x_2,x_3) = (1{,}0{,}0)$

$$1 = c_1$$

(*vi*) $(x_1,x_2,x_3) = (1{,}0{,}1)$

$$\begin{aligned} 0 &= c_{13} \oplus c_1 \oplus c_3 \\ 0 &= c_{13} \oplus 1 \oplus 0 \\ 1 &= c_{13} \end{aligned}$$

(*vii*) $(x_1,x_2,x_3) = (1{,}1{,}0)$

$$\begin{aligned} 1 &= c_{12} \oplus c_1 \oplus c_2 \\ 1 &= c_{12} \oplus 1 \oplus 1 \\ 1 &= c_{12} \end{aligned}$$

(*viii*) $(x_1,x_2,x_3) = (1{,}1{,}1)$

$$\begin{aligned} 1 &= c_{123} \oplus c_{12} \oplus c_{13} \oplus c_{23} \oplus c_1 \oplus c_2 \oplus c_3 \\ 1 &= c_{123} \oplus 1 \oplus 1 \oplus 0 \oplus 1 \oplus 1 \oplus 0 \\ 1 &= c_{123} \end{aligned}$$

$$(x_1 \vee x_2)(x_2 \vee \bar{x}_3) = x_1x_2x_3 \oplus x_1x_2 \oplus x_1x_3 \oplus x_1 \oplus x_2$$

△

A Boolean function f on $\mathcal{B}^n$ is *linear* if there are coefficients $c_1,\ldots,c_n \in \{0,1\}$ such that

$$f(x_1,\ldots,x_n) = c_0 \oplus \bigoplus_{i=1}^{n} c_i x_i$$

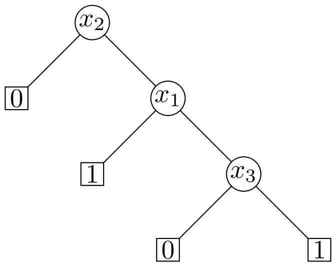
That is, each term is at most linear (i.e. it is an affine combination of variables).

29.1.12.3 Decision Trees

A *decision tree* is a rooted directed binary tree in which every non-leaf vertex v is labelled by a variable $x_{j(v)}$ for some labelling function $j : V \to [n]$, and every leaf vertex is labelled by the constants 0 or 1.

Every decision tree D corresponds to a Boolean function $\phi_D : \mathcal{B}^n \to \mathcal{B}$ as follows. Let $x = (x_1,\ldots,x_n)$ be a binary vector. Starting from the root, we move between vertices on the tree, following the left arc out of v if $x_{j(v)} = 0$, and the right arc otherwise, and stop when we arrive at a leaf, in which case we say that x is *classified* into this leaf. The label of the leaf then defines the value of $\phi_D(x)$.

Example. The decision tree



represents the function:

x_1,x_2,x_3	$f(x_1,x_2,x_3)$
000	0
001	0
010	1
011	1
100	0
101	0
110	0
111	1

△

A decision tree can be converted into a DNF as follows:

- For each leaf vertex v with label 1, construct the term corresponding to the path from the root to v, where a left arc corresponds to a complemented variable and a right arc to an uncomplemented variable;
- Take the disjunction of all such terms.

Example. For the tree above, the left leaf node with label 1 has a left arc from x_1, and a right arc from x_2, so the corresponding term is $\overline{x}_1x_2$. The rightmost leaf node similarly has term $x_1x_2x_3$. So, the DNF is given by

$$\overline{x}_1x_2 \vee x_1x_2x_3$$

△

We can also easily find the DNF for $\overline{f}$ by performing this algorithm on the leaves with label 0 instead.

Similar to DNF representations, decision tree representations are not unique.

29.2 Duality Theory

Recall that the dual f^d of a Boolean function $f(x_1,\dots,x_n)$ is the function

$$f^d(x_1,\dots,x_n) = \overline{f(\overline{x}_1,\dots,\overline{x}_n)}$$

Lemma 29.2.1. *$g = f^d$ if and only if $f(X) \vee g(\overline{X}) = 1$ and $f(X) \wedge g(\overline{X}) = 0$ for all $X \in \mathcal{B}^n$.*

Lemma 29.2.2. *$f \le g$ if and only if $g^d \le f^d$.*

Theorem 29.2.3. *Let $\phi = \bigvee_{k=1}^{m} \left(\bigwedge_{i \in P_k} x_i \bigwedge_{j \in N_i} \overline{x}_j \right)$ be a DNF of a Boolean function f, and let $C_{PN} = \bigwedge_{i \in P} x_i \bigwedge_{j \in N} \overline{x}_j$ be an elementary conjunction. Then,*

(i) C_{PN} is an implicant of f^d if and only if

$$(P \cap P_k) \cup (N \cap N_i) \neq \varnothing$$

for all $k \in [m]$;

(ii) C_{PN} is a prime implicant of f^d if and only if (i) holds , for every $P' \subseteq P$ and $N' \subseteq N$ with $P' \cup N' \neq P \cup N$, there exists an index $k \in [m]$ such that $(P' \cap P_k) \cup (N' \cap N_k) = \varnothing$.

Proof.

(i) By the definition of a dual function, $C_{PN} = \bigwedge_{i \in P} x_i \bigwedge_{j \in N} \overline{x}_j$ is an implicant of f^d if and only if $C_{NP} = \bigwedge_{i \in P} \overline{x}_i \bigwedge_{j \in N} x_j$ is an impcilant of $\overline{f}$. Since $f \wedge \overline{f} = 0$, the identity $C_{P_i N_j} \wedge C_{NP} = 0$ must hold, and hence $(P \cap P_i) \cup (N \cap P_i) \neq \varnothing$ for all $i \in [m]$.

Conversely, if this intersection is non-empty, then $f \wedge C_{NP} = 0$ holds identicallyh, so C_{NP} is an implicant of $\overline{f}$.

(ii) This follows from the defition of prime implicants.

■

Note that in the previous theorem, the conjunctions $C_{P_i N_i}$ could have been taken to be prime implicants, rather than arbitrary implicants of f.

29.2.1 Dual-comparable Functions

A Boolean function f is *dual-minor* if $f \le f^d$, *dual-major* if $f \ge f^d$, and *self-dual* if $f = f^d$.

Example. The function $f = x_1 x_2 x_3$ is dual-minor, with dual $f^d = x_1 \vee x_2 \vee x_3$:

x_1,x_2,x_3	f	f^d
000	0	0
001	0	1
010	0	1
011	0	1
100	0	1
101	0	1
110	0	1
111	1	1

Equivalently, $f^d \geq (f^d)^d = f$ is dual-major.

The function $g = x_1x_2\overline{x}_3 \vee x_1\overline{x}_2x_3 \vee \overline{x}_1x_2x_3 \vee \overline{x}_1\overline{x}_2\overline{x}_3$ has dual

$$\begin{aligned} g^d &= (x_1 \vee x_2 \vee \overline{x}_3)(x_1 \vee \overline{x}_2 \vee x_3)(\overline{x} \vee x_2 \vee x_3)(\overline{x}_1 \vee \overline{x}_2 \vee \overline{x}_3) \\ &= x_1x_2\overline{x}_3 \vee x_1\overline{x}_2x_3 \vee \overline{x}_1x_2x_3 \vee \overline{x}_1\overline{x}_2\overline{x}_3 \\ &= g \end{aligned}$$

so g is self-dual. △

Theorem 29.2.4. *Suppose that a Boolean function f has a prime implicant of degree 1. Then, f is dual-major. Moreover, f is dual-minor (and hence self-dual) if and only if it has no other prime implicants.*

Proof. Without loss of generality, suppose that x_1 is a prime implicant of f, so $f(x_1,\dots,x_n) = x_1 \vee g(x_2,\dots,x_n)$. Then, $f^d = x_1g^d$. Since $f = 0$ requires $x_1 = 0$, we have $f^d = 0$ whenever $f = 0$, so f is dual-major.

Now, suppose that f has no other prime implicant. Since x_1 is an implicant, $f = 1$ when $x_1 = 1$; and conversely, no point with $x_1 = 0$ is covered by any implicant, so $f = 0$ if $x_1 = 0$. Hence, $f = x_1$ is a projection, and is in particular self-dual.

Conversely, if f has another prime implicant, then there exists a point $X = (x_1^*,\dots,x_n^*) \in \mathcal{B}^n$ such that $x_1^* = 0$ and $f(x_1^*,\dots,x_n^*) = 1$. But $x_1^* = 0$ implies $f^d(x_1^*,\dots,x_n^*) = 0$, so f is not dual-minor. ■

The following result can be viewed as a restatement of the definition of dual comparisons:

Theorem 29.2.5. *Let f be a Boolean function on $\mathcal{B}^n$. Then,*

(i) f is dual-minor if and only if the complement of every true point of f is a false point of f. That is, for all $X \in \mathcal{B}^n$, $f(X) = 1$ implies $f(\overline{X}) = 0$, or equivalently, $f(X)f(\overline{X}) = 0$.

(ii) f is dual-major if and only if the complement of every false point of f is a true point of f. That is, for all $X \in \mathcal{B}^n$, $f(X) = 0$ implies $f(\overline{X}) = 1$, or equivalently, $f(X) \vee f(\overline{X}) = 0$.

(iii) f is self-dual if and only if every pair of complementary points contains exactly one true point and one false point of f. That is, for every $X \in \mathcal{B}^n$, $f(X) = 1$ if and only if $f(\overline{X}) = 0$.

A dual-minor function f is *maximally dual-minor* if there does not exist a distinct dual-minor function $g \neq f$ such that $f \leq g$.

Theorem 29.2.6. *A Boolean function is self-dual if and only if it is maximally dual-minor.*

Proof. If f is self-dual and g is a dual-minor function such that $f \leq g$, then,

$$g^d \leq f^d = f \leq g \leq g^d$$

so $g^d = f = f^d$ and hence $f = g$ is maximally dual-minor.

Conversely, suppose that f is not self-dual. If f is not dual-minor, we are done. Otherwise, suppose that f is dual-minor, so there exists a point $X^* = (x_1^*, \dots, x_n^*)$ with $f(X^*) = 0$ and $f^d(X^*) = 1$. Without loss of generality, suppose that $x_1^* = 1$, and consider the function $g = f \vee f^d x_1$ (if $X^* = 0$, take $g = f \vee f^d \overline{x}_1$ instead). Clearly, $f \leq g$, and $f \neq g$ since $g(X^*) = 1$. Moreover, g is dual-minor (actually, self-dual):

$$g^d = f^d(f \vee x_1) = f^d f \vee f^d x_1 = g$$

so f is not maximally dual-minor. ■

The dual result followss similarly:

Theorem 29.2.7. *A Boolean fuction is self-dual if and only if it is minimally dual-major.*

The construction in the proof above can be generalised to yield a simple standand way of associating a self-dual function to any arbitrary Boolean function.

Given a Boolean function f on $\mathcal{B}^n$, the *self-dual extension* of f is the function f^{SD} on $\mathcal{B}^{n+1}$ defined by

$$f^{\mathrm{SD}}(x_1, \dots, x_{n+1}) \coloneqq f(x_1, \dots, x_n)\overline{x}_{n+1} \vee f^d(x_1, \dots, x_n)x_{n+1}$$

Lemma 29.2.8. *For every Boolean function f, the function f^{SD} is self-dual.*

Proof. The dual of the f^{SD} is:

$$\begin{aligned}(f^{\mathrm{SD}})^d &= \big(f^d(X) \vee \overline{x}_{n+1}\big)\big(f(X) \vee x_{n+1}\big)\\ &= f^d(X)f(X) \vee f(X)\overline{x}_{n+1} \vee f^d(X)x_{n+1} \vee x_{n+1}\overline{x}_{n+1}\\ &= f(X)\overline{x}_{n+1} \vee f^d(X)x_{n+1}\end{aligned}$$

and hence f^{SD} is self-dual. ■

Theorem 29.2.9. *The mapping $(-)^{\mathrm{SD}} : f \mapsto f^{\mathrm{SD}}$ is a bijection from the set of Boolean functions of n variables and the set of self-dual functions of $n+1$ variables.*

Proof. The mapping $(-)^{\mathrm{SD}}$ is injective, since the restriction of f^{SD} to $x_{n+1} = 0$ is precisely f. Moreover, $(-)^{\mathrm{SD}}$ has an inverse given by $g \mapsto g|_{x_{n+1}=0}$ for every self-dual function g on $\mathcal{B}^n$:

$$\begin{aligned}(g|_{x_{n+1}=0})^{\mathrm{SD}} &= g|_{x_{n+1}=0}\overline{x}_{n+1} \vee (g|_{x_{n+1}=0})^d x_{n+1}\\ &= g|_{x_{n+1}=0}\overline{x}_{n+1} \vee g^d|_{x_{n+1}=0}x_{n+1}\\ &= g|_{x_{n+1}=0}\overline{x}_{n+1} \vee g|_{x_{n+1}=1}x_{n+1}\end{aligned}$$

which is precisely the Shannon expansion of g. ■

When applied to dual-minor functions, the definition of a self-dual extension takes a simpler form:

Theorem 29.2.10. *If f is dual-minor, then $f^{\mathrm{SD}} = f \vee f^d x_{n+1}$*

Proof. This holds since for all $a,b,x \in \mathcal{B}$, $a \leq b$ implies $a\overline{x} \vee bx = a \vee bx$. ■

29.2.2 Duality Properties of Positive Functions

Recall that a Boolean function f is positive if and only if $X \leq Y$ implies $f(X) \leq f(Y)$ for all $X,Y \in \mathcal{B}^n$, and if and only if f can be represented by a positive expression, i.e. an expression without complemented variables.

We have also seen that the complete DNF of a positive Boolean function is positive and irredundant, and since the dual of a positive expression is positive, we have:

Theorem 29.2.11. *A function f is positive if and only if its dual f^d is positive.*

Recall that for a positive function f, we denote by $\min T(f)$ the set of minimal true points of f, and by $\max F(f)$ the set of maximal false points of f. We have seen that an elementary conjunction $C_A = \bigwedge_{i\in A} x_i$ is a prime implicant of f if and only if its characteristic vector e_A is a minimal true point of f. This can dualised for maximal false points as follows:

Theorem 29.2.12. *Let f be a positive Boolean function on $\mathcal{B}^n$, let $C_A = \bigwedge_{i\in A}$ be an elementary conjunction, and let $e_{[n]\setminus A}$ be the characteristic vector of $[n] \setminus A$. Then, C_A is a prime implicant of f^d if and only if $e_{[n]\setminus A}$ is a maximal false point of f.*

Example. Let $f = x_1 \vee x_2x_3$ be a positive function. Its dual is given by $f^d = x_1(x_2 \vee x_3) = x_1x_2 \vee x_1x_3$, with prime implicants x_1x_2 and x_1x_3. So, the maximal false point of f are given by the complementary indicator vectors 001 and 010. △

We can also characterise dual prime implicants of positive Boolean functions in terms of hypergraphs.

Let $\mathcal{H} = (V,\mathcal{E})$ be a hypergraph. Then, a set $S \subseteq V$ of vertices is:

- *stable* if it does not contain any edge of $\mathcal{H}$;
- a *transversal* of $\mathcal{H}$ if it intersects every edge of $\mathcal{H}$.

A transversal is furthermore *minimal* if it is minimal with respect to inclusion of transversals. A set $E \subseteq \mathcal{E}$ of pairwise disjoint edges is a *matching.*

A hypergraph $\mathcal{H} = (V,\mathcal{E})$ is a *clutter*, *Sperner family*, or a *simple hypergraph* if no edge of $\mathcal{H}$ is a subset of any other edge.

For a positive Boolean function f on $\mathcal{B}^n$, we associate the hypergraph $\mathcal{H}_f = \big([n],\mathcal{P}\big)$ where $\mathcal{P}$ is the collection of sets $P \subseteq [n]$ of indices such that $\bigwedge_{i\in P} x_i$ is a prime implicant of f. $\mathcal{H}_f$ is necessarily a clutter since the implicants are prime.

Theorem 29.2.13. *Let $f = \bigvee_{P\in\mathcal{P}} \bigwedge_{i\in P} x_i$ and $g = \bigvee_{T\in\mathcal{T}} \bigwedge_{i\in T} x_i$ be the complete DNFs of two positive functions on $\mathcal{B}^n$. Then, the following are equivalent:*

(*i*) *$g = f^d$;*

(*ii*) *For every partition of $[n]$ into two disjoint sets A and $\overline{A}$, there is either a member of $\mathcal{P}$ contained in A, or a member of $\mathcal{T}$ containied in $\overline{A}$, but not both.*

(*iii*) *$\mathcal{T}$ is precisely the family of minimal transversals of $\mathcal{P}$.*

Example. Again, consider the positive function $f = x_1 \vee x_2x_3$ and its dual $f^d = x_1x_2 \vee x_1x_3$. The hypergraph $\mathcal{H}_f$ has the edge set $\mathcal{E} = \big\{\{1\},\{2,3\}\big\}$, and $\{1,2\}$ and $\{1,3\}$ are exactly the minimal transversals of $\mathcal{H}_f$. △

29.3 Complexity Measures of Boolean Functions

In this section, let f be a Boolean function on $\mathcal{B}^n$ and $X = (x_1,\ldots,x_n) \in \mathcal{B}^n$ be a Boolean point.

29.3.1 Certificate Complexity

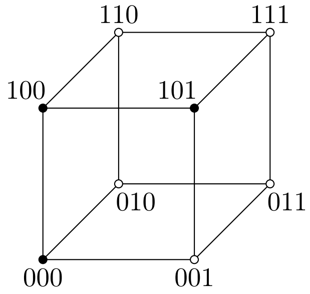
A *certificate* of f on X is a set $S \subseteq [n]$ of indices such that $f(Y) = f(X)$ for every Boolean point Y with $y_i = x_i$ for all $i \in S$. That is, a certificate is a collection of indices sufficient to determine the value of f.

The *size* of a certificate S is its cardinality $|S|$. The *certificate complexity* of f on X, denoted by $C(f,X)$, is the size of a smallest certificate of f on X. The *certificate complexity* $C(f)$ of f is the maximum certificate complexity of f over all Boolean points $X \in \mathcal{B}^n$. The 0-*certificate complexity* $C_0(f)$ of f is the maximum certificate complexity of f over all false points of f, and the 1-*certificate complexity* $C_1(f)$ of f is the maximum taken over all true points.

$$C(f) = \max_{X \in \mathcal{B}^n} C(f,X)$$
$$C_0(f) = \max_{X \in F(f)} C(f,X)$$
$$C_1(f) = \max_{X \in T(f)} C(f,X)$$

Informally, the certificate complexity of a Boolean function f on $\mathcal{B}^n$ at a point X is the codimension of the largest subcube containing X that defines a constant function; the 0-certificate complexity is the maximum codimension taken over all 0 points, and the 1-certificate complexity is the maximum codimension taken over all 1 points.

Example. Consider the function f represented by:



The certificate complexity at 111 is 1, since the largest constant subcube containing 111 is the entire back face, which is of codimension 1. The certificate complexity at 000 is 2 since the largest constant subcube is the edge 000-100 of codimension 2. We also have $C(f) = C_0(f) = C_1(f) = 2$ since every point is contained in a constant edge of codimension 2. △

Lemma 29.3.1. *The intersection of any* 0*-certificate and any* 1*-certificate is empty.*

Proof. Let S_0 be a 0-certificate on a false point X and let S_1 be a 1-certificate on a true point Y. Suppose $S_0 \cap S_1 = \varnothing$, and let Z be a point that coincides with X in all the S_0-positions and coincides with Y in all the S_1-positions. Then, Z is simultaneously true and false, which is impossible. So $S_0 \cap S_1 \neq \varnothing$. ■

Lemma 29.3.2. *A Boolean function f can be written as a k-DNF (a DNF where every term has at most k literals) if and only if $C_1(f) \leq k$. Similarly, a Boolean function f can be written as a k-CNF if and only if $C_0(f) \leq k$.*

Proof. Let ϕ_f be a k-DNF representing f. For every true point X of f, there is a term T of ϕ_f with $T(X) = 1$. Observe that the set of indices of literals in T is a certificate of X. Since ϕ_f is a k-DNF, then $C_1(f) \leq k$.

Conversely, let $C_1(f) \leq k$. For every true point X, a minimal certificate on X correspont s to a maximal subcube containing X, all of whose points are true. This subcube corresponds to a term with at most k

literals. The disjunction of all such terms taken over all true points then represents f. The proof for C_0 is similar. ■

29.3.2 Sensitivity and Block Sensitivity

Given a set $S \subseteq [n]$ of indices, we denote by X^S the Boolean point obtained from X by complementing all the components x_i with $i \in S$. In particular, we abbreviate $X^{\{i\}}$ to X^i.

The *sensitivity* $s(f,X)$ of f on X is the number of indices i such that $f(X) \neq f(X^i)$. The *sensitivity* $s(f)$ of f is the maximum sensitivity over all Boolean points $X \in \mathcal{B}^n$; the 0-*sensitivity* $s_0(f)$ is the maximum sensitivity over all false points of f; and the 1-*sensitivity* $s_1(f)$ is the maximum taken over all true points:

$$s(f) = \max_{X \in \mathcal{B}^n} s(f,X)$$
$$s_0(f) = \max_{X \in F(f)} s(f,X)$$
$$s_1(f) = \max_{X \in T(f)} s(f,X)$$

In terms of the hypercube, the sensitivity of f at a given vertex is the number of neighbouring vertices with a different colour.

Example. For the function f from the previous example, the sensitivity of f at 011 is 0 since it has zero neigbours of differing colours, i.e. complementing any of the bits does not change the value of f. On the other hand, the sensitivity of f at 000 is 2, since flipping the second or third bit changes f from 0 to 1, and flipping the first bit leaves f unchanged. △

The *block sensitivity* $bs(f,X)$ of f on X is the maximum number of disjoint non-empty sets of indices $B_1,\dots,B_b \subseteq [n]$ called *sensitivity blocks* such that $f(X) \neq f(X^{B_i})$ for all i. The *block sensitivity* $bs(X)$ of f is the maximum block sensitivity over all Boolean points $X \in \mathcal{B}^n$; the 0-*block sensitivity* $s_0(f)$ is the maximum block sensitivity over all false points of f; and the 1-*block sensitivity* $s_1(f)$ is the maximum taken over all true points:

$$bs(f) = \max_{X \in \mathcal{B}^n} bs(f,X)$$
$$bs_0(f) = \max_{X \in F(f)} bs(f,X)$$
$$bs_1(f) = \max_{X \in T(f)} bs(f,X)$$

Example. For the same function f as in previous example, the sensitivity of f at 100 is 1 since f only changes if we flip the middle bit, but the block sensitivity of f at 100 is 2, since we have the blocks $B_1 = \{1,3\}$ and $B_2 = \{2\}$ (yielding 001 and 110, respectively). △

Lemma 29.3.3. *For any Boolean function f,*

$$s(f) \leq bs(f) \leq C(f)$$

Proof. For any point $X \in \mathcal{B}^n$, the sensitivity of f on X coincides with the block sensitivity of f on X if we do not allow blocks of size more than 1. Therefore, by allowing blocks of arbitrary size, we cannot decrease the sensitivity and hence $s(f) \leq bs(f)$.

For each Boolean point X, a certificate on X must contain at least one index from each sensitivity block, and hence $bs(f,X) \leq C(f,X)$, so $bs(f) \leq C(f)$. ■

Theorem 29.3.4. *For any Boolean function f,*

$$C(f) \leq s(f)bs(f)$$

Proof. Consider a point $X \in \mathcal{B}^n$. Finst, note that if B is a minimal sensitivity block for X, then $|B| \leq s(f)$, since of we flip one of the bits in X^B indexed by B, then the function value must flip from $f(X^B)$ to $f(X)$ since B is minimal, so every coordinate in B is sensitive on X^B. Therefore, $|B| \leq s(f,X^B) \leq s(f)$.

Now, let $B_1,\ldots,B_b$ be disjoint minimal blocks that achieve the block sensitivity $b = bs(f,X) \leq bs(f)$, and let $C = \bigcup_i B_i$. If C is not a certificate for f on X, then there is an index $i \notin C$ such that $f(X) \neq f(X^i)$. But then, $\{i\}$ is a sensitivity block for f on X disjoint from $B_1,\ldots,B_b$, contradicting that $b = bs(f,X)$. Thus, C is a certificate for f on X. By the same argument as above, $|B_i| \leq s(f)$ for all $i \in [b]$, and hence $|C| = |\bigcup_i B_i| \leq s(f)bs(f)$.

Since for each X there is a certificate of size at most $s(f)bs(f)$, we have $C(f) \leq s(f)bs(f)$. s(f)bs(f). ■

It was a long standing open problem, known as the *sensitivity conjecture*, whether there is a constant c such that $bs(f) \leq s(f)^c$ for any Boolean function f. The conjecture was eventually resolved positively in the following:

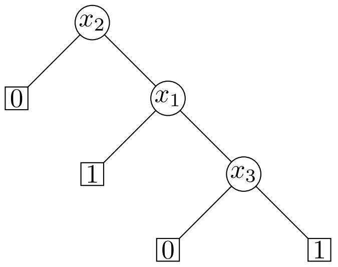
Theorem 29.3.5 (Huang)**.** *For every Boolean function f,*

$$bs(f) \leq s(f)^4$$

29.3.3 Decision Tree Complexity

Let t be a decision tree for a Boolean function f. For each point $X \in \mathcal{B}^n$, the number of bits of X examined by t before it is classified is called the *cost* of t on X, denoted $\text{cost}(t,X)$.

Example. Let t be the following decision tree:



The cost of t on 000 is 1, since only x_2 is examined before 000 is immediate classified with 0, while the cost of t on 101 is 3, since x_2 then x_1 must be examined before 101 is classified with 1. △

We denote by $\mathcal{T}$ the set of all decision trees that represent f. Then, the *decision tree complexity* of f is defined as:

$$D(f) = \min_{t\in\mathcal{T}} \max_{X\in\mathcal{B}^n} \text{cost}(t,X)$$

Equivalently, the decision tree complexity of f is the depth of an optimal decision tree that represents f.

Theorem 29.3.6. *For any Boolean function f,*

$$bs(f) \leq D(f)$$

Proof. Consider a point $X \in \mathcal{B}^n$ with maximally many sensitivity blocks $B_1,\ldots,B_{bs(f)}$. To evaluate f on X, a decision tree must examine at least one index from each block B_i, since other wise we could flip that block without the tree being able to detect this, i.e. the tree would be unable to distinguish $f(X) \neq f(X^{B_i})$. Thus, the tree must make at least $bs(f)$-many queries on X. ■

Theorem 29.3.7. *For any Boolean function f,*

$$D(f) \leq C_1(f)C_0(f) \leq C(f)^2$$

Proof. Let $k = C_1(f)$ and $\ell = C_0(f)$. Using Lemma 29.3.2, let D be a k-DNF representation of f and C be an ℓ-DNF representation of f. Take a term T in D, and examine the values of (the at most k) variables in T. Once the values of the variables in T are fixed, we are left with a function f' with fewer variables. Since every clause in C has a variable in common with T by Lemma 29.3.1, after fixing the values of the variables in T, C transforms into an $(\ell - 1)$-CNF C' representing f'. By induction and Lemma 29.3.2, $D(f') \leq C_1(f')C_0(f') \leq k(\ell-1)$, and hence $D(f) \leq k+k(\ell-1) = k\ell = C_1(f)C_0(f)$. ■

29.4 Functional Completeness

So far, we have considered the notion of a Boolean expression as compositions defined inductively over the set of three functions; namely conjunction, disjunction, and negation. We have also considered the notion of a Zhegalkin polynomials as expressions defined inductively over the set of functions $\{0,1,\oplus,\wedge\}$.

We now consider expressions defined inductively over arbitrary sets of function, and not necessarily of two variables.

A set of Boolean functions $\{f_1,f_2,\ldots\}$ is (*functionally*) *complete* if any Boolean function can be written as an expression over the functions in the set.

The set of all Boolean functions is trivially complete, but we have also seen that the sets $\{\overline{x},x_1\wedge x_2,x_1\vee x_2\}$ and $\{0,1,x_1\oplus x_2,x_1\wedge x_2\}$ are complete. Obviously, not every set of functions is complete, as, for instance, the set $\{0,1\}$ is not complete. The following theorem allows us to reduce the question of completeness of some sets of Boolean functions to the same question for other sets of Boolean functions.

Theorem 29.4.1. *Suppose we have two sets of Boolean functions $\mathcal{F} = \{f_1,f_2,\ldots\}$ and $\mathcal{G} = \{g_1,g_2,\ldots\}$. If the set $\mathcal{F}$ is complete and every function in $\mathcal{F}$ can be represented as an expression over the functions in $\mathcal{G}$, then $\mathcal{G}$ is also complete.*

Proof. Let h be a Boolean function represented as an expression $C[f_1,f_2,\ldots]$ in $\mathcal{F}$. By assumption, every function f_i in $\mathcal{F}$ can be represented as an expression $C_i[g_1,g_2,\ldots]$ over the functions in $\mathcal{G}$. Then, $C[C_1,C_2,\ldots]$ expresses h over the functions in $\mathcal{G}$, so $\mathcal{G}$ is complete. ■

Example.

- The set $\{\overline{x},x_1 \wedge x_2\}$ is complete, since $x_1 \vee x_2 = \overline{\overline{x}_1 \wedge \overline{x}_2}$, and $\{\overline{x},x_1 \wedge x_2,x_1 \vee x_2\}$ is complete. Dually, the set $\{\overline{x},x_1 \vee x_2\}$ is complete.
- The set $\{x_1 \uparrow x_2\}$ is famously complete, enabling most modern computer hardware to be built from only NAND gates. To see this, observe that $x \uparrow x = \overline{x}$ and $(x_1 \uparrow x_2) \uparrow (x_1 \uparrow x_2) = x_1 \wedge x_2$. Dually, $\{x_1 \downarrow x_2\}$ is also complete.

△

Let $\mathcal{F}$ be a set of Boolean functions. The *closure* $[\mathcal{F}]$ of $\mathcal{F}$ is the set of all Boolean functions that can be represented as expressions over the functions in $\mathcal{F}$.

Theorem 29.4.2. *For any sets $\mathcal{F},\mathcal{F}_1,\mathcal{F}_2$ of functions,*

- *$\mathcal{F} \subseteq [\mathcal{F}]$;*
- *$[[\mathcal{F}]] = [\mathcal{F}]$;*
- *If $\mathcal{F}_1 \subseteq \mathcal{F}_2$, then $[\mathcal{F}_1] \subseteq [\mathcal{F}_2]$;*
- *$[\mathcal{F}_1] \cup [\mathcal{F}_2] \subseteq [\mathcal{F}_1 \cup \mathcal{F}_2]$.*

A set $\mathcal{F}$ of functions is (*functionally*) *closed* if $\mathcal{F} = [\mathcal{F}]$.

Example. The set of all Boolean functions is closed, while the set $\{1, x_1 \oplus x_2\}$ is not. Conversely, the set of all linear functions is closed, since a linear expression of linear expressions is linear. △

We can also characterise completeness in terms of closedness: a set $\mathcal{F}$ is complete if and only if $[\mathcal{F}]$ contains the set of all Boolean functions.

29.4.1 Important Closed Classes

29.4.1.1 Functions Preserving Constants

We denote by T_0 the class of Boolean functions that map the constant zero vector $000\ldots0$ to 0. That is, the functions $f(x_1,\ldots,x_n)$ such that

$$f(0,0,\ldots,0) = 0$$

Example. The functions 0, x, $x_1 \wedge x_2$, $x_1 \vee x_2$, and $x_1 \oplus x_2$ belong to T_0, while the functions 1 and $\overline{x}$ do not. △

Lemma 29.4.3. *The class T_0 contains 2^{2^n-1} functions on $\mathcal{B}^n$.*

Proof. We fix the value of the function at 0, and there are $2^n - 1$ other Boolean points, each of which can take 2 values. ■

Lemma 29.4.4. *The class T_0 is closed.*

Proof. For any functions $f, f_1, \ldots, f_n \in T_0$, the function $F = f(f_1,\ldots,f_n)$ belongs to T_0:

$$\begin{aligned} F(0,\ldots,0) &= f\big(f_1(0,\ldots,0),\ldots,f_n(0,\ldots,0)\big) \\ &= f(0,\ldots,0) \\ &= 0 \end{aligned}$$

so F preserves 0. ■

Similarly, we denote by T_1 the class of Boolean functions that send the constant one vector $111\ldots1$ to 1. That is, the functions $f(x_1,\ldots,x_n)$ such that

$$f(1,1,\ldots,1) = 1$$

Example. The functions 1, x, $x_1 \wedge x_2$, and $x_1 \vee x_2$, belong to T_1, while the functions 0, $\overline{x}$ do not. △

Since T_1 consists of functions dual to the functions of T_0, all theorems immediately dualise to T_1:

Corollary 29.4.4.1. *The class T_1 contains 2^{2^n-1} functions on $\mathcal{B}^n$.*

Corollary 29.4.4.2. *The class T_1 is closed.*

29.4.1.2 Self-Dual Boolean Functions

We denote by S the class of all self-dual Boolean functions.

Example. x and $\overline{x}$ are self-dual functions. △

Lemma 29.4.5. *The class S contains $2^{2^{n-1}}$ Boolean functions on $\mathcal{B}^n$.*

Proof. Self-dual Boolan functions must take opposite values on complementary Boolean points, so a self-dual Boolean function need only be defined on half the Boolean points. ■

Lemma 29.4.6. *The class S is closed.*

Proof. For any functions $f,f_1,\dots,f_n \in S$, the function $F = f(f_1,\dots,f_n)$ belongs to S:

$$\begin{aligned} F^d &= f^d(f_1^d,\dots,f_n^d) \\ &= f(f_1,\dots,f_n) \\ &= F \end{aligned}$$

so F is self-dual. ■

Lemma 29.4.7. *If f does not belong to S, then by substituting functions x and $\overline{x}$, it can be transformed into a constant* 0 *or* 1*.*

Proof. Since f is not self-dual, there exists a point $X \in \mathcal{B}^n$ such that $f(X) = f(\overline{X})$. For each i, define the functions

$$\phi_i(x) = \begin{cases} x & x_i = 1 \\ \overline{x} & x_i = 0 \end{cases}$$

and consider the function

$$F(x) = f\big(\phi_1(x),\dots,\phi_n(x)\big)$$

That is, we complement the components corresponding to the non-zero entries of X. Then,

$$\begin{aligned} F(0) &= f\big(\phi_1(0),\dots,\phi_n(0)\big) \\ &= f(\overline{x}_1,\dots,\overline{x}_n) \\ &= f(x_1,\dots,x_n) \\ &= f\big(\phi_1(1),\dots,\phi_n(1)\big) \\ &= F(1) \end{aligned}$$

so F is a constant. ■

29.4.1.3 Positive Functions

Recall that a Boolean function f is positive if any of the following equivalent conditions hold:

- The restrictions in every variable x_i satisfy $f|_{x_i=0} \le f|_{x_i=1}$;
- Whenever $X \le Y$, $f(X) \le f(Y)$;
- f has a DNF representation without complemented variables.

We denote by M the class of positive Boolean functions.

Example. 0, 1, x, $x_1 \wedge x_2$, and $x_1 \vee x_2$ are positive functions. △

Lemma 29.4.8. *The class M is closed.*

Proof. For any functions $f,f_1,\dots,f_n \in M$, let $F = f(f_1,\dots,f_n)$. Denote by p_i the number of variables of the function f_i, and by m the number of variables of F, and suppose without loss of generality that F depends only on the variables that appear in the function $f_1,\dots,f_n$.

For a Boolean point $X \in \mathcal{B}^n$, denote by X^i the projection of X into $\mathcal{B}^{p_i}$ along the variables corresponding to f_i. Note that if $X \le Y$, then the projections also satisfy $X^i \le Y^i$. Since the functions $f_1,\dots,f_n$ are positive, $f_i(X^i) \le f_i(Y^i)$, so

$$\big(f_1(X^1),\dots,f_n(X^n)\big) \le \big(f_1(Y^1),\dots,f_n(Y^n)\big)$$

and since f is positive, we have

$$F(X) = f\big(f_1(X^1),\dots,f_n(X^n)\big) \leq f\big(f_1(Y^1),\dots,f_n(Y^n)\big) = F(Y)$$

so F is positive. ■

Let us call two Boolean points $X,Y \in \mathcal{B}^n$ *neighbouring* if they differ in precisely one coordinate.

Lemma 29.4.9. *If f does not belong to M, then by substituting the constants* 0 *and* 1 *and the function x, it can be transformed into the function $\overline{x}$.*

Proof. First, we claim that there exist two neighbouring points $X^*,Y^* \in \mathcal{B}^n$ such that $X^* \leq Y^*$, and $f(X^*) > f(Y^*)$. Indeed, since f is not positive, there exist two Boolean points X' and Y' such that $X' \leq Y'$ and $f(X') > f(Y')$, and if X' and Y' are not neighbours and differ in $t > 1$ coordinates, then there is a sequence of Boolean points

$$X' = Z^1 \leq Z^2 \leq \cdots \leq Z^t = Y'$$

where Z^{i+1} is obtained from Z^i by flipping the first bit of Z^i that disagrees with Y, i.e. Z^{i+1} and Z^i are neighbours. Since $f(X') > f(Y')$, there is a pair of consecutive points X^* and Y^* in the sequence above such that $X^* \leq Y^*$, and $f(X^*) > f(Y^*)$. Suppose that $X^* = X^i$, so X^* and Y^* differ in the ith coordinate, and consider the function

$$\phi(x) = f(x_1^*,\dots,x_{i_1}^*,x,x_{i+1}^*,\dots,x_n^*)$$

Then,

$$\begin{aligned}\phi(0) &= f(x_1^*,\dots,x_{i-1}^*,0,x_{i+1}^*,\dots,x_n^*)\\ &= f(X^*)\\ &> f(Y^*)\\ &= f(x_1^*,\dots,x_{i-1}^*,1,x_{i+1}^*,\dots,x_n^*)\\ &= \phi(1)\end{aligned}$$

so $\phi(0) = 1$ and $\phi(1) = 0$, i.e. $(x) = \overline{x}$. ■

29.4.1.4 Linear Functions

Recall that a Boolean function f is linear if it can be expressed in the form

$$f(x_1,\dots,x_n) = c_0 \oplus \bigoplus_{i=1}^{n} c_n x_n$$

That is, it is a GL(2)-affine combination of variables.

We denote the class of linear Boolean functions by L.

Example. 0, 1, x, $\overline{x} = x \oplus 1$, and $x_1 \oplus x_2$ are linear functions, but $x_1 \wedge x_2$ and $x_1 \vee x_2$ are not. △

Lemma 29.4.10. *If f does not belong to L, then by substituting the constants* 0 *and* 1 *and the functions x and $\overline{x}$, and possibly by negating f, it can be transformed into the function $x_1 \wedge x_2$.*

Proof. Let

$$f(x_1,\dots,x_n) = \bigoplus_{A\in\mathcal{P}\left([n]\right)} c(A) \prod_{i\in A} x_i$$

be a Zhegalkin polynomial for f. Since f is not linear, there is a non-zero term in this polynomial of at least quadratic order, involving at least, say, x_1 and x_2. Then, the polynomial can be transformed as

$$\bigoplus_{A\in\mathcal{P}\left([n]\right)} c(A)\prod_{i\in A} x_i = x_1x_2f_1(x_3,\dots,x_n)\oplus x_1f_2(x_3,\dots,x_n)\oplus x_2f_3(x_3,\dots,x_n)\oplus f_4(x_3,\dots,x_n)$$

where $f_1\neq \mathbf{0}_{n-2}$, since the polynomial is unique. That is, there exist $(a_3,\dots,a_n)\in\mathcal{B}^{n-2}$ such that $f_1(a_3,\dots,a_n)=1$. Now, consider the function

$$f(x_1,x_2)=f(x_1,x_2,a_3,\dots,a_n)=x_1x_2\oplus\alpha x_1\oplus\beta x_2\oplus\gamma$$

for some coefficients $\alpha,\beta,\gamma\in\{0,1\}$. Now, let $\psi(x_1,x_2)$ be the function defined as

$$\psi(x_1,x_2)=\phi(x_1\oplus\beta,x_2\oplus\alpha)\oplus\alpha\beta\oplus\gamma$$

Then,

$$\begin{aligned}\phi(x_1\oplus\beta,x_2\oplus\alpha)\oplus\alpha\beta\oplus\gamma &= (x_1\oplus\beta)(x_2\oplus\alpha)\oplus\alpha(x_1\oplus\beta)\oplus\beta(x_2\oplus\alpha)\oplus\gamma\oplus\alpha\beta\oplus\gamma\\ &= x_1x_2\oplus\alpha x_1\oplus\beta x_2\oplus\alpha\beta\oplus\alpha x_1\oplus\alpha\beta\oplus\beta x_2\oplus\alpha\beta\oplus\gamma\oplus\alpha\beta\oplus\gamma\\ &= x_1x_2\end{aligned}$$

so $\alpha=\beta=\gamma=0$. To complete the proof it suffices to observe that $x\oplus 0=x$ and $x\oplus 1=\overline{x}$. ■

29.4.2 Post's Theorem

Theorem 29.4.11 (Post)**.** *A set $F=\{f_1,f_2,\dots\}$ of Boolean functions is complete if and only if it is not a subset of any of the following five closed classes: T_0, T_1, S, M, L.*

Proof. Suppose F is complete, so $[F]$ is the class of all Boolean functions. Now, suppose for a contradiction that $F\subseteq X$ for X being one of the forbidden classes T_0, T_1, S, M, L. But then, $[F]\subseteq[X]=X$, which is a contradiction, since none of the five classes contain all Boolean functions.

Conversely, suppose that F is not contained in any of the forbidden classes. Then, F contains a subset $F'=\{f_0,f_1,f_s,f_m,f_\ell\}$ of 5 (not necessarily distinct) functions witnessing this non-containment, i.e. $f_0\notin T_0$, $f_1\notin T_1$, $f_s\notin S$, $f_m\notin M$, $f_\ell\notin L$. Without loss of generality, suppose that these functions all depend on the same set of variables $x_1,\dots,x_n$. We claim that F' is complete.

First, the constants 0 and 1 can be obtained from f_1, f_0, and f_s. If $f_0(1,\dots,1)=1$, then $\phi(x)=f_0(x,\dots,x)$ is the constant 1, and the function $f_1\bigl(\phi(x),\dots\phi(x)\bigr)=f_1(1,\dots,1)=0$ is the constant 0 function. Otherwise, if $f_0(1,\dots,1)=0$, then $\phi(x)=f_0(x,\dots,x)=\overline{x}$, and hence by Lemma 29.4.7 we can use ϕ and f_s to obtain a constant. The second constant can then be obtained from the first by using ϕ.

Now, we can apply Lemma 29.4.9 using the two constants 0 and 1 and the function f_m to obtain the function $\overline{x}$.

Finally, we can apply Lemma 29.4.10 using the two constants 0 and 1, and the functions $\overline{x}$ and f_ℓ to constuct the function $x_1\wedge x_2$.

Since $\{\overline{x},x_1\wedge x_2\}$ is complete, the set F', and hence F, is also complete. ■

Corollary 29.4.11.1. *Every closed set of Boolean functions, different from the set of all Boolean functios is contained in one of the classes T_0, T_1, S, M, L.*

A set F of Boolean functions is *precomplete* if F is not complete, but for any Boolean function $f\notin F$, the set $F\cup\{f\}$ is complete. It follows that any precomplete set is closed.

Corollary 29.4.11.2. *There exist precisely* 5 *precomplete sets of Boolean functions: T_0, T_1, S, M, L.*

Theorem 29.4.12. *Every complete set F of Boolean functions contains a complete subset of at most* 4 *functions.*

Proof. We have seen in the proof of Post's theorem that F contains a complete subset of at most 5 functions. Moreover, we have seen that the function $f_0 \notin T_0$ does not belong either to S (if $f_0(0,\dots,0) = f_0(1,\dots,1) = 1$) or to $T_1 \cup M$ (if $f_0(0,\dots,0) = 1$, $f_0(1,\dots,1) = 0$). Therefore, either the set $\{f_0,f_1,f_m,f_\ell\}$ or the set $\{f_0,f_s,f_\ell\}$ is complete. ■

This bound is sharp, since the set of functions

$$\{f_1 = x_1x_2, \quad f_2 = 0, \qquad f_3 = 1, \quad f_4 = x_1 \oplus x_2 \oplus x_3\}$$

satisfies $f_3 \notin T_0$, $f_2 \notin S$, $f_4 \notin M$, and $f_1 \notin L$, and is thus complete, but any proper subset is incomplete, since $\{f_2,f_3,f_4\} \subseteq L$, $\{f_1,f_3,f4\} \subseteq T_1$, $\{f_1,f_2,f_4\} \subseteq T_0$, and $\{f_1,f_2f_3\} \subseteq M$.

A subset F' of a closed set F is *complete in* F if $[F'] = F$. That is, every function in F can be represented as an expression over the functions in F'. A *basis* of F is a minimal subset F' complete in F.

Example. From the previous example,

$$\{f_1 = x_1x_2, \quad f_2 = 0, \qquad f_3 = 1, \quad f_4 = x_1 \oplus x_2 \oplus x_3\}$$

is a basis for the set of all Boolean functions.

It is also possible to show that the set $\{0,1,x_1x_2,x_1 \vee x_2\}$ is a basis of M. △

In addition to the main theorem characterising functionally complete sets, Post also proved the following results:

Theorem 29.4.13. *Every closed class of Boolean functions has a finite basis.*

Theorem 29.4.14. *The set of all closed classes of Boolean functions is countable.*

We observe that the first of these two theorems implies the second one. However, originally Post proved the second theorem before the first.

29.5 Quadratic Functions

A DNF

$$\phi(x_1,\dots,x_n) = \bigvee_{k=1}^{m} \bigwedge_{i \in P_k} x_i \bigwedge_{j \in N_k} \overline{x}_j$$

is *quadratic* if althe its terms are quadratic. That is, if they are conjunctions of at most two literals. A term is called *linear* or *purely quadratic* if it has exactly one or exactly two literals, respectively. Similarly, a CNF is *quadratic* if all its clauses are disjunctions of at most two literals.

A Boolean function f is *quadratic* if it admits a quadratic DNF representation. The function f is *dually quadratic* if it admits a quadratic CNF representation. This is equivalent to f^d being quadratic.

A quadratic Boolean function f is *purely quadratic* if it is not constant and has no linear prime implicant. Equivalently, f is purely quadratic if no linear term appears in any DNF of f.

The next result follows immediately from the definition.

Lemma 29.5.1. *If f is purely quadratic, then in every quadratic DNF of f, every term is a prime implicant.*

Note that it is possible for a quadratic function to be represented by a DNF of higher degree.

Example. The function

$$f = x_1x_2\overline{x}_3\overline{x}_4 \vee x_1x_2\overline{x}_3x_4 \vee \overline{x}_1x_2\overline{x}_3x_4 \vee x_1x_2x_3 \vee \overline{x}_2\overline{x}_3x_4 \vee \overline{x}_1x_3 \vee \overline{x}_2\overline{x}_4$$

is quadratic, since it also admits the DNF

$$f = x_1x_2 \vee \overline{x}_1x_3 \vee \overline{x}_2\overline{x}_4 \vee \overline{x}_3x_4$$

$\triangle$

29.5.1 Quadratic Boolean Functions and Graphs

There are many connections between certain classes of quadratic functions and graphs.

Given any undirected graph $G = (V,E)$, its *stability function* is the quadratic Boolean function given by

$$f_G = \bigvee_{ij \in E} x_ix_j$$

Note that the prime implicants of f are precisely the terms x_ix_j of this DNF, which is also the unique irredundant DNF of f. It follows that this mapping from undirected graphs to positive purely quadratic Boolean functions is a a bijection.

29.5.1.1 The Matched Graph

Another graph that can be conveniently associated with a quadratic DNF ϕ is the *matched graph* G_ϕ. This undirected graph has vertex set $\{x_1,\dots,x_n,\overline{x}_1,\dots,\overline{x}_n\}$ with edges given by

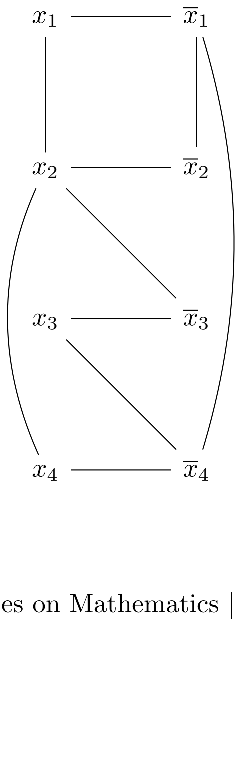
$$\big((x_i,\overline{x}_i) : i \in [n]\big) \cup \big((\alpha,\beta) : \alpha\beta \text{ is a term of } \phi\big)$$

That is, start with the bipartite graph with parts $\{x_1,\dots,x_n\}$ and $\{\overline{x}_1,\dots,\overline{x}_n\}$ with complementary vertices matched, then connect vertices that are paired up in terms. Note that if ϕ contains linear terms, a loop (α,α) is added for each such term α.

Example. The matched graph associated to the DNF

$$\phi = x_1x_2 \vee \overline{x}_1\overline{x}_2 \vee \overline{x}_1\overline{x}_4 \vee x_2\overline{x}_3 \vee x_2x_4 \vee x_3\overline{x}_4$$

is given by:



$\triangle$

The edges of G_ϕ are classified as:

- *positive*, (x_i,x_j);
- *negative*, $(\overline{x}_i,\overline{x}_j)$;
- *mixed*, $(x_i,\overline{x}_j)$;
- *null*, $(x_i,\overline{x}_i)$.

Example. The positive edges are the edges within the left part; the negative edges are the edges within the right part; the mixed edges are the non-horizontal edges between the two parts; and the null edges are the horizontal edges. △

The consistency of the quadratic Boolean equation $\phi = 0$ has a nice graph-theoretic counterpart for G_ϕ as follows.

Let $\mu(G)$ be the maximum cardinality of a matching on G, and $\tau(G)$ be the minimum cardinality of a vertex cover in G. Note that

$$\tau(G) \leq \tau(G)$$

since we need at least one vertex in a minimum vertex cover for every edge in a maximum matching.

The graph G is said to have the *Kőnig–Egerváry* (KE) *property* if this the above expression is in fact an equality.

Theorem 29.5.2. *The quadratic Boolean equation $\phi = 0$ in n variables is consistent if and only if the matched graph G_ϕ has the Kőnig–Egerváry property.*

Proof. The null edges form a maximum matching in G_ϕ, so G_ϕ has the KE property if and only if there is a vertex cover C in G_ϕ with cardinality n.

Suppose first that G_ϕ has the KE property, and let C be a vertex cover with cardinality n. As every null edge has exactly one endpoint in C, we define the Boolean point $Z = (z_1,\dots,z_n) \in \mathcal{B}^n$ as $z_i = 0$ if and only if $x_i \in C$, and $z_i = 1$ otherwise. Since C is a vertex cover, Z is a solution of the equation $\phi = 0$.

Conversely, let Z be a solution of $\phi = 0$, and let C be the set of vertices x_i for which $z_i = 0$ and $\overline{x}_i$ for which $x_i = 1$. Then, C is a vertex cover of cardinality n, and so G_ϕ has the KE property. ■

That is, we can find a solution to $\phi = 0$ by finding a vertex cover with cardinality n, and taking the complementary indicator vector of the vertex cover.

Example. In the matched graph from the previous example, $\{\overline{x}_1,x_2,\overline{x}_3,x_4\}$ is a vertex cover of cardinality $4 = n$, so $\phi = 0$ is consistent, and in particular, the Boolean point 1010 is a solution. △

29.5.1.2 The Implication Graph

As an alternative to the matched graph G_ϕ, we can also associate with a quadratic DNF ϕ a directed graph D_ϕ called the *implication graph* of ϕ, and again characterise the consistency of $\phi = 0$ in terms of a simple property of this graph.

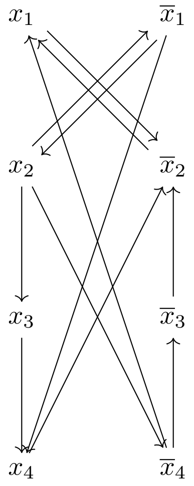
The definition of an implication graph arises from the observation that the relation $\alpha\beta = 0$ is equivalent to the implication $\alpha \Rightarrow \overline{\beta}$, as well as to the implication $\beta \Rightarrow \overline{\alpha}$.

As in the matched graph G_ϕ, the vertices of the implication graph D_ϕ has vertex set $\{x_1,\dots,x_n,\overline{x}_1,\dots,\overline{x}_n\}$. For each quadratic term $\alpha\beta$, we add the arcs $(\alpha,\overline{\beta})$ and $(\beta,\overline{\alpha})$. Either of these arcs is called the *mirror arc* of the other, and the simultaneous presence of these two arcs is called the *mirror property* (MP). Then, for each linear term α, we add the single arc $(\alpha,\overline{\alpha})$.

Example. The graph associated with the DNF

$$\phi = x_1x_2 \vee \overline{x}_1\overline{x}_2 \vee \overline{x}_1\overline{x}_4 \vee x_2\overline{x}_3 \vee x_2x_4 \vee x_3\overline{x}_4$$

is given by:



△

Recall that a *strongly connected component*, or just *strong component*, of a graph $G = (V,E)$ is a maximal subset $C \subseteq V$ of vertices such that every two vertices of C are connected by a path in C in both directions. The strong components of G form a partition of V.

By replacing each strong component of an implication graph D_ϕ by a single vertex, we obtain an acyclic digraph $\hat{D}_\phi$ called the *condensed implication graph* of ϕ. Notice that, because of the mirror propertly, the strong components of D_ϕ come in pairs: if C is a strong component of D_ϕ, then the set $\overline{C}$ of the negation of all literals in C is also a strong component.

Example. The strong components of the implication graph in the previous example are $\{\overline{x}_1, x_2\}$, $\{\overline{x}_2,x_1\}$, and the singletons of every other vertex. △

Lemma 29.5.3. *An assignment of binary values to the vertices of D_ϕ is a solution of $\phi = 0$ if and only if*

(i) x_i and $\overline{x}_i$ receive complementary values;

(ii) no arc (and hence no direct path) connects from a 1-vertex to a 0-vertex.

Theorem 29.5.4. *The quadratic Boolean equation $\phi = 0$ in n variables is consistent if and only if no strongly connected component of the implication graph D_ϕ simultaneously contains a literal and its negation.*

Proof. ■

Example. The strong components found in the previous example are either $\{\overline{x}_1, x_2\}$, $\{\overline{x}_2,x_1\}$, or singletons, none of which include a literal and its negation, so ϕ is consistent (agreeing with the previous result from the matched graph). △

The implication graph not only allows us to determine the consistency of the corresponding quadratic Boolean equation, but also, if it is consistent, to infer further properties of its solutions.

A literal α is *forced to value a* for $a \in \{0,1\}$ if either $\phi = 0$ is inconsistent, or if α takes the value a in all possible solutions.

Theorem 29.5.5. *Suppose that $\phi = 0$ is consistent. Then, the literal α is forced to 0 if and only if there is a directed path from α to $\overline{\alpha}$ in D_ϕ.*

Proof. ■

Example. In the previous example, x_1 is not forced to 0, since there is no directed path from x_1 to $\overline{x}_1$ in D_ϕ. Conversely, x_2 is forced to 0 since there is a directed path x_2,$\overline{x}_4$,$\overline{x}_3$,$\overline{x}_2$ from x_2 to $\overline{x}_2$. △

Theorem 29.5.6. *Let α be a literal not forced to 0 and β be a literal not forced to 1. Then, the relation $\alpha \leq \beta$ holds in all solutions of $\phi = 0$ if and only if there is a directed path from α to β in D_ϕ.*

Proof. ■

Two literals α and β are said to be *twins* if $\alpha = \beta$ in every solution to $\phi = 0$.

Corollary 29.5.6.1. *Suppose that two literals α and β are not forced. Then, they are twins if and only if they are in the same strong component of D_ϕ.*

29.5.1.3 More Relations Between Quadratic Equations and Graphs

Recall that an *independent set* in a graph $G = (V,E)$ is a set of vertices such that no two are adjacent, and a *clique* is a set of vertices such that every pair are adjacent, i.e. induces a complete subgraph. A graph is *bipartite* if its vertex set V can be partitioned into two independent sets $V = L \sqcup R$, and is *split* if V can be partitioned into an independent set and a clique $V = I \sqcup C$.

Given a graph $G = (V,E)$, introduce a variable x_i for each vertex $i \in V$. Then, G is bipartite if and only if the quadratic Boolean equation

$$\bigvee_{ij \in E} (x_i x_j \vee \overline{x}_i \overline{x}_j) = 0$$

is consistent. Also, G is split if and only if the quadratic Boolean equation

$$\bigvee_{ij \in E} \overline{x}_i \overline{x}_j \vee \bigvee_{ij \notin E} x_i x_j \vee \overline{x}_i \overline{x}_j = 0$$

is consistent.

29.6 Horn Functions

An elementary conjunction is a *Horn term* if it contains at most one negated variable. A Horn term is *pure Horn* if it contains precisely one negated variable, and is *positive* otherwise. A DNF is *Horn* if all of its terms are Horn, and a Boolean fnction is a *Horn function* if it can be represented by a Horn DNF.

Lemma 29.6.1. *The consensus of two Horn terms is Horn. Specifically, the consensus of two pure Horn terms is pure Horn, while the consensus of a positive and a pure Horn term is positive.*

Proof. Let xC and $\overline{x}D$ be two Horn terms that have a consensus. Then, D must only contain positive literals, and C can contain at most one negated variable, which cannot belong to A. Hence their consensus CD contains at most one negated variable, i.e. is Horn, and is positive (respectively, pure Horn) if xC is positive (respectively, pure Horn). ■

Lemma 29.6.2. *All prime implicants of a Horn function are Horn.*

Proof. Let f be a Horn function and ϕ be a (pure) Horn DNF representing f. Then, we may compute all prime implicants of f by applying the consensus procedure to ϕ. Thus, all prime implicants of f may be obtained by a sequence of consensus operations, starting with the (pure) Horn terms in ϕ, and by the previous lemma, consensus operations preserve (pure) Horn terms, so every prime implicant must also be (pure) Horn. ■

Theorem 29.6.3. *A Boolean function is Horn if and only if the set of its false points is closed under conjunction.*

Proof. ■

Corollary 29.6.3.1. *A Boolean function f on $\mathcal{B}^n$ is Horn if and only if $f(X \wedge Y) \leq f(X) \vee f(Y)$ for all X,$Y \in \mathcal{B}^n$.*

Proof. Suppose $f(X \wedge Y) \leq f(X) \vee f(Y)$ for all X,$Y \in \mathcal{B}^n$. Then, $f(X) \vee f(Y) = 0$ if and only if X and Y are false points of f, and $f(X \vee Y) \leq f(X) \vee f(Y) = 0$, so $X \vee Y$ is also a false point of f, so $F(f)$ is closed under conjunction, and hence f is Horn. The same argument in reverse proves the reverse implication. ■

29.6.1 Horn Boolean Functions and the Union-Closed Sets Conjecture

Let $(U,\mathcal{F})$ be a set system. The family $\mathcal{F}$ is *union-closed* if for any two sets A,$B \in \mathcal{F}$, we have $A \cup B \in \mathcal{F}$. The following conjecture is known as the *union-closed sets conjecture* or *Frankl's conjecture.*

Conjecture 29.6.1. *Any finite union-closed family $\mathcal{F} \neq \{\varnothing\}$ of finite sets contains an element that belongs to at least of half the sets in the family.*

The family $\mathcal{F}$ is *intersection-closed* if for any two sets A,$B \in \mathcal{F}$, we have $A \cap B \in \mathcal{F}$. Without loss of generality, suppose that every element of the universe appears in at least one set of $\mathcal{F}$. Then, $\mathcal{F}$ is intersection closed if and only if the family of relative complements $\{U \setminus A : A \in \mathcal{F}\}$ is union-closed. So, Frankl's conjecture can be equivalently stated as:

Conjecture 29.6.2. *Any finite intersection-closed family of at least two finites sets contains an element that belongs to at most half of the sets in the family.*

The conjecture admits many other equivalent formulations, in particular, in the language of lattice and graph theory.

In spite of its simple formulation, the conjecture remains open and has been verified only for special classes of sets, lattices, or graphs. Here, we develop a Boolean approach to the conjecture and verify it for submodular functions.

Let F be an intersection-closed family over the universe $U = x_1, \ldots ,x_n$ and let $A \in \mathcal{F}$. We represent A by its characteristic vector c_A, i.e. a binary vector with 1 in the ith coordinate if $x_i \in A$, and 0 otherwise. In doing so, we can interpret $\mathcal{F}$ as a Boolean function f over the variables $x_1, \ldots ,x_n$ whose false points are precisely the elements of $\mathcal{F}$. Then, $\mathcal{F}$ being intersection-closed is equivalent to the false points of f being closed under conjunction, i.e. f is Horn.

We say that a variable x_i *belongs* to a Boolean point X if the ith component of X is 1. Frankl's conjecture can then be restated as follows:

Conjecture 29.6.3. *Any Horn Boolean function f with at least two false points contains a variable that belongs to at most half of the false point of f.*

Given a Horn Boolean function f, we associate to its set of true points $T = T(f)$ a set system $\mathcal{T}$ over the same universe U such that $A \subseteq U$ is an element of $\mathcal{T}$ if and only if the characteristic vector c_A is a true point of f.

Note that a variable belongs to at most half the false points if and only if it belongs to at least half of the true points of the function, which suggests that the relation between $\mathcal{F}$ and $\mathcal{T}$ is similar to the relation between intersection-closed and union-closed families. However, in general, $\mathcal{T}$ is neither intersection-closed nor union-closed.

In the terminology of set systems, an element that appears in at least half the subsets is *abundant*, and an element that appears in at most half the subsets is *rare*. In the terminology of Boolean functions, every variable that is abundant for true points is rare for false points, and vice versa. In the following results, we will frequently switch between the two roles of the same variable. To avoid ambiguities, we will call a variable abundant in false points, or equivalently, rare in true points, *good*. In this terminology, Frankl's theorem can be restated as:

Conjecture 29.6.4. *Any Horn Boolean function f with at least two false points contains good variable.*

We will say that a Horn Boolean function *satisfies Frankl's conjecture* if it satisfies this last characterisation of the conjecture. We will now verify Frankl's conjecture for a certain subclass of Horn functions.

A Boolean function $f(X)$ on $\mathcal{B}^n$ is *co-Horn* if $g(X) := f(\overline{X})$ is Horn. In other words, a function is co-Horn if it admits a DNF representation in which every term contains at most one positive literal.

Previous theorems about Horn functions then transform to results about co-Horn functions as follows:

Theorem 29.6.4. *A Boolean function is co-Horn if and only if the set of its false points is closed under disjunction.*

Corollary 29.6.4.1. *A Boolean function f on $\mathcal{B}^n$ is co-Horn if and only if $f(X \vee Y) \leq f(X) \vee f(Y)$ for all X,$Y \in \mathcal{B}^n$.*

A Boolean function $f(X)$ is *submodular* if $f(X \vee Y) \vee f(X \wedge Y) \leq f(X) \vee f(Y)$.

Theorem 29.6.5. *A Boolean function is submodular if and only if it is both Horn and co-Horn. All prime implicants of a submodular function are either linear or quadratic pure Horn.*

Proof. Since A,$B \leq A \vee B$ holds for any Boolean points A,B, $f(X \vee Y) \vee f(X \wedge Y) \leq f(X) \vee f(Y)$ (f is submodular) if and only if $f(X \wedge Y) \leq f(X) \vee f(Y)$ (f is Horn) and $f(X \vee Y) \leq f(X) \vee f(Y)$ (f is co-Horn).

So, if f is submodular, all prime implicants of f are both Horn and co-Horn, so each contains at most one positive and one negative literal, and is thus is either a single variable x_i or its complement $\overline{x}_i$ (i.e. is linear), or is of the form $x_i\overline{x}_j$ (i.e. is quadratic pure Horn). ■

Lemma 29.6.6. *Let f be a Horn function represented by a Horn DNF D_f. If a variable x_i of f does not appear in D_f negatively, then x_i is a good variable for f.*

Proof. If $f|_{x_i=0}$ does not have true points, then the number of false points of x_i containing x_i is at most equal to the number of false points that do not contain x_i (i.e. if f is identically zero, and is less in any other case), so x_i is a good variable for f.

Conversely, let X be a true point of $f|_{x_i=0}$, i.e. a true point of f with $x_i = 0$, and let t be a term of D_f with $t(X) = 1$. Since x_i does not appear negatively in t and $x_i = 0$, x_i must not appear in t. So, changing the x_i to 1 in X yields a true point of f with $x_i = 1$. This injects the set of true points of $f|_{x_i=0}$ into the set of true points of $f|_{x_i=1}$, so x_i belongs to at least half of the true points of f and is hence good. ■

Theorem 29.6.7. *Submodular Boolean functions satisfy Frankl's conjecture.*

Proof. ■

29.7 Threshold Functions

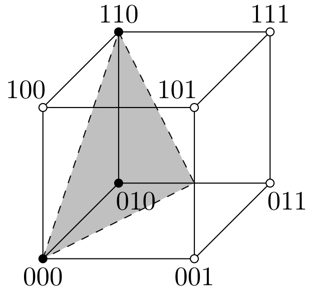
A Boolean function f on $\mathcal{B}^n$ is called a *threshold* or *linearly separable* function if there exist coefficients $w_1,\ldots,w_n \in \mathbb{R}$ called *weights* and a *threshold* value $t \in \mathbb{R}$ such that

$$f(x_1,\ldots,x_n) = 0 \qquad \Longleftrightarrow \qquad \sum_{i=1}^{n} w_i x_i \leq t$$

The hyperplane $\{X \in \mathbb{R}^n : \sum_{i=1}^{n} w_i x_i \leq t\}$ is called a *separator* of f, and the tuple of weights and threshold value $(w_1,\ldots,w_n,t)$ is called a *separating structure* of f. We say that the separator and the separating structure *represent* f.

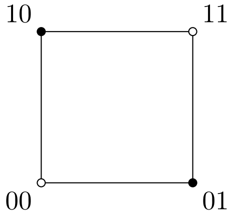
Geometrically, a function is threshold precisely if the set of its true points can be separated from the set of its false points by a hyperplane (where the hyperplane may contain false points).

Example. The function $f(x,y,z) = x\overline{y} \vee z$ is a threshold function with separator $\big\{(x,y,z) \in \mathbb{R}^3 : x-y+2z = 0\big\}$ and structure $(1,-1,2,0)$.



Note that the separator of a threshold function is not unique, and in fact f in this case admits infinitely many separators.

The function $f(x,y) = xy \vee \overline{x}\,\overline{y}$ is not a threshold function:



The convex hulls of $T(f)$ and $F(f)$ intersect, so they cannot be separated by the hyperplane separation theorem. △

Theorem 29.7.1. *Every threshold function has an integral separating structure. That is, a separating structure $(w_1,\ldots,w_n,t)$ with $w_1,\ldots,w_n,t \in \mathbb{Z}$.*

29.7.1 Basic Properties of Threshold Functions

Theorem 29.7.2. *Elementary conjunctions and elementary disjunctions represent threshold functions.*

Proof. Given an elementary conjunction $C_{AB} = \bigwedge_{i\in A} x_i \bigwedge_{j\in B} \overline{x}_j$, the equation $\sum_{i\in A} x_i + \sum_{j\in B}(1 - x_j) = |A| - |B| - 1$ defines a separator for C_{AB}. Similarly, given an elementary disjunction $D_{AB} = \bigvee_{i\in A} x_i \bigvee_{j\in B} \overline{x}_j$, the equation $\sum_{i\in A} x_i + \sum_{j\in B}(1 - x_j) = 0$ defines a separator for D_{AB} ■

Geometrically, elementary conjunctions define a single true point of the hypercube, and elementary disjunctions define single false points, which can both clearly be separated from the opposite kind of points.

Another important property of thresfold functions is that they constitute a class of functions closed under restriction.

Theorem 29.7.3. *If f is a threshold function on $\mathcal{B}^n$ with separating structure $(w_1,\dots,w_n,t)$, then $f_{x_i=1}$ is a threshold function on $\mathcal{B}^{n-1}$ with separating structure $(w_1,\dots,\widehat{w_k}\dots,w_n,t-w_i)$, and $f_{x_i=0}$ is a threshold function on $\mathcal{B}^{n-1}$ with separating structure $(w_1,\dots,\widehat{w_i}\dots,w_n,t)$.*

Proof. Since f is threshold, $f|_{x_i=1}(x_1,\dots,\widehat{x_i},\dots,x_n) = f(x_1,\dots,1,\dots,x_n) = 0$ if and only if

$$\sum_{j\neq i} w_j x_j + w_i \leq t$$

$$\sum_{j\neq i} w_j x_j \leq t - w_i$$

so $f|_{x_i=1}$ is threshold with separating structure $(w_1,\dots,\widehat{w_i}\dots,w_n,t-w_i)$.

Similarly, $f|_{x_i=0}(x_1,\dots,\widehat{x_i},\dots,x_n) = f(x_1,\dots,0,\dots,x_n) = 0$ if and only if

$$\sum_{j\neq i} w_j x_j \leq t$$

so $f|_{x_i=0}$ is threshold with separating structure $(w_1,\dots,\widehat{w_i}\dots,w_n,t)$. ■

Our next observation is that every threshold function is monotone, and hence can be turned into a positive function by "switching" some of its variables. Moreover, the negativity and positivity of each variable is captured in the sign of the corresponding weight.

Theorem 29.7.4. *Every threshold function is monotone. More precisely, if f is a threshold function with separating structure $(w_1,\dots,w_h,t)$, then for each $i\in[n]$,*

(i) If $w_i = 0$, then f does not depend on x_i;

(ii) If x_i does not depend on x_i, then $(x_1,\dots,w_{i-1},0,w_{i+1},\dots,x_n,t)$ is a separating structure of f;

(iii) If $w_i > 0$, then f is positive in x_i;

(iv) If f is positive in x_i and f depends on x_i, then $w_i > 0$;

(v) If $w_i < 0$, then f is negative in x_i;

(vi) If f is negative in x_i and f depends on x_i, then $w_i < 0$;

(vii) If $w_j \geq 0$ for $j = 1,\dots,k$, and $w_j < 0$ for $j = k+1,\dots,n$, then the function

$$g(x_1,\dots,x_n) \coloneqq f(x_1,\dots,x_k,\overline{x}_{k+1},\dots,\overline{x}_n)$$

is a positive threshold function with separating structure

$$\left(w_1,\dots,w_k,-w_{k+1},\dots,-w_n,t-\sum_{j=k+1}^{n} w_j\right)$$

Example. As seen previously, the function $f(x,y,z) = x\overline{y}\vee z$ is a threshold function with separating structure $(1,-1,2,0)$. The associated function $g(x,y,z) = xy\vee z$ with all negations removed is then also a threshold function with separating structure $(1,2,3,2)$. △

We emphasise that a variably may have a non-zero weight in the separating structure of a threshold function even if the function dos not depend on the variable.

Example. The function $f(x,y,z,w) = xy \vee z$ is a threshold function with separating structure (2,4,6,1,5). The variable w is inessential, but has positive weight in this separating structure. △

For three or fewer variables, monotonicity is equivalent to thresholdness. However, this fails in general for functions of more variables.

Example. The functions

$$\begin{aligned} f(x,y,z,w) &= xy \vee zw \\ g(x,y,z,w) &= xy \vee yz \vee zw \\ h(x,y,z,w) &= xy \vee yz \vee zw \vee xw \end{aligned}$$

are positive but not threshold. Up to permutation of their variables, these are the only positive non-threshold functions of four variables. △

Theorem 29.7.5. *If f is a threshold function on $\mathcal{B}^n$ and $(w_1,\ldots,w_n,t)$ is in integral separating structure of f, then f^d is a threshold function with separating structure*

$$\left(w_1,\ldots,w_n,\left(\sum_{i=1}^n w_i\right) - t - 1\right)$$

Furthermore,

(i) If $t \leq \frac{1}{2}\sum_{i=1}^n w_i - 1$, then f is dual-major;

(ii) If $t \geq \frac{1}{2}\sum_{i=1}^n w_i - 1$, then f is dual-minor;

Proof. Let $t' = \sum_{i=1}^n w_i - t - 1$. Since the threshold t and weights $w_1,\ldots,w_n$ are integra, the following equivalences hold for all $X \in \mathcal{B}^n$:

$$\begin{aligned} f^d(X) = 0 \quad &\iff \quad f(\overline{X}) = 1 \\ &\iff \quad \sum_{i=1}^n w_i(1-x_i) > t \\ &\iff \quad \sum_{i=1}^n w_i x_i \leq t' \end{aligned}$$

So f^d is threshold with separating structure $(w_1,\ldots,w_n,(\sum_{i=1}^n w_i) - t - 1)$.

The last two parts follow from the observation that $f^d \leq f$ if $t \leq t'$ and $f \leq f^d$ if $t' \leq t$. ■

Example. The function $f(x,y,z,w) = xy \vee xz \vee xw \vee yxw$ admits the separating structure (4,2,2,2,5), so the dual f^d is threshold with separating structure $(4,2,2,2,4+2+2+2-5-1) = (4,2,2,2,4)$. Since the new threshold is smaller, f^d is dual-minor.

However, another separating structure for f is (2,1,1,1,2), which yields the dual separating structure $(2,1,1,1,2+1+1+1-2-1) = (2,1,1,1,2)$, so f is in fact self-dual. △

Theorem 29.7.6. *A function $f(x_1,\ldots,x_n)$ is a threshold function if and only if its self-dual extension $f^{\mathrm{SD}}(x_1,\ldots,x_{n+1}) = f\overline{x}_{n+1} \vee f^d x_{n+1}$ is a threshold function.*

Proof. Suppose that f is a threshold function with integral separating structure $(w_1,\ldots,w_n,t)$. Then, by the previous theorem,

$$\left(w_1,\ldots,w_n,2t+1-\sum_{i=1}^n w_i,t\right)$$

is a separating structure for f^{SD}. Conversely, if f^{SD} is a threshold function with separating structure $(w_1,\ldots,w_{n+1},t)$, then $(x_1,\ldots,x_n,t)$ is a separating structure for f. ■

29.7.2 Characterisation of Threshold Functions

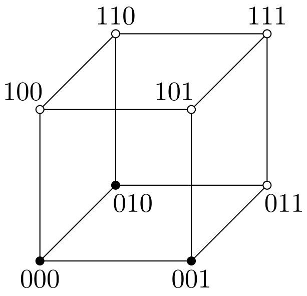
The first characterisation is a simple linear programming formulation which provides a useful computational tool for the recognition of threshold functions. For the sake of simplicity, we only state it for positive functions: since every threshold function is monotone, this restriction does not entail any essential loss of generality.

Theorem 29.7.7. *A positive Boolean function f with maximal false point $X^1,X^2,\ldots,X^p$ and minimal true points $Y^1,Y^2,\ldots,Y^m$ is a threshold function if and only if the system of inequalities*

$$\begin{aligned} \sum_{i=1}^{n} w_i x_i^j &\le t && j=1,\ldots,p \\ \sum_{i=1}^{n} w_i y_i^j &\ge t+1 && j=1,\ldots,m \\ w_i &\ge 0 && j=1,\ldots,m \end{aligned}$$

has a solution $(w_1,\ldots,w_n,t)$. When this is the case, every solution of the system is a separating structure for f.

Example. Let $f(x,y,z) = x \vee yz$.



The maximal false points are 010 and 001, and the minimal true points are 100 and 011, so the system of inequalities is given by

$$\begin{aligned} w_2 &\le t && \text{(010 false)} \\ w_3 &\le t && \text{(001 false)} \\ w_1 &\le t+1 && \text{(100 true)} \\ w_2 + w_3 &\ge t+1 && \text{(011 true)} \\ w_i &\ge 0 \end{aligned}$$

This system has solution $(w_1,w_2,w_3,t) = (2,1,1,1)$, so f is a threshold function with separating structure $(2,1,1,1)$. △

Let $k \ge 2$ be a natural number. A Boolean function f on $\mathcal{B}$ is *k-summable* if for some $r \in \{2,3,\ldots,k\}$, there exist r-many not-necessarily distinct false points of f, say $X^1,\ldots,X^r$ and r-many not-necessarily distinct true points $Y^1,Y^2,\ldots,Y^r$ such that

$$\sum_{i=1}^{r} X^i = \sum_{i=1}^{r} Y^i$$

A function is *k-asummable* if it is not k-summable, and is *asummable* if it is k-asummable for all $k \ge 2$.

Example. The function $f(x_1,x_2) = x_1x_2 \vee \overline{x}_1\overline{x}_2$ has true points 00 and 11, and false points 01 and 10. Then, f is 2-summable since

$$00 + 11 = 01 + 10$$

△

Theorem 29.7.8. *A Boolean function is a threshold function if and only if it is asummable.*

Proof. ■

29.7.3 Threshold Functions and Chow Parameters

Recall that the Chow parameters of a Boolean function f on $\mathcal{B}^n$ are the $n+1$ integers $(\omega_1,\ldots,\omega_n,\omega)$ where $\omega = \omega(f)$ is the number of true points of f and ω_i is the number of true points $X^* = (x_1^*,\ldots,x_n^*)$ of f with $x_i^* = 1$.

Note that

$$(\omega_1,\ldots,\omega_n) = \sum_{j=1}^{\omega} Y^j$$

where $Y^1,\ldots,Y^\omega$ are the true points of f.

A Boolean function f is a *Chow function* if no other function has the same Chow parameters as f.

Example. The function $f(x_1,x_2) = x_1x_2 \vee \overline{x}_1\overline{x}_2$ is not a Chow function since it has the same Chow parameters as $g(x_1,x_2) = x_1\overline{x}_2 \vee \overline{x}_1x_2$, namely (1,1,2). △

Theorem 29.7.9. *Every threshold function is a Chow function.*

Proof. Let f be a threshold function on $\mathcal{B}^n$, and let g be a function on $\mathcal{B}^n$ with the same Chow parameters. Let $Y^1,\ldots,Y^\omega$ be the true points of f, and $X^1,\ldots,X^k,Y^{k+1},\ldots,Y^\omega$ be the true points of g, where the X^i are false points of f.

Since f and g have the same Chow parameters,

$$\sum_{j=1}^{\omega} Y^j = \sum_{j=1}^{k} X^j + \sum_{j=k+1}^{\omega} Y^j$$

or equivalently,

$$\sum_{j=1}^{k} Y^j = \sum_{j=1}^{k} X^j$$

Now, if $k \geq 1$, this contradicts the asummability of f. So $k = 0$, and f and g have the same set of true points, i.e. $f = g$. ■

Note that all the points occuring in this final sum are distinct. This motivates the following definition.

A Boolean function f is *weakly asummable* if for all $k \geq 1$, thexe do not exist k-many *distinct* false points $X^1,\ldots,X^k$ and k-many *distinct* true points $Y^1,\ldots,Y^k$ such that

$$\sum_{i=1}^{r} X^i = \sum_{i=1}^{r} Y^i$$

Clearly, every asummable (and hence threshold) function is weakly asummable. Moreover, the previous proof actually establishes that every weakly asummable function is a Chow function. In fact, the converse implication folds as well.

Theorem 29.7.10. *A Boolean function is weakly asummable if and only if it is a Chow function.*

Proof. The forward implication is shown above. For the reverse implication, let $X^1, \ldots, X^q$ denote the false points of a function f, and let $Y^1, \ldots, Y^p$ denote its true points.

If f is not weakly assumable, then without loss of generality by reordering the points, we have

$$\sum_{i=1}^{k} X^i = \sum_{j=1}^{k} Y^j$$

for some $k \geq 1$. Let g be the Boolean function whose true points are precisely $X^1, \ldots X^k, Y^{k+1}, \ldots, Y^p$. Then, f and g have distinct true points and are hence distinct functions, but f and g share the same Chow parameters, and hence f is not a Chow function. ■

It is natural to expect some sort of relationship between the Chow parameters of a threshold function and the separating structure defining the function, since both types of coefficients provide a "measure" of the "influence" of each variable on the function. This relationship is most natural expressed in terms of the so-called *modified Chow parameters* of the function.

The *modified Chow parameters* of a Boolean function $f(x_1, \ldots, x_n)$ are the $n+1$ numbers $(\pi_1, \ldots, \pi_n, \pi)$ defined as $\pi = \omega - 2^{n-1}$ and $\pi_k = 2\omega_k - \omega$, where $(\omega_1, \ldots, \omega_n, \omega)$ are the Chow parameters of f.

Since there is a bijection between Chow parameters and modified Chow parameters, every threshold function is uniquely determined by its modified Chow parameters, or by Chow parameters, or by any of its separating structures.

Theorem 29.7.11. *If f is a Boolean function with modified Chow parameters $(\pi_1, \ldots, \pi_n, \pi)$, then for all $i \in [n]$,*

(*i*) *If f is positive in x_i and f depends on x_i, then $\pi_i > 0$;*

(*ii*) *If f is negative in x_i and f depends on x_i, then $\pi_i < 0$;*

(*iii*) *If f does not depend on x_i, then $\pi_i = 0$;*

(*iv*) *The modified Chow parameters of f^d are $(\pi_1, \ldots, \pi_n, -\pi)$;*

(*v*) *If f is dual-major ($f^d \leq f$) then $\pi \geq 0$;*

(*vi*) *If f is dual-minor ($f \leq f^d$) then $\pi \leq 0$;*

Proof. ■

Theorem 29.7.12. *If $f(x_1, \ldots, x_n)$ is a threshold Boolea function give by the integral separating structure $(\omega_1, \ldots, \omega_n, t)$, then the number of true points of f can be computed in $O(nt)$ arithmetic operations.*

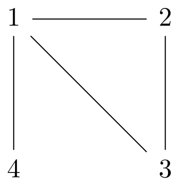
Proof. ■

29.7.4 Threshold Graphs

In this section, we specialise some of these previous results to the case of graphic (i.e. purely quadratic and positive) functions. Recall that such a function $f(x_1, \ldots, x_n) = \bigvee_{ij \in E} x_i x_j$ can be identified with an undirected graph $G_f = \big([n], E\big)$. Conversely, if $G = (V, E)$ is an arbitrary undirected graph, we define its corresponding stability function f_G by the expression $\bigvee_{ij \in E} x_i x_j$.

A graph G is a *threshold graph* if its stability function f_G is threshold, and we say that $(\omega_1, \ldots, \omega_n, t)$ is a *separating structure* of G if it is a separating structure of f_G.

Example. The function $f(x_1,x_2,x_3,x_4) = x_1x_2 \vee x_1x_3 \vee x_1x_4 \vee x_2x_3$ is graphic with associated graph:



This function is threshold with separating structure, say (3,2,2,1,3). So G_f is a threshold graph with separating structure (3,2,2,1,3). △

Is there an easy way to determine which graphic functions are threshold, or equivalently, which graphs are threshold?

Recall that an independent set in a graph is a set of vertices of which no two are adjacent.

Theorem 29.7.13. *A graph $G = (V,E)$ is a threshold graph if and only if there exists a structure $(\omega_1,\ldots,\omega_n,t)$ such that for every subset S of vertices, S is an independent set if and only if*

$$\sum_{i\in S} w_i \leq t$$

We also recall that for a graph $G = (V,E)$ and a vertex $i \in V$, the *neighbourhood* $N(i)$ of i is the set of vertices adjacent to i. We say that i is *isolated* if $N(i) = \varnothing$, and that i is dominating if $N(i) = V \setminus \{i\}$. Note that isolated vertices of G correspond to inessential variables of f_G, since they don't appear in any term.

Theorem 29.7.14. *A graph G is threshold if and only if it is C_4, P_4, and $2K_2$-free.*

Proof. ■

Example. The graph above is threshold since it does not contain C_4, P_4, nor $2K_2$ as induced subgraphs. △

Theorem 29.7.15. *A graphic function $f(x_1,\ldots,x_n)$ is threshold if and only if there is a permutation $\sigma : [n] \to [n]$ such that for every $i \in [n]$, σ_i is either isolated or dominating in the subgraph of G_f induced by $\{i,\ldots,n\}$.*

29.8 Read-Once Functions

A Boolean function f is *read-once* if it can be represented by a Boolean expression over $\{\overline{x}, \wedge, \vee\}$ such that every variable appears exactly once. Such an expression is called a *read-once expression* for f.

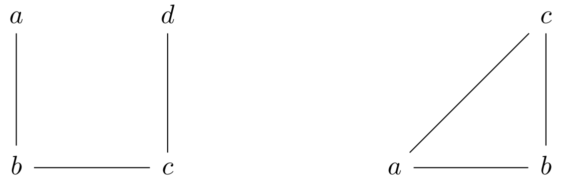
Example. The function $f_0(a,b,c,w,x,y,z) = ay \vee cxy \vee bw \vee bz$ is a read-once function since it can be factored into the expression $f_0 = y(a \vee cx) \vee b(w \vee z)$, where every variable appears exactly once. △

Note that read-once functions are necessarily monotone since every variable appears either in its positive or negative form in the read-once expression, an thus cannot contribute in conflicting directions. However, we will make a stronger assumption that a read-once function is positive, simply by renaming any negative variables $\overline{x}_i$ as new positive variables x_i'.

Consider the two simple functions $f_1 = ab \vee bc \vee cd$ and $f_2 = ab \vee bc \vee ac$. Neither of these functions are read-once. As we will see, these illustrate the two types of forbidden functions that characterise read-once functions.

Let f be a positive Boolean function over the variables $x_1,\dots,x_n$. The *co-occurence graph* $G(f)$ of f is the undirected graph with vertex set $V = \{x_1,\dots,x_n\}$ and edge set defined by $(x_i,x_j) \in E$ if and only if x_i and x_y occur together at least once in some prime implicant of f.

Example. The co-occurence graphs of f_1 and f_2 are:



△

A Boolean function is *normal* if every clique of its co-occurence graph is contained in the set of variables of a prime implicant of f.

Example. f_2 is not normal, since $\{a,b,c\}$ is a clique, and the prime implicants of are ab, bc, and ac, which all contain only 2 variables. △

Theorem 29.8.1. *A positive Boolean function f is read-once if and only if it is normal and its co-occurence graph $G(f)$ is P_4-free.*

Before proving this theorem, we review a few properties of the dual of a Boolean function and prove an important result on positive Boolean functions.

29.8.1 Dual Implicants

Recall that the dual f^d of a Boolean function f is the function defined by

$$f^d(X) = \overline{f(\overline{X})}$$

An expression for f^d can be obtained from any expression for f by interchanging the operators $\wedge$ and $\vee$ as well as the constants 0 and 1. In particular, given a DNF expression for f, this exchange yields a CNF expression for f^d. This shows that the dual of a read-once function is also read-once.

Let $\mathcal{P}$ be the set of prime implicants of a Boolean function f over the variables $x_1,\dots,x_n$, and let $\mathcal{D}$ be the collection of prime implicants for the dual function f^d. We assume throughout that all of the variables for f (and hence for f^d) are essential.

We use the term "dual (prime) implicant" of f to mean a (prime) implicant of f^d. For positive functions, the prime implicants of f correspond precisely to the set of minimal true points $\min T(f)$, and the dual prime implicants of f correspond precisely to the set of maximal false points $\max F(f)$.

We have also seen that the implicants and dual implicants of a Boolean function f, viewed as sets of literals, have pairwise non-empty intersections. In particular, this holds for the prime implicants and the dual prime implicants, and moreover, the prime implicants and the dual prime implicants are minimal with this property. That is, for every proper subsets S of a dual prime implicant of f, there is a prime implicant P such that $P \cap S = \varnothing$.

In terms of hypergraph theory, the prime implicants $\mathcal{P}$ form a clutter (i.e. a collection of sets, or hyperedges, such that no set contains another set), as does the collection of dual prime implicants $\mathcal{D}$.

Theorem 29.8.2. *Let f and g be positive Boolean functions over the varibles $x_1,\dots,x_n$, and let $\mathcal{P}$ and $\mathcal{D}$ be the collections of prime implicants of f and g, respectively. Then, the following are equivalent:*

(*i*) *$g = f^d$;*

(ii) *For every partition of $\{x_1,\dots,x_n\}$ into two disjoint sets A and $\overline{A}$, there is either a member of $\mathcal{P}$ contained in A, or a member of $\mathcal{D}$ containied in $\overline{A}$, but not both;*

(iii) *$\mathcal{D}$ is precisely the family of minimal transversals of $\mathcal{P}$;*

(iv) *$\mathcal{P}$ is precisely the family of minimal transversals of $\mathcal{D}$;*

(v) *For all $P \in \mathcal{P}$ and $D \in \mathcal{D}$, we have $P \cap D \neq \varnothing$, and for every set $B \subset \{x_1,\dots,x_n\}$ of variables, there exists $D \in \mathcal{D}$ such that $D \subseteq B$ if and only if $P \cap B \neq \varnothing$ fot every $P \in \mathcal{P}$.*

Theorem 29.8.3. *A set of variables B is a dual implicant of the function f if and only if $P \cap B \neq \varnothing$ for all prime implicants P of f.*

A subset T of the variables is called a *dual sub-implicant* of f if T is a subset of a dual prime implicant of f. That is, there exists a prime implicant D of f^d such that $T \subseteq D$. A *proper* dual sub-implicant is a non-empty proper subset of a dual prime implicant.

Example. Let $f = x_1x_2 \vee x_2x_3x_4 \vee x_4x_5$. Its dual is $f^d = x_1x_3x_5 \vee x_1x_4 \vee x_2x_4 \vee x_2x_5$. The proper dual sub-implicants of f are the pairs $\{x_1,x_3\}$, $\{x_3,x_5\}$, $\{x_1,x_5\}$, and the singletons of each variable. △

Note that if T is a proper dual sub-implicant o f, then there exists a prime implicant $P \in \mathcal{P}$ such that $T \cap P = \varnothing$.

29.9 Characterising Read-Once Functions

A positive Boolean expression over the operation of conjunction and disjunction may be represented as a rooted parse tree whose leaves are labeled by the variables $\{x_1,\dots,x_n\}$, and whose internal nodes are labeled by the Boolean operations $\vee$ and $\wedge$. The parse tree represents the computation of the associated Boolean function according to the given expression, and each internal node is the root of a subtree corresponding to a part of the expression. If the expression is read-once, then each variable appears on exactly one leaf of the tree, and there is a unique path from the root to the variable.

Lemma 29.9.1. *Let T be a parse tree for a read-once expression for a positive Boolean function f over the variables $x_1,\dots,x_n$. Then (x_i,x_j) is an edge in the co-occurence graph $G(f)$ if and only if the lowest common ancestor of x_i and x_j in the tree T is labeled by a conjunction $\wedge$.*

Proof. ■

Theorem 29.9.2. *Let f be a positive Boolean function over the variables $x_1,\dots,x_n$. Then, the following are equivalent:*

(i) *f is a read-once function;*

(ii) *The co-occurence graphs $G(f)$ and $G(f^d)$ are complementary, i.e. $\overline{G(f)} = G(f^d)$;*

(iii) *The co-occurence graphs $G(f)$ and $G(f^d)$ have no edges in common, i.e. $E\big(G(f)\big) \cap E\big(G(f^d)\big) = \varnothing$;*

(iv) *For all $p \in \mathcal{P}$ and $D \in \mathcal{D}$, $|P \cap D| = 1$;*

(v) *f is normal and the co-occurence graph $G(f)$ is P_4-free.*

Proof. ■

29.10 Linear Read-Once Functions

29.10.1 Specifying Sets and Specification Number

29.10.2 Essential Points

29.10.3 The Number of Essential Points and the Number of Extremal Points

29.10.4 Positive Functions and the Number of Extremal Points

29.10.4.1 A Property of Extremal Points

29.10.4.2 Canalysing Functions

29.10.4.3 Non-Canalysing Functions with Canalysing Restrictions

29.10.4.4 Non-Canalysing Functions Containing Non-Canalysing Restrictions

29.10.5 Chow and Read-Once Functions

29.10.6 Threshold Functions and Specification Number

29.10.6.1 Minimal Non-LRO Functions

29.10.6.2 Non-LRO Threshold Functions with Minimum Specification Number

29.11 Partially-Defined Boolean Functions and Logical Analysis of Data

29.11.1 Extensions of PDBFs

29.11.2 Patterns and Theories of PDBFs

29.11.3 Roles of Theories and Co-Theories

29.11.4 Decision Trees and PDBFs

29.12 Pseudo-Boolean Functions

29.12.1 Pseudo-Boolean Optimisation

29.12.2 Posiform Transformations and Conflict Graphs

29.12.2.1 The Struction

Chapter 30

Computability Theory

"The purpose of abstraction is not to be vague, but to create a new semantic level in which one can be absolutely precise."

— Edsger Dijkstra

Chapter 31

Program Verification

“Why repeat the old errors, if there are so many new errors to commit?”
— Bertrand Russell, *Unpopular Essays*

Chapter 32

Digital Signal Processing

Chapter 33

Linear Algebra

> *"It is my experience that proofs involving matrices can be shortened by* 50% *if one throws the matrices out."*
>
> — Emil Artin, *Geometric Algebra*

Linear algebra is the study of systems of linear equations, vector spaces and linear maps. Linear algebra is essential to almost every area of mathematics, due to how abstractly vector spaces are defined. If your problem/model/whatever has some notion of scaling things, and adding two things together to get a third, linear algebra theorems probably apply to it. In this document, we focus mainly on matrices and vector spaces.

Notes on formatting conventions:

- Scalars are written in lowercase italics, c, or using greek letters.
- Vectors are written in lowercase bold, $\mathbf{v}$, or rarely overlined, $\overrightarrow{v}$, where more contrast or clarity is required.
- Matrices are written in uppercase bold, $\mathbf{A}$.

  Note: transformations represented by matrices may be written in just italics, as functions often are: i.e. $s(\mathbf{v}) = \mathbf{A}\mathbf{v}$.

33.1 Vectors

33.1.1 Mathematical Interpretation

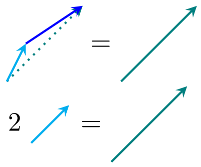
In physics, vectors are often treated as *arrows* pointing in space – some kind of quantity which has a *magnitude* and a *direction.* As long as the length and direction of a vector are the same, it's the same vector, no matter where it is. For example, you might model the velocity of an object as a vector, and consider the velocity as staying the same if the length and direction remain constant.

Sets of vectors that all lie within a plane are two-dimensional, and those in the space we live in are three-dimensional. Picturing an arbitrary n-dimensional vector in this context can be rather tricky, due to the limitations of our reality.

On the other hand, in computer science, vectors are ordered lists of numbers, or *tuples.* For example, you might model the population of two species of animals, say, foxes and rabbits, in a given area with a pair of numbers, the first representing the number of foxes, and the second representing the number of rabbits. Note that order matters; two vectors are not equal if the numbers are swapped around.

In this context, we'd be modelling the populations as a two-dimensional vector. What makes the vector two-dimensional is that the list has two elements within it.

In maths, we are much more general. A vector is anything where we have some kind of notion of adding two objects, our *vectors*, and multiplying those vectors by a number, called a *scalar*. A *vector space* is just a set whose elements are vectors.



$$\overrightarrow{v} + \overrightarrow{w}$$
$$2\overrightarrow{v}$$

$$\begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} a+c \\ b+d \end{bmatrix}$$
$$2\begin{bmatrix} e \\ f \end{bmatrix} = \begin{bmatrix} 2e \\ 2f \end{bmatrix}$$

But in a sense, what makes a vector space, a vector space, are these fundamental operations, independent of how the vectors themselves are represented – it doesn't matter whether you think about vectors as fundamentally being arrows which happen to have a nice numerical representation as lists; or as lists of numbers which happen to have a nice visual representation as arrows. The usefulness of linear algebra is less to do with specific representations of vectors, and more to do with the ability to translate between and equate these different views.

This very general view encompasses both the arrows and ordered lists, and more, but in exchange, is very abstract, and can be possibly more difficult to pick up.

For now, we will first focus on a geometric interpretation of vectors, before moving on to more abstract vector spaces.

When we say a vector, for now, picture an arrow within a coordinate system, with the tail rooted at the origin. Note that this is somewhat distinct from the physics viewpoint discussed above, as vectors in that sense aren't tied to a specific coordinate system, and are free to move about.

This specific view is very helpful as we can then use matrix algebra in our calculations, and changing to a computer science tuple view is just as easy as reading off the coordinates of the head of the vector.

33.1.2 Basis Vectors, Span & Linear Independence

When we write a vector as a pair of coordinates, say,

$$\mathbf{v} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

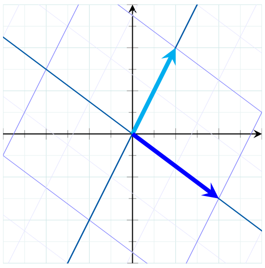
you can think of these coordinates as scalars scaling two vectors.

In the Cartesian coordinate system, there are two very special vectors we often use; the vector pointing to the right with length 1, denoted $\hat{\mathbf{\imath}}$, and the vector pointing up with length 1, denoted $\hat{\mathbf{\jmath}}$.

Thinking of the coordinates as scalars, we scale $\hat{\mathbf{\imath}}$ by 2, and $\hat{\mathbf{\jmath}}$ by 3, before adding them together to give $\mathbf{v}$, so $\mathbf{v}$ is the sum of two scaled vectors. Though this is an extremely simple example, this concept of adding two scaled vectors is worth keeping in mind, as it will soon come up, a lot. Any time we scale up vectors and add them together, it's called a *linear combination* of the vectors.

Together, $\hat{\mathbf{\imath}}$ and $\hat{\mathbf{\jmath}}$ have a special name. They are the *basis vectors* of the Cartesian coordinate system. Specifically, we call them the *canonical* or *standard* basis vectors.

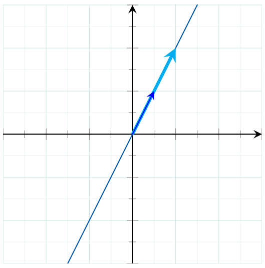
Informally, what it means to form a basis is that, when you use coordinates as scalars, the basis vectors are what the scalars act on. But this brings up a pretty interesting question. What if we picked other basis vectors?



Think about all the arrows you can get by picking two scalars, using them to scale these two arrows, then adding the results together. For these particular two arrows, the answer is that you can reach every possible two-dimensional arrow. You can see this by the fact that the transformed grid covers all of the 2D plane.

A new pair of basis vectors like this also gives us a valid way to translate between pairs of numbers and arrows in the plane. But notice that this translation is different from the canonical basis. [1,1] in this new basis certainly points to a different place than in the canonical basis. We will go into more detail later on, on how coordinates in different bases are related, but for now, just appreciate that any time we describe vectors numerically, it depends on some implicit arbitrary choice of basis vectors.

Now, if we allow the scalars to vary through all possible pairs of values, considering the linear combination given by each pair, we have three possible situations. For most pairs of vectors, we can reach every point in the plane, like in the example above. But if your two vectors line up and are parallel, then the resulting vector is also forced onto the line passing through the origin, parallel to the vectors.



Compared to the previous case, here, the transformed grid is compressed onto a single line.

Additionally, if both vectors are the zero vector, you're just stuck on the origin. The set of all possible vectors you can reach with a linear combination of a set of vectors is the *span* of the vectors. So, we can say that the span of the first pair of vectors above is the entire Cartesian plane (or equivalently, we say that the Cartesian plane is *spanned by* those two vectors, or that those two vectors form a *spanning set* of the Cartesian plane), while the span of the second pair of vectors is just a line, and the span of two zero vectors is just the single point on the origin.

In this second case, we note that one of the vectors is somewhat redundant. We can still access the full line, just using one of the vectors. In this case, we say that the vectors are *linearly dependent* – one of the vectors in the set can be expressed as a linear combination of the others, since it already lies within the span of the others.

Conversely, if each new vector adds a new dimension to the span, we say that the vectors are *linearly independent.*

More specifically, we say that a set of vectors, $\mathbf{v}_1$,$\mathbf{v}_2$,$\mathbf{v}_3, \ldots$,$\mathbf{v}_n$ are linearly independent if the equation,

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n = \mathbf{0}$$

holds only if $a_1 = a_2 = \cdots = a_n = 0$. In other words, if you can find a way to add up scaled versions of your vectors to get back to the origin, they are not linearly independent.

Now, we can more formally define a basis of a vector space as a set of linearly independent vectors that span the space, and the *dimension* of a vector space is the number of vectors in its basis.

If a linearly independent set of vectors span a space, then every vector in that space can be written as a unique linear combination of those vectors. If this spanning set is not linearly independent, then this linear combination representation of vectors will not be unique (which is why such a set is not usable as a basis – coordinates are not unique). Any two bases of the same vector space contain the same number of vectors. (Take a moment to think about why these properties are true, given the definitions we have just seen.)

33.2 Linear Transformations

33.2.1 Transformations as Matrix-Vector Multiplication

As the name suggests, in linear algebra, we only consider transformations that are *linear*.

Let V and W be vector spaces over a field, K (we will discuss what a field is later). A function $f : V \to W$, is linear if for any two vectors, $\mathbf{u}$,$\mathbf{v} \in V$, and any scalar, $c \in K$,

- $f(\mathbf{u} + \mathbf{v}) = f(\mathbf{u}) + f(\mathbf{v})$ (*additivity* or *operation of vector addition*)
- $f(c\mathbf{u}) = cf(\mathbf{u})$ (*degree 1 homogeneity* or *operation of scalar multiplication*)

In other words, it does not matter whether the linear map is applied before or after the operations of vector addition and scalar multiplication. In particular, linear maps preserves linear combinations. This means that a linear transformation is really a vector space homomorphism – a map that is compatible with and preserves the vector space structure.

Geometrically, a transformation is linear if the origin is fixed in place, and all lines remain lines under the transformation.

Although this is a rather restrictive condition, there are still a vast range of linear transformations. So, how do we represent these transformations numerically? Given a pair of numbers – a point, a coordinate – how do we find the image of that pair under any given transformation?

Looking back at the definition of a linear transformation, it doesn't matter whether we apply the map before or after the operations of vector addition and scalar multiplication, and, as discussed earlier, every vector can be seen as scaling and adding up the basis vectors – so, if we keep track of where the basis vectors are mapped under the transformation, everything else immediately follows on.

For example, if we know that,

$$\hat{\mathbf{i}} \mapsto \begin{bmatrix} 1 \\ -3 \end{bmatrix} \qquad \hat{\mathbf{j}} \mapsto \begin{bmatrix} -2 \\ 4 \end{bmatrix}$$

then we can easily tell where any arbitrary vector,

$$\mathbf{v} = \begin{bmatrix} x \\ y \end{bmatrix}$$

is mapped, by using the linear properties of these maps and breaking it down into its constituent parts, $\mathbf{v} = x\hat{\mathbf{i}} = y\hat{\mathbf{j}}$, so,

$$\begin{bmatrix} x \\ y \end{bmatrix} \mapsto x \begin{bmatrix} 1 \\ -3 \end{bmatrix} + y \begin{bmatrix} -2 \\ 4 \end{bmatrix} = \begin{bmatrix} 1x - 2y \\ -3x + 4y \end{bmatrix}$$

In doing so, we see that every two-dimensional linear map is completely determined by just 4 numbers – the coordinates of the image of $\hat{\mathbf{\imath}}$ and $\hat{\mathbf{\jmath}}$. Or more generally, the coordinates of the image of the basis vectors of the relevant space. But sticking with $\hat{\mathbf{\imath}}$ and $\hat{\mathbf{\jmath}}$ for now, we often like to package these coordinates into an array of numbers – a *matrix.*

We do this in such a way that the first column contains the coordinates of where $\hat{\mathbf{\imath}}$ lands, and the second, the coordinates for $\hat{\mathbf{\jmath}}$.

$$\begin{bmatrix} 1 & -2 \\ -3 & 4 \end{bmatrix}$$
$$\underbrace{}_{\hat{\mathbf{\imath}}} \quad \underbrace{}_{\hat{\mathbf{\jmath}}}$$

If you have the matrix for some linear transformation, and you want to know the image of any given vector, you take the coordinates of that vector, multiply them by the respective columns of the matrix, and sum the results. In other words, we are adding up the scaled versions of the new basis vectors.

For an arbitrary matrix and vector,

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}, \begin{bmatrix} x \\ y \end{bmatrix}$$

the image of the vector is given by,

$$\begin{bmatrix} x \\ y \end{bmatrix} \mapsto x \begin{bmatrix} a \\ c \end{bmatrix} + y \begin{bmatrix} b \\ d \end{bmatrix} = \begin{bmatrix} ax+by \\ cx+dy \end{bmatrix}$$

Since the matrix really represents a linear map – a kind of function – let's write it to the left of the vector like we normally do with functions, and give the vector as the function variable.

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \left(\begin{bmatrix} x \\ y \end{bmatrix} \right) = \begin{bmatrix} ax+by \\ cx+dy \end{bmatrix}$$

But the brackets are somewhat clumsy, so we often drop them from this expression, and read the function as multiplication,

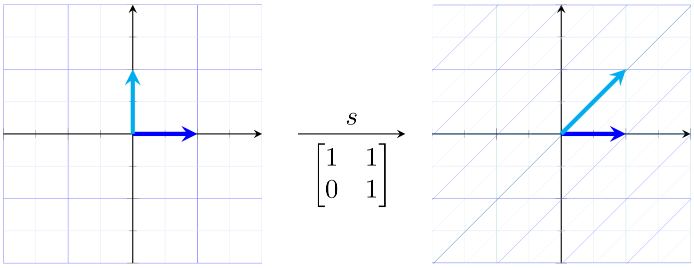
$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax+by \\ cx+dy \end{bmatrix}$$

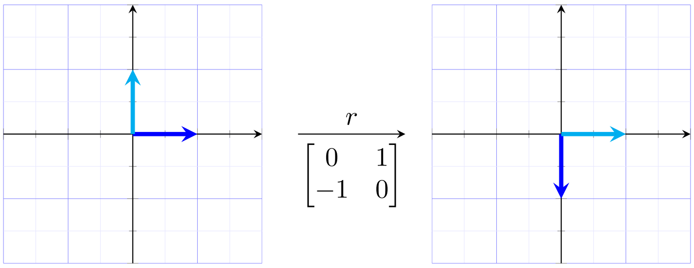
and, we've just discovered matrix-vector multiplication. If you've ever wondered why matrix-vector multiplication is what it is, this is why: it stems from linear transformations being applied to vectors.

33.2.2 Composition as Matrix-Matrix Multiplication

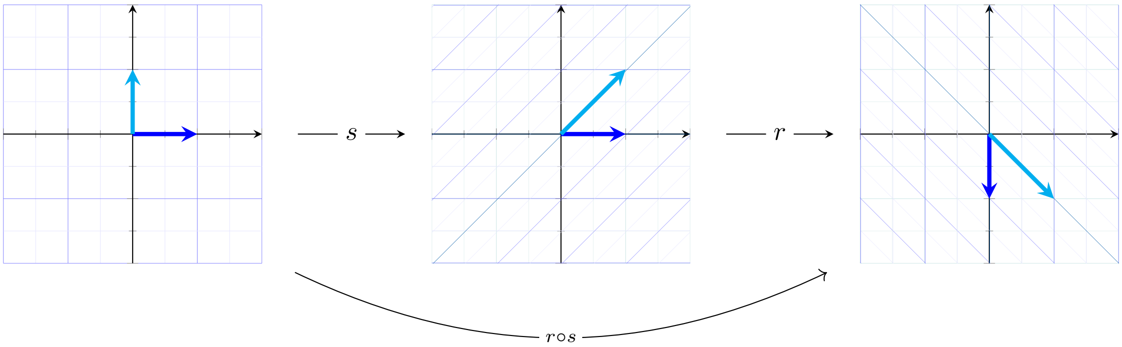
Now, we often don't study transformations in isolation: what happens if we apply two transformations to a vector, one after another?

For example, consider the transformations given by shearing parallel to the horizontal axis, and rotating by 90° clockwise.





Because of linearity, the overall effect of applying the shear then rotation is another linear transformation, distinct from both the shear and rotation alone. The new transformation is the *composition* of the two original transformations.



(We read right to left for composition. This notation stems from function notation; $(r \circ s)(\hat{\mathbf{\imath}}) = r(s(\hat{\mathbf{\imath}}))$, so we apply s first.)

Now, being a linear transformation, this composition also has a matrix representation. Above, we see that,

$$\hat{\mathbf{\imath}} \mapsto \begin{bmatrix}0\\-1\end{bmatrix} \qquad \hat{\mathbf{\jmath}} \mapsto \begin{bmatrix}1\\-1\end{bmatrix}$$

so the composition matrix is given by,

$$\begin{bmatrix}0 & 1\\-1 & -1\end{bmatrix}$$

This matrix gives the effect of shearing, then rotating, in a single transformation – one action, instead of two successive ones.

We can otherwise write the composition out in terms of the original transformations by multiplying a vector on the left by the shear, to give the image under the shear, then multiplying again by the rotation, to apply it after.

$$\underbrace{\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}}_{\text{Rotation}} \left(\underbrace{\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}}_{\text{Shear}} \left(\begin{bmatrix} x \\ y \end{bmatrix} \right) \right)$$

Given our definition of matrix-vector multiplication, this is exactly what it means to apply linear transformations as matrices to a vector.

But, given that this pair of transformations has the same overall effect as the composition matrix on any vector, it seems sensible to write

$$\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix}$$

More generally, two arbitrary transformation matrices will give another transformation matrix. You probably know how matrix multiplication is defined, but put that knowledge aside for a second, and we will rederive that definition.

$$\underbrace{\begin{bmatrix} a & b \\ c & d \end{bmatrix}}_{M_2} \underbrace{\begin{bmatrix} A & B \\ C & D \end{bmatrix}}_{M_1} = \begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}$$

To figure out the overall matrix, we need to follow where $\hat{\mathbf{\imath}}$ goes. By how we construct matrices in the first place, the image of $\hat{\mathbf{\imath}}$ is just the first column of M_1. To see where that column is mapped, we then multiply that column by M_2:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} A \\ C \end{bmatrix}$$

Using our definition of matrix-vector multiplication we defined earlier, this gives

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} A \\ C \end{bmatrix} = A \begin{bmatrix} a \\ c \end{bmatrix} + C \begin{bmatrix} b \\ d \end{bmatrix} = \begin{bmatrix} Aa + Cb \\ Ac + Cd \end{bmatrix}$$

which is the first column of the composition matrix. Similarly, for $\hat{\mathbf{\jmath}}$, we have,

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} B \\ D \end{bmatrix} = B \begin{bmatrix} a \\ c \end{bmatrix} + D \begin{bmatrix} b \\ d \end{bmatrix} = \begin{bmatrix} Ba + Db \\ Bc + Dd \end{bmatrix}$$

so the composition matrix is,

$$\begin{bmatrix} Aa + Cb & Ba + Db \\ Ac + Cd & Bc + Dd \end{bmatrix}$$

This is where the quite arbitrary-feeling “rows into columns” definition of matrix multiplication actually comes from. It’s just how the numbers work out when we compose linear transformations together.

Furthermore, seeing matrix multiplication as composition of transformations makes the various properties of matrix multiplication much easier to understand.

For example, rotating then shearing, and, shearing then rotating, clearly give different results, so matrix multiplication is not commutative. This is a trivial property you can verify in your head, without having to compute anything at all.

Similarly, matrix multiplication is clearly assocative: applying transformation A, then (B then C) is clearly the same thing as applying transformation (A then B), then C. There’s nothing to prove here; it’s the same three transformations being applied in the same order both ways.

Trying to prove these properties symbolically is a nightmare, but, as transformation compositions, they’re trivial. Not only are these valid proofs, they’re good intuitive explanations as to why these properties should be true.

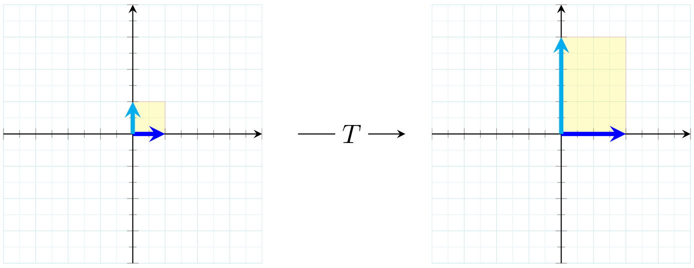
33.2.3 The Determinant

Consider the transformation given by the matrix,

$$\mathbf{T} = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$$

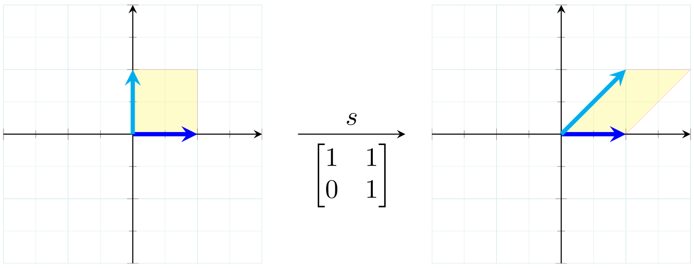
and look at how it transforms the unit square in the first quadrant (the square with sides $\hat{\mathbf{\imath}}$ and $\hat{\mathbf{\jmath}}$).

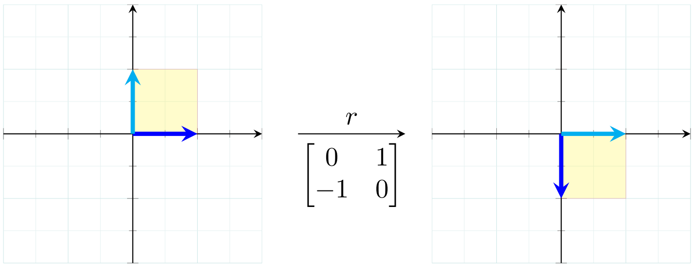
Following $\hat{\mathbf{\imath}}$ and $\hat{\mathbf{\jmath}}$, we find that this square is transformed into a 2 by 3 rectangle.



As the square started with an area of 1, and the resulting rectangle has area 6, this transformation has scaled the area of the square by a factor of 6.

But not all transformations will scale this square. Of course, the identity transformation trivially leaves the square unchanged, but more interestingly, the two transformations we explored in the previous section also do not affect the area of this square:





Although the shapes themselves are distorted (possibly moreso in the shear than the rotation), these transformations seem to leave areas unchanged, at least, in the case of the unit square.

However, because the transformation is linear, these transformation scale the area of *any* shape in the 2D plane by the same factor, and not just the unit square.

Recalling the geometric interpretation of a linear transformation, the origin is fixed in place, and all lines remain lines – so any square that lies within the grid containing the axes is transformed similarly to the unit square, and we can approximate any arbitrary shape that isn't a grid square as closely as we'd like with smaller and smaller squares, each of which are scaled by this same factor.

So, if we know how much the area of the unit square changes, we know how any other shape changes under that transformation.

This scaling factor is called the *determinant* of the transformation, and can variously be written as,

$$\det \mathbf{T} = \det(\mathbf{T}) = |\mathbf{T}| = \det \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} = \begin{vmatrix} 2 & 0 \\ 0 & 3 \end{vmatrix} = 6$$

This idea extends to three dimensions with scaling volume, and in arbitrary dimensions with scaling something called "*measure*" (area and volume are the specific 2D and 3D variants of measure).

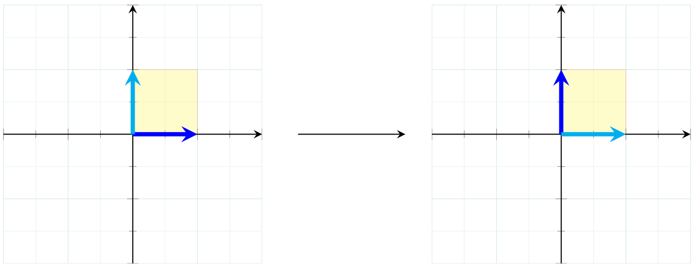
Because the shear and rotation leave areas unchanged, we also have $\det(s) = 1$ and $\det(r) = 1$.

Determinants don't have to be increases in areas either – the transformation given by,

$$\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

makes areas smaller by a factor of $\frac{1}{2}$, and thus has a determinant of $\frac{1}{2}$.

The full definition of the determinant actually allows for negative values, and it doesn't really make sense to scale an area by a negative amount. This has to do with something called *orientation.*



This transformation is a reflection in the line $y = x$. You can somewhat intuitively see that space is "flipped" in some way under this transformation. Before the transformation, if you stand at the origin, facing in the direction of $\hat{\mathbf{\imath}}$, then $\hat{\mathbf{\jmath}}$ is to your left, but after the transformation, $\hat{\mathbf{\jmath}}$ is now to your right. If this is the case, we say that the orientation of space has been *inverted.*

You can similarly check the orientation of a space in 2D and 3D using a handedness rule, like with fields in physics, but, especially for arbitrary dimensions, it's generally easier to just check if the determinant is negative.

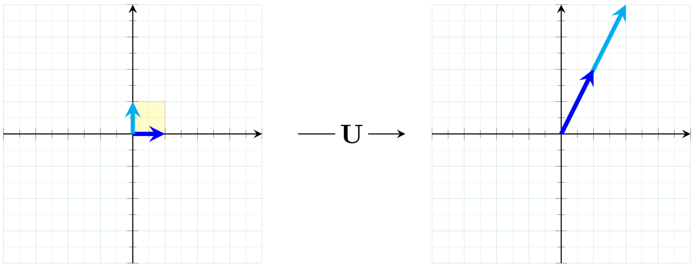
So, more properly, the *magnitude* of the determinant tells us the scaling factor of the transformation, and the *sign* tells us whether the transformation inverts the orientation of space. With this in mind, can you explain why $\det(\mathbf{A})\det(\mathbf{B}) = \det(\mathbf{AB})$?

Now, what happens if the determinant is zero?

If we look at the matrix,

$$\mathbf{U} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

we notice that the image of $\hat{\mathbf{\jmath}}$ is just $\hat{\mathbf{\imath}}$ scaled by a factor of 2, so they are mapped on to the same line.



In other words, the images of $\hat{\mathbf{\imath}}$ and $\hat{\mathbf{\jmath}}$ are not linearly independent.

Because the image of a vector is given by the sum of scaled versions of $\hat{\mathbf{\imath}}$ and $\hat{\mathbf{\jmath}}$, it's clear that the image of any arbitrary vector also ends up stuck on the one-dimensional line spanned by $\hat{\mathbf{\imath}}$ and $\hat{\mathbf{\jmath}}$. In other words, the transformation compresses down all of 2D space onto a 1D line.

The square clearly now has 0 area, so the determinant of this matrix is 0. In general, the determinant of a matrix is zero if the image vectors (of your basis vectors) are linearly dependent.

We can also have another case, as given by the zero matrix,

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

This matrix compresses all of 2D space on to the origin, and also has zero determinant. We call matrices with zero determinant *singular* (and *nonsingular*, *invertible* or *nondegenerate* otherwise).

But notice the distinction between these two cases: one matrix returns a 1D line, and the other, just a single point. These are clearly very different from each other, but both fall under the bracket of "zero determinant".

33.2.4 Column Space & Rank

We have some terminology for this: when the output of a transformation is a line (1 dimensional), the *rank* of the matrix is 1. If all the output vectors form some two-dimensional plane, then the rank of the matrix is 2. We often call this output space the *image*, or the *column space*. This latter name comes from the columns of the matrix being the image of the basis vectors – so the space of all possible vector outputs of the matrix is just the space spanned by the columns (We similarly define the *row space* of a matrix to be the space spanned by its rows, but this is rarely used).

Note that the image of a matrix is a space (the space spanned by the transformed basis vectors), and not a matrix, while the image of a vector is another vector, and not any kind of space.

In general, the rank of a matrix is the number of dimensions in the column space.

So, for the matrices above, we have,

$$\operatorname{rank}\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} = 1, \qquad \text{and} \qquad \operatorname{rank}\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = 0$$

For the 2×2 matrices we've been discussing, rank 2 is the best we can have. There's no way we can map a pair of vectors to span all of 3D space. We simply do not have enough basis vectors.

That isn't to say that the codomain of a linear transformation can't be of higher dimension than the codomain, just that the *image* is at most the dimension of the domain. For example, $T : \mathbf{R}^2 \to \mathbf{R}^3$, $(x,y) \mapsto (x,y,0)$ maps the 2D plane into 3D space, but notice that the image is still just a plane sitting within that 3D space.

When the rank is as high as possible, we say that the matrix is *full rank.* For a 3×3 matrix, we need rank 3 for the matrix to be full rank.

33.2.5 Null Space & Nullity

Note that the zero vector is always within the column space, as linear transformations must keep the origin fixed in place. In particular, for a full rank transformation, the only vector that is mapped to the origin, is the zero vector itself. But for transformations that aren't full rank – transformations that compress space down at least 1 dimension – there will be many more vectors that are mapped to zero.

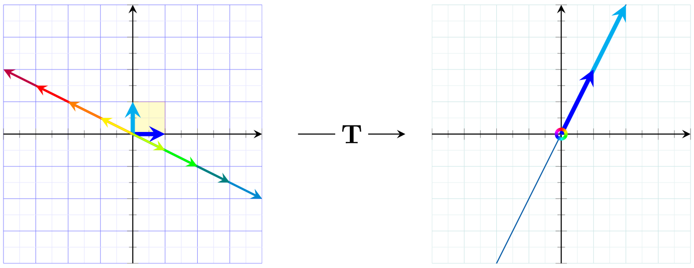
For the matrix we saw earlier, **U**, the vector,

$$\mathbf{v} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$$

is mapped to the origin. You may verify this yourself. But, because images are the sum of scaled basis vectors, we can actually deduce that all vectors that a multiple of this vector will also map to the origin.

$$\mathbf{v} = t \begin{bmatrix} -2 \\ 1 \end{bmatrix}, t \in \mathbb{R}$$

So, we have an entire line of vectors that are mapped to the origin:



The multicoloured vectors are all mapped to the origin, and every other vector will map to somewhere on the line spanned by the transformed basis vectors.

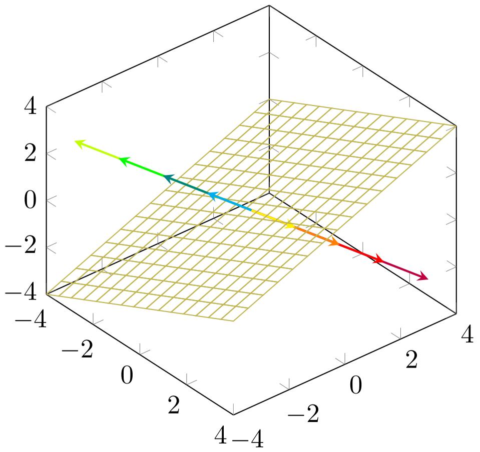
Conversely, the zero matrix maps every vector in the plane to the origin (and the column space would just be the origin, in this case).

This set of vectors that gets mapped to the origin is called the *null space* or *kernel* of the transformation. The dimension of this space is called the *nullity* of the transformation.

For **T**, since we have a line that is mapped to zero, the null space of **T** is that line, and the nullity of **T** is 1. For the zero matrix, the null space is the entire 2D plane, so the nullity is 2.

For any 3 × 3 matrix that maps 3D space to a plane, there will be a line of vectors that are mapped to the origin. Similarly, if 3D space is mapped on to a line, there will be an entire plane of vectors that are mapped to the origin.

Notice that these lines or planes (or volumes/spaces in higher dimensions) will always contain the origin, due to the property discussed above.



You can view nullity as the number of dimensions that are "lost" or "compressed" under a transformation, which leads us to the *rank-nullity theorem*. This theorem effectively states that the sum of the rank and nullity of a transformation is equal to the dimension of the space you are working in.

We will state this theorem more formally once we have defined what spaces even are.

33.2.6 Computational Skills

So, we haven't actually discussed much computation yet. This is predominantly because there is actually very little to do outside of basic matrix algebra. However, we need to cover a little bit now, before moving on to more abstract vector spaces.

33.2.6.1 Elementary Matrix Operations

There are three types of *row operations* we can perform on a matrix:

- Row Switching – swapping two rows;
- Row Scaling – multiplying every element in a row by a non-zero constant;
- Row Addition – replacing a row with the sum of that row and the multiple of another.

Applying a row operation to an identity matrix, then left multiplying by this new matrix is equivalent to performing the row operation on that matrix. A matrix that differs from the identity matrix by a single row operation is an *elementary matrix*.

Column operations are defined similarly, but far more rarely used.

Row operations preserve row space, but not column space. They do, however, preserve the linear independence relationships between columns.

Column operations behave the exact same with with "row" and "column" in the previous paragraph switched.

33.2.6.2 Row Reduction

Using row operations, we can transform matrices into other forms. Some of these forms are particularly useful, and are named.

A matrix is in *row echelon form* if,

- All zero-rows are below all non-zero rows;
- The *pivot* for every non-zero row is strictly to the right of the pivot of the row above.

where a pivot is the first non-zero element of a row.

Furthermore, a matrix is in *reduced row echelon form* if,

- It is in row echelon form;
- Every pivot element is 1;
- The elements above each pivot element are 0.

Example.

$$\begin{bmatrix} 1 & 3 & 6 & 7 & 0 & 5 \\ 0 & 2 & 0 & 2 & 0 & 3 \\ 0 & 0 & 3 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix}$$

is in row echelon form, but *not* reduced row echelon form, as the pivot elements in the second, third and fourth row, have non-zero elements above them, and the pivots are not 1.

$$\begin{bmatrix} 1 & 0 & 9 & 0 & 0 & 7 \\ 0 & 1 & 3 & 0 & 0 & 5 \\ 0 & 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{bmatrix}$$

is in reduced row echelon form. △

The number of pivots in the row reduced matrix is the rank of the matrix, and the nullity is the number of columns, minus the rank.

The row reduced echelon form of a matrix is unique.

Applying both row and column reduction to a matrix brings the matrix into *Smith normal form*, which basically looks like an identity matrix in the top left corner, with zero everywhere else.

Because row operations preserve the linear independence relations between columns, we can find the basis for the image of a matrix by row reducing the matrix, and finding the pivot columns. The vectors formed from the columns that correspond to the pivots in the row reduced matrix form a basis of the image of the matrix

Example. Find a basis for the image of,

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix}$$

A row reduces to,

$$\begin{bmatrix} 1 & 0 & -1 & -2 \\ 0 & 1 & 1 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The first and second columns are pivot columns, so the first and second columns of the original matrix,

$$\begin{bmatrix} 1 \\ 5 \\ 9 \\ 13 \end{bmatrix}, \begin{bmatrix} 2 \\ 6 \\ 10 \\ 14 \end{bmatrix}$$

form a basis of the image of **A**. Because there are only two pivot columns, we only have two basis vectors, so we know the image of **A** is a 2D plane.

Because the image is the same thing as the column space – the space spanned by the columns of the matrix – if you were asked to find a linearly independent set from a set of vectors, you would just augment the vectors together into a matrix, then perform the same procedure above. △

We can also find a basis of the kernel of a matrix through row reduction.

Example. Find a basis for the kernel of $\mathbf{A}$, from the above example.

We once again row reduce the matrix, but now multiply it by an arbitrary vector, and set it equal to the zero vector.

$$\begin{bmatrix}1 & 0 & -1 & -2\\ 0 & 1 & 1 & 3\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}a\\ b\\ c\\ d\end{bmatrix}=\begin{bmatrix}0\\ 0\\ 0\\ 0\end{bmatrix}$$

Now, we rewrite the arbitrary vector in terms of the variables corresponding to non-pivot columns, using the information from the matrix. For example, we know that $a = c+2d$, so 1 and 2 are the first variables in the two vectors.

$$\begin{bmatrix}a\\ b\\ c\\ d\end{bmatrix}=c\begin{bmatrix}1\\ -2\\ 1\\ 0\end{bmatrix}+d\begin{bmatrix}2\\ -3\\ 0\\ 1\end{bmatrix}$$

The two vectors on the right form a basis of the kernel of $\mathbf{A}$.

Once you are more comfortable with this, you can skip out writing the variables, and just read the numbers off of the matrix (but take care with signs!). △

We can also extend a linearly independent set into a basis. Augment the given vectors together, then augment on any basis that spans the desired space. In practice, you can just use the identity matrix, as it is generally the easiest to work with.

Example. The vectors

$$\begin{bmatrix}1\\ 5\\ 9\\ 13\end{bmatrix},\begin{bmatrix}2\\ 6\\ 10\\ 14\end{bmatrix}$$

span $\mathbb{R}^2$, as found above. Extend this set of vectors to a basis of $\mathbb{R}^4$.

Augment the vectors together, along with $\mathbf{I}_4$:

$$\begin{bmatrix}1 & 2 & 1 & 0 & 0 & 0\\ 5 & 6 & 0 & 1 & 0 & 0\\ 9 & 10 & 0 & 0 & 1 & 0\\ 13 & 14 & 0 & 0 & 0 & 1\end{bmatrix}$$

and row reduce,

$$\begin{bmatrix}1 & 0 & 0 & 0 & -\frac{7}{2} & \frac{5}{2}\\ 0 & 1 & 0 & 0 & \frac{13}{4} & -\frac{9}{4}\\ 0 & 0 & 1 & 0 & -3 & 2\\ 0 & 0 & 0 & 1 & -2 & 1\end{bmatrix}$$

The first four columns are pivot columns, so the first four columns of the original matrix,

$$\begin{bmatrix}1\\ 5\\ 9\\ 13\end{bmatrix},\begin{bmatrix}2\\ 6\\ 10\\ 14\end{bmatrix},\begin{bmatrix}1\\ 0\\ 0\\ 0\end{bmatrix},\begin{bmatrix}0\\ 1\\ 0\\ 0\end{bmatrix}$$

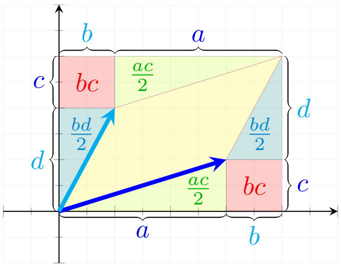
form a basis of $\mathbb{R}^4$. △

33.2.6.3 Determinants

For a 2 × 2 matrix,

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

the determinant is given by $\det(\mathbf{A}) = ad - bc$. This is one place where I think a geometric explanation isn't particularly helpful, but here is a diagram, if you are still interested:



$$\det\begin{bmatrix} a & b \\ c & d \end{bmatrix} = (a+b)(c+d) - 2\left(\frac{ac}{2}\right) - 2\left(\frac{bd}{2}\right) - 2bc = ad - bc$$

For higher dimensional matrices, there are various methods to calculate the determinant – you've probably learnt Laplacian expansion to find determinants before.

However, we will often use Gaussian elimination, or more generally, row reduction, to compute the determinant of larger matrices: Scaling a row by k scales the determinant by k, and swapping two rows multiplies the determinant by -1. If we row reduce our matrix to the identity matrix (which has determinant 1), we can then run the row operations in reverse and keep track of the determinant.

33.2.7 Systems of Linear Equations & Matrix Inverses

One reason why linear algebra is required for such a wide variety of technical disciplines, is that it allows us to solve a certain type of system of equations.

If your system of equations only involves equations which add up multiples of your variables, i.e., linear combinations of variables, then we can apply the tools of linear algebra.

We usually organise this sort of special system of equations by putting all the variables on the left, lining them up vertically, and putting all the constants on the right.

$$\begin{aligned} 4x - 5y + 7z &= 6 \\ 2x + 3y - 2z &= 2 \\ 9x - 7y + 3z &= 5 \end{aligned}$$

This might remind you of matrix-vector multiplication. And indeed, we can wrap this entire system of equations up into a single vector equation:

$$\begin{bmatrix} 4 & -5 & 7 \\ 2 & 3 & -2 \\ 9 & -7 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 6 \\ 2 \\ 5 \end{bmatrix}$$

We've separated out the constant coefficients into a matrix, and packed all of the variables into a vector, and we want their matrix-vector product to be some constant vector.

We often label the matrix $\mathbf{A}$, the variable vector as $\mathbf{x}$, and the constant vector as $\mathbf{v}$.

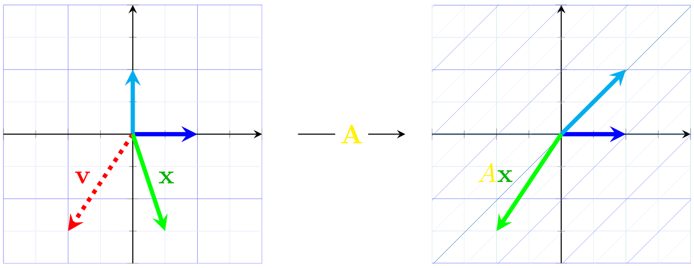
$$\mathbf{A}\mathbf{x} = \mathbf{v}$$

And this isn't just a notational trick to save space – we can now view this algebraic problem as a geometric one. **A** is a matrix, and therefore represents some linear transformation, despite the coefficients and source problem possibly having nothing to do with geometry. So, solving $\mathbf{Ax} = \mathbf{v}$ for **x** means we're trying to find a vector, **x**, which gets mapped to **v** after applying the transformation **A**.

For simplicity, suppose **A** is a 2×2 matrix.

Now, there are a couple of possible cases, depending on whether the transformation given by **A** compresses space down into a lower dimension or not. In other words, we care about whether **A** is singular or not.

If **A** is non-singular, meaning space is not compressed into a zero-area region, then there will be a single unique vector that lands on **v** under **A**, and we can find it by playing the transformation backwards.



Following where **v** goes under this backwards transformation will give us **x**.

Playing **A** in reverse actually gives a separate linear transformation, the *inverse* of **A**, written $\mathbf{A}^{-1}$.

Here, **A** is a rightwards shear that moves $\hat{\mathbf{j}}$ one unit to the right, so **A** inverse would be a leftwards shear that moves $\hat{\mathbf{j}}$ one unit to the left. If **A** is a 90° clockwise rotation, then $\mathbf{A}^{-1}$ would be a 90° anticlockwise rotation.

In general, $\mathbf{A}^{-1}$ is the unique transformation such that, if you apply the transformation **A**, then the transformation **A** inverse, the overall effect is just the identity. Applying transformations sequentially is algebraically expressed as matrix multiplication, so another way to describe this property, is that $\mathbf{A}^{-1}$ is the unique matrix such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.

If you can find this inverse matrix, we can solve the equation by multiplying both sides by the inverse:

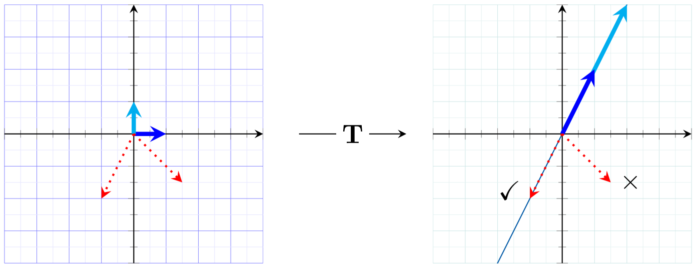
$$\begin{aligned} \mathbf{Ax} &= \mathbf{v} \\ \mathbf{A}^{-1}\mathbf{Ax} &= \mathbf{A}^{-1}\mathbf{v} \\ \mathbf{Ix} &= \mathbf{A}^{-1}\mathbf{v} \\ \mathbf{x} &= \mathbf{A}^{-1}\mathbf{v} \end{aligned}$$

This non-zero determinant case, which, for any random matrix, is almost certainly the case, corresponds with the idea that, if you have two variables and two equations, you'll almost certainly have one unique solution to the system.

This also extends to higher dimensions, when the number of equations equals the number of variables. You can translate these algebraic problems into geometric ones, finding vectors that land on other vectors under some transformation, given that the transformation associated with the coefficient matrix doesn't compress space into a lower dimension.

However, when matrix is singular, and the transformation does collapse down dimensions, then there is no inverse to find, because information is lost when compressing down space. You cannot decompress a line back into a plane – doing so would require transforming a single vector on the line into an entire line of vectors in the plane, which functions cannot do.

Solutions can still exist when the determinant is zero, but only if the constant vector just happens to be in the column space of the matrix:



Clearly, if the target vector is outside of the column space, then there's no way to get there from the basis vectors: the column space is defined as the space you can reach with linear combinations of the basis vectors. Similarly, if the target vector lies inside the column space, it's clearly reachable.

There are several ways to find the inverse of any given matrix. One which you may have learned previously is by using the determinant and matrix of cofactors. You may also have learned Cramer's rule, or the Cayley-Hamilton theorem, which each can also be used to (perhaps somewhat inefficiently, depending on the matrix) find the inverse of a matrix.

Here, we will often use Gaussian elimination, or row reduction, which is much faster,* especially on larger matrices.

If we have a matrix we wish to find the inverse of, we first append or *augment* the matrix with the identity matrix on the right, then row reduce the resulting rectangular matrix. The inverse matrix will then be on the right hand side.

Example. Find the inverse of:

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 1 \\ 0 & 4 & 1 \\ 2 & -1 & 0 \end{bmatrix}$$

First, we augment the matrix with the identity. For clarity, a separating line showing where the augmentation happened is included.

$$\left[\begin{array}{ccc|ccc} 1 & 3 & 1 & 1 & 0 & 0 \\ 0 & 4 & 1 & 0 & 1 & 0 \\ 2 & -1 & 0 & 0 & 0 & 1 \end{array}\right]$$

Then row reduce, until the identity is now on the left.

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -1 & 1 & 1 \\ 0 & 1 & 0 & -2 & 2 & 1 \\ 0 & 0 & 1 & 8 & -7 & -4 \end{array}\right]$$

with $\underbrace{}_{\mathbf{A}^{-1}}$ under the right-hand block.

The inverse is then on the right. △

This is very useful for getting inverses, but when we are solving systems of linear equations, we often don't need the inverse itself, and are just looking for the solution vector, $\mathbf{x}$.

In this case, we can similarly rewrite a system of linear equations as an *augmented matrix* by writing the coefficients of the variables into the matrix as usual, but this time, we augment on the vector containing the constants.

* For an $n \times n$ matrix, Gaussian elimination takes about $O(n^3)$ time (assuming multiplication and addition takes constant time, which, they don't, as the intermediate matrix entries tend to grow exponentially in size). The cofactor matrix method requires finding the determinants of n^2 distinct $(n-1) \times (n-1)$ matrices, and finding determinants is superpolynomial in complexity. Cramer's rule takes $O(n!)$ time. In any case, even with idealisations about the speed of multiplication and addition, row reduction is generally a lot faster.

Row reducing the matrix, we can immediately read off what the solutions should be. This is useful where the inverse is not needed and/or is very complicated.

Example.

$$\begin{aligned} 4x - 5y + 7z &= 6 \\ 2x + 3y - 2z &= 2 \\ 9x - 7y + 3z &= 5 \end{aligned}$$

$$\left[\begin{array}{ccc|c} 4 & -5 & 7 & 6 \\ 2 & 3 & -2 & 2 \\ 9 & -7 & 3 & 5 \end{array}\right]$$

Row reducing this matrix, we have,

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & \frac{9}{11} \\ 0 & 1 & 0 & \frac{8}{11} \\ 0 & 0 & 1 & \frac{10}{11} \end{array}\right]$$

so $x = \frac{9}{11}$, $y = \frac{8}{11}$, and $z = \frac{10}{11}$ are solutions. △

You can try finding the inverse of the matrix above as an exercise, but the entries will not be very nice to work with (they are all rational, but with very large denominators).

This method also allows us to determine whether solutions exist for any given system of linear equations:

Example.

$$\begin{aligned} 3x + 6y - 6z &= -6 \\ -6x + 3y + 3z &= 2 \\ -3x - y + 3z &= -2 \end{aligned}$$

$$\left[\begin{array}{ccc|c} 3 & 6 & -6 & -6 \\ -6 & 3 & 3 & 2 \\ -3 & -1 & 3 & -2 \end{array}\right]$$

which row reduces to,

$$\left[\begin{array}{ccc|c} 1 & 0 & -\frac{4}{5} & 0 \\ 0 & 1 & -\frac{3}{5} & 0 \\ 0 & 0 & 0 & 1 \end{array}\right]$$

but the last line implies that $0 = 1$, so we know the system is inconsistent and has no solutions.

We can say that the coefficient matrix is of lower rank than the augmented matrix, so the system is inconsistent and has no solutions. △

33.3 Scalars & Fields

So far, we've been saying that we can multiply or scale vectors by certain numbers called scalars. But what numbers can these scalars actually be? So far, we've mostly been using the integers or rationals, but, we can pick other numbers too.

In general, for any given vector space, the scalars must all come from a *field*.

Fields have a dedicated chapter in which they are explored, so here, we only give a short summary of a definition in terms of rings and groups (§12.3), and a second axiomatic definition. For a more detailed discussion, see §11.2.

33.3.1 Fields from Groups and Rings

A *ring* is a triple, $(R, +, \times)$, where R is a set and $+$ and $\times$ are binary operations such that

- R is an abelian group under $+$;
- R is closed under $\times$;
- R contains an identity under $\times$
- $\times$ is associative on R
- $\times$ left and right distributes over $+$

We call the operation denoted by $+$ *addition*, and the operation denoted by $\times$ *multiplication* or *product*, regardless of what the operations actually are. We also call the additive identity, 0_R or the *ring zero*, as it is also the zero element for the multiplication operation. We also denote the multiplicative identity, 1_R.

Furthermore, $(R, +, \times)$ is a *commutative ring* if $\times$ is commutative on R.

Note that the "commutative" part of "commutative ring" refers to multiplication, as commutativity of addition is required regardless.

However, rings notably do **not** require multiplicative inverses.

Let R be a ring and $a,b \in R$. Then,

- $a \times 0_R = 0_R \times a = 0_R$
- $-(a \times b) = (-a) \times b = a \times (-b)$

An element, a, of a ring R is a *unit* if there exists some $b \in R$ such that $ab = ba = 1_R$. Essentially, a is a unit of R if a has a multiplicative inverse in R. In any non-zero ring, 0_R is a non-unit.

Example. In $\mathbb{R}$, $\mathbb{Q}$ and $\mathbb{C}$, every non-zero element, k, has a multiplicative inverse, $\frac{1}{k}$, so the units are the non-zero elements.

However, in $\mathbb{Z}$, $\frac{1}{k}$ is an integer only for $k = \pm 1$, so the units in $\mathbb{Z}$ are ± 1. △

A *field*, $(F, +, \times)$, is a commutative ring such that every non-zero element is a unit, and $0_F \neq 1_F$.

Equivalently, $(F, +, \times)$ is a field if $(F,+)$ is an abelian group with additive identity 0_F, $(F \setminus \{0_F\},\times)$ is an abelian group with multiplicative identity 1_F, $0_F \neq 1_F$ and multiplication distributes over addition.

This $0_F \neq 1_F$ condition is called the *non-degeneracy condition*, and is basically there just to exclude the trivial set, $\{0\}$, from being a field.

33.3.2 Field Axioms

Given a set S, a *binary operation* on S is a function that takes two elements of S, called the *operands* or *arguments* of the operation, and returns another element of S: it is *closed* over S. That is, it is a binary function $S \times S \to S$.

A *field* is a set, K, together with two elements, $0_K \neq 1_K \in K$, and two binary operations, $\cdot : K \times K \to K$ and $+ : K \times K \to K$, called *multiplication* and *addition*, respectively, that satisfies the following axioms:

(A1) $\forall a,b \in K, a + b = b + a$ (commutativity of addition);

(A2) $\forall a,b,c \in K, a + (b + c) = (a + b) + c$ (associativity of addition);

(A3) $\exists 0_K \in K$ such that $\forall a \in K, a + 0_K = 0_K + a = a$ (existence of additive identity);

(A4) $\forall a \in K, \exists(-a) \in K$ such that $a + (-a) = (-a) + a = 0_K$ (existence of additive inverses).

(M1) $\forall a,b \in K, a \cdot b = b \cdot a$ (commutativity of multiplication);

(M2) $\forall a,b,c \in K, a \cdot (b \cdot c) = (a \cdot b) \cdot c$ (associativity of multiplication);

(M3) $\exists 1_K \in G$ such that $\forall a \in K, a \times 1_K = 1_K \times a = a$ (existence of additive identity);

(M4) $\forall a \in K, \exists (a^{-1}) \in K \setminus \{0\}$ such that $a * (a^{-1}) = (a^{-1}) * a = 1_K$ (existence of multiplicative inverses);

(D) $\forall a,b,c \in K, (a+b)c = ac + bc$ (distributivity of multiplication over addition);

(ND) $0_K \neq 1_K$ (non-degeneracy).

Where there is no room for confusion, we write ab for $a \cdot b$, and 0 and 1 for 0_K and 1_K, respectively. We often denote general fields with the letter, K (originally from the German word *Körper*, meaning "corpus" or "body", suggesting a closed entity) or otherwise with F. The symbol $\mathbb{F}$ is reserved for a certain type of finite field.

Informally, a set is a field if it equipped with operations analogous to those of addition, subtraction, multiplication, and division on the real numbers.

33.4 Vector Spaces

A *vector space* over a field, K, is a set, V, along with two maps, $+ : V^2 \to V$ and $\cdot : K \times V \to V$, called *vector addition* and *scalar multiplication*, respectively, that satisfies the following axioms for all $\mathbf{u},\mathbf{v},\mathbf{w} \in V$ and $a,b \in K$:

(V1) $(V,+)$ is an abelian group.

- (A1) $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$ (commutativity of vector addition);
- (A2) $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$ (associativity of vector addition);
- (A3) $\exists \mathbf{0}_V$ such that $\mathbf{v} + \mathbf{0}_V = \mathbf{0}_V + \mathbf{v} = \mathbf{v}$ (existence of vector additive identity);
- (A4) $\exists (-\mathbf{v}) \in V$ such that $\mathbf{v} + (-\mathbf{v}) = (-\mathbf{v}) + \mathbf{v} = \mathbf{0}_V$ (existence of vector addition inverses);
- (A5) $\mathbf{u} + \mathbf{v} \in V$ (closure of vector addition).

(V2) $a \cdot (\mathbf{u} + \mathbf{v}) = a \cdot \mathbf{u} + b \cdot \mathbf{v}$ (distributivity of scalar multiplication over vector addition);

(V3) $(a+b) \cdot \mathbf{v} = a \cdot \mathbf{v} + b \cdot \mathbf{v}$ (distributivity of scalar multiplication over field addition);

(V4) $(ab) \cdot \mathbf{v} = a \cdot (b\mathbf{v})$ (compatibility of scalar multiplication with field multiplication);

(V5) $1_K \cdot \mathbf{v} = \mathbf{v}$ (existence of scalar multiplicative identity).

33.4.1 Subspaces

Let V be a vector space over a field K and let $W \subseteq V$ be a non-empty set. If W is also a vector space over K, then W is is called a (*linear* or *vector*) *subspace* of V. If $W \neq V$, then W is a *proper* subspace. If $W = \{\mathbf{0_V}\}$, then W is the *trivial* subspace. The trivial subspace has dimension 0.

Theorem 33.4.1. *A subset W of a vector space V over a field K is a subspace if W is closed under vector addition and scalar multiplication. That is, W is a subspace if for every pair of vectors $\mathbf{u},\mathbf{v} \in W$ and every pair of scalars $\alpha,\beta \in K$, we have $\alpha\mathbf{u} + \beta\mathbf{v} \in W$.*

Let V be a vector space over a field K and let W_1, W_2 be subspaces of V. Then, the intersection, $W_1 \cap W_2 = \{\mathbf{w} : \mathbf{w} \in W_1 \cap W_2\}$, and sum, $W_1 + W_2 = \{\mathbf{w_1} + \mathbf{w_2} : \mathbf{w_1} \in W_1, \mathbf{w_2} \in W_2\}$, are also subspaces.

If $U = W_1 + W_2$, then every $\mathbf{u} \in U$ can be written as $\mathbf{u} = \mathbf{w_1} + \mathbf{w_2}$, where $\mathbf{w_1} \in W_1$ and $\mathbf{w_2} \in W_2$. If this representation is unique, then we write $U = W_1 \oplus W_2$, or that U is the *direct sum* of W_1 and W_2.

Two subspaces W_1 and W_2 of a vector space, V, are *complementary* if $W_1 \cap W_2 = \{\mathbf{0}_V\}$ and $W_1 + W_2 = V$, or equivalently, $W_1 \oplus W_2 = V$. However, the union of two subspaces is generally not a subspace.

Example. The sets $U = \big\{(x,x) : x \in \mathbb{R}\big\}$ and $W = \big\{(x, -x) : x \in \mathbb{R}\big\}$ are linear subspaces of $\mathbb{R}^2$, but their union is not closed under vector addition. For instance, $(1,1) \in U$ and $(1, -1) \in W$, but $(1,1) + (1, -1) = (2,0) \notin U \cup W$. △

Let V be a vector space over a field K such that V is finite-dimensional, and let W_1, W_2 be subspaces of V. Then,

$$\dim W_1 + W_2 = \dim W_1 + \dim W_2 - \dim W_1 \cap W_2$$

33.4.2 Quotient Spaces

The *quotient* of a vector space V by a subspace N, denoted V/N is a vector space obtained by identifying the elements in N with zero. This construction is analogous to that of quotient groups (or any other algebraic quotient object).

Let V be a vector space over a field K, and let $N \subseteq V$ be a subspace. Let $\sim$ be the equivalence relation on V defined by $x \sim y$ if and only if $x - y \in N$. That is, x is related to y if their vector difference is an element of N, or equivalently, if one may be obtained from the other by adding an element of N. Consequently, every element of N lies in the same equivalence class, as it is closed under vector addition as a subspace, and moreover, this equivalence class includes the zero $\mathbf{0}_V$.

The equivalence class, or *coset*, of a vector $\mathbf{v} \in V$ is denoted by

$$\begin{aligned}[\mathbf{v}] &= \mathbf{v} + N \\ &= \{\mathbf{v} + \mathbf{n} : \mathbf{n} \in \mathbb{N}\}\end{aligned}$$

The quotient space V/N is then defined to be $V/\!\sim$, the set of equivalence classes on V induced by $\sim$.

This set has a natural vector space structure over K with operations given by

- $\alpha[\mathbf{v}] = [\alpha\mathbf{v}]$;
- $[\mathbf{u}] + [\mathbf{v}] = [\mathbf{u} + \mathbf{v}]$,

for all $\alpha \in K$ and $\mathbf{u},\mathbf{v} \in V$, and the zero element is given by the equivalence class $N = [\mathbf{0}]$. It is not hard to verify that these operations are well-defined (the proof is entirely analogous to that in §12.5.2). The natural map $q : V \to V/N$ defined by $\mathbf{v} \mapsto [\mathbf{v}]$ is then called the *quotient map*.

Intuitively, the quotient space V/N can be viewed geometrically as the set of all affine subsets of V parallel to N.

If U is a subspace of V, then the dimension of the quotient V/U is called the *codimension* of U in V. Since a basis of V may be constructed from a basis A of U and a basis B of V/U by adding a representative of each equivalence class in B into A, it follows that the dimension of V is the sum of the dimensions of U and V/U. If V is finite-dimension, then we have

$$\operatorname{codim} U = \dim(V/U) = \dim V - \dim U$$

If we quotient a space by the trivial subspace, we just obtain the space itself since two elements are identified if and only if they differ by the zero vector: but this just means the equivalence classes are all singletons since two vectors that differ by the zero vector are just the same vector.

$$V/\{\mathbf{0}_V\} \cong V$$

Similarly, if we quotient a space by itself, we obtain the trivial space since all elements of V differ by another element of V

$$V/V \cong \{\mathbf{0}_V\}$$

Example. Let $\mathbb{R}^2$ be the standard Cartesian plane as a vector space over $\mathbb{R}$, and let N be a line through the origin. Then, the quotient space $\mathbb{R}^2/N$ can be identified with the space of all lines in V parallel to N. Points within any such given line satisfy the equivalence relation since their difference vector lies within N.

Any line not parallel to N also intersects each of these lines at exactly one point, so the quotient space can also be identified with the set of points along such a line.

Similarly, the quotient of $\mathbb{R}^3$ by a line N through the origin may be identified as the space of all lines parallel to N. Again, any plane not containing N will also intersect all of these lines at exactly one point, so the quotient space may also be identified with such a plane.

We may also quotient $\mathbb{R}^3$ by a plane Π through the origin. This time, the quotient space may be identified as the space of all planes parallel to Π; or alternatively, as a line not contained within Π. △

Example. Consider the subspace of $\mathbb{R}^n$ spanned by the first m standard basis vectors. Vectors in $\mathbb{R}^n$ consists of n-tuples $[x_1,\dots,x_n]$, while the vectors in the subspace consist of the n-tuples $[x_1,\dots,x_m,0,\dots,0]$ that are 0 in the last $n-m$ coordinates. Thus, two vectors in $\mathbb{R}^n$ are equivalent modulo $\mathbb{R}^m$ if and only if they agree in the last $n-m$ coordinates, so the quotient space $\mathbb{R}^n/\mathbb{R}^m$ is canonically isomorphic to $\mathbb{R}^{n-m}$ in the obvious manner.

This generalises the previous example, where we quotiented $\mathbb{R}^2$ by a line isomorphic to $\mathbb{R}^1$ to obtain a space of lines isomorphic to $\mathbb{R}^1$; or $\mathbb{R}^3$ by a line $N \cong \mathbb{R}^1$ or plane $\Pi \cong \mathbb{R}^2$ to obtain a space isomorphic to $\mathbb{R}^2$ and $\mathbb{R}^1$, respectively. △

Let $V = U \oplus V$ be a direct sum. Then, quotienting by one of the subspaces yields a space naturally isomorphic to the other:

$$V/U \cong W$$
$$V/W \cong U$$

33.4.3 Rank-Nullity Theorem

For a linear transformation, $T : V \to W$, we have a pair of important fundamental subspaces given by the image and kernel, and we define the rank and nullity to be their respective dimensions:

$$\begin{aligned} \operatorname{im} T &:= \{T(\mathbf{v}) : \mathbf{v} \in V\} \\ \ker T &:= \{\mathbf{v} \in V : T(\mathbf{v}) = \mathbf{0}_W\} \\ \operatorname{rank} T &:= \dim(\operatorname{im} T) \\ \operatorname{null} T &:= \dim(\ker T). \end{aligned}$$

An important result linking these notions together is the *rank-nullity theorem*, which is in fact just the first isomorphism theorem (§12.5.4) for vector spaces.

But first, we state and prove an important lemma useful for its proof:

Lemma (Steinitz Exchange Lemma)**.** *Let U and W be finite subsets of a vector space V over a field K such that the vectors in U are linearly independent and the vectors in W span V. Then,*

(i) $|U| \leq |W|$;

(ii) There exists a set $W' \subseteq W$ with cardinality $|W'| = |W| - |U|$ such that $U \cup W'$ spans V.

Proof. Suppose $U = \{\mathbf{u}_i\}_{i=1}^m$ and $W = \{\mathbf{w}_i\}_{i=1}^n$ satisfy the hypotheses above. We induct on m.

Let $P(m)$ be the statement that there is an ordering of the $\mathbf{w}_i$ such that the set

$$S_m := \{\mathbf{u}_1,\dots,\mathbf{u}_m,\mathbf{w}_{m+1},\dots,\mathbf{w}_n\}$$

spans V.

For $m = 0$, $P(0)$ holds, as there are no $\mathbf{u}_i$, and the set $\{\mathbf{u}_1,\ldots,\mathbf{u}_m,\mathbf{w}_{m+1},\ldots,\mathbf{w}_n\} = \{\mathbf{w}_1,\ldots\mathbf{w}_n\} = W$ spans V by assumption.

Now, suppose P holds for some arbitary fixed value $m-1 \geq 0$. By the inductive hypothesis, there is an ordering of the $\mathbf{w}_i$ such that $S_{m-1} = \{\mathbf{u}_1,\ldots,\mathbf{u}_{m-1},\mathbf{w}_m,\ldots,\mathbf{w}_n\}$ spans V. Since $\mathbf{u}_m \in V$, it is expressible as a linear combination of vectors in S_{m-1}. That is, there exist coefficients $\alpha_1,\ldots\alpha_n \in K$ such that

$$u_m = \sum_{i=1}^{m-1} \alpha_1 \mathbf{u}_i + \sum_{j=m}^{n} \alpha_j \mathbf{w}_j$$

At least one of the coefficients α_j in the second sum must be non-zero or this would contradict the linear independence of U, so we must have $m \leq n$, proving (i).

By reordering $\alpha_m\mathbf{w}_m,\ldots,\alpha_n\mathbf{w}_n$ if necessary, we may assume that α_m is non-zero. Then, we have,

$$\begin{aligned}\mathbf{w}_m = \frac{1}{\alpha_m}\left(\mathbf{u}_m - \sum_{i=1}^{m-1} \alpha_1 \mathbf{u}_i - \sum_{j=m}^{n} \alpha_j \mathbf{w}_j\right) \\ \in \operatorname{span}\{\mathbf{u}_1,\ldots,\mathbf{u}_m,\mathbf{w}_{m+1},\ldots,\mathbf{w}_n\}\end{aligned}$$

and since this span contains each of the vectors $\mathbf{u}_1,\ldots,\mathbf{u}_{m-1},\mathbf{w}_m,\mathbf{w}_{m+1},\ldots,\mathbf{w}_n$, by the inductive hypothesis, it contains V, which is exactly $P(m)$. Induction then completes the proof of (ii). ■

The rank-nullity theorem links the dimension of a vector space with the rank and nullity of a linear transformation out from that space:

Theorem (Rank-Nullity)**.** *Let V,W be vector spaces over a field K, $T : V \to W$ a linear transformation, and V be finite-dimensional. Then,*

$$\operatorname{rank} T + \operatorname{null} T = \dim V$$

or,

$$\dim(\operatorname{im} T) + \dim(\ker T) = \dim(\operatorname{dom} T)$$

Proof. Let $n = \dim V$. As $\ker T$ is a subspace of V, it has a basis. Let $k = \dim \ker T$ and let

$$\mathcal{K} \coloneqq \{\mathbf{v}_1,\ldots,\mathbf{v}_k\} \subseteq \ker T$$

be such a basis. By the Steinitz exchange lemma, we may extend $\mathcal{K}$ with $n-k$ linearly independent vectors $\mathbf{w}_1,\ldots,\mathbf{w}_{n-k}$ to form a basis of V. Let

$$\mathcal{S} \coloneqq \{\mathbf{w}_1,\ldots,\mathbf{w}_{n-k}\} \subseteq V \setminus \ker T$$

be such that

$$\mathcal{B} \coloneqq \mathcal{K} \cup \mathcal{S} = \{\mathbf{v}_1,\ldots,\mathbf{v}_k,\mathbf{w}_1,\ldots,\mathbf{w}_{n-k}\} \subseteq V$$

is a basis for V. From this, we have,

$$\begin{aligned}\operatorname{im} T &= \operatorname{span}\big(T(\mathcal{B})\big) \\ &= \operatorname{span}\big\{T(\mathbf{v}_1),\ldots,T(\mathbf{v}_k),T(\mathbf{w}_1),\ldots,T(\mathbf{w}_{n-k})\big\} \\ &= \operatorname{span}\big\{T(\mathbf{w}_1),\ldots,T(\mathbf{w}_{n-k})\big\} \\ &= \operatorname{span}\big(T(S)\big)\end{aligned}$$

So $T(S)$ spans $\operatorname{im} T$. Now, suppose $T(S)$ is not linearly independent, so there exist coefficients $\alpha_1, \dots \alpha_{n-k} \in K$ such that

$$\sum_{i=1}^{n-k} \alpha_i T(\mathbf{w}_i) = \mathbf{0}_W$$

By linearity of T, we have

$$T\left(\sum_{i=1}^{n-k} \alpha_i \mathbf{w}_i\right) = \mathbf{0}_W$$

so $\sum_{i=1}^{n-k} \alpha_i \mathbf{w}_i \in \ker T = \operatorname{span} \mathcal{K} \subseteq V$, contradicting that $\mathcal{B}$ is a basis. It follows that $T(S)$ is linearly independent and hence forms a basis of $\operatorname{im} T$.

Finally, we have,

$$\begin{aligned} \operatorname{rank} T + \operatorname{null} T &= \dim \operatorname{im} T + \dim \ker T \\ &= \left|T(S)\right| + |\mathcal{K}| \\ &= (n-k) + k \\ &= n \\ &= \dim V \end{aligned}$$

■

One corollary of this theorem is that surjectivity, injectivity, and bijectivity are equivalent for linear transformations:

Corollary 33.4.1.1. *Let $T : V \to W$ be a linear transformation, and suppose that $\dim V = \dim W$. Then, the following are equivalent:*

1. *T is injective;*
2. *T is surjective;*
3. *T is bijective (and hence constitutes an isomorphism).*

Proof. Let $n = \dim V = \dim W$. If T is injective, then $\operatorname{null} T = 0$, so by rank-nullity, $\operatorname{rank} T = \dim(\operatorname{im} T) = n = \dim W$, so T is surjective; the same argument applies in reverse, so surjectivity of T implies its injectivity, and hence both are equivalent to bijectivity. ■

Because linear maps can be represented by matrices, the rank-nullity theorem can be restated in terms of matrices. Specifically, an $m \times n$ matrix M represents a linear map $f : K^n \to K^m$, where K is the underlying field, so the dimension of $\operatorname{dom} f$ is the number of columns of M, or n. The rank-nullity theorem then says:

Corollary 33.4.1.2. *For any $m \times n$ matrix M,*

$$\operatorname{rank} M + \operatorname{null} M = m$$

33.4.3.1 Cokernels

Let $T : V \to W$ be a linear transformation with $n = \dim(V)$ and $m = \dim(W)$ finite.

We have looked at images and kernels, but a third fundamental subspace is given by the *cokernel*, which is the quotient space

$$\operatorname{coker} T \coloneqq W\big/\operatorname{im} T$$

and the dimension of this space is called the *corank* of T.

The rank-nullity theorem links the image and kernel of T together in that if T has rank r, then the kernel of T has dimension $n - r$. However, there is a similar *dual* result that connects the image and the *cokernel* – if T has rank r, then the cokernel of T has dimension $m - r$:

Theorem 33.4.2. *Let V,W be vector spaces over a field K, $T : V \to W$ a linear transformation, and W be finite-dimensional. Then,*

$$\operatorname{rank} T + \operatorname{corank} T = \dim W$$

or,

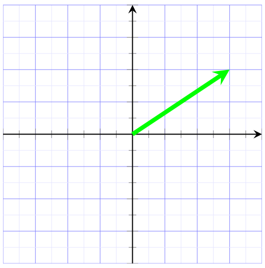
$$\dim(\operatorname{im} T) + \dim(\operatorname{coker} T) = \dim(\operatorname{cod} T)$$

Proof. Formal dual of the rank-nullity theorem. ■

Together with the rank nullity theorem, these two results are called the *fundamental theorem of linear algebra.*

33.5 Change of Basis

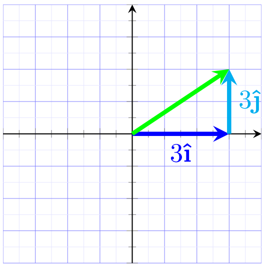
Right at the beginning, we said that assigning numbers to vectors – in the sense of arrows rooted at the origin – depends on some choice of basis vectors to provide a meaningful translation between geometry and algebra.



In our standard system, we would say that this green vector has coordinates,

$$\begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

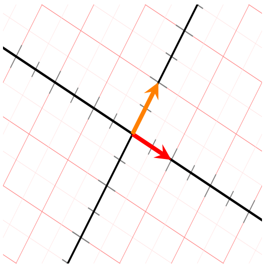
because going from its tail to its tip requires moving 3 units to the right, and 2 units up. We think of these coordinates as scalars – something that scales up a vector. In this case, we implicitly take the first coordinate to scale $\hat{\mathbf{i}}$, and the second to scale $\hat{\mathbf{j}}$, before adding up the result, with all the information about distance and direction tied up in our choice of basis vectors.



We call these ways to translate between these arrows and sets of numbers a *coordinate system.* The choice of $\hat{\mathbf{i}}$ being the target of the first scalar, and $\hat{\mathbf{j}}$ being the target of the second scalar gives us the standard Cartesian coordinate system.

But of course, other basis vectors are available.

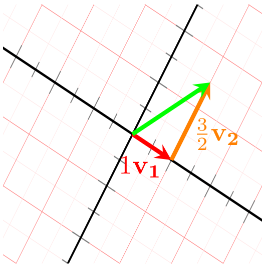
Say we have a friend, Alice, who uses a different set of basis vectors: $\mathbf{v_1}$, which points to the bottom-right and is acted upon by the first scalar of a coordinate, and$\mathbf{v_2}$, that points to the top right and is acted upon by the second scalar.



We can draw the same green vector again – the one we would describe as [3,2] – on to her grid. Alice would then describe this green vector as,

$$\begin{bmatrix} 1 \\ \frac{3}{2} \end{bmatrix}$$

What this means is that, to get to the tip of that vector using her basis vectors, is to scale up $\mathbf{v_1}$ by 1, $\mathbf{v_2}$ by $\frac{3}{2}$, then add up the results



Whenever Alice uses coordinates to describe a vector, she thinks of the first coordinate scaling $\mathbf{v_1}$, and the second, scaling $\mathbf{v_2}$, just like how we scale $\hat{\mathbf{\imath}}$ and $\hat{\mathbf{\jmath}}$, respectively.

We note that, although the two coordinates look different, they actually represent the same vector, just in two different coordinate systems. We're both describing the same things, but in a different language.

$$\begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{3}{2} \end{bmatrix}$$

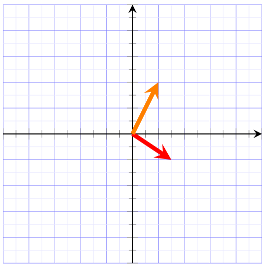
We have been showing the choices of bases using colour (and will continue doing so), but it is helpful to have notation for this as well. To do this, we give a label to each choice of basis, and subscript our vectors with that label. Often, this is done with set brackets (for example, $\{\mathbf{e}_i\}$) to indicate that the basis is a set of vectors, but here, for clarity, We will label the Cartesian coordinate system as E, and Alice's coordinate system as A.

$$\begin{bmatrix} 3 \\ 2 \end{bmatrix}_E = \begin{bmatrix} 1 \\ \frac{3}{2} \end{bmatrix}_A$$

But how did we find that second set of coordinates? More generally, how do we find the coordinates of some vector in some given different coordinate system? Well, we should first look at the basis vectors of the coordinate systems in question.

We can describe the basis vectors of the target coordinate system in terms of our standard one. In E, Alice's basis vectors are,

$$\mathbf{v_1} = \begin{bmatrix} \frac{3}{2} \\ -1 \end{bmatrix}_E, \text{ and } \mathbf{v_2} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}_E$$



But it is important to note that, in her system, these vectors are

$$\mathbf{v_1} = \begin{bmatrix}1\\0\end{bmatrix}_A, \text{ and } \mathbf{v_2} = \begin{bmatrix}0\\1\end{bmatrix}_A$$

since this is exactly what it means to even be a basis vector. They are what *define* the meaning of [1,0] and [0,1] in her system.

Both systems look at the same vector in space, but assign it different coordinates. The point is, the grids both of us use are just artificial constructs. Space does not intrinsically have a grid.

So, if we want to translate from our standard basis to Alice's, we want to find some kind of function that maps $[\frac{3}{2}, -1]$ in our system to [1,0] in Alice's system, and similarly, [1,2] to [0,1]. We also note that [0,0] is exactly the same in both coordinate systems: we both agree on where the origin is, since scaling any vector by 0 should always give the same result, regardless of coordinate system.

Now, on the surface, this seems rather difficult. However, it might be easier to find the translation from the Alice's basis to our standard basis, where we're looking to map [1,0] to $[\frac{3}{2}, -1]$, and [0,1] to [1,2], and you might already see where we're going with this.

If we were given some vector, say $[1, -2]$, given in Alice's coordinates, A, how would we go about translating this into our standard coordinates, E? Well, the first coordinate scales Alice's first basis vector, and similarly to the second, and we know how to express those basis vectors in our coordinate system, so we have,

$$1\begin{bmatrix}\frac{3}{2}\\-1\end{bmatrix} - 2\begin{bmatrix}1\\2\end{bmatrix} = \begin{bmatrix}-2\\4\end{bmatrix}$$

So $[1, -2]_A$ is expressed as $[-2,4]_E$ in our standard coordinate system. But...

Doesn't this look familiar?

Recalling from a long way back (§33.2.1), we've already done this exact same thing before! It's matrix-vector multiplication, with the matrix containing Alice's basis vectors expressed in our coordinate system.

$$1\begin{bmatrix}\frac{3}{2}\\-1\end{bmatrix} - 2\begin{bmatrix}1\\2\end{bmatrix} = \begin{bmatrix}\frac{3}{2} & 1\\-1 & 2\end{bmatrix}\begin{bmatrix}1\\-2\end{bmatrix}$$

This mapping between bases is actually a linear transformation in and of itself, and we often label its associated matrix as $\mathbf{P}$.

$$\mathbf{P} = \begin{bmatrix}\frac{3}{2} & 1\\-1 & 2\end{bmatrix}$$

In general, this matrix is given by the basis vectors of the coordinate system being converted *from*, expressed in the coordinate system being converted *to* ($\mathbf{P}$ here converts from Alice's coordinates to ours, so we use Alice's basis vectors written in our language.)

This matrix, $\mathbf{P}$, is called a *change of basis* matrix. In this case, from A to E. To convert any vector given in F to its representation in E, we left multiply by this matrix.

Since we wanted to find the coordinates for the green vector in A, given that we know its coordinates in E, we simply take the inverse, $\mathbf{P}^{-1}$, and left multiply by that instead.

$$\mathbf{P}^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} \\ \frac{1}{4} & \frac{3}{8} \end{bmatrix}$$

And you can verify yourself that this matrix, multiplied with $[3{,}2]_E$, gives $[\frac{3}{2}{,}1]_A$.

From our geometric interpretation of matrix-vector multiplication from before, this is actually a very reasonable thing to do. The matrix containing the coordinates of the new basis vectors moves our basis vectors, $\hat{\mathbf{\imath}}$ and $\hat{\mathbf{\jmath}}$ – the things we think of as [1,0] and [0,1], – over to Alice's basis vectors – the things she thinks of as [1,0] and [0,1].

For example, if Alice was talking about a vector, say, [1,2], then multiplying [1,2] by $\mathbf{P}$ transforms our basis vectors over to Alice's, where the process of scaling then adding basis vectors by the coordinates [1,2] works in our favour, as we're effectively now working with Alice's basis vectors.

Geometrically, this matrix transforms our grid into Alice's, but numerically, it translates a vector in Alice's system into our system.

$$\mathbf{P} = \underbrace{\begin{bmatrix} a & b \\ c & d \end{bmatrix}}_{\substack{\text{Alice's basis vectors} \\ \text{in our coordinates}}} \qquad \mathbf{P} \overbrace{\begin{bmatrix} x_0 \\ y_0 \end{bmatrix}}^{\substack{\text{Vector in} \\ \text{Alice's coordinates}}} = \underbrace{\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}}_{\substack{\text{The same vector,} \\ \text{in our coordinates}}}$$

33.5.1 Transformations in Different Bases

Now we can translate vectors between bases, how about transformations? If we have the 90° clockwise rotation matrix,

$$\mathbf{U} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

how would Alice represent this same transformation in her own coordinate system? To be clear, we're trying to write down a matrix that takes Alice's grid, and rotates it 90° clockwise.

The columns of the matrix encode information about where our basis vectors, $\hat{\mathbf{\imath}}$ and $\hat{\mathbf{\jmath}}$ go, so just translating the columns into Alice's coordinates is not enough. That would just give a matrix that tells her where our basis vectors would land, written in her coordinate system.

She wants a matrix that gives where *her* basis vectors land, and it needs to describe those landing spots in her coordinate system as well.

Let's first consider what the rotation matrix does to a single specific vector, given in her coordinate system, say,

$$\begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

Since we don't know the rotation matrix in her system, let's first convert this vector into our coordinate system. We do this by using the change of basis matrix – the matrix containing her basis vectors in our coordinate system as columns.

$$\overbrace{\underbrace{\begin{bmatrix} 1 & \frac{3}{2} \\ 2 & -1 \end{bmatrix}}_{\text{Change of basis matrix, }\mathbf{P}} \begin{bmatrix} 3 \\ 4 \end{bmatrix}}^{\substack{\text{The same vector,} \\ \text{but in our language}}}$$

And now we have the vector in a form we can work with. The multiplication has been left unexpanded, but keep in mind that the whole right hand side represents a vector – the exact same vector as before, just described in our language.

Since we know the rotation matrix in our system, we can just multiply this whole thing by it:

$$\overbrace{\underbrace{\begin{bmatrix}0&1\\-1&0\end{bmatrix}}_{\substack{\text{Transformation matrix}\\\text{in our language}}}\begin{bmatrix}1&\frac{3}{2}\\2&-1\end{bmatrix}\begin{bmatrix}3\\4\end{bmatrix}}^{\substack{\text{Transformed vector}\\\text{in our language}}}$$

This tells us where the vector should go, but it's still in our language, so we convert it back into Alice's basis with the inverse change of basis matrix:

$$\overbrace{\underbrace{\begin{bmatrix}1&\frac{3}{2}\\2&-1\end{bmatrix}^{-1}}_{\substack{\text{Inverse change of}\\\text{basis matrix, }\mathbf{P}^{-1}}}\begin{bmatrix}0&1\\-1&0\end{bmatrix}\begin{bmatrix}1&\frac{3}{2}\\2&-1\end{bmatrix}\begin{bmatrix}3\\4\end{bmatrix}}^{\substack{\text{Transformed vector}\\\text{in Alice's language}}}$$

And we've just figured out where some specific vector, given in Alice's language, should go, under a 90° clockwise rotation. But, since the choice of vector was arbitrary, we've found the transformation we wanted!

We apply the change of basis matrix to get the vector into a workable form, then the transformation (which we know in our language), then the inverse change of basis matrix to translate back.

$$\underbrace{\begin{bmatrix}1&\frac{3}{2}\\2&-1\end{bmatrix}^{-1}\begin{bmatrix}0&1\\-1&0\end{bmatrix}\begin{bmatrix}1&\frac{3}{2}\\2&-1\end{bmatrix}}_{\substack{\text{Transformation matrix}\\\text{in Alice's language}}}\mathbf{v}$$

This composition of three matrices, together, gives the rotation matrix in Alice's coordinate system. It takes in a vector, in her language, and returns the transformed version of that vector, in her language.

We can represent all of this with the (mis)use of commutative diagrams:

$$\begin{tikzcd}
V_E \arrow[r, "\mathbf{R}"] & T(V_E) \\
V_A \arrow[u, "\mathbf{P}"', ""{name=L}] \arrow[r, "\mathbf{S}"] & T(V_A) \arrow[u, "\mathbf{P}"', ""{name=M}]
\arrow[from=L, to=M, Rightarrow, "T" description]
\end{tikzcd}$$

where the transformation, $T : V \to V$ is acting on V, with the choice of basis indicated using subscripts. Note that the change of basis matrix works the same before and after the transformation – it still translates between Alice's system and ours. The transformation of space, T (notice that this is not written in bold, as it is a transformation not tied to a specific matrix), can be represented as the matrices, $\mathbf{R}$ and $\mathbf{S}$, specific to the bases E and A.

Note, since we're only dealing with one transformation that is pretty obviously an endomorphism, and only one vector space is being considered, the above diagram could be abbreviated to

$$\begin{tikzcd}
E \arrow[r, "R"] & E' \\
A \arrow[u, "P"'] \arrow[r, "S"] & A' \arrow[u, "P"']
\end{tikzcd}$$

to save time. Some of the diagrams in this section will contain additional detail to aid my explanations, but by no means do you have to include every little extraneous detail when using these diagrams yourself.

Some people define the change of basis matrix to be the opposite way as is defined here, with $\mathbf{P}$ being the change of basis from E to A (so what we would call $\mathbf{P}^{-1}$), but as long as you are consistent with your arrows, the diagram makes everything clear.

On the top diagram, we want S, which directly transforms V_A to $T(V_A)$. An alternative route there, is to take P, then R, then to go along the second P arrow, but against the direction it is pointing, which indicates we should take an inverse of the matrix representing P. So, in terms of matrices, $\mathbf{S} = \mathbf{P}^{-1}\mathbf{R}\mathbf{P}$ (reading right to left, as per function notation), matching the result from before.

Now, let's say we have another friend, Bob, who uses yet another coordinate system, distinct from both Alice's and ours. How might Alice and Bob communicate? How do we find a change of basis from Bob to Alice, and vice versa?

We can use our standard basis as an intermediary:

$$\begin{tikzcd}
A \arrow[r, "P"] \arrow[d, "R"'] & B \arrow[dl, "S"] \\
E &
\end{tikzcd}$$

Where R and S are the change of basis transformations from Alice's and Bob's systems to ours, as found earlier. We want P here, so we travel along R, then backwards along S, so $P = S^{-1}R$.

How would Bob give a transformation to Alice?

$$\begin{tikzcd}
A \arrow[r, "T"] \arrow[d, "R"] & A' \arrow[d, "R"'] \\
E & E' \\
B \arrow[u, "S"'] \arrow[r, "U"] \arrow[uu, bend left=60, "P"] & B' \arrow[u, "S"] \arrow[uu, bend right=60, "P"']
\end{tikzcd}$$

Following the arrows, we have,

$$U = S^{-1}RTR^{-1}S$$

or,

$$U = P^{-1}TP$$

In general, when we give a transformation between vector spaces, $T : V \to W$, we have to be careful when turning this transformation into a matrix. V and W, being abstract spaces, don't intrinsically have grids mapping them out – we have to assign basis vectors to each.

For example, let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be a 90° clockwise rotation. The matrix for $\mathbb{R}^2_E \to \mathbb{R}^2_E$ is just the rotation matrix we're familiar with, and we found the matrix for $\mathbb{R}^2_A \to \mathbb{R}^2_A$ earlier. But a matrix giving the rotation from $\mathbb{R}^2_B \to \mathbb{R}^2_A$ is an equally valid representation of that same transformation. In the diagram above, this matrix could be given as $TR^{-1}S$, or $R^{-1}SU$.

Now, for some more terminology. Going back to the first diagram,

$$\begin{tikzcd}
V_E \arrow[rr, "\mathbf{R}"] & & T(V_E) \\
{} \arrow[rr, Rightarrow, "T" description] & & {} \\
V_A \arrow[uu, "\mathbf{P}"'] \arrow[rr, "\mathbf{S}"] & & T(V_A) \arrow[uu, "\mathbf{P}"]
\end{tikzcd}$$

Because $\mathbf{R}$ and $\mathbf{S}$ represent the same transformation, T, within the same space, V, just with respect to different bases, we call them *similar* matrices. In general, two $n \times n$ matrices, $\mathbf{A}$ and $\mathbf{B}$ are similar if you can write $\mathbf{B} = \mathbf{P}^{-1}\mathbf{AP}$ for some (usually change of basis) matrix $\mathbf{P}$. Two similar matrices must be square.

Recalling from group theory, if we think about $\mathbf{R}$ and $\mathbf{S}$ as elements of the general linear group, $GL_n(\mathbb{R})$, we can see that matrix similarity is just specific case for the conjugacy equivalence relation (§12.6.2).

More generally, for possibly rectangular $m \times n$ matrices, we have an analogous concept of *equivalence* (this equivalence is a type of equivalence relation, if you have done binary relations). Two rectangular matrices, $\mathbf{A}$ and $\mathbf{B}$ are equivalent if you can write $\mathbf{B} = \mathbf{QAP}$ for two invertible matrices $\mathbf{P}$ and $\mathbf{Q}$ (the change of basis matrices for each of the pairs of coordinate systems for each space).

On a diagram, this would be,

$$\begin{tikzcd}
V_A \arrow[rr, "\mathbf{R}"] \arrow[dd, "\mathbf{P}"] & & W_{A'} \arrow[dd, "\mathbf{Q}"] \\
{} \arrow[rr, Rightarrow, "T" description] & & {} \\
V_B \arrow[rr, "\mathbf{S}"] & & W_{B'}
\end{tikzcd}$$

Here, V and W are vector spaces of different dimension, with subscripts indicating choice of basis. There are 4 bases in play here, as Alice and Bob each choose their own bases for both V and W. Note that, unlike the previous diagram, $\mathbf{P} \neq \mathbf{Q}$, since they are change of basis matrices in different spaces. Here, R and S both represent the same transformation of space, then they are equivalent, as you could write $\mathbf{R} = \mathbf{Q}^{-1}\mathbf{SP}$. As a side note, to find $\mathbf{P}$, you would do the same procedure as a few diagrams ago, and use the standard basis as an intermediary, but the diagram was already getting cluttered enough, so this is left as an exercise for the reader.

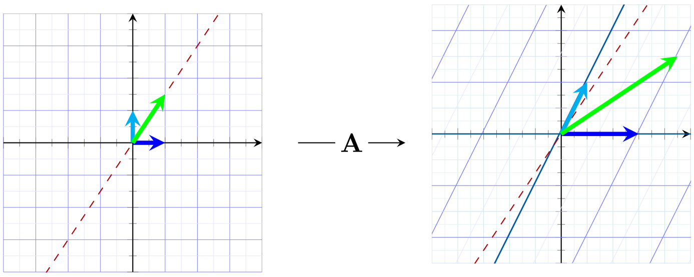
In general, similar matrices are equivalent, but equivalent matrices are not necessarily similar.

33.5.2 Eigenvectors

Consider the linear transformation given by the matrix,

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix}$$

and how it acts on some arbitrary vector. In particular, think about the span of that vector.



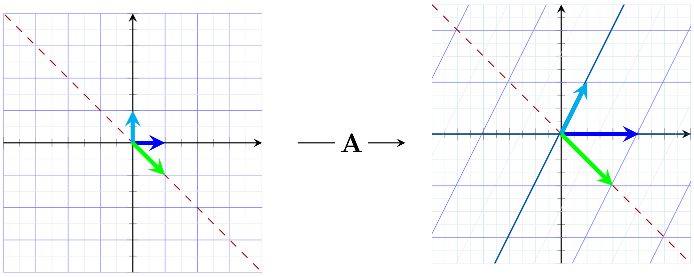
For most vectors in the plane, they get knocked off of their span during the transformation. But some special vectors do remain on their own span, meaning the transformation has no rotational effect on that vector, only scaling it by some amount.

As you might have guessed, such a vector is called an *eigenvector* of the transformation, and the amount by which it is scaled is its associated *eigenvalue*.

For the transformation above, $\hat{\mathbf{i}}$ is one such vector. The span of $\hat{\mathbf{i}}$ is just the horizontal axis, and the image of $\hat{\mathbf{i}}$ clearly remains on that axis after the transformation. From the matrix, we can see that $\hat{\mathbf{i}}$ lands on [3,0], so it is scaled by a factor of 3. We say that $\hat{\mathbf{i}} = [1{,}0]$ is an eigenvector of $\mathbf{A}$, with an eigenvalue of 3.

Furthermore, due to linearity, *any* vector on the horizontal axis is also similarly scaled by a factor of 3, also remaining on their own spans.

But there are more, slightly less obvious, eigenvectors to this particular transformation:



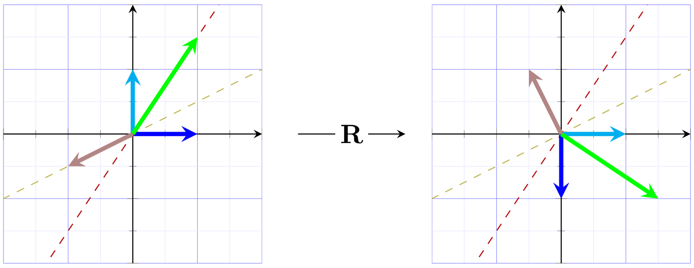
This vector, [1,−1] lands on [2,−2], being scaled by a factor of 2, so [1,−1] is also an eigenvector of $\mathbf{A}$, with eigenvalue 2. And again, due to linearity, any vector on that line will also be an eigenvector with eigenvalue 2.

For this transformation, those are all the eigenvectors there are. Every other vector in the plane will get moved off of their spans under this transformation.

A matrix may also have more or fewer eigenvectors as well; for instance, any rotation matrix

$$\mathbf{R} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

will move every vector off of its span,

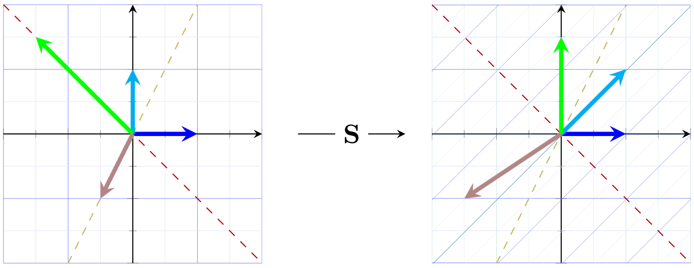


so this transformation has zero eigenvectors.

On the other hand, this shear,

$$\mathbf{S} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

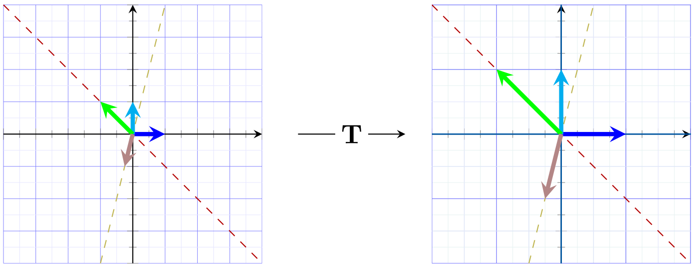
has every vector on the horizontal axis as an eigenvector, with eigenvalue 1, as they are unchanged by the transformation.



Every other vector is moved off of its span, so this transformation just has a single line of eigenvectors – namely, the horizontal axis – with eigenvalue 1.

Eigenvalues don't have to be unique either – there can be multiple distinct lines of eigenvectors that share the same eigenvalue. For instance, we have,

$$\mathbf{T} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$



This scaling matrix just stretches every vector in the plane by a factor of 2, so *every* vector is an eigenvector of this transformation, all with the same eigenvalue of 2.

By definition, the effect of a transformation, $\mathbf{A}$, on an eigenvector, $\mathbf{v}$, is just to scale it by some amount, λ, the eigenvalue. We can write this definition symbolically, as,

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

The left side is matrix-vector multiplication, while the right is scalar multiplication, so we tend to do some rearranging of this expression, by writing the right side as a matrix-vector product.

We want to scale $\mathbf{v}$ by a scalar, λ. The columns of the desired matrix need to scale each basis vector by λ, so this matrix will have λ across the diagonal, and zeros everywhere else.

$$\begin{bmatrix} \lambda & 0 & 0 & \cdots & 0 \\ 0 & \lambda & 0 & \cdots & 0 \\ 0 & 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda \end{bmatrix}$$

Factoring out the λ, this is just the identity matrix.

$$\mathbf{A}\mathbf{v} = (\lambda\mathbf{I})\mathbf{v}$$

And now both sides are a matrix-vector product. We can then subtract the right side, and factor out the $\mathbf{v}$,

$$\mathbf{A}\mathbf{v} - (\lambda\mathbf{I})\mathbf{v} = \mathbf{0}$$

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$$

The expression inside the bracket is just a matrix – the original transformation matrix, $\mathbf{A}$, but with a λ being subtracted from the diagonal, and would look something like,

$$\begin{bmatrix} 3-\lambda & 1 & 4 \\ 1 & 5-\lambda & 9 \\ 2 & 6 & 5-\lambda \end{bmatrix}$$

We're looking for a vector, $\mathbf{v}$, such that this new matrix maps $\mathbf{v}$ to the zero vector.

If $\mathbf{v} = \mathbf{0}$, then this is trivially true and isn't particularly helpful, so we're looking for non-zero solutions for $\mathbf{v}$ – a non-zero eigenvector.

Recalling terminology from earlier (§33.2.5), we're looking for a non-zero vector that lies in the null space of $(\mathbf{A} - \lambda\mathbf{I})$. If the null space of $(\mathbf{A} - \lambda\mathbf{I})$ is non-empty, i.e., there exists an non-zero eigenvector, it follows that $(\mathbf{A} - \lambda\mathbf{I})$ cannot be full rank, and therefore has a zero determinant (if this doesn't make immediate sense, reread the definitions of null space and rank, and take a few moments to consider what it means for a transformation to have a non-empty null space).

In other words, the only way for a non-zero vector to be mapped to the origin, is if the transformation collapses space down into a lower dimension, corresponding to a zero determinant. That is, the goal is to solve the equation

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

for λ. This is called the *characteristic equation* of the matrix, and the left side by itself is called the *characteristic polynomial.*

For example, earlier, we had,

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix} \\ \mathbf{A} - \lambda\mathbf{I} &= \begin{bmatrix} 3-\lambda & 1 \\ 0 & 2-\lambda \end{bmatrix} \\ \det(\mathbf{A} - \lambda\mathbf{I}) &= (3-\lambda)(2-\lambda) - (1\cdot 0) \\ \det(\mathbf{A} - \lambda\mathbf{I}) &= 0 \\ 0 &= (3-\lambda)(2-\lambda) \\ \lambda &= 3{,}2 \end{aligned}$$

matching the results from before. Then, to find the actual eigenvectors, multiply the modified matrix by an arbitrary vector, and set it equal to the zero vector.

For $\lambda = 2$, we have

$$\begin{aligned} \begin{bmatrix} 3-2 & 1 \\ 0 & 2-2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ \begin{bmatrix} x+y \\ 0 \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ x+y &= 0 \\ y &= -x \end{aligned}$$

So any vector of the form,

$$\begin{bmatrix} t \\ -t \end{bmatrix}$$

is an eigenvector with eigenvalue 2. We generally just pick one single eigenvector as a representative for this entire line, so choosing $t = 1$ yields [1,−1], as before.

For a transformation which has multiple eigenvectors with the same eigenvalue, you'll find that the simultaneous equations in the final step will have multiple solutions, corresponding to the multiple eigenvectors.

Doing the same process for the rotation matrix,

$$\begin{aligned} \mathbf{R} - \lambda\mathbf{I} &= \begin{bmatrix} 0-\lambda & 1 \\ -1 & 0-\lambda \end{bmatrix} \\ \det(\mathbf{R} - \lambda\mathbf{I}) &= (-\lambda)(-\lambda) - (1 \cdot (-1)) \\ \det(\mathbf{R} - \lambda\mathbf{I}) &= 0 \\ \lambda^2 + 1 &= 0 \\ \lambda &= \pm i \end{aligned}$$

we find that there are no real eigenvalues for this transformation. The eigenvalues of $\pm i$ correspond to the fact that multiplying by i represents a 90° rotation in the complex plane, and the magnitude of the eigenvalues being 1 corresponds to the fact that vectors aren't scaled under this transformation. In general, imaginary components of eigenvalues correspond to some kind of rotation. We can still solve for eigenvectors, but they will have complex components.

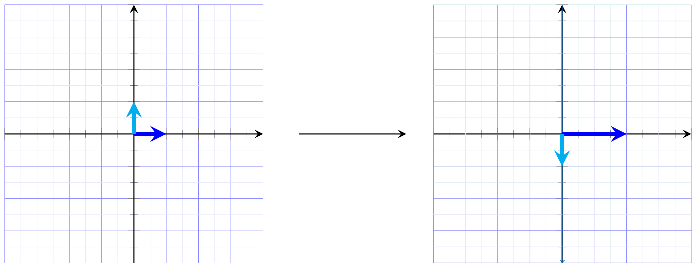
As one very basic application of eigenvectors, if you can find an eigenvector of a 3D rotation, you've found the axis of rotation – and it's much easier to think of rotations in 3D as an angle around an axis, rather than the entire 3×3 rotation matrix.

This is a common theme throughout linear algebra – with any linear transformation given as a matrix, we can interpret what it is doing by looking at its columns and seeing where the basis vectors are mapped. But this puts a lot of emphasis on coordinate systems – another way, less dependent on coordinate systems, is to look at the eigenvectors and eigenvalues.

Two similar matrices – two matrices representing the same linear transformations, but in different coordinate systems – will have the same characteristic equation, and the same eigenvalues. Changing the coordinate system doesn't change the eigenvalues of a transformation – regardless of how you label space, eigenvectors are scaled the same way.

For another, much more general application, consider what happens if our basis vectors both happen to be eigenvectors. Let's start in the canonical coordinate system, and say we have a linear transformation such that,

$$\hat{\mathbf{i}} \mapsto \begin{bmatrix} 2 \\ 0 \end{bmatrix} = 2\hat{\mathbf{i}} \qquad \hat{\mathbf{j}} \mapsto \begin{bmatrix} 0 \\ -1 \end{bmatrix} = -1\hat{\mathbf{j}}$$



So the matrix associated with that transformation would be,

$$\begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix}$$

Notice how the eigenvalues of the basis vectors lie along the diagonal of the matrix, and every other entry is zero. Any matrix with this property is called a *diagonal matrix*, and we've met one before – the identity matrix is a diagonal matrix.

The way to interpret a diagonal matrix, is that all the basis vectors are eigenvectors, with the eigenvalues written along the diagonals.

There are many reasons why diagonal matrices are much nicer to work with. One application is in taking powers of matrices, or equivalently, applying a transformation to a vector many times. Since a diagonal matrix only scales each basis vector by some eigenvalue, applying that matrix to a vector n times, just means you scale each basis vector by the eigenvalue to the power of n.

$$\begin{aligned}
\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} &= \begin{bmatrix} 2x \\ 3y \end{bmatrix} \\
\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} &= \begin{bmatrix} 2^2x \\ 3^2y \end{bmatrix} \\
\underbrace{\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \cdots \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}}_{100} \begin{bmatrix} x \\ y \end{bmatrix} &= \begin{bmatrix} 2^{100}x \\ 3^{100}y \end{bmatrix} \\
&= \begin{bmatrix} 2^{100} & 0 \\ 0 & 3^{100} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}
\end{aligned}$$

Just looking at the transformation overall, we can write an exceedingly simple formula for the nth power of a diagonal matrix:

$$\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}^n = \begin{bmatrix} a^n & 0 \\ 0 & b^n \end{bmatrix}$$

In contrast, try calculating the 100th power of a non-diagonal matrix. There is no simple pattern to find.

For a few more nice properties of diagonal matrices, the determinant of a diagonal matrix is just the product of the diagonal. This is because each entry on the diagonal tells us how much the basis vector is scaled in that direction, so the product of all of these entries gives us how much measure is scaled overall.

Of course, all of this is only useful when the matrix we're working with is diagonal – when our basis vectors just happen to both be eigenvectors.

However, if your transformation has a lot of eigenvectors, enough so we can choose a set that spans the space the transformation is acting on, then we could use a change of basis matrix to change those eigenvectors to be our basis.

For the matrix earlier,

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix}$$

we found two eigenvectors,

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \text{and } \mathbf{v}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

with eigenvalues $\lambda_1 = 3$, and $\lambda_2 = 2$, respectively.

We use the eigenvectors as the columns of a change of basis matrix, and change the transformation matrix into our new basis, as before.

$$\begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix}^{-1} \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix}$$

Because we've chosen the basis vectors to be eigenvectors, we know that resulting matrix will be a diagonal matrix, with the corresponding eigenvalues along the diagonal, without even doing any calculations.

$$\begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$$

This is because we're now working in a basis, where the effect of this specific transformation on these basis vectors is just to scale them by these eigenvalues.

A basis where every basis vector is an eigenvector is called an *eigenbasis*, and this process of changing a matrix to an eigenbasis is called *diagonalisation.*

If we wanted to calculate the 100th power of **A**, we could change to an eigenbasis, calculate the power there, using our simple formula, then change back.

But not every transformation admits an eigenbasis. The shear we saw earlier, for example, only has a single line of eigenvectors, which isn't enough to span all of 2D space.

33.6 Abstract Vector Spaces

At the very beginning of this document, we asked the question, "What are vectors?".

Is a vector fundamentally an arrow, which we can describe with coordinates, or are they fundamentally lists of numbers, which just happen to have a nice visualisation. Or are both of these views manifestations of something deeper?

Thinking of vectors as primarily being lists of numbers makes them very straightforward and intuitive. Things like four-dimensional or n-dimensional vectors are very easy to think of in this context – they're just longer lists of numbers. Working in 2 dimensions is just as easy as working in 200. Otherwise, four-dimensional space is just some strange, geometric idea that can't be easily visualised or described.

On the other hand, as you get more used to working in linear algebra – particularly with changing your basis – you'll find that many concepts are inherently to do with a space that exists independently from any choice of coordinates. Core ideas like determinants and eigenvectors don't care about the coordinate systems. The determinant tells you how much a transformation scales area, or volume, or measure, and eigenvectors are the ones which stay on their own span under a transformation. Both of these ideas are inherently spatial, and their algebraic equivalents seem completely arbitrary when seen alone.

If vectors are neither arrows nor lists of numbers, and are some abstract concept to do with space, we still haven't really answered the question. We've just stated things that aren't fundamental to a vector.

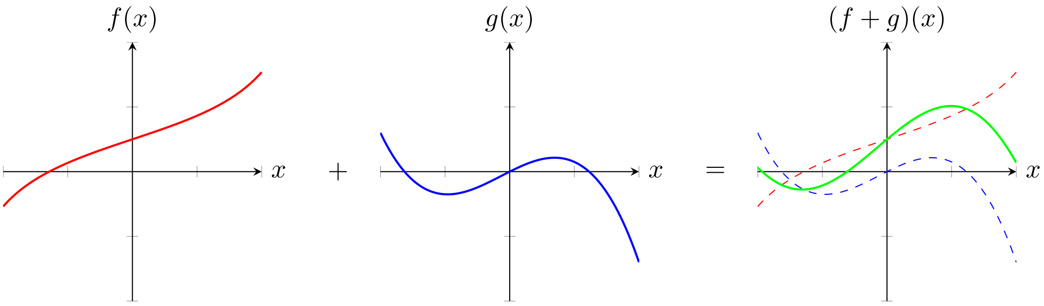
To answer this question more deeply, let's discuss something that is neither arrows, nor lists of numbers: functions.

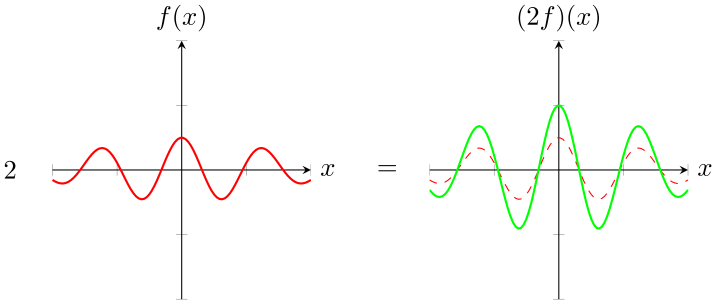
In the same way we can add two vectors together and multiply them by a scalar,

$$\begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} a+c \\ b+d \end{bmatrix}$$

$$2\begin{bmatrix} e \\ f \end{bmatrix} = \begin{bmatrix} 2e \\ 2f \end{bmatrix}$$

we can similarly add two functions, f and g to get a new function, $(f+g)$, or multiply a function by a number to scale it, $2f(x) = (2f)(x)$:





The value of the sum function, $(f+g)$ at any given input x is the sum of the values of $f(x)$ and $g(x)$, so $(f+g)(x) = f(x) + g(x)$.

This is similar to adding vectors together, coordinate by coordinate, just that, we now have uncountably infinitely many coordinates – one for each possible input x.

For scaling a function, we just multiply each output by the scalar, just as we do for vector components, though again, here we have uncountably infinitely many coordinates.

Functions appear to be some kind of infinite-dimensional vector-ish thing. A reasonable question is, what kind of transformations of functions are there that are linear? What does it even mean for such a function to be linear?

Although functions don't look like vectors, we can still use the symbolic definition of linearity from before (§33.2.1). In the context of functions, these transformations are called *operators* instead, but they're really the same thing.

One example of an operator you'll be familiar with, is the derivative – it's something that transforms one function, into another.

If you add two functions, then take the derivative, it's the same as first taking the derivative of the two functions, then adding them:

$$L(\mathbf{u} + \mathbf{v}) = L(\mathbf{u}) + L(\mathbf{v})$$

$$\Updownarrow$$

$$\frac{d}{dx}(x^3 + x^2) = \frac{d}{dx}(x^3) + \frac{d}{dx}(x^2)$$

Similarly, scaling a function, then taking the derivative is the same as taking the derivative, then scaling the result:

$$L(c\mathbf{v}) = cL(\mathbf{v})$$

$$\Updownarrow$$

$$\frac{d}{dx}(cx^2) = c\frac{d}{dx}(x^2)$$

We can see that the differential operator is linear – in fact, the linearity requirements are exactly the sum rule and the constant factor rule from differential calculus.

One of the most important consequences of linearity, is that a linear transformation is completely described by its action on a basis. Since any vector can be expressed by scaling and adding the basis vectors in some way, finding the image of a vector under a transformation is simplified down to finding the image of the basis vectors. This is just as true for functions, as it is for arrows, or lists of numbers.

Because the differential operator is a linear transformation, we should be able to express it as a matrix. For now, let us restrict our space to the space of polynomials, so each element in our space is a polynomial with finitely many terms, but the whole space includes polynomials of arbitrarily large degree.

First, we need to pick a basis for the space of functions. Since polynomials are already written as a linear combination of powers of x, we can just choose monomials in x to be our basis.

So, the polynomial, $1 + 5x + 4x^3$ would be written as,

$$\begin{bmatrix} 1 \\ 5 \\ 0 \\ 4 \\ \vdots \end{bmatrix}$$

with infinitely many zeros following on. You can read this as 1 times the first basis function, which is just 1, plus 5 times the next basis function, x, plus 0 times x^2, 4 times x^3, plus zero times all the other basis functions.

Since polynomials only have finitely many terms, every polynomial will be represented by a finite string of numbers at the top of the vector, followed by infinitely many zeros.

In this coordinate system, the derivative is described by the infinite matrix,

$$\frac{d}{dx} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 2 & 0 & \cdots \\ 0 & 0 & 0 & 3 & \cdots \\ 0 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

which is mostly full of zeros, but has the natural numbers running down the superdiagonal. To get a feel of why it works, cut the matrix off at a finite point, and multiply some vector by it.

This is all possible because the derivative operator is linear. If you had wanted to construct this matrix yourself, you could consider what the derivative operator does to each basis function, and put the coordinates of the results into each column.

For example, the derivative of the first basis function, 1, is 0, which has a vector representation of $[0{,}0{,}0,\cdots]$, corresponding to the first column of the matrix. Then, the derivative of the next basis function, x, is 1, which is $[1{,}0{,}0,\cdots]$, then x^2 is $2x$, which is $[0{,}2{,}0,\cdots]$.

So, it turns out that taking a derivative and matrix-vector multiplication are really members of the same family. In fact, the vast majority of concepts in linear algebra have direction analogues with respect to functions:

Linear transformations ⇔ Linear operators

Dot products ⇔ Inner products

Eigenvectors ⇔ Eigenfunctions

There are a lot of vector-ish things in maths, all of which have their own analogues to these concepts. The answerwe gave to “what is a vector” before was, “anything where we have some kind of notion of scaling and adding”, whether that’s a set of arrows in space or lists of numbers. As long as those two requiremnts hold, all of the tools of linear algebra regarding vectors – linear transformations, eigenvectors, determinants – will apply.

If you’re a mathematician developing new theory, you want all of your new definitions to fully apply to all of these vector-ish things, and not just some specific sub-cases. So, we use the axioms we defined before, as an interface between different vector spaces. You, as the mathematician, never have to think about all the vector spaces that could possibly exist – you just have to prove your results in terms of these axioms, and anyone who can prove that their new crazy space follows those axioms, can apply your results, even if you’ve never even thought about their space before. As a consequence, all of our results tend to be expressed extremely abstractly – only in terms of the axioms, rather than on a specific view, like arrows, lists of numbers, or functions.

So, the mathematician’s answer to “what is a vector?” is just to ignore the question. In modern theory, it doesn’t matter what the vectors are themselves – it’s the fact that they obey the vector space axioms that matter.

On the topic of abstraction, what’s really the difference between a linear transformation and transformations between other algebraic structures? And why do we have to work with scalars from fields? Why not rings? Or groups? Or even things that aren’t sets? These things do generally have names – for example, replacing the field of scalars with a ring gives a structure called a *module.*

It turns out that modules also happen to be a generalisation of abelian groups, as abelian groups are exactly the modules over the ring of integers, and in fact, we will explore some theory behind them later on in this chapter. All these structures are wonderfully interlinked, and the theories behind them are unified under the mathematical field of abstract or universal algebra.

33.7 Exercises

These questions are very roughly ordered in difficulty. For many of these questions, it may be helpful to keep the epigraph of this chapter in mind.

1. Give an example of a matrix $\mathbf{A}$ such that,
   (a) $\mathbf{A}$ is nonsingular.
   (b) $\mathbf{A}$ is diagonalisable and singular.
   (c) $\mathbf{A}$ is nondiagonalisable and singular.
2. Explain (briefly) why 5 vectors in $\mathbb{R}^4$ cannot be linearly independent.
3. Prove that, if $\mathbf{u}$ and $\mathbf{v}$ are linearly independent, then $\mathbf{u}$, $\mathbf{v}+\mathbf{u}$ are linearly independent.
4. Prove that, if $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ are linearly independent, then $\mathbf{u}-\mathbf{v}$, $\mathbf{v}-\mathbf{w}$, and $\mathbf{w}-\mathbf{u}$ are linearly independent.
5. Prove that the zero vector is linearly dependent with every other vector.
6. Give an example a pair of vectors in $\mathbb{R}^3$ which are,
   (a) Linearly dependent.
   (b) Linearly independent.
   (c) In both cases, extend the set of vectors to a basis of $\mathbb{R}^3$.
7. Explain (briefly) why non-square matrices do not have determinants.
8. Let K be a field with zero 0_K and unity 1_K, and let $(V, +, \cdot)$ be a vector space over K.
   (a) Prove that for all $\mathbf{v} \in V$, $-\mathbf{v} = -1_K \cdot \mathbf{v}$. That is, the additive inverse of a vector is its negative.
   (b) Prove that this inverse is unique.
   (c) Prove that for all $\mathbf{v} \in V$, $0_K \cdot \mathbf{v} = \mathbf{0}$. That is, the additive identity of the field is the annihilator of the vector space.
   (d) Prove that the zero vector is unique.
9. Let $M \in GL_n(\mathbb{R})$, and suppose $\det M < 0$. Prove that M^3 cannot be the identity matrix.
10. Let $T : \mathbb{R}^3 \to \mathbb{R}^3$ be a reflection in the plane $x + 4y - 2z = 0$. Prove that all eigenvalues of T are real. (Hint: consider the eigenvalues of T^2.)
11. Let $A \in GL_n(\mathbb{R})$, and suppose that for every vector $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{A}\mathbf{v}$ and $\mathbf{v}$ are linearly dependent. Prove that $\mathbf{A} = k\mathbf{I}_n$ for some $k \in \mathbb{R}$.
12. Prove that if W is a subspace of V, then $\dim W \leq \dim V$.
13. Let $V \subseteq \mathbb{R}^4$ be the set of vectors perpendicular to the vectors,
$$\mathbf{a} = \begin{bmatrix} 1 \\ 2 \\ 0 \\ -1 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} -2 \\ 0 \\ 1 \\ 2 \end{bmatrix}$$
    That is, $V = \left\{ \mathbf{x} \in \mathbb{R}^4 : \mathbf{a}^\top \cdot \mathbf{x} = 0, \mathbf{b}^\top \cdot \mathbf{x} = 0, \mathbf{c}^\top \cdot \mathbf{x} = 0 \right\}$.
   (a) Prove that V is a subspace of $\mathbb{R}^4$. (Hint: recall the relation between dot products and matrix-vector multiplication.)

(b) Find a basis for V, and hence find $\dim V$.

14. Consider the vector space $V = \mathbb{R}[x]_{\leq 2}$ of polynomials with real coefficients of at most quadratic degree.

    (a) Prove that the polynomials 1, x, x^2 form a basis of V.

    (b) Prove that the polynomials $1 - x^2$, $x^2 - x$, $x^2 - 2x$ form a basis of V.

    (c) Find the change of basis matrices to convert between these bases in both directions.

    (d) Consider the transformation $T : V \to V$ defined by $T(f(x)) = f(x+1)$. Prove that T is linear.

    (e) Find matrices to represent T in the two previous bases.

15. Let $\mathbf{A}$ be a 4×4 matrix, and suppose that the diagonal entries of $\mathbf{A}$ are zero, and every other entry is an odd integer. Prove that $\mathbf{A}$ is nonsingular. (Hint: consider parities, or equivalently, work in $\mathbb{F}_2$.)

16. Let $\mathbf{A} \in GL_4(\mathbb{C})$, and suppose that the diagonal entries of $\mathbf{A}$ are zero, and every other entry is the imaginary unit, i. Find all eigenvalues of $\mathbf{A}$, and determine their algebraic and geometric multiplicities.

17. Let V be a vector space over a field, K, and let $U \subseteq V$ be a non-empty subset of V. Prove that if for all $\mathbf{u}$,$\mathbf{v} \in U$ and all $k \in K$, $\mathbf{u} + k\mathbf{v} \in U$, then U is also a vector space.

18. Let U and V be vector spaces over a field, K, and let $T : U \to V$ be a linear map.

    (a) Prove that T cannot be injective if $\dim U > \dim V$.

    (b) Prove that T is injective if and only if $\operatorname{null} T = 0$.

19. Let $T : V \to V$ be an endomorphism, and let A and B be distinct bases of V. Using geometric considerations, explain why the matrices representing T in bases A and B must have the same eigenvalues.

20. Let $B = \{\mathbf{v}_1,\mathbf{v}_2,\mathbf{v}_3\}$ be a basis of a vector space V over a field K. Prove that every vector $\mathbf{v} \in V$ can be uniquely represented as a linear combination of the vectors in B. That is,

$$\forall \mathbf{v} \in V : \exists! c_1,c_2,c_3 : v = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3$$

21. Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be an idempotent transformation, so $T = T \circ T$. Prove that $\ker T$ and $\operatorname{im} T$ are complementary subspaces such that their direct sum is $\mathbb{R}^n$. That is,

$$\ker T \oplus \operatorname{im} T = \mathbb{R}^n$$

22. Let V be a n-dimensional vector space over a field, K, where n is finite, and let $\operatorname{End}(V)$ be the space of linear transformations from V to itself.

    (a) Prove that $\operatorname{End}(V)$ is a vector space.

    Denote the direct sum of V with itself n times as $V^{\oplus n}$. That is,

$$V^{\oplus n} = \underbrace{V \oplus V \oplus \cdots \oplus V}_{n}$$

    Let $\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_n$ be a basis of V, and consider the map $\phi : \operatorname{End}(V) \to V^{\oplus n}$ defined by

$$\phi(T) = (T(\mathbf{v}_1),T(\mathbf{v}_2),\ldots,T(\mathbf{v}_n))$$

    (b) Prove that ϕ is a linear transformation.

(c) Prove that $\ker\phi$ is trivial.

(d) Prove that $\operatorname{im}\phi = V^{\oplus n}$.

(e) Hence deduce that $\operatorname{End}(V)$ and $V^{\oplus n}$ are isomorphic.

23. Let $c_1,c_2,\ldots,c_n$ be distinct real numbers.

(a) Prove that $e^{c_1 x},e^{c_2 x},\ldots e^{c_n x}$ are linearly independent over $\mathbb{R}$

(b) Let $C[-1,1]$ be the vector space of continuous real-valued functions defined on the interval $[-1,1]$. Prove that

$$V = \left\{f(x) \in C[-1,1] : f(x) = ae^{x} + be^{2x} + ce^{3x}\right\}$$

is a subspace of $C[-1,1]$.

24. Let $\mathcal{F}[-\pi,\pi]$ be the vector space of all real-valued functions defined on the interval $[-\pi,\pi]$.

(a) Prove that $\cos x$ and $\sin x$ are linearly independent in $\mathcal{F}[-\pi,\pi]$.

Let the class of functions $f_{(-,-)} : \mathbb{R}^2 \to \mathcal{F}[-\pi,\pi]$ be defined by,

$$f_{(a,b)}(x) = a\cos x + b\sin x$$

(b) Prove that $f_{(a,b)}$ is a linear map for any choice of $(a,b) \in \mathbb{R}^2$.

(c) Prove that

$$\operatorname{im} f = \left\{f_{(a,b)}(x) \in \mathcal{F}(-\pi,\pi)\right\}$$

is a vector space.

(d) Prove that $\cos x$ and $\sin x$ form a basis of $\operatorname{im} f$

(e) Prove that $\operatorname{null} f = 0$, and hence deduce that $\mathbb{R}^2$ and $\operatorname{im} f$ are isomorphic.

(f) Consider the map $g : \operatorname{im} f \to \operatorname{im} f$ defined by

$$g(f(x)) = \frac{d}{dx} f(x)$$

Prove that g is linear.

(g) Find a matrix representation of g with respect to the basis $\{\sin x, \cos x\}$

33.8 Jordan Canonical Form

In this section, we will take V to be an n-dimensional vector space over a field K. We will take $T : V \to V$ to be an linear map from V to V (an *endomorphism*) and $\mathbf{A}$ will be the matrix representing T with respect to a fixed ordered basis $E = (\mathbf{e}_i)_{i=1}^n$.

Our goal is to find a new basis $e = (\mathbf{e}_i)_{i=1}^n$ such that the matrix of T with respect to this new basis is as simple as possible (or equivalently, a change of basis matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{AP}$ is as simple as possible). One particularly simple form of a matrix is that of a diagonal matrix, but as mentioned previously, not every linear transformation admits a diagonal matrix representation.

Theorem 33.8.1. *Let $T : V \to V$ be a linear map. Then, the matrix of T is diagonalisable if and only if V has an eigenbasis.*

However, if K is $\mathbb{C}$ (or is a field extension of $\mathbb{C}$), then every matrix $\mathbf{A}$ is similar to a form that is almost as good as diagonal: the *Jordan canonical form* or *Jordan normal form.*

33.8.1 Generalised Eigenspaces

Theorem 33.8.2. *Let $(\lambda_i)_{i=1}^r$ be distinct eigenvalues of $T : V \to V$, and let $(\mathbf{v}_i)_{i=1}^r$ be corresponding eigenvectors – that is, $T(\mathbf{v}_i) = \lambda_i \mathbf{v}_i$ for all $1 \leq i \leq r$. Then, $(\mathbf{v}_i)_{i=1}^r$ are linearly independent.*

Theorem 33.8.3. *Let $\mathbf{A} \in K^{n\times n}$ be a $n \times n$ matrix over K. Then, there is some non-zero polynomial $p \in K[x]$ of degree at most n^2 such that $p(\mathbf{A}) = \mathbf{0}_n$.*

A polynomial is *monic* if the coefficient of the highest degree term is 1.

Theorem 33.8.4. *Let $\mathbf{A} \in K^{n\times n}$ represent the linear transformation $T : V \to V$. Then,*

- *There is a unique monic non-zero polynomial p with minimal degree and coefficients in K such that $p(\mathbf{A}) = \mathbf{0}_n$.*
- *If q is any polynomial with $q(\mathbf{A}) = \mathbf{0}_n$, then p divides q.*

This unique polynomial is called the *minimal polynomial* of $\mathbf{A}$, and is denoted $\mu_{\mathbf{A}}$.

Theorem 33.8.5. *Similar matrices have the same minimal polynomial.*

Theorem 33.8.6. *Let $\mathbf{D}$ be a diagonal matrix with distinct diagonal entries $(\delta_i)_{i=1}^r$. Then,*

$$\begin{aligned}\mu_{\mathbf{D}}(x) &= \prod_{i=1}^{r}(x - \delta_i) \\ &= (x - \delta_1)(x - \delta_2)\cdots(x - \delta_r)\end{aligned}$$

Corollary 33.8.6.1. *If $\mathbf{A}$ is diagonalisable, then $\mu_{\mathbf{A}}$ is a product of linear factors.*

33.8.2 Cayley-Hamilton Theorem

Recall the *characteristic equation* of a matrix $\mathbf{A}$ is defined as follows:

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

The left side by itself is called the *characteristic polynomial* of $\mathbf{A}$, denoted $c_{\mathbf{A}}$.

Theorem 33.8.7 (Cayley-Hamilton). *Let $A \in K^{n\times n}$, and let $c_{\mathbf{A}}$ be the characteristic polynomial of A. Then, $c_{\mathbf{A}}(\mathbf{A}) = \mathbf{0}_n$.*

Corollary 33.8.7.1. *For any $\mathbf{A} \in K^{n\times n}$, $\mu_{\mathbf{A}}$ divides $c_{\mathbf{A}}$, and in particular,* $\deg(\mu_{\mathbf{A}}) \leq n$.

33.8.3 Calculating Minimal Polynomials

Lemma 33.8.8. *Let λ be any eigenvalue of* $\mathbf{A}$*. Then,* $\mu_A(\lambda) = 0$.

Using this lemma with the Cayley-Hamilton theorem lets us reduce the possibilities for the minimal polynomial.

Algorithm 3 Top Down Algorithm

```
Calculate the characteristic polynomial, $c_{\mathbf{A}}$.
Factorise the characteristic polynomial by inspection, or using the factor and remainder theorem with polynomial division.
Evaluate each possible combination of factors that include all eigenvalues as roots in order of increasing degree. The first to return $\mathbf{0}_n$ is the minimal polynomial.
```

As an example for the last step, if $c_{\mathbf{A}}(x) = (x-1)(x-2)^2(x-3)^3$, then,

$$\mu_{\mathbf{A}}(x) \in \begin{cases} (x-1)(x-2)(x-3) \\ (x-1)(x-2)(x-3)^2 \\ (x-1)(x-2)^2(x-3) \\ (x-1)(x-2)^2(x-3)^2 \\ (x-1)(x-2)(x-3)^3 \\ (x-1)(x-2)^2(x-3)^3 \end{cases}$$

Then, evaluate the polynomials in this list at $\mathbf{A}$ from top to bottom (the list is sorted in degree order), and the first one to return the zero matrix is the minimal polynomial.

Algorithm 4 Bottom Up Algorithm

```
Pick some simple non-zero vector, $\mathbf{v}$ (the standard basis vector $\mathbf{e}_1$ is often a good choice).
Apply $\mathbf{A}$ to $\mathbf{v}$ repeatedly to form a chain of vectors
    $\mathbf{v}_0 \overset{\mathbf{A}}{\mapsto} \mathbf{v}_1 \overset{\mathbf{A}}{\mapsto} \mathbf{v}_2 \overset{\mathbf{A}}{\mapsto} \cdots$
At some point, these image vectors will become linearly dependent, say, after $d$ applications of $\mathbf{A}$, so there exists coefficients $(\alpha_i)$ such that
    $\sum_{i=0}^{d} \alpha_i \mathbf{v}_i = 0$
  with $\alpha_d = 1$.
Then, the monic polynomial
    $\sum_{i=0}^{d} \alpha_d x^d$
  divides the minimal polynomial.
Repeat this process with different starting vectors that do not lie in the image of previous generated chains, until all the generated chains span $V$. The minimal polynomial is then the least common multiple of these polynomials.
```

33.8.4 Jordan Chains

Recall that a non-zero vector $\mathbf{v}$ that satisfies $(\mathbf{A} - \lambda\mathbf{I}_n)\mathbf{v} = \mathbf{0}$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ. We weaken this notion to classify a more general type of vector.

A non-zero vector $\mathbf{v}$ that satisfies $(\mathbf{A} - \lambda\mathbf{I}_n)^i\mathbf{v} = \mathbf{0}$ for some $i > 0$ is a *generalised eigenvector* of $\mathbf{A}$ with eigenvalue λ, and, for a fixed $i > 0$ and fixed λ, the collection of these generalised eigenvectors,

$$N_i(\mathbf{A},\lambda) := \{\mathbf{v} \in V : (\mathbf{A} - \lambda\mathbf{I}_n)^i\mathbf{v} = 0\}$$

is the nullspace of $(\mathbf{A} - \lambda\mathbf{I}_n)^i$, and is called the *generalised eigenspace of index* i with respect to λ.

The *full* generalised eigenspace of $\mathbf{A}$ with respect to λ is defined as

$$\{\mathbf{0}\} \cup \bigcup_{i\in\mathbb{N}} N_i(\mathbf{A},\lambda)$$

That is, it is the union of all generalised eigenspaces with respect to λ, along with the zero vector.

A *Jordan chain* of length k is a sequence of non-zero vectors $(\mathbf{v}_i)_{i=1}^k \subset K^{n,1}$ such that, for some eigenvalue λ of $\mathbf{A}$,

$$\mathbf{A}\mathbf{v}_1 = \lambda\mathbf{v_1}, \quad \mathbf{A}\mathbf{v}_i = \lambda\mathbf{v}_i + \mathbf{v}_{i-1}, \quad 2 \le i \le k$$

or equivalently,

$$(\mathbf{A} - \lambda\mathbf{I}_n)\mathbf{v}_1 = \mathbf{0}, \quad (\mathbf{A} - \lambda\mathbf{I}_n)\mathbf{v}_i = \mathbf{v}_{i-1}, \quad 1 \le i \le k$$

thus all vectors in a Jordan chain are generalised eigenvectors with $\mathbf{v}_i \in N_i(\mathbf{A},\lambda)$.

Lemma 33.8.9. *The vectors in a Jordan chain are linearly independent.*

Theorem 33.8.10. *The dimensions of corresponding generalised eigenspaces of similar matrices are the same.*

We define the *Jordan block* of degree k with eigenvalue λ to be the $k \times k$ matrix $\mathbf{J}_{\lambda,k}$ given by

$$\mathbf{J}_{\lambda,k} = (J_{i,j}) = \begin{cases} \lambda & \text{if } j = i \\ 1 & \text{if } j = i + 1 \\ 0 & \text{otherwise} \end{cases}$$

That is, the main diagonal has values λ, and the superdiagonal has values 1.

For example,

$$\mathbf{J}_{2,3} = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{bmatrix}, \qquad \mathbf{J}_{i,2} = \begin{bmatrix} i & 1 \\ 0 & i \end{bmatrix}, \qquad \mathbf{J}_{-2,4} = \begin{bmatrix} -2 & 1 & 0 & 0 \\ 0 & -2 & 1 & 0 \\ 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & -2 \end{bmatrix}$$

A matrix $\mathbf{A}$ of a transformation T with respect to the basis $(\mathbf{v}_i)_{i=1}^n \subset K^n$ is a Jordan block of degree n if and only if $(\mathbf{v}_i)_{i=1}^n$ is a Jordan chain for $\mathbf{A}$.

The minimal and characteristic polynomials of $\mathbf{J}_{\lambda,k}$ are given by,

$$\mu_{\mathbf{J}_{\lambda,k}}(x) = (x - \lambda)^k$$
$$c_{\mathbf{J}_{\lambda,k}}(x) = (\lambda - x)^k$$

We denote the $m \times n$ zero matrix by $\mathbf{0}_{m,n}$. If $\mathbf{A}$ is an $m \times m$ matrix, and $\mathbf{B}$ is an $n \times n$ matrix, we define their *direct sum* $\mathbf{A} \oplus \mathbf{B}$ to be the $(m+n) \times (m+n)$ matrix with block form

$$\left[\begin{array}{c|c} \mathbf{A} & \mathbf{0}_{m,n} \\ \hline \mathbf{0}_{n,m} & \mathbf{B} \end{array}\right]$$

A *Jordan basis* for $\mathbf{A}$ is a basis of K^n consisting of one or more Jordan chains. The matrix of a transformation with respect to a Jordan basis is the direct sum of the corresponding Jordan blocks.

Lemma 33.8.11. *Suppose that* $\mathbf{M} = \mathbf{A} \oplus \mathbf{B}$*. Then,* $c_{\mathbf{M}} = c_{\mathbf{A}} \times c_{\mathbf{B}}$ *and* $\mu_M = \operatorname{lcm}(\mu_{\mathbf{A}}, \mu_{\mathbf{B}})$*.*

Theorem 33.8.12. *Let* $\mathbf{A}$ *be an* $n \times n$ *matrix over* $\mathbb{C}$*. Then, there exists a Jordan basis for* $\mathbf{A}$*, and hence* $\mathbf{A}$ *is similar to a matrix* $\mathbf{J} = \bigoplus \mathbf{J}_{\lambda,k}$*, where the Jordan blocks* $\mathbf{J}_{\lambda,k}$ *are uniquely determined by* $\mathbf{A}$*.*

The matrix $\mathbf{J}$ in the above is called the *Jordan canonical form* or *Jordan normal form* of $\mathbf{A}$, and is determined uniquely up to the order of the blocks. The field has to be at least $\mathbb{C}$ (or an extension of $\mathbb{C}$) so that $\mathbf{A}$ has at least one eigenvalue, since $\mathbb{C}$ is algebraically closed.

Theorem 33.8.13. *Let* $A \in \mathbb{C}^{n \times n}$*, and suppose* $\{\lambda_i\}_{i=1}^r$ *are the eigenvalues of* A*. Then,*

$$c_{\mathbf{A}}(x) = (-1)^n \prod_{i=1}^{r} (x - \lambda_i)^{a_i}$$

where a_i *is the sum of the degrees of the Jordan blocks of* $\mathbf{A}$ *of eigenvalue* λ_i*;*

$$\mu_{\mathbf{A}}(x) = \prod_{i=1}^{r} (x - \lambda_i)^{b_i}$$

where b_i *is the largest among the degrees of the Jordan blocks of* $\mathbf{A}$ *of eigenvalue* λ_i*;*

- $\mathbf{A}$ *is diagonalisable if and only if* $\mu_{\mathbf{A}}(x)$ *has no repeated factors.*

33.8.5 Computing the Jordan Canonical Form

Suppose a matrix $\mathbf{A} \in \mathbb{C}^{10,10}$ has a characteristic polynomial,

$$c_{\mathbf{A}}(x) = (x-1)^3 (x-2)^4 (x-3)^2 (x-4)$$

and thus has eigenvalues $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = 3$ and $\lambda_4 = 4$.

We say that the eigenvalue 1 has *algebraic multiplicity* $\alpha(\lambda_1) = 3$, because it is repeated as a root of the characteristic polynomial 3 times. The other algebraic multiplicities are then $\alpha(\lambda_2) = 4$, $\alpha(\lambda_3) = 2$, and $\alpha(\lambda_4) = 1$. The sum of the algebraic multiplicities over all eigenvalues is equal to the dimension of the matrix:

$$\sum_{i=1}^{r} \alpha(\lambda_i) = n$$

The *geometric multiplicity* $\gamma(\lambda_i)$ of an eigenvalue λ_i is the dimension of the kernel of $\mathbf{A} - \lambda_i \mathbf{I}_n$, or, $\operatorname{null}(\mathbf{A} - \lambda_i \mathbf{I}_n)$.

The *generalised geometric multiplicity* $\gamma_k(\lambda_i)$ of an eigenvalue λ_i is the dimension of the kernel of $(\mathbf{A} - \lambda_i \mathbf{I}_n)^k$, or, $\operatorname{null}\big((\mathbf{A} - \lambda_i \mathbf{I}_n)^k\big)$.

Now, the JCF of $\mathbf{A}$ will have the eigenvalues along the diagonal:

$$\mathbf{J} = \begin{bmatrix} \lambda_1 &&&&&&&&& \\ & \lambda_1 &&&&&&&& \\ && \lambda_1 &&&&&&& \\ &&& \lambda_2 &&&&&& \\ &&&& \lambda_2 &&&&& \\ &&&&& \lambda_2 &&&& \\ &&&&&& \lambda_2 &&& \\ &&&&&&& \lambda_3 && \\ &&&&&&&& \lambda_3 & \\ &&&&&&&&& \lambda_4 \end{bmatrix}$$

with each eigenvalue λ_i appearing $\alpha(\lambda_i)$ times. Note that there are many different possibilities for the orderings of these eigenvalues, but our convention will be to group the same eigenvalues together, and (where possible) to order these groups in increasing order. If the eigenvalues are complex, just pick any sensible ordering.

We will call these groups *Jordan boxes* (not to be confused with Jordan blocks), highlighted below:

$$\mathbf{J} = \begin{bmatrix} \begin{bmatrix} \lambda_1 && \\ & \lambda_1 & \\ && \lambda_1 \end{bmatrix} &&& \\ & \begin{bmatrix} \lambda_2 &&& \\ & \lambda_2 && \\ && \lambda_2 & \\ &&& \lambda_2 \end{bmatrix} && \\ && \begin{bmatrix} \lambda_3 & \\ & \lambda_3 \end{bmatrix} & \\ &&& \begin{bmatrix} \lambda_4 \end{bmatrix} \end{bmatrix}$$

A Jordan box is like a Jordan block, but we don't necessarily know where the 1s on the superdiagonal are yet. That is, we need to fill a Jordan box with Jordan blocks, and once we have done so for all boxes, we will have determined a Jordan canonical form for $\mathbf{A}$.

To begin with, the geometric multiplicity tells us how many blocks are in each box. For instance,

$$\begin{bmatrix} \lambda_3 & \\ & \lambda_3 \end{bmatrix} \longrightarrow \underbrace{\begin{bmatrix} [\lambda_3] & \\ & [\lambda_3] \end{bmatrix}}_{\gamma(\lambda_3)=2} \text{ or } \underbrace{\begin{bmatrix} \begin{bmatrix} \lambda_3 & 1 \\ & \lambda_3 \end{bmatrix} \end{bmatrix}}_{\gamma(\lambda_3)=1}$$

Here, we can have either two 1×1 blocks, if the geometric multiplicity of λ_3 is 2, or a single 2×2 block, if $\gamma(\lambda_3) = 1$.

For 3×3,

$$\begin{bmatrix} \lambda_1 && \\ & \lambda_1 & \\ && \lambda_1 \end{bmatrix} \longrightarrow \underbrace{\begin{bmatrix} [\lambda_1] && \\ & [\lambda_1] & \\ && [\lambda_1] \end{bmatrix}}_{\gamma(\lambda_1)=3} \text{ or } \underbrace{\begin{bmatrix} \begin{bmatrix} \lambda_1 & 1 \\ & \lambda_1 \end{bmatrix} & \\ & [\lambda_1] \end{bmatrix}}_{\gamma(\lambda_1)=2} \text{ or } \underbrace{\begin{bmatrix} \begin{bmatrix} \lambda_1 & 1 & \\ & \lambda_1 & 1 \\ && \lambda_1 \end{bmatrix} \end{bmatrix}}_{\gamma(\lambda_1)=1}$$

the Jordan box can contain three 1×1 Jordan blocks, one 2×2 Jordan block and one 1×1 Jordan block, or a single 3×3 Jordan block, if the geometric multiplicity is 3, 2, or 1, respectively.

However, for 4×4 boxes or larger, the geometric multiplicity alone is not sufficient to determine the blocks within the box. For instance,

$$\begin{bmatrix} \lambda & & & \\ & \lambda & & \\ & & \lambda & \\ & & & \lambda \end{bmatrix} \longrightarrow \underbrace{\begin{bmatrix} \begin{bmatrix} \lambda & 1 & \\ & \lambda & 1 \\ & & \lambda \end{bmatrix} & \\ & [\lambda] \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \begin{bmatrix} \lambda & 1 \\ & \lambda \end{bmatrix} & \\ & \begin{bmatrix} \lambda & 1 \\ & \lambda \end{bmatrix} \end{bmatrix}}_{\gamma(\lambda_1)=2}$$

are both consistent with a geometric multiplicity of 2. So, we have to calculate generalised geometric multiplicities to gain more information.

The generalised geometric multiplicities of index k tell us how many chains exist in each generalised eigenspace of index k, allowing us to determine the lengths of the Jordan chains. For instance, suppose $\alpha(\lambda) = 7$, so we have a 7×7 Jordan box. If $\gamma_1(\lambda) = 4$, $\gamma_2(\lambda) = 6$, $\gamma_3(\lambda) = 7 = \alpha(\lambda)$, then the chains would be:

$$\begin{array}{r|llll} i = 3 & \bullet & & & \\ & \downarrow & & & \\ i = 2 & \bullet & \bullet & & \\ & \downarrow & \downarrow & & \\ i = 1 & \bullet & \bullet & \bullet & \bullet \; \}\gamma_1(\lambda) \end{array} \left.\vphantom{\begin{array}{c}a\\a\\a\end{array}}\right\}\gamma_2(\lambda) \quad \left.\vphantom{\begin{array}{c}a\\a\\a\\a\\a\end{array}}\right\}\gamma_3(\lambda)$$

and the lengths of the chains indicate the dimensions of the Jordan blocks within the Jordan box for λ. We have one chain of length 3, one chain of length 2, and two chains of length 1, so we have,

$$\begin{bmatrix} \begin{bmatrix} \lambda & & \\ & \lambda & \\ & & \lambda \end{bmatrix} & & & \\ & \begin{bmatrix} \lambda & \\ & \lambda \end{bmatrix} & & \\ & & [\lambda] & \\ & & & [\lambda] \end{bmatrix}$$

in this Jordan box (the ordering of the blocks within the box is arbitrary). We repeat this process for every Jordan box of dimensions 4 or higher.

To find the transformation matrix $\mathbf{P}$ such that $\mathbf{A} = \mathbf{P}^{-1}\mathbf{J}\mathbf{P}$, we calculate actual vectors for these chains (usually by calculating the generalised eigenbases), then augment them together in the same order as we arranged the boxes and blocks in $\mathbf{J}$.

Algorithm 5 JCF Decomposition

```
Calculate the eigenvalues of $\mathbf{A} \in \mathbb{C}^{n\times n}$, $(\lambda_i)_{i=1}^{k}$.
For each eigenvalue $\lambda_i$, determine,
    • the algebraic multiplicity $\alpha(\lambda_i)$ = number of times repeated as a root in $c_{\mathbf{A}}$;
    • the geometric multiplicity $\gamma(\lambda_i) = \mathrm{null}(\mathbf{A} - \lambda_i \mathbf{I}_n)$.
If, for an eigenvalue $\lambda_i$, we have $\alpha(\lambda_i) \geq 4$, additionally calculate $\gamma_k(\lambda_i) = \mathrm{null}\big((\mathbf{A} - \lambda_i \mathbf{I}_n)^k\big)$ for $k = 2,3,\ldots$ as until $\gamma_k(\lambda_i) = \alpha(\lambda_i)$.
Use these generalised geometric multiplicities to determine the lengths of the chains, and hence the size of the Jordan blocks in each Jordan box. Repeat for each eigenvalue.
To further find the transformation matrix $\mathbf{P}$, calculate the bases of $\ker\big((\mathbf{A} - \lambda_i \mathbf{I}_n)^k\big)$ for each eigenvalue $\lambda_i$, where the basis for each generalised eigenspace is a subset of the basis for the next generalised eigenspace.
Calculate the Jordan chains for each $\lambda_i$ by picking vectors $\mathbf{v}_j \in \ker\big((\mathbf{A}-\lambda_i\mathbf{I}_n)^k\big) \setminus \ker\big((\mathbf{A}-\lambda_i\mathbf{I}_n)^{k-1}\big)$, and recursively computing $\mathbf{v}_{j-1} = (\mathbf{A} - \lambda_i \mathbf{I}_n)\mathbf{v}_j$ until the chain is complete, starting with the longest chain. Then, compute the next chain, ensuring that the vectors picked at every step is linearly independent with those already selected in previous chains to ensure the chains do not converge.
Augment the Jordan chains together in order corresponding to boxes and blocks to form $\mathbf{P}$.
Note: this procedure yields the right transformation matrix. That is, we obtain $\mathbf{P}$ such that $\mathbf{A} = \mathbf{P}^{-1}\mathbf{J}\mathbf{P}$ and not $\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$.
```

Example. Determine a JCF for the matrix,

$$\mathbf{A} = \begin{bmatrix} 5 & -2 & 1 & -7 & 1 & 5 \\ 0 & 3 & 4 & -4 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & 2 \\ 1 & -1 & -2 & 2 & 1 & 2 \\ 0 & 0 & 1 & -1 & 2 & 0 \\ 0 & 0 & -1 & 1 & 0 & 3 \end{bmatrix}$$

You may use that $c_{\mathbf{A}}(x) = (x-2)^2(x-3)^4$ without proof.

From the characteristic polynomial, we have eigenvalues $\lambda_1 = 2$, and $\lambda_2 = 3$, with $\alpha(\lambda_1) = 2$, and $\alpha(\lambda_2) = 4$, so we have the Jordan box form,

$$\mathbf{J} = \begin{bmatrix} \begin{bmatrix} \lambda_1 & \\ & \lambda_1 \end{bmatrix} & \\ & \begin{bmatrix} \lambda_2 & & & \\ & \lambda_2 & & \\ & & \lambda_2 & \\ & & & \lambda_2 \end{bmatrix} \end{bmatrix}$$

Next, we calculate the geometric multiplicity $\gamma(\lambda_1)$:

$$\begin{aligned} \gamma(\lambda_1) &= \dim\ker(\mathbf{A} - \lambda_1 \mathbf{I}_6) \\ &= \dim\ker \begin{bmatrix} 3 & -2 & 1 & -7 & 1 & 5 \\ 0 & 1 & 4 & -4 & -1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 2 \\ 1 & -1 & -2 & 0 & 1 & 2 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 1 \end{bmatrix} \end{aligned}$$

$$\xrightarrow{\substack{\text{row}\\\text{reduce}}} \dim\ker \begin{bmatrix} 1 & 0 & 0 & -2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

There are 5 pivot columns, so $\dim\ker(\mathbf{A} - \lambda_1\mathbf{I}_6) = 6 - 5 = 1 = \gamma(\lambda_1)$. This is sufficient information to determine a 2 × 2 Jordan box, so we may stop here, but we will continue to compute the generalised geometric multiplicity $\gamma_2(\lambda_1)$, as it will be helpful later for computing the transformation matrix.

$$\begin{aligned} \gamma_2(\lambda_1) &= \dim\ker\big((\mathbf{A} - \lambda_1\mathbf{I}_6)^2\big) \\ &= \dim\ker \begin{bmatrix} 3 & -2 & 1 & -7 & 1 & 5 \\ 0 & 1 & 4 & -4 & -1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 2 \\ 1 & -1 & -2 & 0 & 1 & 2 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 1 \end{bmatrix}^2 \\ &= \dim\ker \begin{bmatrix} 3 & -2 & 4 & 10 & -1 & 6 \\ 0 & 1 & 6 & -6 & -1 & 2 \\ 1 & -1 & -1 & -1 & 0 & 2 \\ 1 & -1 & -2 & 0 & 0 & 2 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & -2 & 2 & 0 & 1 \end{bmatrix} \\ &\xrightarrow{\substack{\text{row}\\\text{reduce}}} \dim\ker \begin{bmatrix} 1 & 0 & 0 & -2 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned}$$

Important: when calculating $(\mathbf{A} - \lambda_1\mathbf{I}_6)^2$, do not square the row reduced form we found earlier. You must use the $\underline{\text{original}}$ non-reduced matrix.

There are 4 pivot columns, so $\gamma_2(\lambda_1) = 6 - 4 = 2$, and we have reached the algebraic multiplicity $\alpha(\lambda_1)$, so we may stop here.

Next, we similarly calculate the geometric multiplicity $\gamma(\lambda_2)$:

$$\begin{aligned} \gamma(\lambda_2) &= \dim\ker(\mathbf{A} - \lambda_2\mathbf{I}_6) \\ &= \dim\ker \begin{bmatrix} 2 & -2 & 1 & -7 & 1 & 5 \\ 0 & 0 & 4 & -4 & -1 & 1 \\ 1 & -1 & -2 & -1 & 1 & 2 \\ 1 & -1 & -2 & -1 & 1 & 2 \\ 0 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \end{bmatrix} \\ &\xrightarrow{\substack{\text{row}\\\text{reduce}}} \dim\ker \begin{bmatrix} 1 & -1 & 0 & -3 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned}$$

There are 4 pivot columns, so $\dim\ker(\mathbf{A}-\lambda_2\mathbf{I}_6) = 6-4 = 2 = \gamma(\lambda_2)$. This is insufficient information to determine the Jordan box of λ_2, so we calculate higher index generalised geometric multiplicities:

$$\begin{aligned}
\gamma_2(\lambda_2) &= \dim\ker\big((\mathbf{A}-\lambda_2\mathbf{I}_6)^2\big) \\
&= \dim\ker \begin{bmatrix} 2 & -2 & 1 & -7 & 1 & 5 \\ 0 & 0 & 4 & -4 & -1 & 1 \\ 1 & -1 & -2 & -1 & 1 & 2 \\ 1 & -1 & -2 & -1 & 1 & 2 \\ 0 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \end{bmatrix}^2 \\
&= \dim\ker \begin{bmatrix} -2 & 2 & 2 & 4 & -3 & -4 \\ 0 & 0 & -2 & 2 & 1 & 0 \\ -1 & 1 & 2 & 1 & -2 & -2 \\ -1 & 1 & 2 & 1 & -2 & -2 \\ 0 & 0 & -1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\
&\xrightarrow{\substack{\text{row}\\\text{reduce}}} \dim\ker \begin{bmatrix} 1 & -1 & 0 & -3 & 0 & 2 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}
\end{aligned}$$

Again, ensure you use the original matrix, and not the row reduced matrix we found before.

There are 3 pivot columns, so $\gamma_2(\lambda_2) = 6-3 = 3$. We still have not reached the algebraic multiplicity of λ_2, so we continue to the next generalised eigenspace:

$$\begin{aligned}
\gamma_3(\lambda_2) &= \dim\ker\big((\mathbf{A}-\lambda_2\mathbf{I}_6)^3\big) \\
&= \dim\ker \begin{bmatrix} 2 & -2 & 1 & -7 & 1 & 5 \\ 0 & 0 & 4 & -4 & -1 & 1 \\ 1 & -1 & -2 & -1 & 1 & 2 \\ 1 & -1 & -2 & -1 & 1 & 2 \\ 0 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \end{bmatrix}^3 \\
&= \dim\ker \begin{bmatrix} 2 & -2 & -5 & -1 & 5 & 4 \\ 0 & 0 & 1 & -1 & -1 & 0 \\ 1 & -1 & -3 & 0 & 3 & 2 \\ 1 & -1 & -3 & 0 & 3 & 2 \\ 0 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\
&\xrightarrow{\substack{\text{row}\\\text{reduce}}} \dim\ker \begin{bmatrix} 1 & -1 & 0 & -3 & 0 & 2 \\ 0 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}
\end{aligned}$$

There are 2 pivot columns, so $\gamma_2(\lambda_2) = 6-2 = 4$ and we have reached the algebraic multiplicity of λ_2, so we stop here.

At this point, we will draw our Jordan chains for λ_2 to keep track of our results:

$$\begin{array}{r|ll} i=3 & \mathbf{w}_3 & \\ & \downarrow & \\ i=2 & \mathbf{w}_2 & \\ & \downarrow & \\ i=1 & \mathbf{w}_1 & \mathbf{u}_1 \end{array}$$

This indicates that we have one 3×3 Jordan block and one 1×1 Jordan block within the 4×4 Jordan box for λ_2.

For λ_1, the Jordan chains would be:

$$\begin{array}{r|l} i=2 & \mathbf{v}_2 \\ & \downarrow \\ i=1 & \mathbf{v}_1 \end{array}$$

as $\gamma_1(\lambda_1) = 1$ and $\gamma_2(\lambda_1) = 2$. (Doing this is unnecessary as $\gamma_1(\lambda_1)$ alone is sufficient to determine a 2×2 Jordan box, but this generalised procedure will work for boxes of any size.)

Thus, the blocks are:

$$\mathbf{J} = \begin{bmatrix} \begin{bmatrix}\begin{bmatrix}\lambda_1 & 1 \\ & \lambda_1\end{bmatrix}\end{bmatrix} & \\ & \begin{bmatrix} [\lambda_2] & \\ & \begin{bmatrix}\lambda_2 & 1 & \\ & \lambda_2 & 1 \\ & & \lambda_2\end{bmatrix}\end{bmatrix} \end{bmatrix} = \begin{bmatrix} 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 & 3 \end{bmatrix}$$

Next, we will compute the transformation matrix. To calculate the Jordan chains, we will begin by finding bases for each of the generalised eigenspaces we have found so far.

For λ_1, we have,

$$(\mathbf{A} - \lambda_1\mathbf{I}_6)\mathbf{v} = \mathbf{0} \quad \overset{\substack{\text{row}\\\text{reduce}}}{\longrightarrow} \quad \begin{bmatrix} 1 & 0 & 0 & -2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$a = 2d$$

$$c = d$$

$$\begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = d\begin{bmatrix} 2 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \longrightarrow \ker(\mathbf{A} - \lambda_1\mathbf{I}_6) = \text{span}\left(\begin{bmatrix} 2 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}\right)$$

$$\left((\mathbf{A} - \lambda_1\mathbf{I}_6)^2\right)\mathbf{v} = \mathbf{0} \quad \overset{\substack{\text{row}\\\text{reduce}}}{\longrightarrow} \quad \begin{bmatrix} 1 & 0 & 0 & -2 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{aligned} a &= 2d + e \\ b &= e \\ c &= d \end{aligned}$$

$$\begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = d \begin{bmatrix} 2 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + e \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \longrightarrow \ker(\mathbf{A} - \lambda_1 \mathbf{I}_6) = \operatorname{span}\left(\begin{bmatrix} 2 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \right)$$

For λ_2, we have,

$$(\mathbf{A} - \lambda_2 \mathbf{I}_6)\mathbf{v} = \mathbf{0} \quad \xrightarrow{\text{row reduce}} \quad \begin{bmatrix} 1 & -1 & 0 & -3 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{aligned} a &= b + 3d \\ c &= d \end{aligned}$$

$$\begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = b \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + d \begin{bmatrix} 3 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \longrightarrow \ker(\mathbf{A} - \lambda_2 \mathbf{I}_6) = \operatorname{span}\left(\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \right)$$

$$\big((\mathbf{A} - \lambda_2 \mathbf{I}_6)^2\big)\mathbf{v} = \mathbf{0} \quad \xrightarrow{\text{row reduce}} \quad \begin{bmatrix} 1 & -1 & 0 & -3 & 0 & 2 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{aligned} a &= b + 3d - 2f \\ c &= d \end{aligned}$$

$$\begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = b \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + d \begin{bmatrix} 3 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + f \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \longrightarrow \ker\big((\mathbf{A} - \lambda_2 \mathbf{I}_6)^2\big) = \operatorname{span}\left(\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right)$$

$$\big((\mathbf{A} - \lambda_2 \mathbf{I}_6)^3\big)\mathbf{v} = \mathbf{0} \quad \xrightarrow{\text{row reduce}} \quad \begin{bmatrix} 1 & -1 & 0 & -3 & 0 & 2 \\ 0 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{aligned} a &= b + 3d - 2f \\ c &= d + e \end{aligned}$$

$$\begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \end{bmatrix} = b\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + d\begin{bmatrix} 3 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + f\begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} + e\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \longrightarrow \ker\big((\mathbf{A}-\lambda_2\mathbf{I}_6)^3\big) = \operatorname{span}\left(\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}\right)$$

The basis vectors have been coloured according to which index they first appeared in, as this will be helpful for the next step.

Recall the Jordan chains we have found:

$$\begin{tikzcd}
 & \lambda_1 & \lambda_2 & \\
i=3 & & \mathbf{w}_3 \arrow[d] & \\
i=2 & \mathbf{v}_2 \arrow[d] & \mathbf{w}_2 \arrow[d] & \\
i=1 & \mathbf{v}_1 & \mathbf{w}_1 & \mathbf{u}_1
\end{tikzcd}$$

We begin by choose a vector to be $\mathbf{v}_2$. From the diagram above, it lies in the generalised eigenspace of index 2 for λ_1, but not of 1. We have one obvious option,

$$\mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

We then compute $\mathbf{v}_1$:

$$\begin{aligned} \mathbf{v}_1 &= (\mathbf{A}-\lambda_1\mathbf{I}_n)\mathbf{v}_2 \\ &= \begin{bmatrix} 2 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \end{aligned}$$

completing the chain, and the eigenvalue.

Moving on to λ_2, we select the largest chain, so we have to choose $\mathbf{w}_3 \in \ker\big((\mathbf{A}-\lambda_2\mathbf{I}_n)^3\big) \setminus \ker\big((\mathbf{A}-\lambda_2\mathbf{I}_n)^2\big)$. Again, we have an obvious option,

$$\mathbf{w}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

Then we compute $\mathbf{w}_2$ and $\mathbf{w}_1$:

$$\mathbf{w}_2 = (\mathbf{A}-\lambda_2\mathbf{I}_n)\mathbf{w}_3$$

$$= \begin{bmatrix} 2 \\ 3 \\ -1 \\ -1 \\ 0 \\ -1 \end{bmatrix}$$

$$\mathbf{w}_1 = (\mathbf{A} - \lambda_2 \mathbf{I}_n)\mathbf{w}_2$$

$$= \begin{bmatrix} -1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

completing the chain. We begin the next chain, and need to choose $\mathbf{u}_1 \in \operatorname{span}(\ker(\mathbf{A} - \lambda_2 \mathbf{I}_6) \setminus \{\mathbf{w}_1\})$. There is again only one remaining basis vector for our choice:

$$\mathbf{u}_1 = \begin{bmatrix} 3 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

Now, recall our choice of ordering of the boxes and blocks in $\mathbf{J}$:

$$\mathbf{J} = \begin{bmatrix} \left[\begin{bmatrix} \lambda_1 & 1 \\ & \lambda_1 \end{bmatrix}\right] & & \\ & \left[\begin{matrix} [\lambda_2] & \\ & \begin{bmatrix} \lambda_2 & 1 & \\ & \lambda_2 & 1 \\ & & \lambda_2 \end{bmatrix} \end{matrix}\right] \end{bmatrix}$$

We have the single λ_1 block, followed by the 1×1 then 3×3 λ_2 blocks. The corresponding chains are the $\mathbf{v}$ chain, $\mathbf{u}$ chain, and $\mathbf{w}$ chain. Augmenting these vectors together in ascending order within each chain gives the required transformation matrix:

$$\mathbf{P} = [\overbrace{\underbrace{\mathbf{v}_1|\mathbf{v}_2}_{2\times 2}}^{\lambda_1 \text{ box}} | \overbrace{\underbrace{\mathbf{u}_1}_{1\times 1} | \underbrace{\mathbf{w}_1|\mathbf{w}_2|\mathbf{w}_3}_{3\times 3}}^{\lambda_2 \text{ box}}]$$

$$\text{Jordan blocks}$$

$$= \begin{bmatrix} 2 & 1 & 3 & -1 & 2 & 0 \\ 0 & 1 & 0 & -1 & 3 & 0 \\ 1 & 0 & 1 & 0 & -1 & 1 \\ 1 & 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{bmatrix}$$

$\triangle$

33.8.6 Review

Theorem 33.8.14. *Let $\mathbf{A} \in \mathbb{C}^{n \times n}$ be a square matrix with complex entries, and let $p \in \mathbb{C}[x]$ be any polynomial. Then if λ is an eigenvalue of $\mathbf{A}$, then $p(\lambda)$ is an eigenvalue of $p(\mathbf{A})$, and any eivenvalue of $p(\mathbf{A})$ is of this form.*

Theorem 33.8.15. *For $A \in \mathbb{C}^{n\times n}$, define*

$$N_i(\mathbf{A},\lambda) := \ker\big((\mathbf{A} - \lambda_n)^i\big)$$

Suppose $\mathbf{A}$,$\mathbf{B} \in \mathbb{C}^{n\times n}$ *are similar. That is, there exists an invertible matrix* $\mathbf{S} \in \mathbb{C}^{n\times n}$ *such that* $\mathbf{B} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}$*. Then,* $\mathbf{A}$ *and* $\mathbf{B}$ *share the same set of eigenvalues,*

$$\lambda_1 = \mu_1, \ldots, \lambda_k = \mu_k$$

and moreover,

$$\dim N_i(\mathbf{A},\lambda_j) = \dim N_i(\mathbf{B},\mu_j) \quad \textit{for all } i,j$$

Or, using our earlier notation,

$$\gamma_i(\lambda_j) = \gamma(\mu_j) \quad \textit{for all } i,j$$

The converse of this result also holds. That is, if $\mathbf{A}$,$\mathbf{B} \in \mathbb{C}^{n\times n}$ *share the same eigenvalues and satisfy the equations above, then* $\mathbf{A}$ *and* $\mathbf{B}$ *are similar.*

33.9 Matrix Functions

33.9.1 Matrix Powers

Suppose we wish to compute $\mathbf{A}^n$ for a general matrix $\mathbf{A}$ and large exponent $n \gg 1$.

Ideally, $\mathbf{A}$ is diagonalisable, and we may compute,

$$\begin{aligned}\mathbf{A}^n &= (\mathbf{P}^{-1}\mathbf{D}\mathbf{P})^n \\ &= (\mathbf{P}^{-1}\mathbf{D}\mathbf{P})(\mathbf{P}^{-1}\mathbf{D}\mathbf{P})\cdots(\mathbf{P}^{-1}\mathbf{D}\mathbf{P}) \\ &= \mathbf{P}^{-1}\mathbf{D}(\mathbf{P}\mathbf{P}^{-1})\mathbf{D}(\mathbf{P}\cdots\mathbf{P}^{-1})\mathbf{D}\mathbf{P} \\ &= \mathbf{P}^{-1}\mathbf{D}^n\mathbf{P}\end{aligned}$$

and powers of diagonal matrices are trivial to compute:

$$\mathbf{D}^n = \begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_k\end{bmatrix}^n = \begin{bmatrix}\lambda_1^n & & & \\ & \lambda_2^n & & \\ & & \ddots & \\ & & & \lambda_k^n\end{bmatrix}$$

But, as we have already seen, not every matrix admits an eigenbasis. On the other hand, every matrix does have a Jordan canonical form, and by a similar telescoping sum, we have,

$$\mathbf{A}^n = \mathbf{P}^{-1}\mathbf{J}^n\mathbf{P}$$

The problem is now to efficiently compute powers of a matrix in Jordan canonical form.

Theorem 33.9.1. *For any square matrices* $\mathbf{A}$,$\mathbf{B}$*,*

$$(\mathbf{A} \oplus \mathbf{B})^n = \mathbf{A}^n \oplus \mathbf{B}^n$$

and more generally, for any collection of square matrices $(\mathbf{A}_i)_{i=1}^k$

$$\left(\bigoplus_{i=1}^k \mathbf{A}_i\right)^n = \bigoplus_{i=1}^k \mathbf{A}_i^n$$

Recall that $\mathbf{J}$ is the direct sum of Jordan blocks, so if these blocks are sufficiently small or simple, this can simplify the calculation greatly. We also have a general formula for larger Jordan blocks:

Theorem 33.9.2. *For any Jordan block $\mathbf{J}_{\lambda,k}$,*

$$\mathbf{J}_{\lambda,k}^n = \begin{bmatrix} \binom{n}{0}\lambda^n & \binom{n}{1}\lambda^{n-1} & \cdots & \binom{n}{k-2}\lambda^{n-k+2} & \binom{n}{k-1}\lambda^{n-k+1} \\ 0 & \binom{n}{0}\lambda^n & \cdots & \binom{n}{k-3}\lambda^{n-k+3} & \binom{n}{k-2}\lambda^{n-k+2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \binom{n}{0}\lambda^n & \binom{n}{1}\lambda^{n-1} \\ 0 & 0 & \cdots & 0 & \binom{n}{0}\lambda^n \end{bmatrix}$$

noting that $\binom{n}{k} = 0$ wherever $k > n$.

33.9.2 Lagrange Interpolation

Another way to compute arbitrary powers of matrices is to use *Lagrange interpolation.*

Theorem 33.9.3 (Lagrange Interpolation)**.** *Suppose $\psi(\mathbf{A}) = \mathbf{0}_n$ for a polynomial $\psi \in \mathbb{C}[x]$, and furthermore suppose ψ has roots $(\alpha_i)_{i=1}^k$ with corresponding (algebraic) multiplicities $(m_i)_{i=1}^k$. (In practice, we would choose $\psi = c_{\mathbf{A}}$ or $\psi = \mu_{\mathbf{A}}$.)*

Then, for any sufficiently well-behaved function $f : \mathbb{C} \to \mathbb{C}$, there exists a function q such that*

$$f = q\psi + r$$

where r is a polynomial of degree strictly lower than ψ and

$$f^{(t)}(\alpha_i) = r^{(t)}(\alpha_i)$$

for all $1 \leq j \leq k$, $0 \leq t < m_j$, and furthermore, $f(\mathbf{A}) = r(\mathbf{A})$.

The idea is that we compute a polynomial, r, that acts effectively like a Taylor polynomial near the roots of ψ.

Example. Find a general formula for $f(\mathbf{A}) = \mathbf{A}^n$, where

$$\mathbf{A} = \begin{bmatrix} 3 & 1 & 0 & 1 \\ -1 & 5 & 4 & 1 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

You may use that $\mu_{\mathbf{A}}(x) = (2-x)(4-x)^2$ without proof.

Here, $f(z) = z^n$, which is certainly a well-behaved function with easily computable derivatives so we may attempt Lagrange interpolation.

$\mu_{\mathbf{A}}$ is a cubic, so we may choose $r(z) = \alpha z^2 + \beta z + \gamma$.

We know that f and $\mu_{\mathbf{A}}$ agree at the roots of $\mu_{\mathbf{A}}$, so we compute,

$$\begin{aligned} f(2) = 2^n \quad &= r(2) \ = 4\alpha + 2\beta + \gamma \\ f(4) = 4^n \quad &= r(4) \ = 16\alpha + 4\beta + \gamma \\ f'(4) = n4^{n-1} &= r'(4) = 8\alpha + \beta \end{aligned}$$

* f must be analytic in a neighbourhood around every $(\alpha_i)_{i=1}^k$.

Then,

$$\begin{aligned}
r(\mathbf{A}) &= \alpha\mathbf{A}^2 + \beta\mathbf{A} + \gamma\mathbf{I_4} \\
&= \alpha\begin{bmatrix} 8 & 8 & 4 & 8 \\ -8 & 24 & 28 & 8 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 16 \end{bmatrix} + \beta\begin{bmatrix} 3 & 1 & 0 & 1 \\ -1 & 5 & 4 & 1 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix} + \gamma\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 8\alpha+3\beta+\gamma & 8\alpha+\beta & 4\alpha & 8\alpha+\beta \\ -8\alpha-\beta & 24\alpha+5\beta+\gamma & 28\alpha+4\beta & 8\alpha+\beta \\ 0 & 0 & 4\alpha+2\beta+\gamma & 0 \\ 0 & 0 & 0 & 16\alpha+4\beta+\gamma \end{bmatrix} \\
&= \begin{bmatrix} 4^n - n4^{n-1} & n4^{n-1} & 2^n - 4^n + 2n4^{n-1} & n4^{n-1} \\ -n4^{n-1} & 4^n + n4^{n-1} & 4^n - 2^n + 2n4^{n-1} & n4^{n-1} \\ 0 & 0 & 2^n & 0 \\ 0 & 0 & 0 & 4^n \end{bmatrix}
\end{aligned}$$

Note that we did not have to calculate the individual values of α, β and γ, (which are

$$\begin{aligned}
\alpha &= \tfrac{1}{4}\cdot 2^n - \tfrac{1}{4}\cdot 4^n + \tfrac{1}{2}n4^{n-1} \\
\beta &= -2\cdot 2^n - 2\cdot 4^n + 5n4^{n-1} \\
\gamma &= 4\cdot 2^n + 5\cdot 4^n - 12n4^{n-1}
\end{aligned}$$

for those interested) as the entries in $r(\mathbf{A})$ are simple linear combinations of the values of $r(2)$, $r(4)$, and $r'(4)$. For instance, in the first entry, we have,

$$\begin{aligned}
8\alpha + 3\beta + \gamma &= (16\alpha + 4\beta + \gamma) - (8\alpha + \beta) \\
&= r(4) - r'(4) \\
&= 4^n - n4^{n-1}
\end{aligned}$$

with every other entry being computed similarly. △

We've only been using Lagrange interpolation to calculate powers of matrices here, but the technique works identically for any sufficiently well-behaved function $f : \mathbb{C} \to \mathbb{C}$.

33.9.3 Matrix Exponentials

33.9.3.1 Recurrence Relations

Consider a vector-valued first-order recurrence relation,

$$\mathbf{x}_n = \mathbf{A}\mathbf{x}_{n-1}$$

where $(\mathbf{x}_i)_{i=1}^{\infty} \subset K^m$ is a sequence of vectors. We will only be considering autonomous recurrence relations – that is, the matrix $\mathbf{A}$ is not a function of n. If a value for $\mathbf{x}_0$ is given, then these equations are also called *(discrete) initial value problems.*

These recurrence relations can be solved analogously to the scalar-valued case with back substitution:

$$\begin{aligned}
\mathbf{x}_n &= \mathbf{A}\mathbf{x}_{n-1} \\
&= \mathbf{A}^2\mathbf{x}_{n-2} \\
&= \mathbf{A}^3\mathbf{x}_{n-3} \\
&\ \ \vdots
\end{aligned}$$

$$= \mathbf{A}^n \mathbf{x}_0$$

However, in this case, we now have to calculate an arbitrary power of a matrix, but we can use techniques from the last section to do so.

One application of this is in solving higher-order scalar-valued autonomous recurrence relations. For example, consider the Fibonacci sequence, given by the second-order recurrence relation,

$$F_n = F_{n-1} + F_{n-2}$$

We can easily rewrite this as,

$$\begin{aligned} \begin{bmatrix} F_n \\ F_{n+1} \end{bmatrix} &= \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} F_{n-1} \\ F_n \end{bmatrix} \\ \mathbf{v}_n &= \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \mathbf{v}_{n-1} \end{aligned}$$

and now we have a single first-order vector-valued recurrence relation, and this works more generally – we can transform an nth-order recurrence relation into a first-order vector-valued recurrence relation in n dimensions.

33.9.3.2 Differential Equations

Now, suppose we have a system of first-order linear autonomous simultaneous differential equations, say,

$$\begin{aligned} x' &= 7x - 2y + 9z \\ y' &= 7x + 3y - 5z \\ z' &= 2x + 5y + 6z \end{aligned}$$

We can alternatively interpret these separate variables as a single vector:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} 7x - 2y + 9z \\ 7x + 3y - 5z \\ 2x + 5y + 6z \end{bmatrix}$$

and, factoring out the matrix, we can represent this system as a single first-order vector-valued differential equation:

$$\mathbf{v}' = \mathbf{A}\mathbf{v}$$

where,

$$\mathbf{v} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \quad \text{and} \quad \mathbf{A} = \begin{bmatrix} 7 & -2 & 9 \\ 7 & 3 & -5 \\ 2 & 5 & 6 \end{bmatrix}$$

Compare this to the case of an ordinary first-order differential equation,

$$x' = ax$$

where a is some constant. The solution to this differential equation is given by,

$$x(t) = e^{at} x(0)$$

Similarly, the vector-valued differential equation,

$$\mathbf{v}' = \mathbf{A}\mathbf{v}$$

has a solution given by

$$\mathbf{v}(t) = e^{\mathbf{A}t}\mathbf{v}(0)$$

But, what does $e^{\mathbf{A}t}$ mean? Clearly,

$$e^{\begin{bmatrix} 3 & 1 & 4 \\ 1 & 5 & 9 \\ 2 & 6 & 5 \end{bmatrix}} = \underbrace{e \times e \times \cdots \times e}_{\begin{bmatrix} 3 & 1 & 4 \\ 1 & 5 & 9 \\ 2 & 6 & 5 \end{bmatrix} \text{ times?}}$$

is meaningless.

Instead, recall the Taylor series of $f(x) = e^x$ for real inputs $x \in \mathbb{R}$:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

Unlike the expression on the left, it does make sense for us to input things other than real numbers into the series on the right, even if those objects do not immediately make sense as exponents. For instance, we could input complex numbers, or even matrices to this expression.

While the equation above is a *theorem* for real numbers, it's a *definition* for more exotic inputs, like complex numbers or matrices,

$$e^x := \sum_{n=0}^{\infty} \frac{x^n}{n!}, \quad x \in \mathbb{C}, K^{n\times n}, \dots$$

and we sometimes prefer using the notation $\exp(x)$ instead of e^x to emphasise this point more. There are some issues of convergence – after all, why should we expect this series to converge for matrix inputs just because it converges for real inputs – but that is a relatively easy exercise in analysis.

We can use similar techniques from calculating matrix powers before. In particular, exp is entire, so Lagrange interpolation also applies, and is, in general (for non-diagonalisable matrices), simpler than using a JCF decomposition.

Example. Solve the system of differential equations,

$$\begin{aligned} x' &= x - 3z \\ y' &= x - y - 6z \\ z' &= -x + 2y + 5z \end{aligned}$$

with initial conditions,

$$\begin{aligned} x(0) &= 1 \\ y(0) &= 1 \\ z(0) &= 0 \end{aligned}$$

First, we write the system as a vector-valued differential equation:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}' = \underbrace{\begin{bmatrix} 1 & 0 & -3 \\ 1 & -1 & -6 \\ -1 & 2 & 5 \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

Then, we find the characteristic polynomial:

$$\begin{aligned} c_{\mathbf{A}}(z) &= \det(\mathbf{A} - z\mathbf{I}_3) \\ &= -z^3 + 5z^2 - 8z + 4 \\ &= (1-z)(2-z)^2 \end{aligned}$$

so we have roots $\lambda_1 = 1$, and $\lambda_2 = 2$, with multiplicities $\alpha(\lambda_1) = 1$ and $\alpha(\lambda_2) = 2$.

We interpolate $f(z) = e^{zt}$ with $r(z) = \alpha z^2 + \beta z + \gamma$:

$$\begin{aligned} f(1) &= e^t & &= r(1) & &= \alpha + \beta + \gamma \\ f(2) &= e^{2t} & &= r(2) & &= 4\alpha + 2\beta + \gamma \\ f'(2) &= te^{2t} & &= r'(2) & &= 4\alpha + \beta \end{aligned} \begin{cases} \alpha &= (t-1)e^{2t} + e^t \\ \beta &= (4-3t)e^{2t} - 4e^t \\ \gamma &= (2t-3)e^{2t} + 4e^t \end{cases}$$

$$\begin{aligned} r(\mathbf{A}) &= \alpha \mathbf{A}^2 + \beta \mathbf{A} + \gamma \mathbf{I}_3 \\ &= \alpha \begin{bmatrix} 4 & -6 & -18 \\ 6 & -11 & -27 \\ -4 & 8 & 16 \end{bmatrix} + \beta \begin{bmatrix} 1 & 0 & -3 \\ 1 & -1 & -6 \\ -1 & 2 & 5 \end{bmatrix} + \gamma \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 4\alpha + \beta + \gamma & -6\alpha & -18\alpha - 3\beta \\ 6\alpha + \beta & -11\alpha - \beta - \gamma & -27\alpha - 6\beta \\ -4\alpha - \beta & 8\alpha + 2\beta & 16\alpha + 5\beta + \gamma \end{bmatrix} \\ &= \begin{bmatrix} (3t-3)e^{2t} + 4e^t & (6-6t)e^{2t} - 6e^t & (6-9t)e^{2t} - 6e^t \\ (3t-2)e^{2t} + 2e^t & (4-6t)e^{2t} - 3e^t & (3-9t)e^{2t} - 3e^t \\ -te^{2t} & 2te^{2t} & (3t+1)e^{2t} \end{bmatrix} \end{aligned}$$

and so,

$$\begin{aligned} \begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix} &= e^{\mathbf{A}t} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} (3-3t)e^{2t} - 2e^t \\ (2-3t)e^{2t} - e^t \\ te^{2t} \end{bmatrix} \end{aligned}$$

△

33.10 Bilinear Maps

Let V and W be vector spaces over a field K. A *bilinear map* on V and W is a map $\tau : V \times W \to K$ such that for all $\mathbf{v}$,$\mathbf{v}_1$,$\mathbf{v}_2 \in V$, $\mathbf{w}$,$\mathbf{w}_1$,$\mathbf{w}_2 \in W$, and α,$\beta \in K$,

1. $\tau(\alpha\mathbf{v}_1 + \beta\mathbf{v}_2,\mathbf{w}) = \alpha\tau(\mathbf{v}_1,\mathbf{w}) + \beta\tau(\mathbf{v}_1,\mathbf{w})$;
2. $\tau(\mathbf{v},\alpha\mathbf{w}_1 + \beta\mathbf{w}_2) = \alpha\tau(\mathbf{v},\mathbf{w}_1) + \beta\tau(\mathbf{v},\mathbf{w}_2)$.

That is, for a fixed $\mathbf{v}$, $\tau(\mathbf{v},\mathbf{w})$ is linear in $\mathbf{w}$, and for a fixed $\mathbf{w}$, $\tau(\mathbf{v},\mathbf{w})$ is linear in $\mathbf{v}$.

So, if we fix bases of V and W, a bilinear map is completely determined by its actions on the basis vectors. Let $(\mathbf{e}_i)_{i=1}^n$ and $(\mathbf{f}_i)_{i=1}^m$ be bases of V and W, respectively. Then, the $n \times m$ matrix $\mathbf{A} = (\alpha_{i,j})$ defined by $\alpha_{i,j} = \tau(\mathbf{e}_i,\mathbf{f}_j)$ is said to be the matrix of τ with respect to the bases $(\mathbf{e}_i)_{i=1}^n$ and $(\mathbf{f}_i)_{i=1}^m$.

Then, for any vectors,

$$\mathbf{v} = \sum_{i=1}^n a_i\mathbf{e}_i, \quad \mathbf{w} = \sum_{i=1}^m b_i\mathbf{f}_i$$

we have by bilinearity,

$$\begin{aligned}
\tau(\mathbf{v},\mathbf{w}) &= \tau\left(\sum_{i=1}^{n} a_i\mathbf{e}_i, \sum_{j=1}^{m} b_j\mathbf{f}_j\right) \\
&= \sum_{i=1}^{n} a_i\tau\left(\mathbf{e}_i, \sum_{j=1}^{m} b_j\mathbf{f}_j\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i\tau(\mathbf{e}_i,\mathbf{f}_j)b_j \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i\alpha_{i,j}b_j \\
&= \mathbf{v}^\top\mathbf{A}\mathbf{w}
\end{aligned}$$

So, for any fixed bases of V and W, every bilinear map on V and W corresponds to a unique $n \times m$ matrix, and conversely, every $n \times m$ matrix determines a bilinear map.

Example. Write down the matrix corresponding to the bilinear map τ defined by

$$\tau(\mathbf{v},\mathbf{w}) := v_1w_1 - v_1w_2 + 2v_2w_1$$

First, expand out the formula above with a general matrix:

$$\begin{aligned}
\tau(\mathbf{v},\mathbf{w}) &= \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}^\top \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \\
&= \begin{bmatrix} v_1 & v_2 \end{bmatrix} \begin{bmatrix} aw_1 + bw_2 \\ cw_1 + dw_2 \end{bmatrix} \\
&= av_1w_1 + bv_1w_2 + cv_2w_1 + dv_2w_2
\end{aligned}$$

Equating coefficients, we have,

$$\begin{aligned}
a &= 1 \\
b &= -1 \\
c &= 2 \\
d &= 0
\end{aligned}$$

$$\mathbf{A} = \begin{bmatrix} 1 & -1 \\ 2 & 0 \end{bmatrix}$$

△

33.10.1 Bilinear Forms

Theorem 33.10.1. *Let $\mathbf{A}$ be the matrix of the bilinear map $\tau : V \times W \to K$ with respect to the bases $(\mathbf{e}_i)_{i=1}^n$ and $(\mathbf{f}_i)_{i=1}^m$ of V and W, and let $\mathbf{B}$ be its matrix with respect to the bases $(\mathbf{e}'_i)_{i=1}^n$ and $(\mathbf{f}'_i)_{i=1}^m$ of V and W. Let P and Q be the change of basis matrices. Then,*

$$\mathbf{B} = \mathbf{P}^\top\mathbf{A}\mathbf{Q}$$

Now, we consider the case where $W = V$. Then, a bilinear map $\tau : V \times V \to K$ is called a *bilinear form* on V.

The previous theorem then becomes,

Theorem 33.10.2. *Let* $\mathbf{A}$ *be the matrix of the bilinear form* τ *on* V *with respect to the basis* $(\mathbf{e}_i)_{i=1}^n$ *of* V*, and let* $\mathbf{B}$ *be its matrix with respect to the basis* $(\mathbf{e}'_i)_{i=1}^n$*, and let* P *be the change of basis matrix. Then,*

$$\mathbf{B} = \mathbf{P}^\top \mathbf{A}\mathbf{P}$$

If $\mathbf{A}$ and $\mathbf{B}$ satisfy this relation, they are said to be *congruent* matrices.

Note that congruence is distinct from similarity in that, if τ is a bilinear form on V and T is a linear operator on V, it might be the case that τ and T have the same matrix in some specific basis of V, but they do not necessarily have the same matrix in any other basis of V.

The *rank* of a bilinear form τ is the rank of its matrix $\mathbf{A}$.

A vector $\mathbf{v} \in K^n$ is zero if and only if $\mathbf{v}^\top \mathbf{w} = \mathbf{0}$ for all vectors $\mathbf{w} \in K^n$. Since

$$\tau(\mathbf{v},\mathbf{w}) = \mathbf{v}^\top \mathbf{A}\mathbf{w}$$

the kernel of $\mathbf{A}$ is given by

$$\operatorname{span}\{\mathbf{w} \in V : \forall \mathbf{v} \in V, \tau(\mathbf{v},\mathbf{w}) = 0\}$$

This set is also called the *right radical* of τ.

Similarly, the kernel of $\mathbf{A}^\top$ is given by

$$\operatorname{span}\{\mathbf{v} \in V : \forall \mathbf{w} \in V, \tau(\mathbf{v},\mathbf{w}) = 0\}$$

and is also called the *left radical* of τ.

Since $\mathbf{A}$ and $\mathbf{A}^\top$ have the same rank, the left and right radicals both have dimension $n - r$ where r is the rank of τ. In particular, the rank of τ is n if and only if the left and right radicals have dimension 0, and we say that τ is nondegenerate. That is, τ is nondegenerate if and only if its matrix (in any basis) is nonsingular.

A bilinear form τ on V is *symmetric* if $\tau(\mathbf{w},\mathbf{v}) = \tau(\mathbf{v},\mathbf{w})$ for all $\mathbf{v},\mathbf{w} \in V$. τ is *antisymmetric* or *alternating* if $\tau(\mathbf{v},\mathbf{v}) = 0$ for all $\mathbf{v} \in V$.

The antisymmetry condition implies that for all $\mathbf{v},\mathbf{w} \in V$,

$$\tau(\mathbf{v} + \mathbf{w},\mathbf{v} + \mathbf{w}) = \tau(\mathbf{v},\mathbf{w}) + \tau(\mathbf{w},\mathbf{v}) = 0$$

and hence,

$$\tau(\mathbf{v},\mathbf{w}) = -\tau(\mathbf{w},\mathbf{v})$$

If $2 \neq 0$ in K, then the converse of this result holds: that is, $\tau(\mathbf{v},\mathbf{w}) = -\tau(\mathbf{w},\mathbf{v})$ for all $\mathbf{v} \in V$ implies that $\tau(\mathbf{v},\mathbf{v}) = 0$ for all $\mathbf{v} \in V$.

An $n \times n$ matrix $\mathbf{A}$ is *symmetric* if $\mathbf{A}^\top = \mathbf{A}$, and *antisymmetric* if $\mathbf{A}^\top = -\mathbf{A}$ and $\mathbf{A}$ has zeros along the diagonal.

Theorem 33.10.3. *The bilinear form* τ *is symmetric (resp. antisymmetric) if and only if its matrix (with respect to any basis) is symmetric (resp. antisymmetric).*

One important example of a symmetric bilinear form is when $V = \mathbb{R}^n$ and τ is defined by

$$\tau(\mathbf{v},\mathbf{w}) = \sum_{i=1}^{n} v_i w_i$$

$$= \mathbf{v} \cdot \mathbf{w}$$

This form is called the *dot product* or *scalar product.* This bilinear form has matrix form equal to the identity matrix $\mathbf{I}_n$ with respect to the standard basis of $\mathbb{R}^n$.

Theorem 33.10.4. *Suppose that $2 \neq 0$ in K. Then, any bilinear form τ can be written uniquely as $\tau_1 + \tau_2$, where τ_1 is symmetric and τ_2 is antisymmetric.*

Proof. For existence, take $\tau_1(\mathbf{v},\mathbf{w}) = \frac{1}{2}\big(\tau(\mathbf{v},\mathbf{w}) + \tau(\mathbf{w},\mathbf{v})\big)$ and $\tau_2(\mathbf{v},\mathbf{w}) = \frac{1}{2}\big(\tau(\mathbf{v},\mathbf{w}) - \tau(\mathbf{w},\mathbf{v})\big)$.

For uniqueness, suppose τ also decomposes into $\tau_1' + \tau_2'$, with τ_1' symmetric and τ_2' antisymmetric. Then, by symmetry and antisymmetry,

$$\begin{aligned}
\tau_1(\mathbf{v},\mathbf{w}) &= \frac{1}{2}\big(\tau_1'(\mathbf{v},\mathbf{w}) + \tau_1'(\mathbf{w},\mathbf{v}) + \tau_2'(\mathbf{v},\mathbf{w}) + \tau_2'(\mathbf{w},\mathbf{v})\big) \\
&= \frac{1}{2}\big(\tau_1'(\mathbf{v},\mathbf{w}) + \tau_1'(\mathbf{v},\mathbf{w}) + \tau_2'(\mathbf{v},\mathbf{w}) - \tau_2'(\mathbf{v},\mathbf{w})\big) \\
&= \frac{1}{2}\big(\tau_1'(\mathbf{v},\mathbf{w}) + \tau_1'(\mathbf{v},\mathbf{w})\big) \\
&= \tau_1'(\mathbf{v},\mathbf{w})
\end{aligned}$$

so $\tau_1 = \tau_1'$, and hence,

$$\begin{aligned}
\tau_2 &= \tau - \tau_1 \\
&= \tau - \tau_1' \\
&= \tau_2
\end{aligned}$$

so the decomposition is unique.

Note that $\frac{1}{2}$ has to exist in K for the first chain of equations to be meaningful, so we require that $2 \neq 0$ in K. ■

33.10.2 Quadratic Forms

Let V be a vector space over a field K. A *quadratic form* on V is a function $q : V \to K$ such that,

$$q(\lambda\mathbf{v}) = \lambda^2 q(\mathbf{v})$$

for all $\mathbf{v} \in V$ and $\lambda \in K$, and the function $\tau_q : V \times V \to K$ defined by,

$$\tau_q(\mathbf{v},\mathbf{w}) := q(\mathbf{v} + \mathbf{w}) - q(\mathbf{v}) - q(\mathbf{w})$$

is a symmetric bilinear form on V.

Given a symmetric bilinear form τ, we can also define a quadratic form by,

$$q_\tau(\mathbf{v}) := \tau(\mathbf{v},\mathbf{v})$$

These processes are almost inverse to each other, in that, given a quadratic form q and a bilinear form τ, we have,

$$q_{\tau_q} = 2q, \qquad \tau_{q_\tau} = 2\tau$$

so, if $2 \neq 0$ in K, there is a bijection between quadratic forms and symmetric bilinear forms given by,

$$q \mapsto \frac{1}{2}\tau_q, \quad \tau \mapsto q_\tau$$

If $2 = 0$ in K, then this correspondence does not hold, and indeed there exist quadratic forms that are not of the form $\tau(-,-)$ for any symmetric bilinear form τ on V. In general, the standard forms of quadratic and bilinear forms on vector spaces where $2 = 0$ in the underlying field is quite different from the case where $2 \neq 0$.

From this point onwards, we will assume that $2 = 1 + 1 \neq 0$ in the field K.

33.10.3 Bases for Quadratic Forms

Let $(\mathbf{e}_i)_{i=1}^n$ be a basis of V, and let $\mathbf{A} = (\alpha_{i,j})$ be the matrix of a symmetric bilinear form τ with respect to this basis. $\mathbf{A}$ is then also said to be the matrix of the quadratic form $q := q_\tau$ with respect to this basis.

Note that $\mathbf{A}$ is symmetric, as τ is symmetric. Then,

$$q(\mathbf{v}) = \mathbf{v}^\top \mathbf{A}\mathbf{v}$$

just like in the case for bilinear maps, so we can easily write out the matrix

Example. Write down the matrix corresponding to the quadratic form q defined by

$$q\big([x,y,z]\big) := 8x^2 - 7y^2 + 8z^2 + 8xy - 2xz + 8yz$$

As we did for bilinear forms, we again expand $q(\mathbf{v}) = \mathbf{v}^\top \mathbf{A}\mathbf{v}$ for a general matrix $\mathbf{A}$ and vector $\mathbf{v}$, then compare coefficients:

$$\begin{aligned}\begin{bmatrix} x & y & z \end{bmatrix} \begin{bmatrix} a & d & e \\ d & b & f \\ e & f & c \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} &= ax^2 + by^2 + cz^2 + 2dxy + 2exz + 2fyz \\ &= 8x^2 - 7y^2 + 8z^2 + 8xy - 2xz + 8yz \\ \mathbf{A} &= \begin{bmatrix} 8 & 4 & -1 \\ 4 & -7 & 4 \\ -1 & 4 & 8 \end{bmatrix}\end{aligned}$$

△

Theorem 33.10.5. *Let V be a vector space of dimension n equipped with a symmetric bilinear form τ (or equivalently, with a quadratic form q).*

Then, there exist a basis $(\mathbf{b}_i)_{i=1}^n$ of V and constants $(\beta_i)_{i=1}^n$ such that

$$\tau(\mathbf{b}_i,\mathbf{b}_j) = \begin{cases} \beta_i & i = j \\ 0 & i \neq j \end{cases}$$

Equivalently,

- *Give any quadratic form q on V, there exist a basis $(\mathbf{b}_i)_{i=1}^n$ of V and constants $(\beta_i)_{i=1}^n$ such that*

$$q\left(\sum_{i=1}^n x_i \mathbf{b}_i\right) = \sum_{i=1}^n \beta_i x_i^2$$

- *Any symmetric matrix $\mathbf{A}$ is congruent to a diagonal matrix. That is, there exists an invertible matrix $\mathbf{P}$ such that $\mathbf{A} = \mathbf{P}^\top \mathbf{D}\mathbf{P}$, where $\mathbf{D}$ is a diagonal matrix.*

We give an algorithm to find the matrix $\mathbf{P}$.

Algorithm 6 Orthogonal Diagonalisation

```
1: Determine the eigenvalues of $\mathbf{A}$.
2: For each eigenvalue $\lambda_i$, find the corresponding eigenspace, $\ker(\mathbf{A} - \lambda_i \mathbf{I}_n) = \operatorname{span}\big((\mathbf{v}_j)_{j=1}^k\big)$.
3: Write out a normal diagonalisation of $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$, recalling that $\mathbf{D}$ is the matrix with the eigenvalues of $\mathbf{A}$ along the diagonal, and $\mathbf{P}$ is the matrix with the corresponding eigenvectors as columns.
4: Check if the columns of $\mathbf{P}$ are orthogonal by checking if their scalar product is zero or not. If all columns are orthogonal, we are done.
5: Otherwise, apply the Gram-Schmidt process (§33.10.4) to the columns of $P$.
```

Theorem 33.10.6. *For a symmetric real matrix, eigenvectors with distinct eigenvalues are always orthogonal.*

This allows us a slight shortcut in the algorithm above: we only need to check the scalar product of eigenvectors that share the same eigenvalue.

Theorem 33.10.7. *A quadratic form q over $\mathbb{C}$ has the form*

$$q(\mathbf{v}) = \sum_{i=1}^{r} x_i^2$$

with respect to a suitable basis, where $r = \operatorname{rank}(q)$.

Equivalently, given a symmetric matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$, there is an invertible matrix $\mathbf{P} \in \mathbb{C}^{n\times n}$ such that $\mathbf{P}^\top \mathbf{A}\mathbf{P} = \mathbf{B}$, where $\mathbf{B} = (\beta_{i,j})$ is a diagonal matrix with

$$\beta_{i,i} = \begin{cases} 1 & i \in [1,r] \\ 0 & i \in (r,n] \end{cases}$$

where $r = \operatorname{rank}(\mathbf{A})$.

In particular, up to a change of basis, a quadratic form on $\mathbb{C}^n$ is uniquely determined by its rank, and we say that the rank is the only *invariant* of a quadratic form over $\mathbb{C}$.

Theorem 33.10.8 (Sylvester)**.** *A quadratic form q over $\mathbb{R}$ has the form*

$$q(\mathbf{v}) = \sum_{i=1}^{t} x_i^2 - \sum_{i=1}^{u} x_{t+i}^2$$

with respect to a suitable basis, where $t + u = \operatorname{rank}(q)$.

Equivalently, given a symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, there is an invertible matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ such that $\mathbf{A} = \mathbf{P}\mathbf{B}\mathbf{P}^\top$, where $\mathbf{B} = (\beta_{i,j})$ is a diagonal matrix with

$$\beta_{i,i} = \begin{cases} 1 & i \in [1,t] \\ -1 & i \in (t,t+u] \\ 0 & i \in (t+u,n] \end{cases}$$

where $t + u = \operatorname{rank}(\mathbf{A})$.

The numbers t and u of positive and negative terms are invariants of q, and the pair of integers (t,u) is the *signature* of q.

Theorem 33.10.9 (Sylvester's Law of Inertia)**.** *Suppose that q is a quadratic form on the vector space V over $\mathbb{R}$, and that $(\mathbf{e}_i)_{i=1}^n$ and $(\mathbf{e}'_i)_{i=1}^n$ are two bases of V such that*

$$q\left(\sum_{i=1}^{n} x_i \mathbf{e}_i\right) = \sum_{i=1}^{t} x_i^2 - \sum_{i=1}^{u} x_{t+i}^2$$

and

$$q\left(\sum_{i=1}^{n} x_i \mathbf{e}'_i\right) = \sum_{i=1}^{t'} x_i^2 - \sum_{i=1}^{u'} x_{t'+i}^2$$

Then, $t = t'$ and $u = u'$.

33.10.4 The Gram-Schmidt Process

In this section, we will take $K = \mathbb{R}$.

Let V be a vector space over K of dimension n, and let q be a quadratic form on V, with associated symmetric bilinear form τ.

The quadratic form q is *positive definite* if $q(\mathbf{v}) > 0$ for all non-zero $\mathbf{v} \in V$. τ is also called positive definite if q is positive definite.

A quadratic form q is positive definite if and only if $t = n$ and $u = 0$ in Sylvester's theorem. That is, if q has signature $(n,0)$.

A vector space V over $\mathbb{R}$ equipped with a positive definite symmetric bilinear form τ is called a *Euclidean space.* In this case, Sylvester's theorem just states that there is a basis $(\mathbf{e}_i)_{i=1}^{n}$ of V with respect to which the matrix of q is the identity matrix $\mathbf{I}_n$.

In other words, the basis vectors are all unit vectors, and they are all orthogonal to each other. Such a basis is called an *orthonormal basis*, and more generally, any set of vectors such that the vectors are all unit length and orthogonal to each other is called *orthonormal.*

If V is a Euclidean space with an orthonormal basis, then any positive definite symmetric bilinear form τ is equivalent to the dot product, and we will write $\mathbf{v} \cdot \mathbf{w}$ for $\tau(\mathbf{v},\mathbf{w})$.

Given a (finite) set of linearly independent vectors $S = (\mathbf{v}_i)_{i=1}^{k}$, the *Gram-Schmidt process* generates an orthogonal set $S' = (\mathbf{u}_i)_{i=1}^{k}$ that spans the same k-dimensional subspace of V as S.

Algorithm 7 Gram-Schmidt Process

1. Define the projection operator by

$$\mathrm{proj}_{\mathbf{u}}(\mathbf{v}) := \frac{\mathbf{v} \cdot \mathbf{u}}{\mathbf{u} \cdot \mathbf{u}}\mathbf{u}$$

   And if $\mathbf{u} = \mathbf{0}$, then we define $\mathrm{proj}_{\mathbf{u}}(\mathbf{v}) = \mathbf{0}$. This operator projects $\mathbf{v}$ orthogonally onto the line spanned by $\mathbf{u}$.

2. Recursively calculate the sequence of orthogonal vectors $(\mathbf{u}_i)_{i=1}^{n}$ using,

$$\begin{aligned}
\mathbf{u}_1 &= \mathbf{v}_1 \\
\mathbf{u}_2 &= \mathbf{v}_2 - \mathrm{proj}_{\mathbf{u}_1}(\mathbf{v}_2) \\
\mathbf{u}_3 &= \mathbf{v}_3 - \mathrm{proj}_{\mathbf{u}_1}(\mathbf{v}_3) - \mathrm{proj}_{\mathbf{u}_2}(\mathbf{v}_3) \\
\mathbf{u}_4 &= \mathbf{v}_4 - \mathrm{proj}_{\mathbf{u}_1}(\mathbf{v}_4) - \mathrm{proj}_{\mathbf{u}_2}(\mathbf{v}_4) - \mathrm{proj}_{u_3}(\mathbf{v}_4) \\
&\vdots \\
\mathbf{u}_n &= \mathbf{v}_n - \sum_{i=1}^{n-1} \mathrm{proj}_{\mathbf{u}_i}(\mathbf{v}_n)
\end{aligned}$$

3. Normalise each vector to obtain the orthonormal sequence $(\mathbf{e}_i)_{i=1}^{n}$:

$$\mathbf{e}_i = \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|}$$

The calculation of the orthogonal sequence $(\mathbf{u}_i)_{i=1}^{n}$ is more specifically called *Gram-Schmidt orthogonalisation*, while the total calculation of the orthonormal sequence $(\mathbf{e}_i)_{i=1}^{n}$ is called *Gram-Schmidt orthonormalisation* as the vectors are normalised.

Geometrically, to compute the next vector $\mathbf{u}_k$, we project $\mathbf{v}_k$ onto the subspace $U = \operatorname{span}\big((\mathbf{u}_i)_{i=1}^{k-1}\big)$ spanned by the previous orthogonal vectors, which by construction, is the same as the subspace spanned by the first $k-1$ original vectors, $\operatorname{span}\big((\mathbf{v}_i)_{i=1}^{k-1}\big)$. The vector $\mathbf{u}_k$ is then defined to be the difference between $\mathbf{v}_k$ and this projection, guaranteed to be orthogonal to all the vectors in U. Because of this projection action, the vectors $(\mathbf{u}_i)_{i=1}^{n}$ are sometimes also denoted $(\mathbf{v}_i^{\perp})_{i=1}^{n}$.

Example. In the last example, we found that the matrix corresponding to the quadratic form q defined by

$$q\big([x,y,z]\big) \coloneqq 8x^2 - 7y^2 + 8z^2 + 8xy - 2xz + 8yz$$

is given by

$$\mathbf{A} = \begin{bmatrix} 8 & 4 & -1 \\ 4 & -7 & 4 \\ -1 & 4 & 8 \end{bmatrix}$$

Now, find an orthonormal basis of V such that the matrix of q is diagonal. You may use that $c_{\mathbf{A}}(x) = -(x-9)^2(x+9)$ without proof.

We first find an ordinary diagonalisation of $\mathbf{A}$. From the characteristic equation, $\mathbf{A}$ has eigenvalues $\lambda_1 = 9$ and $\lambda_2 = -9$, so,

$$(\mathbf{A} - \lambda_1 \mathbf{I}_3)\mathbf{v} = \mathbf{0} \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \begin{bmatrix} 1 & -4 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$a = 4b - c$$

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = b \begin{bmatrix} 4 \\ 1 \\ 0 \end{bmatrix} + c \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \longrightarrow \ker(\mathbf{A} - \lambda_1 \mathbf{I}_3) = \operatorname{span}\left(\begin{bmatrix} 4 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \right)$$

$$(\mathbf{A} - \lambda_2 \mathbf{I}_3)\mathbf{v} = \mathbf{0} \quad \xrightarrow[\text{reduce}]{\text{row}} \quad \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 4 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$a = c$$
$$b = -4c$$

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = c \begin{bmatrix} 1 \\ -4 \\ 1 \end{bmatrix} \longrightarrow \ker(\mathbf{A} - \lambda_2 \mathbf{I}_3) = \operatorname{span}\left(\begin{bmatrix} 1 \\ -4 \\ 1 \end{bmatrix} \right)$$

So, $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$ with

$$\mathbf{D} = \begin{bmatrix} 9 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & -9 \end{bmatrix}$$

$$\mathbf{P} = \begin{bmatrix} -1 & 4 & 1 \\ 0 & 1 & -4 \\ 1 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 1 \\ 0 \end{bmatrix} = 4$$

$$\neq 0$$

so the first two columns of $\mathbf{P}$ are not orthogonal. The last column must be orthogonal to the first two by Theorem 33.10.6

$$\begin{aligned}
\mathbf{v}_1 &= \mathbf{p}_1 \\
\mathbf{v}_2 &= \mathbf{p}_2 - \frac{\mathbf{p}_2 \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1}\mathbf{v}_1 \\
&= \mathbf{p}_2 - \frac{-4}{2}\mathbf{v}_1 \\
&= \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}
\end{aligned}$$

$$\mathbf{P} = \begin{bmatrix} -1 & 2 & 1 \\ 0 & 1 & -4 \\ 1 & 2 & 1 \end{bmatrix}$$

Normalising:

$$\mathbf{P} = \begin{bmatrix} -\frac{1}{\sqrt{2}} & \frac{2}{3} & \frac{1}{3\sqrt{2}} \\ 0 & \frac{1}{3} & -\frac{4}{3\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{2}{3} & \frac{1}{3\sqrt{2}} \end{bmatrix}$$

So, an orthonormal basis of V with q diagonal is given by,

$$\left\{ \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix}, \begin{bmatrix} \frac{2}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix}, \begin{bmatrix} \frac{1}{3\sqrt{2}} \\ -\frac{4}{3\sqrt{2}} \\ \frac{1}{3\sqrt{2}} \end{bmatrix} \right\}$$

△

33.10.5 Orthogonal Transformations

In a Euclidean space V, the scalar product gives us a notion of length of a vector and angles between vectors, so we might be interested in what kind of transformations preserve these quantities.

A linear map $T : V \to V$ is *orthogonal* if it preserves the scalar product on V. That is, $T(\mathbf{v}) \cdot T(\mathbf{w}) = \mathbf{v} \cdot \mathbf{w}$ for all $\mathbf{v},\mathbf{w} \in V$.

A $n \times n$ matrix $\mathbf{A}$ is *orthogonal* if $\mathbf{A}^\top \mathbf{A} = \mathbf{A}\mathbf{A}^\top = \mathbf{I}_n$, or equivalently, $\mathbf{A}^\top = \mathbf{A}^{-1}$.

Lemma 33.10.10. *A linear map $T : V \to V$ is orthogonal if and only if its matrix $\mathbf{A}$ is orthogonal.*

Theorem 33.10.11. $\det(\mathbf{A}) = \pm 1$ *for any orthogonal matrix $\mathbf{A}$.*

Proof.

$$\begin{aligned}
\det(\mathbf{A})^2 &= \det(\mathbf{A}) \det(\mathbf{A}^\top) \\
&= \det(\mathbf{A}^\top \mathbf{A}) \\
&= \det(\mathbf{I}_n) \\
&= 1
\end{aligned}$$

so $\det(\mathbf{A}) = \sqrt{1} = \pm 1$. ■

Theorem 33.10.12. *A linear map $T : V \to V$ is orthogonal if and only if $\big(T(\mathbf{e}_i)\big)_{i=1}^{n}$ is an orthonormal basis of V.*

Theorem 33.10.13 (QR Decomposition)**.** *Any invertible matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ can be written as $\mathbf{A} = \mathbf{QR}$ where $\mathbf{Q}$ is orthogonal and $\mathbf{R}$ is upper-triangular.*

Algorithm 8 QR Decomposition

1: Perform the Gram-Schmidt process on the columns of $\mathbf{A} = [\mathbf{a}_1|\mathbf{a}_2|\dots|\mathbf{a}_n]$ to obtain the orthonormalised set $(\mathbf{e}_i)_{i=1}^{n}$.

2: Express the columns of $\mathbf{A}$ in the orthonormal basis:

$$\begin{aligned}
\mathbf{a}_1 &= \langle \mathbf{e}_1,\mathbf{a}_1\rangle \mathbf{e}_1\\
\mathbf{a}_2 &= \langle \mathbf{e}_1,\mathbf{a}_2\rangle \mathbf{e}_1 + \langle \mathbf{e}_2,\mathbf{a}_2\rangle \mathbf{e}_2\\
\mathbf{a}_3 &= \langle \mathbf{e}_1,\mathbf{a}_3\rangle \mathbf{e}_1 + \langle \mathbf{e}_2,\mathbf{a}_3\rangle \mathbf{e}_2 + \langle \mathbf{e}_3,\mathbf{a}_3\rangle \mathbf{e}_3\\
\mathbf{a}_4 &= \langle \mathbf{e}_1,\mathbf{a}_4\rangle \mathbf{e}_1 + \langle \mathbf{e}_2,\mathbf{a}_4\rangle \mathbf{e}_2 + \langle \mathbf{e}_3,\mathbf{a}_4\rangle \mathbf{e}_3 + \langle \mathbf{e}_4,\mathbf{a}_4\rangle \mathbf{e}_4\\
&\ \ \vdots\\
\mathbf{a}_k &= \sum_{i=1}^{k} \langle \mathbf{e}_i,\mathbf{a}_k\rangle \mathbf{e}_i
\end{aligned}$$

noting that $\langle \mathbf{e}_i,\mathbf{a}_i\rangle = \|\mathbf{v}_i^{\perp}\|$ (which you had to calculate before during the normalisation step).

3: the above set of equations can be packaged into matrix form,

$$\mathbf{A} = \mathbf{QR}$$

where,

$$\begin{aligned}
\mathbf{Q} &= [\mathbf{e}_1|\mathbf{e}_2|\dots|\mathbf{e}_n]\\
\mathbf{R} &= \begin{bmatrix}
\|\mathbf{v}_1^{\perp}\| & \langle \mathbf{e}_1,\mathbf{a}_2\rangle & \langle \mathbf{e}_1,\mathbf{a}_3\rangle & \langle \mathbf{e}_1,\mathbf{a}_4\rangle & \cdots & \langle \mathbf{e}_1,\mathbf{a}_n\rangle\\
0 & \|\mathbf{v}_2^{\perp}\| & \langle \mathbf{e}_2,\mathbf{a}_3\rangle & \langle \mathbf{e}_2,\mathbf{a}_4\rangle & \cdots & \langle \mathbf{e}_2,\mathbf{a}_n\rangle\\
0 & 0 & \|\mathbf{v}_3^{\perp}\| & \langle \mathbf{e}_3,\mathbf{a}_4\rangle & \cdots & \langle \mathbf{e}_3,\mathbf{a}_n\rangle\\
0 & 0 & 0 & \|\mathbf{v}_4^{\perp}\| & \cdots & \langle \mathbf{e}_4,\mathbf{a}_n\rangle\\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\
0 & 0 & 0 & 0 & \cdots & \|\mathbf{v}_n^{\perp}\|
\end{bmatrix}
\end{aligned}$$

Example. Find a QR decomposition of,

$$\mathbf{B} = \begin{bmatrix} 2 & 1 & -1\\ 1 & 0 & 2\\ 2 & -1 & 3 \end{bmatrix}$$

We perform the Gram-Schmidt process on the columns of $\mathbf{B} = [\mathbf{b}_1|\mathbf{b}_2|\mathbf{b}_3]$.

$$\begin{aligned}
\mathbf{e}_1 &= \frac{\mathbf{b}_1}{\|\mathbf{b}_1\|}\\
&= \frac{1}{3}\begin{bmatrix}2\\1\\2\end{bmatrix}
\end{aligned}$$

$$
\begin{aligned}
&= \begin{bmatrix} \frac{2}{3} \\ \frac{1}{3} \\ \frac{2}{3} \end{bmatrix} \\
\mathbf{b}_2^{\perp} &= \mathbf{b}_2 - \langle \mathbf{b}_2, \mathbf{e}_1 \rangle \mathbf{e}_1 \\
&= \mathbf{b}_2 - 0\mathbf{e}_1 \\
&= \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \\
\mathbf{e}_2 &= \frac{\mathbf{b}_2^{\perp}}{\|\mathbf{b}_2^{\perp}\|} \\
&= \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ -\frac{1}{\sqrt{2}} \end{bmatrix} \\
\mathbf{b}_3^{\perp} &= \mathbf{b}_3 - \langle \mathbf{b}_3, \mathbf{e}_1 \rangle \mathbf{e}_1 - \langle \mathbf{b}_3, \mathbf{e}_2 \rangle \mathbf{e}_2 \\
&= \mathbf{b}_3 - 2\mathbf{q}_1 - (-2\sqrt{2})\mathbf{e}_2 \\
&= \begin{bmatrix} -\frac{1}{3} \\ \frac{4}{3} \\ -\frac{1}{3} \end{bmatrix} \\
\mathbf{e}_3 &= \frac{\mathbf{b}_3^{\perp}}{\|\mathbf{b}_3^{\perp}\|} \\
&= \frac{1}{\sqrt{2}} \begin{bmatrix} -\frac{1}{3} \\ \frac{4}{3} \\ -\frac{1}{3} \end{bmatrix} \\
&= \begin{bmatrix} -\frac{\sqrt{2}}{6} \\ \frac{2\sqrt{2}}{3} \\ -\frac{\sqrt{2}}{6} \end{bmatrix} \\
\mathbf{Q} &= [\mathbf{e}_1 | \mathbf{e}_2 | \mathbf{e}_3] \\
&= \begin{bmatrix} \frac{2}{3} & \frac{1}{\sqrt{2}} & -\frac{\sqrt{2}}{6} \\ \frac{1}{3} & 0 & \frac{2\sqrt{2}}{3} \\ \frac{2}{3} & -\frac{1}{\sqrt{2}} & -\frac{\sqrt{2}}{6} \end{bmatrix} \\
\mathbf{R} &= \begin{bmatrix} 3 & 0 & 2 \\ 0 & \sqrt{2} & -2\sqrt{2} \\ 0 & 0 & \sqrt{2} \end{bmatrix} \\
\mathbf{B} &= \begin{bmatrix} \frac{2}{3} & \frac{1}{\sqrt{2}} & -\frac{\sqrt{2}}{6} \\ \frac{1}{3} & 0 & \frac{2\sqrt{2}}{3} \\ \frac{2}{3} & -\frac{1}{\sqrt{2}} & -\frac{\sqrt{2}}{6} \end{bmatrix} \begin{bmatrix} 3 & 0 & 2 \\ 0 & \sqrt{2} & -2\sqrt{2} \\ 0 & 0 & \sqrt{2} \end{bmatrix}
\end{aligned}
$$

△

33.10.6 Orthonormal Bases for Bilinear Forms

Suppose we have a Euclidean space V, and a linear operator $T : V \to V$ or a quadratic form q on V (not necessarily the same quadratic form as the one providing the Euclidean structure). Is it always possible to find an orthonormal basis of V such that the matrix of q has a simple form? Notice that we are now handling two different matrices simultaneously: we're trying to optimise the matrix of q, while still keeping the matrix of the original quadratic form as the identity.

It turns out that this is also a question about linear operators: given any bilinear form τ on V, we have

$$\tau(\mathbf{v},\mathbf{w}) = \mathbf{v}^\top \mathbf{A}\mathbf{w}$$

and we can interpret matrix-vector multiplication as a linear map, so,

$$= \mathbf{v} \cdot T(\mathbf{w})$$

So, every bilinear form τ on V uniquely determines a linear operator T on V such that,

$$\tau(\mathbf{v},\mathbf{w}) = \mathbf{v} \cdot T(\mathbf{w})$$

where T is the linear operator corresponding to the matrix $\mathbf{A}$ of τ with entries $\mathbf{A}_{i,j} = \tau(\mathbf{e}_i,\mathbf{e}_j)$ for the standard basis $(\mathbf{e}_i)_{i=1}^n$ of V. Conversely, any linear operator T similarly determines a bilinear form τ, where bilinearity follows from the bilinearity of the scalar product and linearity of T.

So, once we have a fixed bilinear form providing the Euclidean structure (i.e., the scalar product), any other bilinear form τ on V can be obtained from applying a linear transformation to one of the arguments of the scalar product, so there is a bijection between bilinear forms and linear operators.

In particular, if T is any linear operator, then $(\mathbf{v},\mathbf{w}) \mapsto (T\mathbf{v}) \cdot \mathbf{w}$ is certainly a bilinear form, so there exists a unique linear operator S such that,

$$(T\mathbf{v}) \cdot \mathbf{w} = \mathbf{v} \cdot (S\mathbf{w})$$

for all $\mathbf{v},\mathbf{w} \in V$. Such a linear operator S is called the *adjoint* of T, and is alternatively denoted T^*.

If we have chosen an orthonormal basis, then the matrix of T^* is the transpose of the matrix of T. It follows that a linear operator is orthogonal if and only if $T^* = T^{-1}$.

A linear operator T is *selfadjoint* if $T^* = T$, or equivalently, if the bilinear form $\tau(\mathbf{v},\mathbf{w}) = \mathbf{v} \cdot (T\mathbf{w})$ is symmetric.

So, if V is a Euclidean space of dimension n, then the following problems are all equivalent:

- Given a quadratic form q on V, find an orthonormal basis of V that makes the matrix of q as simple as possible;
- Given a selfadjoint linear operator T on V, find an orthonormal basis of V that makes the matrix of T as simple as possible;
- Given an $n \times n$ symmetric real matrix $\mathbf{A}$, find an orthogonal matrix P such that $\mathbf{P}^\top \mathbf{A}\mathbf{P}$ is as simple as possible.

Lemma 33.10.14. *Let $\mathbf{A}$ be a $n \times n$ symmetric real matrix. Then, $\mathbf{A}$ has an eigenvalue in $\mathbb{R}$, and moreover, all complex eigenvalues of $\mathbf{A}$ lie in $\mathbb{R}$.*

Theorem 33.10.15 (Spectral Theorem). *Let V be a Euclidean space of dimension n. Then,*

- *Given any quadratic form q on V, there is an orthonormal basis $(\mathbf{f}_i)_{i=1}^n$ of V and constants $(\alpha_i)_{i=1}^n$ uniquely determined up to reordering, such that,*

$$q\left(\sum_{i=1}^n x_i \mathbf{f}_i\right) = \sum_{k=1}^n \alpha_i (x_i)^2$$

for all $x_1, \ldots, x_n \in \mathbb{R}$.

- *Given any selfadjoint linear operator $T : V \to V$, there is an orthonormal basis $(\mathbf{f}_i)_{i=1}^n$ of V consisting of eigenvectors of T.*
- *Given any $n \times n$ symmetric real matrix $\mathbf{A}$, there is an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^\top \mathbf{A} \mathbf{P} = \mathbf{P}^{-1} \mathbf{A} \mathbf{P}$ is a diagonal matrix.*

Example. TO DO △

33.10.7 Reduction of Second Degree Polynomial Equations

The general equation of a second degree polynomial in n variables $(x_i)_{i=1}^n$ is given by,

$$\sum_{i=1}^{n} \alpha_{i,i} x_i^2 + \sum_{i=1}^{n} \sum_{j=1}^{i-1} \alpha_{i,j} x_i x_j + \sum_{i=1}^{n} \beta_i x_i + \gamma = 0$$

for an $n \times n$ lower triangular matrix $\mathbf{A} = (\alpha_{i,j})$ of constants and n-dimensional vector $\mathbf{b} = (\beta_i)$ of constants. That is, there is a term for every variable squared, a term for every product of a pair of variables, a term for every variable alone, and a constant term. For any set of fixed coefficients, this equation describes a *quadric* (hyper)surface in n-dimensional Euclidean space.

For example, for the $n = 3$ case, the general polynomial is given by,

$$Ax^2 + By^2 + Cz^2 + Dxy + Exz + Fyz + Gx + Hy + Iz + J = 0$$

We can simplify these equation by applying various isometries to the coordinate basis. By the spectral theorem, we can apply orthogonal basis changes to eliminate the quadratic terms in mixed variables (the terms with coefficients D, E, and F in the example above). We do this by completing the square in the

$$\sum_{i=1}^{n} \alpha_{i,i} x_i^2 + \sum_{i=1}^{n} \sum_{j=1}^{i-1} \alpha_{i,j} x_i x_j$$

term to see what coordinate changes we should effect. For instance, suppose we have the equation,

$$x^2 + xy + y^2 + x = 0$$

We complete the square on $x^2 + xy + y^2$ to obtain,

$$\left(x + \frac{1}{2}y\right)^2 + \frac{3}{4}y^2 + x = 0$$

Now, we apply the linear transformation,

$$\begin{bmatrix} x \\ y \end{bmatrix} \mapsto \begin{bmatrix} x + \frac{1}{2}y \\ y \end{bmatrix}$$

giving,

$$x^2 + \frac{3}{4}y^2 + x - \frac{1}{2}y = 0$$

and we no longer have any mixed quadratic terms.

Now, whenever $\alpha_{i,i} \neq 0$, we can perform the translation isometry,

$$x_i \mapsto x_i - \frac{\beta_i}{2\alpha_{i,i}}$$

thus eliminating the term $\beta_i x_i$.

From the example above, we would then have,

$$\begin{bmatrix} x \\ y \end{bmatrix} \mapsto \begin{bmatrix} x - \frac{1}{2\cdot 1} \\ y - \frac{-\frac{1}{2}}{2\cdot\frac{3}{4}} \end{bmatrix} = \begin{bmatrix} x - \frac{1}{2} \\ y + \frac{1}{3} \end{bmatrix}$$

giving,

$$\begin{aligned} \left(x - \frac{1}{2}\right)^2 + \frac{3}{4}\left(y + \frac{1}{3}\right)^2 + \left(x - \frac{1}{2}\right) - \frac{1}{2}\left(y + \frac{1}{3}\right) &= 0 \\ \left(x^2 - x + \frac{1}{4}\right) + \frac{3}{4}\left(y^2 + \frac{2}{3}y + \frac{1}{9}\right) + \left(x - \frac{1}{2}\right) - \left(\frac{1}{2}y + \frac{1}{6}\right) &= 0 \\ x^2 - x + \frac{1}{4} + \frac{3}{4}y^2 + \frac{1}{2}y + \frac{1}{12} + x - \frac{1}{2} - \frac{1}{2}y - \frac{1}{6} &= 0 \\ x^2 + \frac{3}{4}y^2 &= \frac{1}{3} \end{aligned}$$

and we can now see that the original equation $x^2 + xy + y^2 + x = 0$ describes an ellipse, a fact that may not be obvious from the original expression.

However, if $\alpha_{i,i} = 0$ for some i, then we cannot eliminate the term $\beta_i x_i$. That is, we can only eliminate linear terms if there is a corresponding quadratic term. Instead, we permute the coordinates such that $\alpha_{i,i} \neq 0$ for $1 \leq i \leq r$ and $\beta_i \neq 0$ for $r < i \leq r + s$.

If $s > 1$, then we don't change x_i for $1 \leq i \leq r$, but replace $\sum_{i=1}^{s} \beta_{r+i} x_{r+1}$ by βx_{r+1}. To show that this transformation is orthogonal, suppose our orthonormal basis is $(e_i)_{i=1}^{n}$. Then, we can extend,

$$e_1, \ldots, e_r, \frac{1}{\sqrt{\sum_{i=1}^{s} \beta_{r+i}^2}} \sum_{i=1}^{s} \beta_{r+i} e_{r+i}$$

to an orthonormal basis of our Euclidean space using the Gram-Schmidt process. Note that the $(r+1)$th vector of the basis is chosen such that our equation will have just the term $\left(\sqrt{\sum_{i=1}^{s} \beta_{r+i}^2}\right) x_{r+1}$, so the equation has at most one non-zero β_i; either there are no linear terms at all, or there is just β_{r+1}.

Finally, if there is a linear term, then $\beta_{r+1} \neq 0$, and, by dividing through by a constant, we can force $\beta_{r+1} = -1$, then perform the translation,

$$x_{r+1} \mapsto x_{r+1} - \frac{\gamma}{\beta_{r+1}}$$

to eliminate the constant γ. If there is no linear term, then we again divide the equation through by a constant to force $\gamma = 0$ or $\gamma = -1$, and move it to the right side in the latter case.

We have proved:

Theorem 33.10.16. *Any second degree polynomial equation can be transformed through isometries of Euclidean space into an equation with one of the following forms:*

$$\sum_{i=1}^{r} \alpha_i x_i^2 = 0$$

$$\sum_{i=1}^{r} \alpha_i x_i^2 = 1$$

$$\sum_{i=1}^{r} \alpha_i x_i^2 - x_{r+1} = 0$$

where $0 \leq r \leq n$ *and* $(\alpha_i)_{i=1}^{r}$ *are non-zero constants, and in the third case,* $r < n$*.*

The sets of solutions defined by the first two cases are called *central* quadrics, as they have central symmetry; that is, if a vector $\mathbf{v}$ satisfies the equation, then so does $-\mathbf{v}$.

33.10.8 Singular Value Decomposition

In this section, we will study linear maps $T : V \to W$ between Euclidean spaces V and W. Again, we wish to find bases of V and W such that the matrix of T is as simple as possible.

From row and column operations, we know it is always possible to choose bases of V and W such that the matrix of T has Smith normal form,

$$\left[\begin{array}{c|c} \mathbf{I}_r & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]$$

where r is the rank of T. However, while simple, this form isn't very useful as it does not respect the Euclidean structure of V and W. The problem now is to choose *orthonormal* bases of V and W such that the matrix of T has a simple form.

Theorem 33.10.17 (Singular Value Decomposition for Linear Maps)**.** *Suppose* $T : V \to W$ *is a linear map of rank* r *between Euclidean spaces* V *and* W*. Then, there exist unique positive numbers* $\gamma_1 \geq \gamma_2 \geq \dots \geq \gamma_r > 0$ *called the singular values of* T*, and orthonormal bases of* V *and* W *such that the matrix of* T *with respect to these bases is,*

$$\Sigma = \left[\begin{array}{c|c} \mathbf{D} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]$$

where $\mathbf{D} = \operatorname{diag}(\gamma_1, \dots, \gamma_r)$*.*

Corollary 33.10.17.1 (Singular Value Decomposition for Matrices)**.** *Given any real* $m \times n$ *matrix* $\mathbf{A}$ *of rank* $r \leq \min\{m,n\}$*, there exist unique singular values* $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_r > 0$ *and (non-unique) orthogonal matrices* $\mathbf{P}$ *and* $\mathbf{Q}$ *such that,*

$$\Sigma = \left[\begin{array}{c|c} \mathbf{D} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right] = \mathbf{P}^\top \mathbf{A} \mathbf{Q}$$

Equivalently, we say that the SVD of $\mathbf{A}$ is,

$$\mathbf{A} = \mathbf{P} \Sigma \mathbf{Q}^\top$$

Theorem 33.10.18. *The matrices* $\mathbf{A}$ *and* $\mathbf{A}^\top$ *share the same singular values.*

Proof.

$$\begin{aligned} \mathbf{A} &= \mathbf{P} \Sigma \mathbf{Q}^\top \\ \mathbf{A}^\top &= (\mathbf{P} \Sigma \mathbf{Q}^\top)^\top \\ &= \mathbf{Q}^\top \Sigma^\top \mathbf{P} \end{aligned}$$

■

We present two algorithms for computing the singular value decomposition of a real $m \times n$ matrix $\mathbf{A}$.

Algorithm 9 Singular Value Decomposition

```
Compute the matrices $\mathbf{A}\mathbf{A}^\top$ and $\mathbf{A}^\top\mathbf{A}$.
The eigenvectors of $\mathbf{A}^\top\mathbf{A}$ form the columns of the $n \times n$ orthogonal matrix $\mathbf{Q}$, and the eigenvectors
  of $\mathbf{A}\mathbf{A}^\top$ form the columns of the $m \times m$ orthogonal matrix $\mathbf{P}$.
Perform the Gram-Schmidt process on the columns of $\mathbf{Q}$ and $\mathbf{P}$ if necessary.
The square roots of the eigenvalues of either matrix form the singular values.
```

This requires finding eigenvectors for a pair of matrices, namely, $\mathbf{A}\mathbf{A}^\top$ and $\mathbf{A}^\top\mathbf{A}$. But, because $\mathbf{A} = \mathbf{P}\Sigma\mathbf{Q}$, once we have one of $\mathbf{P}$ or $\mathbf{Q}$, it is possible to compute the other by multiplying the columns of the matrix we have by $\mathbf{A}$.

Algorithm 10 Singular Value Decomposition Shortcut

```
Let $A$ be an $m \times n$ matrix.
Compute whichever of $\mathbf{A}\mathbf{A}^\top$ and $\mathbf{A}^\top\mathbf{A}$ has higher dimensions (if $\mathbf{A}$ is a "tall" matrix, compute the
  former; if $\mathbf{A}$ is a "wide" matrix, compute the latter).
Order the (orthonormalised) eigenvectors $(\mathbf{q}_i)_{i=1}^n$ of this matrix such that the corresponding eigen-
  vectors are in decreasing order, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$, and $\lambda_i = 0$ for $i > r$, where $r$ is the rank of
  $\mathbf{A}$.
The square roots of the eigenvalues form the singular values.
The $n \times n$ orthogonal matrix $\mathbf{Q}$ is given by

  $$\mathbf{Q} = [\mathbf{q}_1 | \mathbf{q}_2 | \dots | \mathbf{q}_n]$$

Define the sequence of vectors $(\mathbf{p}_i)_{i=1}^r$ by multiplying the corresponding $\mathbf{q}_i$ by $\mathbf{A}$, then normalising:

  $$\mathbf{p}_i = \frac{1}{\|\mathbf{A}\mathbf{q}_i\|}\mathbf{A}\mathbf{q}_i$$

  for $1 \leq i \leq r$. Then, $(\mathbf{p}_i)_{i=1}^r$ is an orthonormal set. Using Gram-Schmidt or otherwise, extend this
  set to an orthonormal basis $(\mathbf{p}_i)_{i=1}^m$ of $\mathbb{R}^m$.
The $m \times m$ orthogonal matrix $\mathbf{P}$ is given by

  $$\mathbf{P} = [\mathbf{p}_1 | \mathbf{p}_2 | \dots | \mathbf{p}_m]$$
```

Example. Find a SVD decomposition of,

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix}$$

We compute the eigenvectors of $\mathbf{A}^\top\mathbf{A}$:

$$\mathbf{A}^\top\mathbf{A} = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$$c_{\mathbf{A}^\top\mathbf{A}}(x) = (x-3)(x-1)x$$

so we have eigenvalues (in descending order) $\lambda_1 = 3$, $\lambda_2 = 1$, and $\lambda_3 = 0$, giving singular values

$\gamma_1 = \sqrt{\lambda_1} = \sqrt{3}$ and $\gamma_2 = \sqrt{\lambda_2} = 1$.

$$\Sigma = \begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Next, we find the eigenvectors:

$$(\mathbf{A}^\top\mathbf{A} - \lambda_1\mathbf{I}_3)\mathbf{v} = \mathbf{0} \quad \overset{\substack{\text{row}\\\text{reduce}}}{\longrightarrow} \quad \begin{bmatrix} 1 & 0 & -2 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$a = 2c$$
$$b = -c$$

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = c \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} \longrightarrow \ker(\mathbf{A} - \lambda_2\mathbf{I}_3) = \operatorname{span}\left(\underbrace{\begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}}_{\mathbf{v}_1} \right)$$

$$(\mathbf{A}^\top\mathbf{A} - \lambda_2\mathbf{I}_3)\mathbf{v} = \mathbf{0} \quad \overset{\substack{\text{row}\\\text{reduce}}}{\longrightarrow} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$b = -c$$

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = c \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} \longrightarrow \ker(\mathbf{A} - \lambda_2\mathbf{I}_3) = \operatorname{span}\left(\underbrace{\begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}}_{\mathbf{v}_2} \right)$$

$$(\mathbf{A}^\top\mathbf{A} - \lambda_3\mathbf{I}_3)\mathbf{v} = \mathbf{0} \quad \overset{\substack{\text{row}\\\text{reduce}}}{\longrightarrow} \quad \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$a = -c$$
$$b = -c$$

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = c \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix} \longrightarrow \ker(\mathbf{A} - \lambda_2\mathbf{I}_3) = \operatorname{span}\left(\underbrace{\begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}}_{\mathbf{v}_3} \right)$$

Normalise the eigenvectors,

$$\mathbf{q}_1 = \frac{1}{\sqrt{6}}\mathbf{v}_1, \qquad \mathbf{q}_2 = \frac{1}{\sqrt{2}}\mathbf{v}_2, \qquad \mathbf{q}_3 = \frac{1}{\sqrt{3}}\mathbf{v}_3$$

These eigenvectors have distinct eigenvalues, so they are orthogonal by Theorem 33.10.6, and we do not have to perform the Gram-Schmidt process. So, we have,

$$\begin{aligned} \mathbf{Q} &= [\mathbf{q}_1|\mathbf{q}_2|\mathbf{q}_3] \\ &= \begin{bmatrix} \frac{2}{\sqrt{6}} & 0 & -\frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \end{bmatrix} \end{aligned}$$

To find $\mathbb{P}$, the first algorithm just repeats all of these calculations again on $\mathbf{A}\mathbf{A}^\top$. For the second algorithm, we calculate,

$$\mathbf{A}\mathbf{q}_1 = \frac{\sqrt{6}}{2}\begin{bmatrix}1\\-1\end{bmatrix}, \qquad \mathbf{A}\mathbf{q}_2 = \frac{\sqrt{2}}{2}\begin{bmatrix}1\\1\end{bmatrix}, \qquad \mathbf{A}\mathbf{q}_3 = \begin{bmatrix}0\\0\end{bmatrix}$$

Ignore the zero vector and normalise $\mathbf{A}\mathbf{q}_1$ and $\mathbf{A}\mathbf{q}_2$:

$$\mathbf{p}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}, \qquad \mathbf{p}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}$$

Here, $\mathbf{p}_1$ and $\mathbf{p}_2$ already form a basis of $\mathbb{R}^2$, so we skip the Gram-Schmidt process, and we have,

$$\begin{aligned}\mathbf{P} &= [\mathbf{p}_1|\mathbf{p}_2]\\ &= \begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix}\end{aligned}$$

So the SVD of $\mathbf{A}$ is given by

$$\begin{aligned}\mathbf{A} &= \mathbf{P}\Sigma\mathbf{Q}^\top\\ &= \begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix}\begin{bmatrix}\sqrt{3} & 0 & 0\\ 0 & 1 & 0\end{bmatrix}\begin{bmatrix}\frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}}\\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}}\end{bmatrix}\end{aligned}$$

△

33.11 Sesqulinear Forms

To do.

33.12 Operators on Hilbert Spaces

To do.

33.13 Finitely Generated Abelian Groups

So far, we have considered vector spaces over *fields*. In this section, we will consider a generalisation of a vector space called a *module* in which this field of scalars is replaced by a ring. Modules also generalise the notion of abelian groups, since the abelian groups are exactly the modules over the ring of integers.

33.13.1 Review

We quickly review some basic group theory from §12. The definitions and theorems here will be stated in terms of abelian groups, but most will hold for general groups.

A *group*, $(G,*)$ is a set, G, equipped with a binary operation, $*$, that obeys the following axioms:

- $\forall a,b \in G, a * b \in G$ (closure);
- $\forall a,b,c \in G, a * (b * c) = (a * b) * c$ (associativity);
- $\exists e \in G$ such that $a * e = e * a = a \, \forall a \in G$ (existence of identity);
- $\forall a \in G, \exists (a^{-1}) \in G$ such that $a * (a^{-1}) = (a^{-1}) * a = e$ (existence of inverses).

Furthermore, if the operation is also commutative, the group is *abelian*. The identity, e, is also written as id_G or 0_G, the latter being used mainly for abelian groups.

For abelian groups, it is common to use additive notation where the binary operation is written as $+$, and we will continue with this notation from this point onwards.

A group G is *cyclic* if there exists an element $x \in G$ such that every element of G is of the form nx for some $n \in \mathbb{Z}$.

Let $(G,+)$ and $(H,*)$ be groups. A function $\phi : G \to H$ is a *group homomorphism* between G and H if

$$\phi(a+b) = \phi(a) * \phi(b)$$

for all $a,b \in G$. Note that this necessarily requires that,

$$\phi(\mathrm{id}_G) = \mathrm{id}_H$$

and

$$\phi(-a) = -\phi(a)$$

An injective homomorphism is called a *monomorphism* and a surjective homomorphism is called an *epimorphism.* If a homomorphism ϕ has an inverse, or equivalently, if ϕ is bijective, then ϕ is furthermore a *group isomorphism* and we write $G \cong H$ if such an isomorphism exists.

Theorem 33.13.1. *Every cyclic group is isomorphic to either* $(\mathbb{Z},+)$ *or to* $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$ *for some* $n > 0$.

The *order* of an element $g \in G$, denoted $|g|$, is the smallest natural n such that $ng = \mathrm{id}_G$. If $ng \neq \mathrm{id}_G$ for all $n \in \mathbb{N}$, then we say that g has *infinite order.*

Theorem 33.13.2. *If* $\phi : G \to H$ *if an isomorphism, then* $|g| = |\phi(g)|$ *for all* $g \in G$.

A group G is *generated* or *spanned* by a subset $X \subseteq G$ if every $g \in G$ can be written as a finite sum,

$$\sum_{i=1}^{|X|} m_i x_i$$

with $m_i \in \mathbb{Z}$ and $x_i \in X$, and we write $G = \langle X \rangle$, or $\langle x_1, \dots x_n \rangle$ if X has finite cardinality. In this latter case, we say that G is *finitely generated.*

In multiplicative notation, $G = \langle X \rangle$ if and only if

$$\prod_{i=1}^{|X|} x_i^{m_i}$$

so every element can be "factored" into elements of X.

A group is cyclic if and only if X is a singleton set.

The *direct sum* of a set of abelian groups $\left(G_i\right)_{i=1}^n$ is defined to be the set,

$$\left\{(g_i)_{i=1}^n : g_i \in G_i\right\}$$

with component-wise addition

$$(g_i)_{i=1}^n + (h_i)_{i=1}^n = (g_i + h_i)_{i=1}^n$$

This forms a group with identity $(\mathrm{id}_{G_i})_{i=1}^n$ and $-(g_i)_{i=1}^n = (-g_i)_{i=1}^n$.

For non-abelian groups, this is more commonly called the *direct product* of groups.

A subset $H \subseteq G$ of a group $(G,+)$ is a *subgroup* of G if $(H,+)$ is also a group, and we write $H \leq G$ to denote this relation. The group itself, G, and the trivial group $\{\text{id}_G\}$ are always subgroups of G. Any subgroup H not equal to G is a *proper* subgroup, and we write $H < G$ to denote this relation. Any subgroup not equal to $\{\text{id}_G\}$ is a *non-trivial* subgroup.

Lemma 33.13.3. *If $H \leq G$, then $\text{id}_H = \text{id}_G$.*

Theorem 33.13.4. *Let $H \subseteq G$. Then, the following statements are equivalent:*

(*i*) $H \leq G$;

(*ii*) (*a*) $H \neq \varnothing$;

(*b*) $a,b \in H \to a + b \in H$;

(*c*) $a \in H \to -a \in H$.

(*iii*) (*a*) $H \neq \varnothing$;

(*b*) $a,b \in H \to a - b \in H$.

(*ii*) and (*iii*) are the *two step* and *one step* subgroup tests (so called because H is often assumed to be non-empty, and hence checking that it is non-empty does not count as a step).

Let G be a group, $H \leq G$ and $g \in G$. The set $g + H = \{g + h : h \in H\}$ is a *left coset* of H in G, and $H + g = \{h + g : h \in H\}$ is a *right coset* of H in G. For abelian groups, left and right cosets coincide, and we just say *coset* alone.

Theorem 33.13.5. *The following statements are equivalent for any $x,g \in G$:*

- $x \in H + g$;
- $H + g = H + x$;
- $x - g \in H$.

Corollary 33.13.5.1. *Two cosets $H + a$ and $H + b$ are either equal or disjoint.*

Corollary 33.13.5.2. *The cosets of H in G partition G.*

Theorem 33.13.6. *If H is finite, then all cosets of H in G have $|H|$ elements.*

The number of distinct left (or right) cosets of H in G is called the *index* of H in G, and is written as $[G : H]$

Theorem 33.13.7 (Lagrange's Theorem)**.** *If $H \leq G$, then,*

$$|G| = [G : H]|H|$$

Corollary 33.13.7.1. *If $H \leq G$, then the order of H divides the order of G.*

Corollary 33.13.7.2. *For any $g \in G$, $|g|$ divides $|G|$.*

Theorem 33.13.8. *Let G be a group of prime order p. Then, G is cyclic and $G \cong \mathbb{Z}_p$.*

If A and B are subsets of a group G, then we define their *sum* by,

$$A + B \coloneqq \{a + b : a \in A, b \in B\}$$

Lemma 33.13.9. *If H is a subgroup of an abelian group G, and $H + a$, $H + b$ are cosets of H in G, then,*

$$(H + g) + (H + k) = H + (g + k)$$

Theorem 33.13.10. *Let H be a subgroup of an abelian group G. Then, the set G/H of cosets $H+g$ of H in G forms a group under addition of sets as defined above.*

Such a group is called a *quotient group* or *factor group* of G by H. Note that if G is finite, then $|G/H| = [G : H] = |G|/|H|$.

Let $\phi : G \to H$ be a group homomorphism. Then, the *kernel* $\ker(\phi)$ of ϕ is defined to be the set of elements of G mapped to the identity id_H. That is,

$$\ker(\phi) = \{g \in G : \phi(g) = \mathrm{id}_H\}$$

Note that $\ker(\phi)$ always contains id_H as group homomorphisms must map identities to identities. If id_H is the only element of $\ker(\phi)$, then the kernel is *trivial*.

The *image* of ϕ is then defined as,

$$\mathrm{im}(\phi) = \{\phi(g) : g \in G\}$$

Theorem 33.13.11. *Let $\phi : G \to H$ be a group homomorphsim. Then, ϕ is a monomorphism (injection) if and only if the kernel $\ker(\phi)$ is trivial.*

Proof. Since $\mathrm{id}_G \in \ker(\phi)$, if ϕ is injective, then we must have $\ker(\phi) = \{\mathrm{id}_G\}$, completing the forward implication.

Conversely, suppose $\ker(\phi) = \{\mathrm{id}_G\}$, and let a,$b \in G$ with $\phi(a) = \phi(b)$. Then, $\phi(a-b) = \phi(a) - \phi(b) = \mathrm{id}_H$, so $a - b \in \ker(\phi)$. But then, $a - b = \mathrm{id}_G$, and hence $a = b$, so ϕ is injective, completing the reverse implication. ■

Theorem 33.13.12. *Let $\phi : G \to H$ be a group homomorphism. Then, $\ker(\phi)$ is a subgroup of G, and $\mathrm{im}(\phi)$ is a subgroup of H.*

Furthermore, if K is a subgroup of G, then the map $\phi : G \to G/K$ defined by $\phi(g) = K + g$ is an epimorphism (surjection) with kernel K.

Proof. The first statement follows from the two step subgroup test. For the second, it is clear that ϕ is surjective, and $\phi(g) = \mathrm{id}_{G/K} \leftrightarrow K + g = K + \mathrm{id}_G \leftrightarrow g \in K$, so $\ker(\phi) = K$. ■

Theorem 33.13.13 (First Isomorphism Theorem)**.** *Let $\phi : G \to H$ be a group homomorphism with kernel K. Then, $G/K \cong \mathrm{im}(\phi)$. More precisely, there is an isomorphism $\bar{\phi} : G/K \to \mathrm{im}(\phi)$ defined by $\bar{\phi}(K+g) = \phi(g)$ for $g \in G$.*

33.13.2 Free Abelian Groups

The direct sum $\mathbb{Z}^n := \underbrace{\mathbb{Z} \oplus \mathbb{Z} \oplus \ldots \oplus \mathbb{Z}}_{n}$ of n copies of $\mathbb{Z}$ is called a (finitely generated) *free abelian group* of *rank* n.

More generally, a finitely generated abelian group is *free abelian* if it is isomorphic to $\mathbb{Z}^n$ for some $n \geq 0$, with the free abelian group $\mathbb{Z}^0$ of rank 0 defined to be the trivial group.

The free abelian groups have many properties in common with vector spaces like $\mathbb{R}^n$, but we would expect some differences, as $\mathbb{Z}$ is not a field. Similarly to vector spaces, we will write elements of $\mathbb{Z}^n$ as column vectors.

We then define the *standard basis* of $\mathbb{Z}^n$ exactly as for the vector space $\mathbb{R}^n$; that is, a set of vectors $(\mathbf{x}_i)_{i=1}^n$ such that $\mathbf{x}_i$ is 0 everywhere apart from the ith component, which has a 1. This basis has the same properties as a basis of a vector space: the vectors are linearly independent, and they span (generate) $\mathbb{Z}^n$, for a modified definition of linear independence and span.

Elements $(x_i)_{i=1}^n$ of an abelian group G are called *linearly independent* if the equation

$$\sum_{i=1}^n \alpha_i x_i = \mathrm{id}_G$$

with integer coefficients $\alpha_i \in \mathbb{Z}$ holds only if $\alpha_i = 0$ for all $1 \leq i \leq n$.

Elements $S = \{x_i\}_{i=1}^n$ of an abelian group G form a *free basis* or *integral basis* of G if and only if they are linearly independent and span (generate) G. That is,

$$G = \langle (x_i)_{i=1}^n \rangle$$

or equivalently, every $g \in G$ can be written as a *unique* linear integer combination of elements in S:

$$g = \sum_{i=1}^n \alpha_i x_i$$

where all the $\alpha_i \in \mathbb{Z}$.

Note that a set of elements in $\mathbb{Z}^n$ that form a basis of $\mathbb{Q}^n$ or $\mathbb{R}^n$ need not be a free basis of $\mathbb{Z}^n$. For instance, the set,

$$\left\{ \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 2 \end{bmatrix} \right\}$$

is a basis of $\mathbb{Q}^2$ and $\mathbb{R}^2$, and are linearly independent in $\mathbb{Z}^2$, but not of $\mathbb{Z}^2$, as we are only allowed integer coefficients in $\mathbb{Z}^n$: there is no way to write, say,

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \alpha_1 \begin{bmatrix} 2 \\ 0 \end{bmatrix} + \alpha_2 \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

with α_1 and α_2 as integers. This also shows that a set of n linearly independent elements of $\mathbb{Z}^n$ does not necessarily form a free basis.

Theorem 33.13.14. *For any set of elements $(g_i)_{i=1}^n$ of an abelian group G, it is possible to extend the assignment $\mathbf{x}_i \mapsto g_i$ to a group homomorphism $\phi : \mathbb{Z}^n \to G$. We define*

$$\phi \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} := \sum_{i=1}^n \alpha_i g_i$$

Then,

- *ϕ is a group homomorphism;*
- *The set (g_i) is linearly independent if and only if ϕ is an monomorphism (injection);*
- *The set (g_i) span G if and only if ϕ is an epimorphism (surjection);*
- *The set (g_i) form a free basis of G if and only if ϕ is a isomorphism.*

Theorem (Universal Property of the Free Abelian Group)**.** *Let G be a free abelian group with a free basis $(g_i)_{i=1}^n$. Let H be an abelian group, and $(h_i)_{i=1}^n$ be elements of H. Then, there exists a unique group homomorphism $\phi : G \to H$ such that $\phi(g_i) = h_i$.*

As for finite-dimensional vector spaces, we have yet to prove that any two free bases of a free abelian group have the same size. Let $(\mathbf{x}_i)_{i=1}^n$ be the standard free basis of $\mathbb{Z}^n$, and let $(\mathbf{y}_i)_{i=1}^m$ be another free basis of $\mathbf{Z}^n$ (expressed in terms of the standard basis). As in linear algebra, we define the associated change of basis matrix $\mathbf{P}$ with respect to the original basis $(\mathbf{x}_i)$ and target basis $(\mathbf{y}_i)$ by,

$$\mathbf{P} = [\mathbf{y}_1|\mathbf{y}_2|\dots|\mathbf{y}_m]$$

That is, if $\mathbf{x}$ and $\mathbf{y}$ are column vectors expressed in terms of the standard basis $(\mathbf{x}_i)_{i=1}^n$ and free basis $(\mathbf{y}_i)_{i=1}^n$, respectively, then,

$$\mathbf{x} = \mathbf{P}\mathbf{y}$$

Theorem 33.13.15. *Let $(\mathbf{y}_i)_{i=1}^m \subset \mathbb{Z}^n$. Then, the following statements are equivalent:*

- *$(\mathbf{y}_i)_{i=1}^n$ is a free basis of $\mathbb{Z}^n$;*
- *$n = m$ and the change of basis matrix $\mathbf{P} \in \mathbb{Z}^{n\times n}$ has an inverse $\mathbf{P}^{-1} \in \mathbb{Z}^{n\times n}$ (that is, the inverse of $\mathbf{P}$ has integer entries);*
- *$n = m$ and $\det(\mathbf{P}) = \pm 1$.*

A square matrix with integer entries and determinant ±1 is called *unimodular.*

For example, if $n = 1$, and we have elements,

$$\mathbf{y}_1 = \begin{bmatrix} 2 \\ 7 \end{bmatrix}, \qquad \mathbf{y}_2 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$$

then we have

$$\mathbf{P} = \begin{bmatrix} 2 & 1 \\ 7 & 4 \end{bmatrix}$$

with $\det(\mathbf{P}) = 2 \cdot 4 - 1 \cdot 7 = 1$, so $\{\mathbf{y}_1,\mathbf{y}_2\}$ is a free basis of $\mathbb{Z}^2$.

In contrast, take our previous example of

$$\mathbf{y}_1 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \qquad \mathbf{y}_2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

Then, $\det(P) = 4 \neq \pm 1$, so this set is not a free basis of $\mathbb{Z}^2$.

33.13.3 Unimodular Smith Normal Form

Recall that, in linear algebra, we may use elementary row and column operations to reduce an $m \times n$ matrix $\mathbf{A}$ of rank r to a matrix,

$$\left[\begin{array}{c|c} \mathbf{I}_r & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]$$

called the *Smith normal form* of $\mathbf{A}$.

For matrices over $\mathbb{Z}$, we can similarly reduce a matrix to a Smith normal form, but now, the non-zero entries will not necessarily be equal to 1.

For matrices over $\mathbb{Z}$, we use *unimodular* row and column operations instead:

(UC1) Replace a column $\mathbf{c}_i$ by $\mathbf{c}_i + \lambda\mathbf{c}_j$, $\lambda \in \mathbb{Z}$.

(UC2) Interchange two columns $\mathbf{c}_i$ and $\mathbf{c}_j$.

(UC3) Replace a column $\mathbf{c}_i$ with $-\mathbf{c}_i$.

(UR1) Replace a row $\mathbf{r}_i$ by $\mathbf{r}_i + \lambda\mathbf{r}_j$, $\lambda \in \mathbb{Z}$.

(UR2) Interchange two rows $\mathbf{r}_i$ and $\mathbf{r}_j$.

(UR3) Replace a row $\mathbf{r}_i$ with $-\mathbf{r}_i$.

Elementary row and column operations on a matrix $\mathbf{A}$ correspond to multiplying $\mathbf{A}$ on the left or right, respectively, by an elementary matrix. These matrices have determinant ±1, and are hence also unimodular matrices. From this, unimodular row and column operations correspond to the following change of bases, where $(\mathbf{e}_i)_{i=1}^n$ is a free basis for $\mathbb{Z}^n$ (the domain of the linear map $\mathbf{A}$ represents) and $(\mathbf{f}_i)_{i=1}^m$ is a free basis of $\mathbb{Z}^m$ (the codomain).

(UC1) $\mathbf{e}_i \mapsto \mathbf{e}_i + \lambda\mathbf{e}_j$;

(UC2) $\mathbf{e}_i \leftrightarrow \mathbf{e}_j$;

(UC3) $\mathbf{e}_i \mapsto -\mathbf{e}_i$;

(UR1) $\mathbf{f}_j \mapsto \mathbf{f}_j - \lambda\mathbf{f}_i$ (note the sign change and the reversal of indices);

(UR2) $\mathbf{f}_i \leftrightarrow \mathbf{f}_j$;

(UR3) $\mathbf{f}_i \mapsto -\mathbf{f}_i$;

Theorem 33.13.16. *Let $\mathbf{A}$ be an $m \times n$ matrix over $\mathbb{Z}$ with rank r. Then, by using a sequence of unimodular elementary row and column operations, we can reduce $\mathbf{A}$ to a matrix*

$$\mathbf{S} = \left[\begin{array}{c|c} \mathbf{D} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right]$$

with $\mathbf{D} = \operatorname{diag}(d_1, d_2, \ldots, d_r)$, where $(d_i)_{i=1}^r$ are positive integers satisfying $d_i | d_{i+1}$ for $1 \leq i < r$.

Subject to these conditions, the d_i are uniquely determined by the matrix $\mathbf{A}$.

The matrix $\mathbf{S}$ is then called the *unimodular Smith normal form* of $\mathbf{A}$, or the $\mathbb{Z}$SNF of $\mathbf{A}$.

Lemma 33.13.17. *Let $\mathbf{A} \in \mathbb{Z}^{m\times n}$ have unimodular Smith normal form $\mathbf{S}$ with r non-zero diagonal entries $(d_i)_{i=1}^r$. Then, the greatest common divisor of all the entries of $\mathbf{A}$ is d_1.*

Algorithm 11 Smith Normal Form Decomposition

```
1: Compute the greatest common divisor $x$ of the entries in the first column, and the greatest common
   divisor $y$ of the entries in the first row. Without loss of generality, suppose $x < y$ (simply reverse
   row and columns in the following, if otherwise).
2: Using the Euclidean algorithm, we can form a row whose first element is $x$.
3: Move this row to the first row such that $x$ is in the first entry. This is the pivot element (and
   row/column).
4: Subtract multiples of the pivot row from every other row until the pivot column has 0s everywhere
   else (possible since $x$ divides everything in this column).
5: Repeat this process for the rows. This will likely undo some of our work on the pivot column, but
   just repeat this process again. This process is guaranteed to terminate since the greatest common
   divisor is reduced with each iteration.
6: Eventually, the pivot column and row will be zero every outside the pivot element. Iterate this
   process on each column/row until the matrix is diagonal.
7: The diagonal entries may not satisfy the divisibility requirements of the unimodular Smith normal
   form, so we again use the Euclidean algorithm to obtain the greatest common divisor of the diagonal
   elements, then add or subtract this value until the divisibility requirements are met.
```

Rows and columns can also be interchanged, if it makes the Bézout coefficients smaller or if the divisor is easier to obtain. If it is easy to spot or calculate the greatest common divisor of all the entries of the matrix, we can shortcut the first few steps somewhat.

Example. Find a $\mathbb{Z}$SNF decomposition of,

$$\mathbf{A} = \begin{bmatrix} -18 & -18 & -18 & 90 \\ 54 & 12 & 45 & 48 \\ 9 & -6 & 6 & 63 \\ 18 & 6 & 15 & 12 \end{bmatrix}$$

It is easy to see that every entry of $\mathbf{A}$ is divisible by 3, but for the sake of illustration, we will not use this shortcut.

Instead, gcd(-18,54,9,18) $= 9$ and gcd(-18,-18,-18,90) $= 9$, so we can choose to work on columns or rows. We already have a 9 in the first column, we will work on the first column:

$$\begin{bmatrix} -18 & -18 & -18 & 90 \\ 54 & 12 & 45 & 48 \\ 9 & -6 & 6 & 63 \\ 18 & 6 & 15 & 12 \end{bmatrix} \xrightarrow{\mathbf{r}_3 \leftrightarrow \mathbf{r}_1} \begin{bmatrix} 9 & -6 & 6 & 63 \\ 54 & 12 & 45 & 48 \\ -18 & -18 & -18 & 90 \\ 18 & 6 & 15 & 12 \end{bmatrix}$$

Now, clear out the rest of the pivot column,

$$\begin{bmatrix} 9 & -6 & 6 & 63 \\ 54 & 12 & 45 & 48 \\ -18 & -18 & -18 & 90 \\ 18 & 6 & 15 & 12 \end{bmatrix} \xrightarrow{\substack{\mathbf{r}_2 \mapsto \mathbf{r}_2 - 6\mathbf{r}_1 \\ \mathbf{r}_3 \mapsto \mathbf{r}_3 + 2\mathbf{r}_1 \\ \mathbf{r}_4 \mapsto \mathbf{r}_4 - 2\mathbf{r}_1}} \begin{bmatrix} 9 & -6 & 6 & 63 \\ 0 & 48 & 9 & -330 \\ 0 & -30 & -6 & 216 \\ 0 & 18 & 3 & -114 \end{bmatrix}$$

gcd(9,-6,6,63) $= 3$, so, we make a 3 in the pivot row.

$$\begin{bmatrix} 9 & -6 & 6 & 63 \\ 0 & 48 & 9 & -330 \\ 0 & -30 & -6 & 216 \\ 0 & 18 & 3 & -114 \end{bmatrix} \xrightarrow{\mathbf{c}_1 \mapsto \mathbf{c}_1 - \mathbf{c}_3} \begin{bmatrix} 3 & -6 & 6 & 63 \\ -9 & 48 & 9 & -330 \\ 6 & -30 & -6 & 216 \\ -3 & 18 & 3 & -114 \end{bmatrix}$$

Clear the pivot row,

$$\begin{bmatrix} 3 & -6 & 6 & 63 \\ -9 & 48 & 9 & -330 \\ 6 & -30 & -6 & 216 \\ -3 & 18 & 3 & -114 \end{bmatrix} \xrightarrow{\substack{\mathbf{c}_2 \mapsto \mathbf{c}_2 + 2\mathbf{c}_1 \\ \mathbf{c}_3 \mapsto \mathbf{c}_3 - 2\mathbf{c}_1 \\ \mathbf{c}_4 \mapsto \mathbf{c}_4 - 21\mathbf{c}_1}} \begin{bmatrix} 3 & 0 & 0 & 0 \\ -9 & 30 & 27 & -141 \\ 6 & -18 & -18 & 90 \\ -3 & 12 & 9 & -51 \end{bmatrix}$$

Clear the pivot column again,

$$\begin{bmatrix} 3 & 0 & 0 & 0 \\ -9 & 30 & 27 & -141 \\ 6 & -18 & -18 & 90 \\ -3 & 12 & 9 & -51 \end{bmatrix} \xrightarrow{\substack{\mathbf{r}_2 \mapsto \mathbf{r}_2 + 3\mathbf{r}_1 \\ \mathbf{r}_3 \mapsto \mathbf{r}_3 - 2\mathbf{r}_1 \\ \mathbf{r}_4 \mapsto \mathbf{r}_4 + \mathbf{r}_1}} \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 30 & 27 & -141 \\ 0 & -18 & -18 & 90 \\ 0 & 12 & 9 & -51 \end{bmatrix}$$

gcd(30,-18,12) $= 6$, and gcd(30,26,-141) $= 3$, so we will work on the row first this time.

$$\begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 30 & 27 & -141 \\ 0 & -18 & -18 & 90 \\ 0 & 12 & 9 & -51 \end{bmatrix} \xrightarrow{\mathbf{c}_2 \mapsto \mathbf{c}_2 - \mathbf{c}_3} \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 3 & 27 & -141 \\ 0 & 0 & -18 & 90 \\ 0 & 3 & 9 & -51 \end{bmatrix}$$

Clear the row out,

$$\begin{bmatrix}3 & 0 & 0 & 0\\ 0 & 3 & 27 & -141\\ 0 & 0 & -18 & 90\\ 0 & 3 & 9 & -51\end{bmatrix} \xrightarrow[\mathbf{c}_4\mapsto\mathbf{c}_4+47\mathbf{c}_1]{\mathbf{c}_3\mapsto\mathbf{c}_3-9\mathbf{c}_1} \begin{bmatrix}3 & 0 & 0 & 0\\ 0 & 3 & 0 & 0\\ 0 & 0 & -18 & 90\\ 0 & 3 & -18 & 90\end{bmatrix}$$

gcd(3,0,3) $= 3$, so our pivot is already the correct divisor. Clear the rest of the pivot column,

$$\begin{bmatrix}3 & 0 & 0 & 0\\ 0 & 3 & 0 & 0\\ 0 & 0 & -18 & 90\\ 0 & 3 & -18 & 90\end{bmatrix} \xrightarrow{\mathbf{r}_4\mapsto\mathbf{r}_4-\mathbf{r}_1} \begin{bmatrix}3 & 0 & 0 & 0\\ 0 & 3 & 0 & 0\\ 0 & 0 & -18 & 90\\ 0 & 0 & -18 & 90\end{bmatrix}$$

From this point, the algorithm would compute gcd(-18,-18) $= 18$, and gcd(-18,90) $= 9$, so we would then work on the row, but at this point, the matrix is small enough that we can obviously just clear the last column, then row:

$$\begin{bmatrix}3 & 0 & 0 & 0\\ 0 & 3 & 0 & 0\\ 0 & 0 & -18 & 90\\ 0 & 0 & -18 & 90\end{bmatrix} \xrightarrow{\mathbf{c}_4\mapsto\mathbf{c}_4-5\mathbf{c}_3} \begin{bmatrix}3 & 0 & 0 & 0\\ 0 & 3 & 0 & 0\\ 0 & 0 & -18 & 0\\ 0 & 0 & -18 & 0\end{bmatrix}$$

$$\begin{bmatrix}3 & 0 & 0 & 0\\ 0 & 3 & 0 & 0\\ 0 & 0 & -18 & 0\\ 0 & 0 & -18 & 0\end{bmatrix} \xrightarrow{\mathbf{r}_4\mapsto\mathbf{r}_4-\mathbf{r}_3} \begin{bmatrix}3 & 0 & 0 & 0\\ 0 & 3 & 0 & 0\\ 0 & 0 & -18 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$$

$$\begin{bmatrix}3 & 0 & 0 & 0\\ 0 & 3 & 0 & 0\\ 0 & 0 & -18 & 0\\ 0 & 0 & 0 & 0\end{bmatrix} \xrightarrow{\mathbf{r}_3\mapsto-\mathbf{r}_3} \begin{bmatrix}3 & 0 & 0 & 0\\ 0 & 3 & 0 & 0\\ 0 & 0 & 18 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$$

In this case, we were lucky in that the diagonal entries already satisfy the divisibility requirements, so we are done.

Otherwise, for each column i where d_i does not divide d_{i+1}, we can fix the divisibility requirement with operations on only colummns i and $i+1$, replacing d_i with $\tilde{d_i} = \gcd(d_i,d_{i+1})$ before diagonalising the matrix again. The new value of d_{i+1} will be a linear combination of the original d_i and d_{i+1} and will thus be divisible by $\tilde{d_i}$.

For illustration, suppose we instead have the matrix:

$$\begin{bmatrix}3 & 0 & 0 & 0\\ 0 & 3 & 0 & 0\\ 0 & 0 & 4 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$$

Here, 3 does not divide 4, so we will aim to form a gcd(3,4) $= 1$ in place of the 3. First, add column $i+1$ to column i:

$$\begin{bmatrix}3 & 0 & 0 & 0\\ 0 & 3 & 0 & 0\\ 0 & 0 & 4 & 0\\ 0 & 0 & 0 & 0\end{bmatrix} \xrightarrow{\mathbf{c}_2\mapsto\mathbf{c}_2+\mathbf{c}_3} \begin{bmatrix}3 & 0 & 0 & 0\\ 0 & 3 & 0 & 0\\ 0 & 4 & 4 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$$

$$\begin{bmatrix}3 & 0 & 0 & 0\\ 0 & 3 & 0 & 0\\ 0 & 4 & 4 & 0\\ 0 & 0 & 0 & 0\end{bmatrix} \xrightarrow{\mathbf{r}_2\mapsto\mathbf{r}_3-\mathbf{r}_2} \begin{bmatrix}3 & 0 & 0 & 0\\ 0 & 1 & 4 & 0\\ 0 & 4 & 4 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$$

$$\begin{bmatrix}3&0&0&0\\0&1&4&0\\0&4&4&0\\0&0&0&0\end{bmatrix}\xrightarrow{\mathbf{r}_3\mapsto\mathbf{r}_3-4\mathbf{r}_2}\begin{bmatrix}3&0&0&0\\0&1&4&0\\0&0&-12&0\\0&0&0&0\end{bmatrix}$$

$$\begin{bmatrix}3&0&0&0\\0&1&4&0\\0&0&-12&0\\0&0&0&0\end{bmatrix}\xrightarrow{\mathbf{c}_3\mapsto4\mathbf{c}_2-\mathbf{c}_3}\begin{bmatrix}3&0&0&0\\0&1&0&0\\0&0&12&0\\0&0&0&0\end{bmatrix}$$

Now, 3 does not divide 1, so we repeat this process again, replacing 3 with gcd(3,1) = 1:

$$\begin{bmatrix}3&0&0&0\\0&1&0&0\\0&0&12&0\\0&0&0&0\end{bmatrix}\xrightarrow{\mathbf{c}_1\mapsto\mathbf{c}_1+\mathbf{c}_2}\begin{bmatrix}3&0&0&0\\1&1&0&0\\0&0&12&0\\0&0&0&0\end{bmatrix}$$

$$\begin{bmatrix}3&0&0&0\\1&1&0&0\\0&0&12&0\\0&0&0&0\end{bmatrix}\xrightarrow{\mathbf{r}_1\mapsto4\mathbf{r}_2-\mathbf{r}_1}\begin{bmatrix}1&4&0&0\\1&1&0&0\\0&0&12&0\\0&0&0&0\end{bmatrix}$$

$$\begin{bmatrix}1&4&0&0\\1&1&0&0\\0&0&12&0\\0&0&0&0\end{bmatrix}\xrightarrow{\mathbf{c}_2\mapsto4\mathbf{c}_1-\mathbf{c}_2}\begin{bmatrix}1&0&0&0\\1&3&0&0\\0&0&12&0\\0&0&0&0\end{bmatrix}$$

$$\begin{bmatrix}1&0&0&0\\1&3&0&0\\0&0&12&0\\0&0&0&0\end{bmatrix}\xrightarrow{\mathbf{r}_2\mapsto\mathbf{r}_2-\mathbf{r}_1}\begin{bmatrix}1&0&0&0\\0&3&0&0\\0&0&12&0\\0&0&0&0\end{bmatrix}$$

△

33.13.4 Subgroups of Free Abelian Groups

Theorem 33.13.18. *Any subgroup of a finitely generated abelian group is finitely generated.*

Let H be a subgroup of the free abelian group $\mathbb{Z}^n$, and suppose that $H = \langle \mathbf{v}_1,\mathbf{v_2},\dots\mathbf{v}_m\rangle$. Then, H can be represented by a $n\times m$ matrix $\mathbf{A}$ defined by

$$\mathbf{A}=[\mathbf{v}_1|\mathbf{v}_2|\dots|\mathbf{v}_m]$$

For example, if $n=3$ and H is generated by,

$$\mathbf{v}_1=\begin{bmatrix}1\\3\\-1\end{bmatrix},\qquad \mathbf{v}_2=\begin{bmatrix}2\\0\\1\end{bmatrix}$$

then,

$$\mathbf{A}=\begin{bmatrix}1&2\\3&0\\-1&1\end{bmatrix}$$

If a different free basis $(\mathbf{y}_i)_{i=1}^n$ of $\mathbb{Z}^n$ with change of basis matrix $\mathbf{P}$ is used, then each column $\mathbf{v}_i$ of $\mathbf{A}$ is replaced by $\mathbf{P}^{-1}\mathbf{v}_i$, and hence $\mathbf{A}$ itself is replaced by $\mathbf{P}^{-1}\mathbf{A}$.

For example, if we have the basis

$$\mathbf{y}_1 = \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix}, \qquad \mathbf{y}_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \qquad \mathbf{y}_3 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

of $\mathbb{Z}^3$, then,

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 1 \\ -1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \qquad \mathbf{P}^{-1} \begin{bmatrix} 1 & -1 & -1 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}, \qquad \mathbf{P}^{-1}\mathbf{A} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \\ 2 & 1 \end{bmatrix}$$

Theorem 33.13.19. *Suppose that a subgroup H of $\mathbb{Z}^n$ is represented by the matrix $\mathbf{A} \in \mathbb{Z}^{n\times m}$. Then, if the matrix $\mathbf{B} \in \mathbb{Z}^{n\times m}$ can be obtained by applying unimodular row and column operations on $\mathbf{A}$, then $\mathbf{B}$ represents the same subgroup H of $\mathbb{Z}^n$ using a (possibly) different free basis of $\mathbb{Z}^n$.*

In particular, we can transform $\mathbf{A}$ to its unimodular Smith normal form, $\mathbf{S}$. So, if $\mathbf{S}$ represents the subgroup H with free basis $(\mathbf{y}_i)_{i=1}^n$ of $\mathbb{Z}^n$, then the r non-zero columns of $\mathbf{S}$ correspond to the elements $(d_i\mathbf{y}_i)_{i=1}^r$. So,

Theorem 33.13.20. *Let H be a subgroup of $\mathbb{Z}^n$. Then, there exists a free basis $(\mathbf{y}_i)_{i=1}^n$ of $\mathbb{Z}^n$ such that $H = \langle (d_i\mathbf{y}_i)_{i=1}^r \rangle$, where each $d_i > 0$ and $d_i | d_{i+1}$ for $1 \leq i < r$.*

By keeping track of the row operations used, we can find the free basis of $\mathbb{Z}^n$ such that the matrix of H has a simple form by applying the operations to the basis vectors.

33.13.5 General Finitely Generated Abelian Groups

A *presentation* is one method of specifying a group. A presentation of a group G is composed of a set S of generators, and set R of relations among those generators. We then say that G has presentation

$$\langle S \mid R \rangle$$

For instance, a cyclic group of order n has the presentation,

$$\langle a \mid a^n = \mathrm{id} \rangle$$

This is also sometimes written as

$$\langle a \mid a^n \rangle$$

under the convention that any terms without an equals is assumed to be equal to the identity element. For instance, the dihedral group D_n has presentation,

$$\langle r{,}f \mid r^n, f^2, (rf)^2 \rangle$$

where r is a rotation and f a reflection; $\mathbb{Z}^2$ has presentation,

$$\langle x{,}y \mid xy = yx \rangle$$

and the free group $F(S)$ on a set S has presentation,

$$\langle S \mid \varnothing \rangle$$

Note that the presentation of a group is not unique.

Let G be a finitely generated abelian group. If G has n generators $(x_i)_{i=1}^n$, then Theorem 33.13.14 gives us a way to define a surjective homomorphism $\phi : \mathbb{Z}^n \to G$, and by the first isomorphism theorem, we can deduce that $G \cong \mathbb{Z}^n/K$, where $K = \ker(\phi)$, so we have proved that every finitely generated abelian group is isomorphic to a quotient group of a free abelian group.

From the definition of ϕ, we see that K is given by,

$$\begin{aligned} K &= \{\mathbf{v} \in \mathbb{Z}^n : \phi(\mathbf{v}) = \mathrm{id}_G\} \\ &= \left\{ [v_1, \dots, v_n]^\top \in \mathbb{Z}^n : \sum_{i=1}^{n} v_i x_i = \mathrm{id}_G \right\} \end{aligned}$$

and, because K is a subgroup of G, which is finitely generated, K is also finitely generated by elements $(\mathbf{v}_i)_{i=1}^m$ of $\mathbb{Z}^m$. The quotient group $\mathbb{Z}^n/K$ then has presentation,

$$\langle \mathbf{x}_1, \dots, \mathbf{x}_n \mid \mathbf{v}_1, \dots, \mathbf{v}_m \rangle$$

(where $(\mathbf{x}_i)_{i=1}^n$ is the standard basis of $\mathbb{Z}^n$) and as before, this group is isomorphic to G,

$$G \cong \langle \mathbf{x}_1, \dots, \mathbf{x}_n \mid \mathbf{v}_1, \dots, \mathbf{v}_m \rangle$$

Now, we can find the unimodular Smith normal form of the matrix of K to find a free basis $(\mathbf{y}_i)_{i=1}^n$ of $\mathbb{Z}^n$ such that $K = \left\langle (d_i \mathbf{y}_i)_{i=1}^r \right\rangle$ for some $r \leq n$ and $d_i > 0$ and $d_i | d_{i+1}$ for $1 \leq i < r$, giving,

$$G \cong \langle \mathbf{y}_1, \dots, \mathbf{y}_n \mid d_1 \mathbf{y}_1, \dots, d_r \mathbf{v}_r \rangle$$

Theorem 33.13.21. *The group,*

$$\langle \mathbf{y}_1, \dots, \mathbf{y}_n \mid d_1 \mathbf{y}_1, \dots, d_r \mathbf{v}_r \rangle$$

is isomorphic to the direct sum of cyclic groups,

$$\mathbb{Z}_{d_1} \oplus \mathbb{Z}_{d_2} \oplus \cdots \oplus \mathbb{Z}_{d_r} \oplus \mathbb{Z}^{n-r}$$

Putting these results together, we obtain,

Theorem 33.13.22 (Fundamental Theorem of Finitely Generated Abelian Groups)**.** *If G is a finitely generated abelian group, then G is isomorphic to a direct sum of cyclic groups. More precisely, if G is generated by n elements, then, for some r with $0 \leq r \leq n$, there exist integers $(d_i)_{i=1}^r$ with $d_i > 0$ and $d_i | d_{i+1}$ for $1 \leq i < r$, such that,*

$$G \cong \mathbb{Z}_{d_1} \oplus \mathbb{Z}_{d_2} \oplus \cdots \oplus \mathbb{Z}_{d_r} \oplus \mathbb{Z}^{n-r}$$

or more compactly,

$$\left(\bigoplus_{i=1}^{r} \mathbb{Z}_{d_i} \right) \oplus \mathbb{Z}^{n-r}$$

That is, G is isomorphic to the direct sum of r finite cyclic groups of orders $d_1, \dots, d_r$, and $n - r$ infinite cyclic groups.

There may be some factors $\mathbb{Z}_1 = \{\mathrm{id}\}$, which can be omitted from the direct sum (unless it is the only factor and $G \cong \mathbb{Z}_1$ is trivial). It could be the case that $n - r = 0$, which occurs if and only if G is finite. We can also have that $d_i = 1$ for all i, which occurs if and only if G is free abelian.

Example. From before, we found that the matrix

$$\mathbf{A} = \begin{bmatrix} -18 & -18 & -18 & 90 \\ 54 & 12 & 45 & 48 \\ 9 & -6 & 6 & 63 \\ 18 & 6 & 15 & 12 \end{bmatrix}$$

has unimodular smith normal form

$$\begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 18 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

This means that the group defined by $\mathbf{A}$, which has presentation,

$$\left\langle \mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4 \;\middle|\; \begin{array}{ll} -18\mathbf{x}_1 + 54\mathbf{x}_2 + 9\mathbf{x}_3 + 18\mathbf{x}_4, & -18\mathbf{x}_1 + 12\mathbf{x}_2 - 6\mathbf{x}_3 + 6\mathbf{x}_4, \\ -18\mathbf{x}_1 + 45\mathbf{x}_2 + 6\mathbf{x}_3 + 15\mathbf{x}_4, & 90\mathbf{x}_1 + 48\mathbf{x}_2 + 63\mathbf{x}_3 + 12\mathbf{x}_4 \end{array} \right\rangle$$

is isomorphic to,

$$\mathbb{Z}_3 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_{18} \oplus \mathbb{Z}^1$$

which has a maximal finite subgroup of order $3 \times 3 \times 18 = 162$. △

33.13.6 Finite Abelian Groups

For any finite abelian group G, we now have $G \cong \bigoplus_{i=1}^{r} \mathbb{Z}_{d_i}$, where $d_i | d_{i+1}$ for $1 \leq i < r$ and $|G| = d_1 d_2 \cdots d_r$. Because the unimodular Smith normal form is unique, this implies that this decomposition is also unique, and so, the isomorphism classes of finite abelian groups of order $n > 0$ are in bijection with the factorisations of n,

$$n = \prod_{i=1}^{r} d_i$$

for which $d_i | d_{i+1}$ for $1 \leq i < r$. This allows us to classify isomorphism classes of finite abelian groups.

Example.

- $n = 4$ – the valid decompositions are 4 and 2×2, so every group of order 4 is isomorphic to $\mathbb{Z}_4$ or $\mathbb{Z}_2 \oplus \mathbb{Z}_2$.
- $n = 15$ – the only valid decomposition is 15, so every group of order 15 is isomorphic to $\mathbb{Z}_{15}$ and is hence necessarily cyclic.
- $n = 36$ – we have 36, 2×18, 3×12, and 6×6, so groups of order 36 are isomorphic to $\mathbb{Z}_{36}$, $\mathbb{Z}_2 \oplus \mathbb{Z}_{18}$, $\mathbb{Z}_3 \oplus \mathbb{Z}_{12}$, or $\mathbb{Z}_6 \oplus \mathbb{Z}_6$.

△

Lemma 33.13.23. *Let $G = \bigoplus_{i=1}^{n} G_i$ be a finite abelian group. Then, the order of $g = (g_1, g_2, \ldots, g_n)$ is the least common multiple of the orders $|g_i|$ of the components of g.*

Chapter 34

Analysis

> *"Calculus required continuity, and continuity was supposed to require the infinitely little; but nobody could discover what the infinitely little might be."*
>
> — Bertrand Russell, *Mysticism and Logic and Other Essays*

Analysis is the study of functions, sequences and series of real or complex numbers. We will begin by investigating the notions of convergence, continuity, and limits before formalising the foundations of differentiation and integration.

34.1 Real Analysis

The theorems of real analysis rely on various properties of the real number system. The properties of the real numbers as a field has already been explored in §12.10, but we will quickly revisit the ordering properties as they will feature prominently in the theory of sequences:

- $\forall a : x < y \leftrightarrow x + a < y + a$ (translational invariance): if $x < y$, then $x + a < y + a$ for all a – adding the same number to each side preserves the inequalities;
- $x < y \wedge u \leq v \rightarrow x + u < y + v$: if $x < y$ and $u \leq v$, then $x + y < y + v$ – adding inequalities that face the same direction preserves the stricter inequality;
- $x < y \wedge a > 0 \rightarrow (x < y \leftrightarrow ax < ay)$ (scaling invariance): if a is a positive number, then $x < y$ if and only if $ax < ay$ – multiplying both sides by a positive number preserves the inequality;
- $x < y \wedge a < 0 \rightarrow (x < y \leftrightarrow ax < ay)$: if a is a negative number, then $x < y$ if and only if $ax > ay$ – multiplying both sides by a negative number reverses the inequality;
- $x < y \wedge y < z \rightarrow x < z$: if $x < y$ and $y < z$, then $x < z$ – inequalities of the same type are transitive.
- $x < y \wedge y \leq z \rightarrow x < z$: if $x < y$ and $y \leq z$, then $x \leq z$ – inequalities facing the same direction but of different types preserve the stricter inequality.
- $x,y \in \mathbb{R}^+,n \in \mathbb{N} \rightarrow x < y \leftrightarrow x^n < y^n$ (power rule): if x and y are positive reals, then $x < y$ if and only if $x^n < y^n$ for all natural n.

For a more axiomatic and proof oriented overview, see §11.3.

The *floor function* $\lfloor x \rfloor$ and *ceiling function* $\lceil x \rceil$ turn an arbitrary real x into an integer: the floor of x is the largest integer less than or equal to x, while the ceiling of x is the smallest integer greater than or equal to x.

- $\lfloor x \rfloor = \sup\{z \in \mathbb{Z} : z \leq x\}$
- $\lceil x \rceil = \inf\{z \in \mathbb{Z} : y \geq x\}$

Because the complex numbers are not an ordered field, the floor and ceiling functions can only take real valued inputs, as non-real complex numbers are not comparable with integers.

The floor and ceiling of a real number is always an integer, with,

$$x - 1 < \lfloor x \rfloor \leq x \leq \lceil x \rceil < x + 1$$

with equality if and only if x is already integer: $x \in \mathbb{Z} \leftrightarrow \lfloor x \rfloor = x = \lceil x \rceil$.

The *fractional* or *decimal part* of x, sometimes denoted $\operatorname{frac}(x)$, can be computed with $x - \lfloor x \rfloor$.

Examples:

- $\lfloor \pi \rfloor = 3$, $\lceil \pi \rceil = 4$.
- $\lfloor 1.5 \rfloor = -1$, $\lceil 1.5 \rceil = 2$.
- $\lfloor -1.5 \rfloor = -2$, $\lceil -1.5 \rceil = -1$.
- $\lfloor 2 \rfloor = \lceil 2 \rceil = 2$.

Some authors denote the floor function with $[x]$, and don't include a ceiling function. Others use $[x]$ to represent the *integer part function*, defined as,

$$[x] = \begin{cases} \lfloor x \rfloor & x \geq 0, \\ \lceil x \rceil & x < 0. \end{cases}$$

so, this function picks the next smallest integer if x is non-negative, and the next largest integer if x is negative – it *rounds towards zero*.

- $[\pi] = 3$
- $[1.5] = 1$
- $[-1.5] = -1$

We will not use the integer part function here.

The *sign* or *signum function*, $\operatorname{sgn}(x)$ returns the sign of its argument, encoded as -1 for negative, 0 for zero, and $+1$ for positive.

$$\operatorname{sgn}(x) = \begin{cases} -1 & x < 0, \\ 0 & x = 0, \\ +1 & x > 0. \end{cases}$$

On the other hand, the *absolute value* of x, written $|x|$, erases the sign of x: $|-3| = |3| = 3$.

$$|x| = \begin{cases} -x & x < 0, \\ x & x \geq 0. \end{cases}$$

We can also define $|x|$ as $x \operatorname{sgn}(x)$.

For all x and y,

1. $|-x| = |x|$ (evenness);
2. $|x| \geq 0$ (non-negativity);
3. $|x| = 0 \leftrightarrow x = 0$ (positive-definiteness);

4. $|x-y| = 0 \leftrightarrow x = y$ (identity of indiscernibles);
5. $||x|| = |x|$ (idempotency);
6. $|xy| = |x||y|$ (multiplicativity);
7. $\left|\frac{x}{y}\right| = \frac{|x|}{|y|}$.

Proof. 1. If x is non-negative, then $|x| = x = |-x|$. Otherwise if x

2. If $x \geq 0$, then $|x| = x \geq 0$. Otherwise, if $x < 0$ then $|x| = -x > 0$. In both cases, $|x| \geq 0$.
3. If $x > 0$, $|x| = x > 0$ so $|x| \neq 0$. If $x < 0$, then $|x| = -x > 0$ so $|x| \neq 0$. If $x = 0$ then $|x| = 0$. By trichotomy, exactly one of the previous cases holds. The first two cases show $x \neq 0 \rightarrow |x| \neq 0$, so by contrapositive, $|x| = 0 \rightarrow x = 0$, completing the forward direction. The third case shows $x = 0 \rightarrow |x| = 0$, completing the backward direction.
4. Follows directly from positive-definiteness.
5. From the definition, $|x| = x$ if x is non-negative, and $|x|$ is always non-negative, so $||x|| = |x|$.
6. We prove this by case analysis.

   If both x and y are positive, then $|x| = x$ and $|y| = y$, and $|x||y| = xy$. As x and y are both positive, xy is positive, so $|xy| = xy = |x||y|$.

   Without loss of generality, suppose x is positive and y is negative. Then, $|xy| = -xy = x(-y) = |x||y|$. By symmetry, this also holds if x is negative and y is positive.

   If both x and y are negative, then $|xy| = xy = (-x)(-y) = |x||y|$.
7. $|1| = \left|x \cdot \frac{1}{x}\right| = |x|\left|\frac{1}{x}\right|$, so $\frac{1}{|x|} = \left|\frac{1}{x}\right|$, and $\left|\frac{x}{y}\right| = |x|\left|\frac{1}{y}\right| = |x|\frac{|1|}{|y|} = \frac{|x|}{|y|}$.

■

The floor and ceiling functions are also idempotent – they can be applied multiple times without changing the result beyond the first application. So, $\lfloor x \rfloor = \lfloor \lfloor x \rfloor \rfloor = \lfloor \lfloor \lfloor x \rfloor \rfloor \rfloor = \ldots$.

Theorem (Interval Property)**.** *If $x \in \mathbb{R}$ and $r \in \mathbb{R}^+$, then $|x| < r$ if and only if $-r < x < r$.*

Proof. Suppose $|x| < r$. If x is non-negative, then $x < r$. Otherwise, $-x < r$, so $-r < x$, proving the forward direction.

Now suppose $-r < x < r$. If $x \geq 0$, then $x = |x|$, so $-r < |x| < r$. If $x < 0$, then $x = -|x|$ and $-r < -|x| < r$. Multiplying by -1, we again have $-r < |x| < r$, completing the backward direction. ■

Corollary 34.1.0.1. *If y,$a \in \mathbb{R}$ and $r \in \mathbb{R}^+$, then $|y-a| < r$ if and only if $a - r < y < a + r$.*

Proof. Substitute $y = x - a$ in the interval property. ■

This corollary justifies the graphical way of thinking of the absolute value $|a-b|$ as the distance along the real line between a and b.

In fact, using the absolute value function in this way is an example of a *metric* on $\mathbb{R}$, which is a generalised way of measuring distances. This topic is explored in more detail in the chapter on topology, §37.

34.1.1 Triangle Inequality

Theorem (Triangle Inequality)**.** *For all real numbers x and y, $|x+y| \leq |x| + |y|$.*

Proof. For all x and y,

$$\begin{aligned} -|x| \leq x \leq |x| \\ -|y| \leq y \leq |y| \end{aligned}$$

Adding the two inequalities, we have,

$$\begin{aligned} -|x| - |y| \leq x + y \leq |x| + |y| \\ -(|x| + |y|) \leq x + y \leq |x| + |y| \end{aligned}$$

By the interval property, $-a \leq b \leq a \leftrightarrow |b| \leq a$.

Set $a = |x| + |y|$ and $b = x + y$, so $-(|x| + |y|) \leq x + y \leq |x| + |y|$, and $|x+y| \leq |x| + |y|$. ■

The triangle inequality is of extreme importance and is used in many different applications.

Corollary 34.1.0.2. $|a-b| \leq |a-c| + |c-b|$

Proof. Let $x = a - b$ and $y = b - c$. Then,

$$\begin{aligned} |x+y| &\leq |x| + |y| \\ |a-b+b-c| &\leq |a-b| + |b-c| \\ |a-c| &\leq |a-b| + |b-c| \end{aligned}$$

■

Corollary (Reverse Triangle Inequality)**.** $|x-y| \geq \big||x| - |y|\big|$

Proof.

$$\begin{aligned} |y| &= |(y-x) + x| \\ |y| &\leq |y-x| + |x| \\ -|y-x| &\leq |x| - |y| \\ -|-1| \cdot |y-x| &\leq |x| - |y| \\ -|x-y| &\leq |x| - |y| \qquad (1) \end{aligned}$$

$$\begin{aligned} |x| &= |(x-y) + y| \\ |x| &\leq |x-y| + |y| \\ |x| - |y| &\leq |x-y| \qquad (2) \end{aligned}$$

Combining (1) and (2), we have,

$$-|x-y| \leq |x| - |y| \leq |x-y|$$

which, by the interval property, gives,

$$\big||x| - |y|\big| \leq |x-y|$$

■

34.1.2 Arithmetic & Geometric Means

The *arithmetic mean* of a set of numbers, X, is the sum of the numbers divided by the cardinality of X, or,

$$\frac{1}{|X|}\sum_{x\in X} x = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

The *geometric mean* of a set of numbers, X, is $|X|$th root of the product of the numbers, or,

$$\sqrt[|X|]{\prod_{x\in X} X} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

Proposition (AM-GM inequality)**.** For any set, X, of cardinality $n \in \mathbb{N}$ containing the non-negative real numbers $x_1, x_2 ... x_{n-1}, x_n$, the inequality

$$\frac{1}{n}\sum_{i=1}^{n} x_i \geq \sqrt[n]{\prod_{i=1}^{n} x_i}$$

holds.

Proof. If all numbers x_i are 0, then the inequality holds with equality, as both sides equal 0. If at least one (but not all) x_i is 0, then the geometric mean is 0, and the inequality is strict. Thus, we may assume that all x_i values are positive from this point onwards.

Let m equal the arithmetic mean of the x_n. That is,

$$\frac{x_1 + x_2 + \cdots + x_{n-1} + x_n}{n} = m$$
$$x_1 + x_2 + \cdots + x_{n-1} + x_n = mn$$

As there are n copies of x_i, it follows that if all values of x_i are equal, then they must also equal m.

Now assume that instead at least one x_i value is strictly greater than m. Now, suppose that all other values of x_i are at least a. The sum of x_i is therefore strictly greater than mn as $x_i > m$ and $x_i n > mn$, contradicting that the sum must be exactly equal to mn as defined above. Thus, the supposition that all other values of x_i are also at least a is false: at least one value of x_i is strictly less than m if at least one x_i value is strictly greater than m.

Similarly, the statement "If a value of $x_i > m$ exists, then a value of $x_i < m$ also exists." can be proven similarly as above with all inequalities reversed.

Combining the two statements above, we have that if all the x_i are not equal to a, then at least one of them is strictly less than a, and at least one another must be strictly greater than a. Let us denote this result as (1).

Suppose not all the x_i are equal, and consider two values of x_i, x_a and x_b such that $x_a < m < x_b$. Replace x_a with m and x_b with $x_a + x_b - m$. $(m) + (x_a + x_b - m) = x_a + x_b$, so the value of the sum of x_i has not changed, and hence the arithmetic mean, m, has also not changed.

Now consider the change to the geometric mean of the x_i caused by this replacement. $x_a x_b$ has been replaced with $m(x_a + x_b - m)$. We will demonstrate that the geometric mean has strictly increased.

$$\begin{aligned} x_a x_b &< m(x_a + x_b - m) \\ x_a x_b &< m x_a + m x_b - m^2 \\ 0 &< -x_a x_b + m x_a + m x_b - m^2 \end{aligned}$$

$$\begin{aligned}0 &< -(x_a - m)(x_b - m)\\ 0 &< (m - x_a)(x_b - m)\end{aligned}$$

Recall that $x_a < m < x_b$. Thus, $(m - x_a) > 0$ and $(x_b - m) > 0$, so the right side is indeed always positive, as required. Therefore, it is possible to replace variables without changing the arithmetic mean, while strictly increasing the geometric mean. Denote this result as (2).

Now, suppose that we have n non-negative real numbers, $x_1, x_2...x_{n-1}, x_n$. As before, let their arithmetic mean be m. If all the x_i are equal, then their arithmetic means and geometric means are both equal to m.

If instead, the x_i are not equal, then we are guaranteed by result (1) that we can find two values of x_i such that $x_a < m < x_b$. Now replace x_a by m, and x_b by $x_a + x_b - m$. By result (2), the arithmetic mean is unchanged (and is equal to m), while the geometric mean is strictly increased.

Every time a replacement is performed, at least one more x_i value is changed to m, so after at most n replacements, all the x_i are equal to m, as are the arithmetic and geometric means.

When the x_i are equal, the arithmetic mean and geometric mean are equal. At all steps where the x_i are not equal, it is possible to increase the value of the geometric mean. Therefore, the geometric mean is at most equal to the arithmetic mean when all the x_i are equal, and is smaller for every other case: the geometric mean is always less than or equal to the arithmetic mean, which is equivalent to the statement that the arithmetic mean is always greater than or equal to the geometric mean. ∎

Another proof of this by the means of backwards-forwards induction is given in §5.

The *harmonic mean* of set of numbers X is $|X|$ divided by the sum of the reciprocals of the numbers, or,

$$n\left(\sum_{x\in X}\frac{1}{x}\right)^{-1} = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}$$

The harmonic mean is used to find the average of rates and ratios, and is also used in physics to find the resistance of parallel resistors.

The arithmetic, geometric and harmonic means are the three *Pythagorean means.*

We also have the *quadratic mean* or *root mean square* of a set of numbers, which is the square root of the mean of squares of the set, or,

$$\sqrt{\frac{1}{|X|}\sum_{x\in X}x^2} = \sqrt{\frac{x_1^2+x_2^2+\cdots+x_n^2}{n}}$$

The quadratic mean is also used in electronics to calculate currents, as well as in statistics as a measure of deviation.

Theorem (HM-GM-AM-QM Inequality)**.** *If $\{x_i\}_{i=1}^n$ is a set of positive real numbers, then,*

$$0 < n\left(\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1} \leq \sqrt[n]{\prod_{i=1}^{n}x_i} \leq \frac{1}{n}\sum_{i=1}^{n}x_i \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2}$$

34.2 Sequences

A sequence is an ordered list of numbers indexed by natural numbers.

We can also call the members of a sequence *terms* of the sequence. The number of elements in a sequence (possibly infinite) is called the *length* of the sequence.

We denote a sequence by (a_n):

$$(a_n) = a_0,a_1,a_2,a_3,\ldots$$

Note the brackets: a_n refers to the term of the sequence (a_n) at position n. We also sometimes exclude a_0 if the formula for generating the a_n terms is problematic at 0, for example, $a_n = \frac{1}{n}$. For our purposes, this loss of a beginning term is unimportant and won't be explicitly mentioned unless relevant.

Formally, a sequence can be defined as a function from the natural numbers to the elements at each position, so a_n is really just syntactic sugar for $a(n)$, with $a : \mathbb{N} \to \mathbb{R}$ being a function.

In this chapter we will only be considering infinite sequences of real numbers.

34.2.1 Monotonicity

A sequence, (a_n), is:

- *strictly increasing* if $\forall n : a_{n+1} > a_n$;
- *increasing* if $\forall n : a_{n+1} \geq a_n$
- *strictly decreasing* if $\forall n : a_{n+1} < a_n$;
- *decreasing* if $\forall n : a_{n+1} \leq a_n$;
- *monotonic* if it is increasing or decreasing or both (i.e. is constant);
- *non-monotonic* if it is neither increasing nor decreasing.

Examples:

- $a_n = n = 0,1,2,3,\ldots$
- $b_n = (-1)^n = 1,-1,1,-1,\ldots$
- $c_n = \frac{1}{n} = \frac{1}{1},\frac{1}{2},\frac{1}{3},\frac{1}{4},\ldots$

1. For all n, $a_n = n < n + 1 = a_{n+1}$, so (a_n) is strictly increasing.
2. $b_1 = -1 < 1 = b_2$ and $b_2 = 1 > -1 = b_3$, so (b_n) is neither increasing nor decreasing, i.e. it is non-monotonic.
3. For all n, $c_n = \frac{1}{n} > \frac{1}{n+1} = c_{n+1}$, so (c_n) is strictly decreasing.

Exercises. Prove whether each of these sequences are monotonic or not:

- $a_n = -1$;
- $a_n = -\frac{1}{n^2}$;
- $a_n = n^2$;
- $a_n = 2^n$;
- $a_n = \sin(n)$;
- $a_n = n^n \sin(n^2\pi)$;
- $a_n = \sqrt{n+1} - \sqrt{n}$;
- $a_n = \frac{1}{n!}$.

34.2.2 Bounded Sequences

A sequence, (a_n) is:

- *bounded above* if $\exists U \in \mathbb{R}$ such that $a_n \leq U \,\forall\, n$. U is then an *upper bound* of the sequence;
- *bounded below* if $\exists L \in \mathbb{R}$ such that $a_n \geq L \,\forall\, n$. L is then a *lower bound* of the sequence;
- *bounded* if it is bounded above *and* bounded below.

Upper and lower bounds are not unique. For example, if M is an upper bound of (a_n), then $M+1$ is clearly also an upper bound.

Every increasing sequence is bounded below by its first term.

Every decreasing sequence is bounded above by its first term.

Exercises.

1. Determine whether each sequence in the previous section is bounded above, bounded below, or bounded, and give examples in the cases where they exist.
2. A sequence is known to be increasing.
   (a) Might it have an upper bound?
   (b) Must it have an upper bound?
3. A sequence is known to be unbounded above.
   (a) Must it have a positive term?
   (b) Must it have an infinite number of positive terms?
   (c) Can it have negative terms?
   (d) Can it have infinitely many negative terms?
   (e) Can it also be unbounded below?

34.2.3 Sequences Tending to Infinity

We say that a sequence *tends to infinity* if its terms become arbitrarily large. We show this by showing that the sequence eventually exceeds any number we select.

Definition 34.2.1. A sequence (a_n) *tends to infinity* if for every $C > 0$, there exists $N \geq 1$ such that $a_n > C$ for all $n > N$.

Similarly, a sequence tends to minus infinity if it becomes arbitrarily negative.

Definition 34.2.2. A sequence (a_n) *tends to minus infinity* if for every $C > 0$, there exists $N \geq 1$ such that $a_n < -C$ for all $n > N$.

If a sequence tends to infinity, we write one of,

$$(a_n) \to \infty$$
$$a_n \to \infty \text{ as } n \to \infty$$
$$\lim_{n\to\infty} a_n = \infty$$

and we say that (a_n) *diverges* to infinity.

Similarly, we write one of,

$$(a_n) \to -\infty$$
$$a_n \to -\infty \text{ as } n \to \infty$$
$$\lim_{n\to\infty} a_n = -\infty$$

if (a_n) diverges to minus infinity.

Example. $a_n = n$ diverges to infinity, because for any $C > 0$, we can pick $N = C + 1$, so, for all $n > N$, $a_n = n > N > C$. △

We can show that a sequence does not tend to infinity by finding an upper bound, and similarly, we can show that a sequence does not tend to minus infinity by finding a lower bound.

Example. $a_n = \frac{1}{n}$ does not diverge to infinity because it is bounded above by, say, 1. It also does not diverge to minus infinity because it is bounded below by, say, 0. △

Theorem 34.2.1. *Let (a_n) and (b_n) be sequences such that $b_n \geq a_n \, \forall \, n$, and suppose that $a_n \to \infty$. Then, $b_n \to \infty$.*

Proof. Suppose $C > 0$. Because $a_n \to \infty$, there exist N such that $a_n > C$ whenever $n > N$. But $b_n \geq a_n$ for all n, so b_n is also greater than C whenever $n > N$, satisfying the definition of divergence to infinity. ■

Suppose (a_n) and (b_n) tend to infinity, then,

- $a_n + b_n \to \infty$;
- $a_n b_n \to \infty$;
- $ca_n \to \infty$ if $c > 0$;
- $ca_n b_n \to -\infty$ if $c < 0$.

Proof. Let $C > 0$ such that $\frac{C}{2} > 0$. Because the sequences both tend to infinity, we know there exists N_1 and N_2 such that $a_n > \frac{C}{2}$ whenever $n > N_1$, and $b_n > \frac{C}{2}$ whenever $n > N_2$.

Let $N = \max(N_1,N_2)$. Because $N \geq N_1$ and $N \geq N_2$, $a_n > \frac{C}{2}$ and $b_{\frac{C}{2}}$ both hold whenever $n > N$.

Then, $a_n + b_n > \frac{C}{2} + \frac{C}{2} = C$ whenever $n > N$, which is the definition of tending to infinity.

The proof of the last three properties is left as an exercise for the reader (though each proof is essentially the exact same proof as above). ■

34.2.4 Convergent Sequences

Consider the sequence,

$$a_n = \frac{1}{2^n} = 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \ldots$$

Each term is getting closer and closer to zero, the difference becoming half the size of each step. You would probably agree that this sequence "tends to zero" or "approaches zero", whatever that means.

What about this sequence?

$$a_n = \frac{1}{n} = 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$$

Sure enough, it doesn't shrink quite as quickly as the previous sequence, but this sequence does also seem to be "tending to zero".

And again,

$$a_n = \frac{(-1)^n}{n} = -1, \frac{1}{2}, -\frac{1}{3}, \frac{1}{4}, -\frac{1}{5}, \ldots$$

this sequence jumps about from both sides of zero, but also seems to close in towards it, so we might say that this sequence also "tends to zero".

Now, this sequence is a little different:

$$a_n = 2 - \frac{1}{2^n} = 1, 1.5, 1.75, 1.875, 1.9375, 1.96875, \ldots$$

This time, it sounds reasonable to say that this sequence is “approaching two”.

But, how do we formalise all of this? We want a precise definition of what it means for a sequence to “tend to” or “approach” a number. Maybe it will help to see a sequence which we can agree *doesn't* tend towards a number.

If we look at this sequence,

$$a_b = (-1)^n = 1, -1, 1, -1, \dots$$

This sequence doesn't seem to tend towards anything. The terms seem to sit at a constant distance of 1 away from 0, so it seems reasonable to say that this sequence doesn't tend towards zero.

Maybe the distances between the term and the target number just has to get smaller with each step. Sounds reasonable enough.

But, consider the sequence given by $a'_n = \frac{1}{2^n} + 1$. This is the same as the first sequence, $a_n = \frac{1}{2^n}$, but we've added 1 to each term. Every term in this new sequence also “gets closer” to 0, but clearly, this sequence never drops below 1, so saying that this sequence also tends to zero seems wrong.

And equivalently, a_n also gets closer to -1 with each term, and also to -2, and -3.

It's not enough for the distance between each term and the target number to get smaller with each step, this distance needs to be able to become *arbitrarily small.*

For now, let's concentrate on the case where the sequence is tending to zero.

34.2.4.1 Null Sequences

Definition 34.2.3. A sequence (a_n) *tends to zero* if, for every $\varepsilon > 0$, there exists a natural N such that $|a_n| < \varepsilon$ for all $n > N$.

To represent this, we write,

$$\begin{aligned}(a_n) &\to 0\\ a_n &\to 0 \text{ as } n \to \infty\\ \lim_{n\to\infty} a_n &= 0\end{aligned}$$

We can also say that (a_n) is a *null sequence*, or that 0 is the *limit* of (a_n).

The idea of this definition is that, for any *error*, ε, we can always find a value of N big enough that the terms of (a_n) always land within ε of 0 whenever $n > N$.

For example, to prove that $a_n = \frac{1}{n}$ converges to 0, we let $\varepsilon > 0$ and let $N = \frac{1}{N}$. Then, whenever $n > N$, $a_n = \frac{1}{n} < \frac{1}{N} < \varepsilon$ so $|a_n| < \varepsilon$

We have actually seen this proof before quite some time ago in §2.3.1.6, if you can still remember.

On the other hand, we can prove that $a_n = (-1)^n$ *doesn't* converge to 0, by letting $\varepsilon = \frac{1}{2}$. For all n, $|a_n| = 1 > \frac{1}{2} > \varepsilon$, so we know that this sequence doesn't converge to zero.

Lemma 34.2.2. *If* $(a_n) \to \infty$*, then* $\left(\frac{1}{a_n}\right) \to 0$.

Proof. Since $\lim_{n\to\infty} a_n \to \infty$, there exists a value $N \in \mathbb{N}$ such that for all $n > N$, $a_n > C$ for any particular value of $C \in \mathbb{R}$.

Let $\varepsilon > 0$ and $C = \frac{1}{\varepsilon}$, such that, for all n, if $0 < a_n < C$, then $0 < \frac{1}{a_n} < \frac{1}{C} = \varepsilon$, so $a_n < \varepsilon$.

So, if a value of N exists such that for all $n > N$, $a_n > C$, then the same N is sufficient for all $n > N$ to have $\frac{1}{a_n} < \varepsilon$. ■

Note that the converse of this statement is not true without an extra condition:

Lemma 34.2.3. *If* $(a_n) \to 0$ *and* (a_n) *is monotonic, then* $\left(\frac{1}{a_n}\right) \to \infty$.

Proof. Exercise. ■

Lemma (Absolute Value Rule)**.** $(a_n) \to 0$ *if and only if* $(|a_n|) \to 0$.

Proof. $\big||a_n|\big| = |a_n|$, so $|a_n| \to 0$ if and only if $\forall\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $\big||a_n|\big| = |a_n| < \varepsilon$ whenever $n > N$, which is exactly the definition of $(a_n) \to 0$, completing the backward direction. The same argument holds in reverse, completing the forward direction. ■

Theorem (Sandwich Theorem for Null Sequences)**.** *If* $(a_n) \to 0$ *and* $0 \leq |b_n| \leq a_n$, *then* $(b_n) \to 0$.

Proof. Let $\varepsilon > 0$. As (a_n) is null, there exists $N \in \mathbb{N}$ such that whenever $n > N$, $|a_n| < \varepsilon$. Because $0 \leq |b_n| \leq a_n$, $|b_n| \leq |a_n| < \varepsilon$, so $(|b_n|)$ is also a null sequence. By the absolute value rule, (b_n) is then also a null sequence. ■

Theorem (Arithmetic of Null Sequences)**.** *Let* $c,d \in \mathbb{R}$ *and* $(a_n) \to 0$ *and* $(b_n) \to 0$. *Then,*

$$ca_n + db_n \to 0 \qquad \textit{Sum rule} \text{ for null sequences}$$
$$a_n b_n \to 0 \qquad \textit{Product rule} \text{ for null sequences}$$

Proof. Sum rule. Let $\varepsilon > 0$, so $\frac{\varepsilon}{2} > 0$.

As a_n is a null sequence, there exists an $N_a \in \mathbb{N}$ such that for all $n > N_a$, $|a_n| < \frac{\varepsilon}{2}$. Similarly, there exists a $N_b \in \mathbb{N}$ such that for all $n > N_b$, $|b_n| < \frac{\varepsilon}{2}$.

Now, let $N = \max(N_a,N_b)$ or $N = N_a + N_b$, such that $N > N_a$ and $N > N_b$. Because $N > N_a$ and $|a_n| < \frac{\varepsilon}{2}$ for all $n > N_a$, we have $|a_n| < \frac{\varepsilon}{2}$ for all $n > N$. $|b_n| < \frac{\varepsilon}{2}$ for all $n > N$ through similar reasoning. It follows that, for all $n > N$, $|a_n| + |b_n| < \varepsilon$, and we have $|a_n + b_n| < \varepsilon$ by the triangle inequality.

Product rule. Let $\varepsilon > 0$, so $\sqrt{\varepsilon} > 0$.

As a_n is a null sequence, there exists an $N_a \in \mathbb{N}$ such that for all $n > N_a$, $|a_n| < \sqrt{\varepsilon}$. Similarly, there exists a $N_b \in \mathbb{N}$ such that for all $n > N_b$, $|b_n| < \sqrt{\varepsilon}$.

Now, let $N = \max(N_a,N_b)$ or $N = N_a + N_b$, such that $N > N_a$ and $N > N_b$. Because $N > N_a$ and $|a_n| < \sqrt{\varepsilon}$ for all $n > N_a$, we have $|a_n| < \sqrt{\varepsilon}$ for all $n > N$. $|b_n| < \sqrt{\varepsilon}$ for all $n > N$ through similar reasoning. It follows that, for all $n > N$, $|a_n||b_n| < \varepsilon$, so $|a_n + b_n| < \varepsilon$ by multiplicativity of absolute values. ■

34.2.4.2 Convergent Sequences

Now that we are comfortable with some simple null sequences, we can extend the definition of convergence to more general non-zero limits.

Definition 34.2.4. A sequence, (a_n) *converges to* or *tends to* a if $\forall\varepsilon > 0\, \exists N \in \mathbb{N}$ such that $|a_n - a| < \varepsilon$ whenever $n > N$.

If such an a exists, the sequence is *convergent*, and a is the *limit* of the sequence. Otherwise, the sequence is *divergent*. Note that sequences di<u>verge</u> to (minus) infinity, not <u>con</u>verge.

If (a_n) converges to a, we write,

$$(a_n) \to a$$

$$
\begin{aligned}
a_n \to a \text{ as } n \to \infty \\
\lim_{n\to\infty} a_n = a
\end{aligned}
$$

So, we can see that null sequences are a special case of convergent sequences – a sequence is null if $a = 0$ in the above.

Corollary 34.2.3.1. *$(a_n) \to a$ if and only if $(a_n - a) \to 0$.*

Above, we said that a is *the* limit of the sequence. Can there be more?

Theorem (Uniqueness of Limits)**.** *A sequence can converge to at most one limit.*

Proof. Let (a_n) be a sequence with two limits, a and b, That is, $\lim_{n\to\infty} a_n = a$, $\lim_{n\to\infty} a_n = b$.

Let $\varepsilon > 0$, and let $\varepsilon_a, \varepsilon_b < \frac{\varepsilon}{2}$. Let $N = \max(N_a, N_b)$ such that $N > N_a$ and $N > N_b$. Because $N > N_a$, and $|a_n - a| < \varepsilon_a$ for all $n > N_a$, we have $|a_n - a| < \varepsilon_a$ for all $n > N$. We also have $|a_n - b| < \varepsilon_b$ for all $n > N$ through similar reasoning. Then,

$$
\begin{aligned}
|a_n - a| + |a_n - b| &< \varepsilon_a + \varepsilon_b \\
&< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\
&< \varepsilon
\end{aligned}
$$

So $|a_n - a| + |a_n - b| < \varepsilon$. Now,

$$
\begin{aligned}
|a - b| &= |a - b + (a_n - a_n)| \\
&= |a - a_n + a_n - b| \\
&\le |a - a_n| + |a_n - b| \\
&\le |-(a_n - a)| + |a_n - b| \\
&\le |a_n - a| + |a_n - b|
\end{aligned}
$$

So $|a - b| \le |a_n - a| + |a_n - b|$. Combined with the first inequality, we have $|a - b| < \varepsilon$.

Suppose $a \neq b$, so $a - b \neq 0$. Then, $|a - b| = k > 0$. If $\varepsilon < k$, then $|a - b| > \varepsilon$, which is a contradiction. It follows that $a = b$. Since a and b were arbitrary, the limit of a sequence is unique. ■

Theorem (Boundedness of Convergent Sequences)**.** *Every convergent sequence is bounded.*

Proof. If a_n converges to k, then there exists $N \in \mathbb{N}$ such that for all $n > N$, $|a_n - k| < \varepsilon$ for any $\varepsilon > 0$. Fix $n = N + 1 > N$. It follows that there are finitely many terms, namely $a_1, a_2, a_3, \cdots a_{N-1}, a_N$, which can be greater than ε away from k. Let $A = \max(a_1, a_2, a_3, \cdots a_{N-1}, a_N)$ and $B = \min(a_1, a_2, a_3, \cdots a_{N-1}, a_N)$. $A \ge a_n$ for $n \in \{1, 2, 3, \cdots N\}$, and similarly, $B \le a_n$ over the same interval. Finally, restrict $\varepsilon < \min(|A - k|, |B - k|)$ so $|a_n - k| < |A - k|$ and $|a_n - k| < |B - k|$ for all $n > N$. Now, $B \le a_n \le A$ for all n, so a_n is bounded by A and B. ■

Theorem (Algebra of Convergent Sequences)**.** *Let $a, b \in \mathbb{R}$, and $(a_n) \to a$ and $(b_n) \to b$ be convergent sequences. Then,*

$$
\begin{aligned}
ca_n + db_n &\to ca + db && \textit{Sum rule} \\
a_n b_n &\to ab && \textit{Product rule} \\
\frac{a_n}{b_n} &\to \frac{a}{b}, \textit{ provided } b \neq 0 && \textit{Quotient rule}
\end{aligned}
$$

Proof. Sum rule. $a_n \to a$ and $b_n \to b$, so $A_n = (a_n - a)$ and $B_n = (b_n - b)$ are null sequences. By the sum rule for null sequences, $cA_n + dB_n \to 0$. $c(a_n - a) + d(b_n - b) = ca_n - ca + db_n - db$, so $ca_n + db_n \to ca + db$.

Product rule. $a_n \to a$ and $b_n \to b$, so $A_n = (a_n - a)$ and $B_n = (b_n - b)$ are null sequences. $a_nb_n - ab = (a_n - a)(b_n - b) + a(b_n - b) + b(a_n - a) = A_nB_n + aB_n + bA_n$. By the product rule for null sequences, $A_nB_n \to 0$. aB_n and bA_n, being scaled null sequences, are also null sequences. Thus the entire right side is the sum of null sequences, so $a_nb_n - ab \to 0$ and $a_nb_n \to ab$.

Quotient rule. Let $b_n \to b$. By the product rule, $bb_n \to b^2$, so, for any $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $n > N$, $|bb_n - b^2| < \varepsilon$.

$$\begin{aligned}\left|\frac{1}{b_n} - \frac{1}{b}\right| &= \frac{|b - b_n|}{|b||b_n|} \\ &= \frac{|-(b_n - b)|}{|b||b_n|} \\ &= \frac{|b_n - b|}{|b||b_n|}\end{aligned}$$

As b_n is convergent, it is bounded. Let A be a lower bound of b_n, such that $0 < A < |b_n|$*.

$$\begin{aligned}&\le \frac{|b_n - b|}{A|b|} \\ &\le \frac{1}{A|b|}|b_n - b|\end{aligned}$$

$\frac{1}{A|b|}$ is a constant, so,

$$\le k|b_n - b|$$

$|b_n - b|$ is null, so $k|b_n - b|$ also converges to 0. It follows that $\left|\frac{1}{b_n} - \frac{1}{b}\right|$ is also null by the sandwich rule for null sequences, so $\frac{1}{b_n} \to \frac{1}{b}$.

$\frac{a_n}{b_n} = a_n \cdot \frac{1}{b_n}$ which, by the product rule, converges to $a \cdot \frac{1}{b} = \frac{a}{b}$. ■

Theorem (Sandwich Theorem for Sequences)**.** *Suppose $(a_n) \to l$ and $(c_n) \to l$. If $a_n \le b_n \le c_n$, then, $(b_n) \to l$.*

Proof. Let b be the limit of b_n. $a_n \le b_n \le c_n$, so $0 \le b_n - a_n \le c_n - a_n$, and apply the sandwich theorem for null sequences, so $(b_n - a_n)$ is a null sequence. But, from the algebra of convergent sequences, we have $b_n - a_n \to 0 = b - l$, so $b - l = 0$ and $b = l$, so $(b_n) \to l$. ■

In many cases, we only care about what happens to a sequence "after a certain point", ignoring what happens at the beginning of a sequence. We can do this by moving the indices of the sequence along, creating a new *shifted* sequence.

Definition 34.2.5. A sequence (a_n) satisfies a certain property *eventually* if there exists an integer $n \ge 0$ such that the sequence (a_{N+n}) satisfies that property.

* We can actually find a more specific value, given that N is large enough:
Suppose $0 < \varepsilon < \frac{b^2}{2}$, so for all $n > N$, $|bb_n - b^2| < \frac{b^2}{2}$. By the interval property, $-\frac{b^2}{2} < bb_n - b^2 < \frac{b^2}{2}$, so $\frac{b^2}{2} < bb_n < \frac{3b^2}{2}$ and $bb_n > \frac{b^2}{2}$ for $n > N$.

Most proofs and properties that make use of "eventually" leverage the fact that finite sets always have suprema and infima, while infinite sets may not.

For example,

Lemma 34.2.4. *If a sequence is eventually bounded, it is bounded.*

Proof. Let $\varepsilon > 0$, and suppose a_n be eventually bounded by u and l with $u > l$ and $\left|\frac{u+l}{2}\right| > \varepsilon$. Let $k = \frac{u+l}{2}$. Then, there exists $N \in \mathbb{N}$ such that for all $n > N$, $|a_n - k| < \varepsilon$. That is, the sequence never moves more than the size of the average of u and l away from the average of u and l; if it did, the sequence would be greater than u or less than l.

Fix $n = N+1 > N$. It follows that there are finitely many terms, namely $a_1,a_2,a_3,\cdots a_{N-1},a_N$, which can be greater than ε away from k. Let $A = \max\left(a_1,a_2,a_3,\cdots a_{N-1},a_N\right)$ and $B = \min\left(a_1,a_2,a_3,\cdots a_{N-1},a_N\right)$. $A \geq a_n$ for $n \in \{1,2,3,\cdots N\}$, and similarly, $B \leq a_n$ for those n. Finally, let $U = \max(A,u)$ and $L = \min(B,l)$ so $a_n < U$ and $a_n > L$ for all n. The sequence is therefore bounded by U and L for all n. ■

The next theorem, the *shift rule*, says that a sequence converges if and only if it converges eventually. Intuitively, this is just saying that convergence is something that happens as the index goes to infinity, it doesn't care about what happens to any finite number of terms at the beginning.

Theorem (Shift Rule). *Let $N \in \mathbb{N}$. Then, $(a_n) \to a$ if and only if $(a_{N+n}) \to a$*

Proof. Let $\varepsilon > 0$, and suppose $(a_n) \to a$. Then, there exists N_1 such that for all $n > N_1$, $|a_n - a| < \varepsilon$. For any $N \in \mathbb{N}$, $N + n > N_1$, so $|a_{N+n} - a| < \varepsilon$ as well, which is the definition of convergence, so $(a_{N+n}) \to a$, completing the forward direction.

Now suppose $(a_{N+n}) \to a$, so there exists N_2 such that for all $n > N_2$, $|a_{N+n} - a| < \varepsilon$. Whenever $n - N > N_2$, $n > N + N_2$, so $|a_{N+(n-N)} - a| = |a_n - a| < \varepsilon$, so $(a_n) \to a$, completing the backward direction. ■

Exercise. Prove that the shift rule also works for sequences which diverge to (minus) infinity.

Corollary (Sandwich Theorem with Shift Rule). *Suppose $(a_n) \to l$ and $(b_n) \to l$. If eventually $a_n \leq c_n \leq b_n$ then $(c_n) \to l$.*

Proof. Exercise. ■

Lemma 34.2.5. *Suppose $(a_n) \to a$. Then, if $a_n \geq 0$ for all n, then $a \geq 0$.*

Proof. Suppose that $a < 0$. As a_n converges to a, there exists $N \in \mathbb{N}$ such that for all $n > N$, $|a_n - a| < \varepsilon$ for any choice of $\varepsilon > 0$. Let $\varepsilon = -a$. It follows that for all $n > N$, $0 < a_n - a < -a$, so $a < a_n < 0$, which is a contradiction as $a_n \geq 0$. The original supposition is therefore false, so $a \geq 0$. ■

Theorem (Inequality rule). *Suppose $(a_n) \to a$ and $(b_n) \to b$. If eventually $a_n \leq b_n$ then $a \leq b$.*

Proof. Consider $c_n = b_n - a_n$. From the sum rule, $c_n \to c$, where $c = b - a$. As $b_n \geq a_n$, $c_n \geq 0$ for sufficiently large n. From the previous lemma, it follows that $c \geq 0$, and $c = b - a \geq 0$, so $b \geq a$. ■

Corollary (Closed Interval Rule). *Suppose $(a_n) \to a$ and $(b_n) \to b$. If eventually $a_n \leq b_n$ then $a \leq b$.*

The above corollary shows that a limit cannot escape from a closed interval. They can, however, escape open intervals, but only as far as the supremum or infimum of the set. We've already seen this happen: $a_n = \frac{1}{n}$ is positive for all n, so the sequence exists in the set, $\{x \in \mathbb{R} : x > 0\}$, but the limit is 0, which is the infimum of the set.

34.2.5 Subsequences

A subsequence of a sequence is a sequence consisting of some (or all) of its terms, without changing the order of those terms. That is, it is a monotonic function $i : \mathbb{N} \to \mathbb{N}$ on the indexing set of the sequence. The relation of a sequence being the subsequence of another is a preorder (§4.4.6).

If $(a_n) = a_1,a_2,a_3,\ldots$ is a sequence, any strictly increasing sequence of natural numbers, $(n_i) = n_1,n_2,n_3,\ldots$, will generate a subsequence, $(a_{n_i}) = a_{n_1},a_{n_2},a_{n_3},\ldots$.

Note that this subsequence is indexed by i, and not n, and in all cases $n_i \geq i$.

There are many useful properties of subsequences, which we'll package together as a lemma.

Lemma (Properties of Subsequences)**.**

- *If (a_n) is convergent if and only if every subsequence is convergent.*
- *If (a_{2n+1}) and (a_{2n}) both converge to the same limit, then (a_n) also converges to that limit.*[*]
- *If (a_n) diverges to $\pm\infty$, then every subsequence diverges to $\pm\infty$.*
- *If (a_n) is bounded above (below), then every subsequence is bounded above (below).*
- *If (n_i) is a strictly increasing sequence of natural numbers, then for all $i \geq 1$, $n_i \geq i$.*

Proof. Exercise. ■

Theorem (Monotonic Subsequence Theorem)**.** *Every sequence has a monotonic subsequence.*

Proof. First, some new terminology is helpful:

Definition 34.2.6. a_f is a *floor term* of (a_n) if $a_n \geq a_f$ for all $n \geq f$.

So each floor term acts as a lower bound for the rest of the sequence that comes after it.

Ceiling terms are defined similarly.

A floor term is eventually a lower bound of the sequence. If there are infinitely many such terms, then these floor terms form a monotonically increasing subsequence as each floor term must be greater than or equal to the last.

If there are finitely many floor terms, then the sequence after the final floor term must be non-increasing, and must therefore contain a monotonically decreasing subsequence. This is because, if a_F is the last floor term, each term, a_A, after a_F must have some following term, a_B with $B > A$, which is less than a_A, or else a_A would be a new floor term. Then some term a_C must exist, and so on. These terms form the montonically decreasing subsequence.

If there are no floor terms, then the same argument above then applies similarly to ceiling terms. Or alternatively, the finite floor terms argument still applies, just with $a_{F+1} = a_0$. ■

[*] More generally, if a set of subsequences cover a sequence (their union is equal to the sequence), and every subsequence in the set converges to the same limit, then the sequence converges to that limit.

34.2.6 Sequences of Roots & Powers

Theorem (Bernoulli's Inequality)**.** *If $x > -1$ is real and $n \in \mathbb{N}$, then,*

$$(1+x)^n \geq 1 + nx$$

with equality if and only if $x = 0$, $n = 0$ or $n = 1$.

The inequality also holds:

- *for all real x for even natural n;*
- *for all real $x > -1$ for real $n \geq 1$;*
- *for all real $x \geq 1$ if n is any real number;*
- *in reverse for real $x \geq -1$ and real $0 \leq n \leq 1$.*

Proof. (first variant)

Let $P(n)$ be the statement that $(1+x)^n \geq 1 + nx$ for $x > -1$,$x \in \mathbb{R}$ and $n \in \mathbb{N}$. We induct (§5) on n.

$P(1)$ holds as $(1+x)^1 = 1 + x \geq 1 + 1x$.

Assume that $P(n)$ holds for some fixed arbitrary value of $n \geq 1$.

$$(1+x)^{n+1} = (1+x)^n(1+x)$$

$x > -1$, so $1 + x > 0$, so direction of inequality is preserved.

$$\begin{aligned}&\geq (1+nx)(1+x)\\&\geq 1 + x + nx + nx^2\\&\geq 1 + (1+n)x + nx^2\end{aligned}$$

$n \geq 0$ and $x^2 \geq 0$, so $nx^2 \geq 0$, so

$$\geq 1 + (1+n)x$$

Thus $P(n) \to P(n+1)$. As the base case has been shown to be true, and the inductive step has been established, the statement $P(n)$ holds for all natural numbers n by the principle of mathematical induction. ■

Bernoulli's inequality is useful for proving a variety of limits. For example, $\sqrt[n]{n} \to 1$ and $\sqrt[n]{x} \to 1$ for $x > 0$.

Lemma 34.2.6. *Let (a_n) be a sequence such that $a_n > 0$ for all n. Suppose $0 < l < 1$ and $\frac{a_{n+1}}{a_n} \leq l$ eventually. Then $(a_n) \to 0$.*

Proof. Suppose $0 < l < 1$ and $\frac{a_{n+1}}{a_n} \leq l$ for all n.

Let $P(n)$ be the statement that $a_n \leq l^n a_0$ for all n and some fixed $0 < l < 1$. We induct on n.

$P(0)$ holds as $a_0 \leq 1 \cdot a_0$.

Assume that $P(n)$ holds for some fixed arbitrary value of $n \geq 1$.

$\frac{a_{n+1}}{a_n} \leq l$ and $a_n > 0$ for all n, so,

$$\begin{aligned}a_{n+1} &\leq la_n\\a_{n+1} &\leq l(l^n a_0)\end{aligned}$$

$$a_{n+1} \leq l^{n+1}a_0$$

So $P(n) \to P(n+1)$. As the base case has been shown to be true, and the inductive step has been established, the statement $P(n)$ holds for all natural numbers n by the principle of mathematical induction.

As (l^n) is a null sequence, $a_0 l^n$ is also a null sequence. As $0 < a_n \leq l^n a_0$ for all n, (a_n) is a null sequence by the sandwich theorem.

Now, suppose $0 < l < 1$ and $\frac{a_{n+1}}{a_n} \leq l$ only eventually.

Then, there exists some $N \in \mathbb{N}$, such that for all $n > N$, $\frac{a_{n+1}}{a_n} \leq l$. So, $\frac{a_{N+n+1}}{a_{N+n}} \leq l$ for all n. It follows that $a_{N+n} \leq l^{N+n}a_0$, so $0 \leq a_{N+n} \leq l^{N+n}a_0$. $l^{N+n} \to 0$ as $n \to \infty$, so (a_{N+n}) is a null sequence by the sandwich theorem. By the shift rule, (a_n) is also a null sequence. ■

Corollary 34.2.6.1. *Let (a_n) be a sequence such that $a_n > 0$ for all n. If $\left(\frac{a_{n+1}}{a_n}\right) \to a$ and $0 \leq a < 1$ then $(a_n) \to 0$.*

Exercises. Find the limits of the following sequences:

- $\left(\frac{x^n}{n^k}\right)$ for $x > 0$ and $k \in \mathbb{N}$;
- $\left(\sqrt[n]{5n^2 + 7^n}\right)$;
- $\left(\frac{n^5 7^n + n^2 9^n}{3^{2n} + 2}\right)$;
- $\left(\frac{n!}{n^n}\right)$;
- $\left(\frac{((2n)!)^3}{(3n)!}\right)$;
- $\left(\frac{5n^5 + \sin^3(n)}{17n^3 + \cos^n(3n^2)}\right)$.

34.3 Completeness

First, we constructed the real numbers through set theory with Dedekind cuts (§4.5.4). We also constructed the real numbers from the rationals using the completeness axiom.

Here, we will explore this idea of completeness further. Previously, we said that the real numbers don't have "gaps", while the rational numbers do.

To formalise this, we first discuss *dense sets.*

34.3.1 Dense Sets

For this section, it is useful for us to have an example of an irrational number.

Lemma 34.3.1. *$\sqrt{2}$ is irrational.*

Proof. Suppose $\sqrt{2} = \frac{p}{q}$, for some integers p and q. Then,

$$2 = \frac{p^2}{q^2}$$
$$2q^2 = p^2$$

so p^2 is even. It follows that p is also even,* so $p = 2n$ for some integer n,

Then,

$$2q^2 = (2n)^2$$

* We can prove this either using case analysis, induction (§5.1.3), or by invoking Euclid's lemma (§10.1.5). Case analysis is the simplest way, but the latter two methods are included as exercises.

$$\begin{aligned} 2q^2 &= 4n^2 \\ q^2 &= 2n^2 \end{aligned}$$

so q^2, and similarly q, are also even. So, $q = 2m$ for some integer m. But then,

$$\frac{p}{q} = \frac{2n}{2m} = \frac{n}{m}$$

So, if $\sqrt{2}$ can be written as a rational, it could then always be written as a rational with smaller parts, which then could also be written with smaller parts, ad infinitum, contradicting the well-ordering principle (see §6.11.4), which states that there is a smallest natural.

It follows that our original assumption that $\sqrt{2}$ can be written as a fraction is false, so $\sqrt{2}$ is irrational. ■

This type of proof is called *proof by infinite descent*, because it shows that the statement being false would imply the existence of an infinitely descending chain of naturals.

This lemma comes in handy whenever we need to assert the existence of at least one irrational. Later in this chapter, we will also prove that e (Euler's number) is irrational.

Theorem 34.3.2. *Between any two distinct rational numbers there is another rational number.*

Proof. If a and b are rational, then so is $\frac{a+b}{2}$.* ■

This proof takes advantage of the fact that the rationals are closed under addition and division, so we can take the arithmetic mean of the two given rationals to get another rational between them. This isn't necessarily the simplest fraction between a and b, in the sense that there exists other fractions between a and b with smaller denominators (see footnote), but it is the fraction that sits halfway between them.

We now show that, despite the fact that irrational numbers exist, we can still always find a rational no matter how closely we zoom in on the number line.

Theorem 34.3.3. *Between any two distinct real numbers there is a rational number.*

Proof. Let $x,y \in \mathbb{R}$ be distinct real numbers. Without loss of generality, suppose $x < y$ so $y - x > 0$. Let $y - x = z$.

By the Archimedean property (Theorem 11.4.1), there exists a natural n such that $nz > 1$, so $ny - nx > 1$. This implies that the interval (nx,ny) has length greater than 1, so there is at least one integer point within that interval. It follows that there exists an integer, m, such that $nx < m < ny$, so $x < \frac{m}{n} < y$, giving a rational. ■

* If $\frac{a}{c}$ and $\frac{b}{d}$ are rationals, then,

$$\frac{a}{c} < \frac{a+b}{c+d} < \frac{b}{d}$$

This middle fraction is called the *mediant*, or sometimes the *freshman sum*, due to being a common mistake in the early stages of learning fractional addition.

Additionally, if the fractions above also satisfy the *determinant* relation $bc - ad = 1$, the mediant is also the *simplest* (smallest denominator) fraction in the interval $\left(\frac{a}{c}, \frac{b}{d}\right)$. Furthermore, this relationship is actually biconditional.

The similarity of $bc - ad$ and determinants in linear algebra is also not a coincidence. Given our set-theoretic definition of rationals, you can view the mediant as a cross product of the two given rational numbers.

Exercises.

- Prove the mediant inequality.
- Prove that $bc - ad = 1$ holds if and only if $\frac{a+b}{c+d}$ is the simplest fraction in $\left(\frac{a}{c}, \frac{b}{d}\right)$ (assuming all fractions are reduced).
- Explore the connection between the cross product and the mediant. Try plotting the two parts of the rational numbers as coordinates.

Corollary 34.3.3.1. *There are (countably) infinitely many rationals in any open interval.*

Proof. Let (a,b) be a non-empty interval. Suppose there are only n rationals in (a,b). Because $<$ is a strict total ordering on the reals, we have,

$$a < r_1 < r_2 < \ldots < r_n < b$$

Applying the above theorem to a and r_1, we know there exists a rational in (a,r_1), which is a subinterval of (a,b). This contradicts our assumption that there are n rationals in (a,b). It follows that there are infinitely many rationals in the interval. ∎

Because the rationals can be arbitrarily close to real numbers, we say that the rational numbers are a *dense subset* of the real numbers.

A subset S of a set X is *dense* in X if every member of X is in S, or is otherwise arbitrarily close* to a member of S.

In this case, any real number can be approximated arbitrarily well by rational numbers, so the rationals are dense in the reals.

Theorem 34.3.4. *Between any two distinct real numbers there is an irrational number.*

Proof. Let $a,b \in \mathbb{R}$, and without loss of generality suppose $a < b$. Then, let $x = \frac{a}{\sqrt{2}}$ and $y = \frac{b}{\sqrt{2}}$. By the density of the rationals, there exists a rational, r, such that $x < r < y$, so

$$\frac{a}{\sqrt{2}} < r < \frac{b}{\sqrt{2}}$$

so $a < \sqrt{2}r < b$, and we have an irrational between a and b. ∎

Corollary 34.3.4.1. *There are (uncountably) infinitely many irrationals in any open interval.*

Proof. Exercise. ∎

This corollary shows that the irrationals are also dense in the reals, but the irrationals, being uncountably infinite in cardinality, are actually far more numerous than the rationals.

34.3.2 Suprema & Infima

We recall the definition of bounds for sets:

Let S be a non-empty set of real numbers. A real number x is an *upper bound* of S if $x \geq s$ for all $s \in S$. A real number y is the *least upper bound* or *supremum* of S if y is an upper bound, and $y \leq x$ for all upper bounds x of S.

We define the *lower bound* and *greatest lower bound* or *infimum* similarly.

We can say *the* least upper bound, because the supremum is unique.

* We haven't really defined what "close" means. X in this definition really has to be a *topological space*, which is a set equipped with some notion of closeness in the form of a *topology*. As a special case, we are working with *metric spaces* here.

For the real numbers, we can use the regular Euclidean distance metric as our measure of closeness, so a and b have distance $|a-b|$ between each other. In $\mathbb{R}^2$, we can use the Euclidean norm, given by $\|a-b\| = \sqrt{(a_1-b_1)^2+(a_2-b_2)^2}$, where $a=(a_1,a_2)$ and $b=(b_1,b_2)$. These are examples of *metrics* or metric functions, and $\mathbb{R}$ and $\mathbb{R}^2$ are *metric spaces.* Metric spaces are a specific type of topological space.

However, in general topological spaces, there may be no concrete formula for measuring distances. The definition of a dense set in a general topological space is given in §37.4.3.

Theorem (Uniqueness of supremum)**.** *Let $(R, \preceq)$ be an ordered set, and let $S \subseteq R$ be a non-empty subset. Then, S has at most one supremum in R.*

Proof. Let s and s' be suprema of S in R. By definition, s and s' are both upper bounds of S in R. So, s is an upper bound and s' is a supremum, so $s' \preceq s$. Similarly, s' is an upper bound and s is a supremum, so $s \preceq s'$. Becaue $s' \preceq s$ and $s' \preceq s$ both hold, $s = s'$ by the antisymmetry (§4.4.7) of $\preceq$. ■

We can also relate the supremum and infimum of a set together. But first, a lemma.

Lemma 34.3.5. *Suppose $A \subseteq \mathbb{R}$ is non-empty and bounded above. For any $\varepsilon > 0$, there exists $a \in A$ such that $\sup A - \varepsilon < a \leq \sup A$.*

Proof. As $\varepsilon > 0$, $\sup(A) - \varepsilon < \sup(A)$, so $\sup(A) - \varepsilon$ is not an upper bound of A. Let $0 < n < \varepsilon$ such that $\sup(A) - \varepsilon < \sup(A) - n < \sup(A)$. As $\sup(A) - n < \sup(A)$, $\sup(A) - n = a$ for some $a \in A$, so $\sup(A) - \varepsilon < a < \sup(A)$. ■

Theorem 34.3.6. *For any non-empty set $A \subseteq \mathbb{R}$, $\inf(A) = -\sup(-A)$, where $-A = \{-a : a \in A\}$.*

Proof. Suppose $A \subseteq \mathbb{R}$ is non-empty and bounded below, so there is at least one element, a, in A. $-A$ therefore contains at least $-a$, and is also non-empty.

For all $a \in A$, $a \geq \inf(A)$ by definition, so $-a \leq -\inf(A)$. $-a$ is the general element of the set $(-A)$, so all elements of $(-A)$ are less than or equal to $-\inf(A)$. It follows that $-\inf(A)$ is the supremum of $(-A)$, so $\inf(A) = -\sup(-A)$. ■

We can use this to show that the least upper bound property implies the greatest lower bounds property, and vice versa.

Theorem 34.3.7. *The least upper bounds property and the greatest lower bounds property are equivalent.*

Proof. Suppose $A \subseteq \mathbb{R}$ is non-empty and bounded below. Then, $-A$ is also a non-empty set of real numbers, but is bounded above. The least upper bound property says that $-A$ has a least upper bound, $\sup(-A)$. But, from the theorem above, $A = -(-A)$ has a greatest lower bound, $\inf A = -\sup(-A)$, completing the forward implication.

The reverse implication is left as an exercise to the reader. ■

34.3.2.1 Bounded Monotonic Sequences

Many results in analysis are based on completeness. Earlier, we proved that convergent sequences are bounded. With completeness, we can go the other way, and prove that certain bounded sequences are convergent:

Theorem (Monotonic Convergence Theorem)**.** *Every bounded monotonic sequence is convergent. In particular, increasing sequences converge to their suprema, and decreasing sequences converge to their infima.*

Proof. Suppose (a_n) is an increasing sequence, and let $\varepsilon > 0$. By Lemma 34.3.5, there exists some a_N for which $\sup(A) - \varepsilon < a_N < \sup(A)$, and as a_n is increasing, the inequality holds for all $n > N$.

For all $n > N$, $\sup(A) - \varepsilon < a_n < \sup A < \sup A + \varepsilon$, so $-\varepsilon < a_n - \sup(A) < \varepsilon$. By the interval property, we have $|a_n - \sup(A)| < \varepsilon$, which is the definition of convergence, so $(a_n) \to \sup(A)$ as $n \to \infty$.

Now suppose a_n is a decreasing sequence. Then $(-a_n)$, is an increasing sequence, so $(-a_n)$ converges to $\sup(-A)$, which, by Theorem Theorem 34.3.6, is equal to $-\inf(A)$, so $(-a_n) \to -\inf(A)$, and by the sum rule, $a_n \to \inf A$. ■

Corollary 34.3.7.1. *Every monotonic sequence is either bounded above or bounded below. In particular, every increasing sequence which is bounded above is bounded, and every decreasing sequence which is bounded below is similarly bounded.*

Proof. Exercise. ■

The above theorems on convergence of bounded monotonic sequences gives us a method to show that monotonic sequences converge, even if we do not necessarily know what the limit is.

We also sometimes say that a bounded monotonic sequence converges "by completeness", as an abbreviation of "by the monotonic convergence theorem".'

34.3.3 General Bounded Sequences

We've previously shown that every subsequence of a bounded sequence is bounded. We have also shown that every sequence has a monotonic subsequence.

Combining these two theorems, we get,

Theorem (Bolzano-Weierstrass Theorem)**.** *Every bounded sequence has a convergent subsequence.*

Proof. By the monotonic subsequence theorem, every bounded sequence has a monotonic subsequence. The bounds of the original sequence are also bounds for the subsequence. This subsequence is then both monotonic and bounded, thus satisfying the conditions for the monotonic convergence theorem, so the subsequence converges. ■

We can also prove the Bolzano-Weierstrass theorem using a technique called *lion hunting.* The name refers to a method of trapping a lion in a square jungle.

First, build a fence dividing the jungle in half, and listen to tell which half the lion is in. Then, divide that half into half again, and repeat. The lion will quickly be trapped in a manageable area, and eventually, into as small as an area as is desired.

Proof. We similarly find a limit point for a sequence on the real line. Suppose a sequence (x_n) is bounded below by a_1 and bounded above by b_1. Divide the interval $[a_1,b_1]$ into two subintervals, $\left[a_1,\frac{a_1+b_1}{2}\right]$ and $\left[\frac{a_1+b_1}{2},b_1\right]$. At least one of the two intervals must contain infinitely many terms of (x_n). We select this half (or select randomly if both halves contain infinitely many terms) and label it as the interval $[a_2,b_2]$. Then, we split this new interval in half, and find the half which contains infinitely many terms of (x_n) and label it $[a_3,b_3]$. Continuing, at the kth step, we start with an interval $[a_k,b_k]$ containing infinitely many terms of (x_n), and subdivide it into two intervals, $\left[a_k,\frac{a_k+b_k}{2}\right]$ and $\left[\frac{a_k+b_k}{2},b_k\right]$, one of which will containing infinitely many terms of (x_n), and we label as $[a_{k+1},b_{k+1}]$.

$a_k < b_k$, so $\frac{a_k+b_k}{2} > \frac{a_k+a_k}{2} = a_k$. a_{k+1} is either equal to a_k, or $\frac{a_k+b_k}{2}$, so in both cases, $a_{k+1} \geq a_k$, so (a_n) is increasing. $a_1 < a_k < b_1$, so (a_n) is also bounded. By the monotonic convergence theorem, (a_n) converges to some limit, l_1.

$\frac{a_k+b_k}{2} < \frac{b_k+b_k}{2} = b_k$, so (b_n) is decreasing through similar reasoning, and $a_1 < b_k < b_1$, so (b_n) is also bounded and similarly convergent to some limit, l_2.

$b_{k+1} - a_{k+1} = b_k - \frac{a_k+b_k}{2} = \frac{b_k-a_k}{2}$, or $b_{k+1} - a_{k+1} = \frac{a_k+b_k}{2} - a_k = \frac{b_k-a_k}{2}$. In both cases, $b_{k+1} - a_{k+1} = \frac{b_k-a_k}{2}$. Because a_1 and b_1 exist and are finite, $b_1 - a_1$ is also finite, so as $k \to \infty$, $b_k - a_k \to 0$, so $l_1 - l_2 \to 0$ and $l_1 = l_2$, so (a_n) and (b_n) converge to the same limit, L.

As (a_n) is increasing and (b_n) is decreasing, $[a_{k+1},b_{k+1}] \in [a_k,b_k]$. Every interval $[a_k,b_k]$ is defined to have infinitely many terms of (x_n), so it is always possible to pick an n_i such that x_{n_i} lies within $[a_k,b_k]$, so $a_k \leq x_{n_i} \leq b_k$. As $(a_n) \to L$ and $(b_n) \to L$, $(x_{n_i}) \to L$ by the sandwich theorem, giving a convergent subsequence of (x_n). ■

34.3.4 Cauchy Sequences

Previously, we have proven that convergent sequences are bounded, and also that bounded monotonic sequences are convergent. However, we can further prove that monotonic sequences converge *if and only if* they are bounded.

Theorem (Convergence Test)**.** *A monotonic sequence converges if and only if it is bounded.*

Proof. Exercise. (This is an easy consequence of the previous two results.) ■

Now, what other conditions are sufficient to prove convergence?

Proposition (Cleverclog's Test)**.** A sequence (a_n) converges if and only if $(a_{n+1} - a_n) \to 0$.

This seems reasonable – a sequence converges if and only if the gaps between the terms go to zero. Unfortunately, it's completely wrong.

Example. Let $(a_n) = \ln n$. Then,

$$\begin{aligned} a_{n+1} - a_n &= \ln(n+1) - \ln n \\ &= \ln\left(\frac{n}{n+1}\right) \end{aligned}$$

As $n \to \infty$, $\frac{n}{n+1} \to 1$, so $\ln\left(\frac{n}{n+1}\right) \to \ln 1 = 0$, thus satisfying Cleverclog's test criterion. However, $\ln n \to \infty$ as $n \to \infty$. △

The gaps between each consecutive terms going to zero doesn't mean that the sequence always converges. The problem is, the gaps can tend to zero slower than the sequence tends towards any finite point. We need a stronger condition.

Definition 34.3.1. A sequence (a_n) has the *Cauchy property* if for each $\varepsilon > 0$, there exists $n \in \mathbb{N}$ such that $|a_n - a_m| < \varepsilon$ for all $n,m > N$.

If a sequence has the Cauchy property, we say it is a *Cauchy sequence.* The Cauchy criterion is stronger than Cleverclog's, because Cauchy requires that *every* term past a certain point is no more than ε away from each other, while Cleverclog only requires consecutive terms to become arbitrarily close.

This time, the Cauchy criterion is sufficient for convergence:

Lemma 34.3.8. *Every Cauchy sequence is convergent.*

Proof. Suppose (a_n) is Cauchy.

Let $\varepsilon = 1$. As (a_n) is Cauchy, there exists $N \in \mathbb{N}$ such that $\forall n,m > N$, $|a_n - a_m| < 1$, so, by the interval property, $-1 < a_n - a_m < 1$ and $a_m - 1 < a_n < a_m + 1$, so (a_n) is eventually bounded for some fixed $m > N$ and all $n > N$. As (a_n) is eventually bounded, it is bounded.

By the Bolzano-Weierstrass theorem, (a_n) contains a subsequence, (a_{n_i}), that converges to a limit, a. Let $\varepsilon > 0$, and let $\varepsilon_1,\varepsilon_2 < \frac{\varepsilon}{2}$.

As (a_n) is Cauchy, there exists $N_1 \in \mathbb{N}$ such that $|a_n - a_{n_i}| < \varepsilon_1$ for all $n,n_i > N_1$. As a_{n_i} converges to a, there exists $N_2 \in \mathbb{N}$ such that $|a_{n_i} - a| < \varepsilon_2$ for all $n_i > N_2$.

Let $N = \max(N_1,N_2)$. Then, $|a_n - a| < |a_n - a_{n_i}| + |a_{n_i} - a|$, so $|a_n - a| < \varepsilon_1 + \varepsilon_2$, and $|a_n - a| < \varepsilon$ for all $n > N$, so a_n converges to a. ■

It turns out, however, that Cauchy is not only sufficient for convergence, it is necessary:

Lemma 34.3.9. *Every convergent sequence is Cauchy.*

Proof. Suppose $(a_n) \to a$. Let $\varepsilon > 0$, and let $\varepsilon_1,\varepsilon_2 < \frac{\varepsilon}{2}$.

$(a_n) \to a$, so $(a_m) \to a$ by shift rule. By the definition of convergence, there exists $N_1 \in \mathbb{N}$ such that $|a_n - a| < \varepsilon_1$ for all $n > N_1$. Similarly, $N_2 \in \mathbb{N}$ exists, such that $|a_m - a| < \varepsilon_2$ for all $m > N_2$. Let $N = \max(N_1,N_2)$. Then, for all $m,n > N$,

$$\begin{aligned}
|a_n - a_m| &= |a_n - a_m + a - a| \\
&= |(a_n - a) + (a - a_m)| \\
&\leq |a_n - a| + |a - a_m| \\
&\leq |a_n - a| + |-(a_m - a)| \\
&\leq |a_n - a| + |a_m - a| \\
&\leq \varepsilon_1 + \varepsilon_2 \\
&< \varepsilon
\end{aligned}$$

So (a_n) is Cauchy. ■

So, combining the two previous results, we have shown that being convergent and having the Cauchy property are equivalent. This gives us a test for convergence:

Theorem (Convergence Test)**.** *A sequence is convergent if and only if it has the Cauchy property.*

This test is incredibly powerful, because it applies to all sequences, not just monotonic ones, and also doesn't depend on us knowing what the limit is.

However, proving that a sequence is Cauchy can be somewhat difficult, so there is an easier test for some cases.

Definition 34.3.2. A sequence is *strictly contracting* if, for some number $0 < l < 1$, called the *contraction factor*,

$$|a_{n+1} - a_n| \leq l|a_n - a_{n-1}|$$

holds for all integer $n \geq 1$.

Lemma 34.3.10. *If (a_n) is a strictly contracting sequence, then $|a_{n+1} - a_n| \leq |a_2 - a_1|l^{n-1}$*

Proof. Let $P(n)$ be the statement that $|a_{n+1} - a_n| \leq |a_2 - a_1|l^{n-1}$, where $n \in \mathbb{N}$. We induct on n.

$P(1)$ holds as $|a_2 - a_1| \leq |a_2 - a_1|l^0$ holds. Now, assume that $P(n)$ holds for some fixed arbitrary value of $n \geq 1$.

$$\begin{aligned}
|a_{n+2} - a_{n+1}| &\leq |a_{n+1} - a_n|l \\
&\leq \left(|a_2 - a_1|l^{n-1}\right) l \\
&\leq |a_2 - a_1|l^n
\end{aligned}$$

Thus, $P(n)$ implies $P(n+1)$. As the base case has been shown to be true, and the inductive step has been established, the statement $P(n)$ holds for all natural numbers n by the principle of mathematical induction. ■

Theorem 34.3.11. *A strictly contracting sequence is Cauchy, and is hence convergent.*

Proof. We begin by applying the triangle inequality repeatedly:

$$|a_n - a_m| \leq |a_n - a_{n-1}| + |a_{n-1} - a_{n-2}| + \cdots + |a_{m+1} - a_m|$$

And now apply the lemma to each term.

$$\begin{aligned}
&\leq |a_2 - a_1|(l^{n-2} + l^{n-3} + l^{n-4} + \cdots + l^{m-1}) \\
&\leq |a_2 - a_1| \sum_{k=m-1}^{n-2} l^k \\
&\leq |a_2 - a_1| \left(\sum_{k=0}^{n-2} l^k - \sum_{k=0}^{m-2} l^k \right) \\
&\leq |a_2 - a_1| \left(\frac{1 - l^{n-1}}{1-l} - \frac{1 - l^{m-1}}{1-l} \right) \\
&\leq |a_2 - a_1| \left(\frac{l^{m-1} - l^{n-1}}{1-l} \right) \\
&\leq |a_2 - a_1| l^{m-1} \left(\frac{1 - l^{n-m}}{1-l} \right)
\end{aligned}$$

$n > m$ and $0 < l < 1$, so $n - m > 0$ and $0 < l^{n-m} < 1$.

$$\leq |a_2 - a_1| l^{m-1} \left(\frac{1}{1-l} \right)$$

As $m \to \infty$, $l^{m-1} \to 0$. $|a_2 - a_1|$ is a constant, as is $\frac{1}{1-l}$, so by product rule, $|a_2 - a_1| l^m \left(\frac{1}{1-l} \right) \to 0$, so there exists an $N \in \mathbb{N}$ such that for all $m > N$, $|a_2 - a_1| l^m \left(\frac{1}{1-l} \right) < \varepsilon$ for any choice of $\varepsilon > 0$, so, for all $n > N$,

$$|a_n - a_m| < \varepsilon$$

Hence, (a_n) is Cauchy. ■

34.3.5 Decimal Sequences

The easiest decimal representations to deal with are finite or *terminating* decimals – ones which only have a finite number of non-zero decimal places, followed by an infinite string of zeros. A positive finite decimal has the form $d_0.d_1d_2 \ldots d_n$, where d_0 is a non-negative integer, and each of d_1,d_2,…,d_n is an integer in [0,9]. Then, $d_0.d_1d_2 \ldots d_n$ is defined to be the number,

$$\sum_{i=0}^{n} \frac{d_i}{10^i} = d_0 + \frac{d_1}{10} + \frac{d_2}{10^2} + \cdots + \frac{d_n}{10^n}$$

Similarly, negative finite decimals have the form $-d_0.d_1d_2 \ldots d_n$, where d_0 is a non-negative integer, and each of d_1,d_2,…,d_n is an integer in [0,9]. Then, $d_0.d_1d_2 \ldots d_n$ is defined to be the number,

$$\sum_{i=0}^{n} -\frac{d_i}{10^i} = -d_0 - \frac{d_1}{10} - \frac{d_2}{10^2} - \cdots - \frac{d_n}{10^n}$$

Infinite or *non-terminating* decimal representations are defined in terms of sequences. A positive real number x has a representation as an infinite decimal if there is a non-negative integer d_0 and a sequence (d_n) with $d_n \in [0,9]$ for all n, such that the sequence (a_n) given by

$$a_n := \sum_{i=0}^{n} \frac{d_i}{10^i} = d_0 + \frac{d_1}{10} + \frac{d_2}{10^2} + \cdots + \frac{d_n}{10^n}$$

converges to x. If this is the case, then we write

$$x = d_0.d_1d_2,d_3\dots$$

Similarly, a negative number requires the sequence

$$a_n = \sum_{i=0}^{n} -\frac{d_i}{10^i} = -d_0 - \frac{d_1}{10} - \frac{d_2}{10^2} - \cdots - \frac{d_n}{10^n}$$

to converge to x, and we write,

$$x = -d_0.d_1d_2,d_3\dots$$

We can easily generate the sequence (d_n) by considering recursively considering sets of numbers:

$$\begin{aligned}
d_1 &\coloneqq \max\left\{i : d_0 + \frac{i}{10} \le x\right\} \\
d_2 &\coloneqq \max\left\{i : d_0 + \frac{d_1}{10} + \frac{i}{10^2} \le x\right\} \\
d_3 &\coloneqq \max\left\{i : d_0 + \frac{d_1}{10} + \frac{d_2}{10^2} + \frac{i}{10^3} \le x\right\} \\
&\vdots \\
d_n &\coloneqq \max\left\{j : \left(\sum_{j=0}^{n-1} \frac{d_j}{10^j}\right) + \frac{i}{10^n} \le x\right\}
\end{aligned}$$

It is easy to verify that each digit is in the required interval. Additionally, after n digits, we have $x - \frac{1}{10^n} < \sum_{i=0}^{n} \frac{d_i}{10^i} \le x$, so the sequence converges to x by the sandwich theorem.

For negative real numbers, we similarly use

$$d_n \coloneqq \max\left\{j : \left(\sum_{j=0}^{n-1} -\frac{d_j}{10^j}\right) - \frac{i}{10^n} \ge x\right\}$$

Theorem (Infinite Decimal Sequences)**.** *Every infinite decimal* $\pm d_0.d_1d_2d_4\dots$ *represents a real number.*

Proof. $b_n = d_0 + \sum_{k=1}^{n}\left(d_k 10^{-k}\right)$. $0 \le d_k \le 9$, so $d_0 + \sum_{k=1}^{n}\left(0 \cdot 10^{-k}\right) \le b_n \le d_0 + \sum_{k=1}^{n}\left(9 \cdot 10^{-k}\right)$, so b_n is bounded within $[d_0, d_0 + 0.\bar{9}]$.

$$\begin{aligned}
b_{n+1} - b_n &= \sum_{k=0}^{n+1}\left(d_k 10^{-k}\right) - \sum_{k=0}^{n}\left(d_k 10^{-k}\right) \\
&= d_{n+1}10^{-(n+1)} \\
&\ge 0
\end{aligned}$$

So b_n is increasing. As b_n is bounded and monotonic, it is convergent by the monotonic convergence theorem. ■

Theorem 34.3.12. $0.99\bar{9} = 1$.

Proof.

$$\begin{aligned}
0.\bar{9} &:= \lim_{n\to\infty}\sum_{i=0}^{n}\frac{9}{10^i}\\
&= \sum_{i=0}^{k}\frac{9}{10^i} + \lim_{n\to\infty}\sum_{i=k+1}^{n}\frac{9}{10^i}\\
&= \sum_{i=0}^{k}\frac{9}{10^i} + \frac{9}{10^{k+1}}\lim_{n\to\infty}\frac{1-(\frac{1}{10^{n-k}})}{1-\frac{1}{10}}\\
&= \sum_{i=0}^{k}\frac{9}{10^i} + \frac{1}{10^k}\\
&= 1
\end{aligned}$$

■

This last example shows one of the annoying features of decimals – there can exist two different decimal representations of same real number. In particular:

Theorem 34.3.13. *If a positive real number has two different representations as an infinite decimal, then one of these representations terminates (or equivalently, ends with an infinite recurring string of zeros), while the other ends with a infinite recurring string of nines.*

Proof. Suppose a real number, x, has two decimal representations, $a_0.a_1a_2a_3\dots$ and $b_0.b_1b_2b_3,\dots$. Suppose that the decimal representations agree until the kth place, where $a_k < b_k$. Then,

$$\begin{aligned}
x &= \lim_{n\to\infty}\sum_{i=0}^{n}\frac{a_i}{10^i}\\
&= \sum_{i=0}^{k}\frac{a_i}{10^i} + \lim_{n\to\infty}\sum_{i=k+1}^{n}\frac{a_i}{10^i}\\
&\leq \sum_{i=0}^{k}\frac{a_i}{10^i} + \lim_{n\to\infty}\sum_{i=k+1}^{n}\frac{9}{10^i}\\
&= \sum_{i=0}^{k}\frac{a_i}{10^i} + \frac{9}{10^{k+1}}\lim_{n\to\infty}\frac{1-\left(\frac{1}{10^{n-k}}\right)}{1-\frac{1}{10}}\\
&= \sum_{i=0}^{k}\frac{a_i}{10^i} + \frac{9}{10^{k+1}}\cdot\frac{1}{1-\frac{1}{10}}\\
&= \sum_{i=0}^{k}\frac{a_i}{10^i} + \frac{9}{10^{k+1}}\cdot\frac{10}{9}\\
&= \sum_{i=0}^{k}\frac{a_i}{10^i} + \frac{1}{10^k}\\
&= \sum_{i=0}^{k-1}\frac{a_i}{10^i} + \frac{a_k}{10^k} + \frac{1}{10^k}\\
&= \sum_{i=0}^{k-1}\frac{a_i}{10^i} + \frac{a_k+1}{10^k}
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{i=0}^{k} \frac{b_i}{10^i} \\
&\leq \lim_{n\to\infty} \sum_{i=0}^{n} \frac{b_i}{10^i} \\
&= x
\end{aligned}
$$

Since we started with x, and, through a series of inequalities, ended up with x, the inequalities must all really be equalities. In particular, if any of the nines in the infinite tail is replaced by any other digit, the third line then only holds strictly, breaking the equality, so one of the decimal representations will always end with an infinite recurring string of nines. ■

We can now more formally categorise decimals into three types:

Definition 34.3.3. An infinite decimal $\pm d_0.d_1d_2d_3\ldots$ is:

- *terminating* if $\exists N : \forall n > N : d_n = 0$ – the decimal ends in an infinite string of zeros;
- *recurring* if $\exists N, r : \forall n > N : d_n = d_{n+r}$ – the decimal repeats;
- *non-recurring* if it is neither terminating nor recurring.

Because we generally don't write out the infinite tail of zeros at the end of terminating decimals, finite and terminating decimals are really the same thing.

Theorem (Characterisation of Terminating Decimals)**.** *A real number x can be represented by a terminating decimal if and only if $x = \frac{p}{q}$, where p and q are integers and the only prime factors of q are* 2 *and* 5.

Proof. If x has a terminating decimal expansion, multiplying by sufficiently large powers of 10 produces an integer. So, if the decimal representation of x terminates after n digits, $10^n x = k$, where k is an integer. It follows that $x = \frac{k}{10^n} = \frac{k}{2^n 5^n}$, proving the forward direction.

Now, suppose $x = \frac{p}{q}$, and that the only prime factors of q are 2 and 5. Then,

$$
\begin{aligned}
x &= \frac{p}{q} \\
x &= \frac{p}{2^n 5^m}
\end{aligned}
$$

Let $k = \max(n,m)$.

$$
\begin{aligned}
x &= \frac{p}{2^k 5^k 2^{n-k} 5^{m-k}} \\
x &= \frac{p}{10^k 2^{n-k} 5^{m-k}} \\
x &= \frac{p 2^{k-n} 5^{k-m}}{10^k}
\end{aligned}
$$

$k \geq n,m$, so $k - n$ and $k - m$ are positive integers, so 2^{k-n} and 5^{k-m} are also integers. So, $p2^{k-n}5^{k-m}$ equals some integer, p'.

$$
x = \frac{p'}{10^k}
$$

which terminates after at most k digits, completing the backward direction. ■

Theorem (Characterisation of Recurring Decimals)**.** *Every recurring decimal represents a rational number.*

Proof. If x has a decimal representation that has recurring blocks of length k after the nth digit, then $d_m = d_{m+k}$ for all $m > n$, so $10^k x$ is still repeating after the nth digit. As $10^k x$ and x have the same decimal representation after n digits, $10^k x - x$ must have a terminating decimal representation. By the previous theorem, $10^k x - x = \frac{p}{q}$, where the only prime factors of q are 2 and 5. Then,

$$\begin{aligned} 10^k x - x &= \frac{p}{q} \\ x(10^k - 1) &= \frac{p}{q} \\ x &= \frac{p}{q(10^k - 1)} \end{aligned}$$

■

Theorem 34.3.14. *Every rational number can be represented by a recurring decimal, or a terminating decimal.*

Proof. When dividing an integer p by another integer q, there are only q possible remainders $(0,1,...,q-1)$. If the remainder is ever 0, the decimal expansion terminates. If the remainder is never 0, then there are only $q - 1$ other values possible. After at most q iterations of division, at least one value is repeated by the pigeonhole principle, at which point the remainder values will begin to repeat, leading to a recurring decimal expansion. As there are only $q - 1$ values possible for the remainder to take, the length of the repeating block is at most $q - 1$. ■

Theorem (Classification of Decimal Representations)**.**

- *Every real number has a decimal representation, and every decimal represents a real number.*
- *The rationals are the set of terminating or recurring decimals.*
- *The irrationals are the set of non-recurring decimals.*
- *If a number has two distinct decimal representations, then one terminates and the other ends with a recurring string of nines.*

34.3.6 Axioms Equivalent to Completeness

We have proved many results that are consequences of the axiom of completeness. It turns out that many of these are not only consequences of completeness, but are logically equivalent to the axiom of completeness.

We have been saying that a sequence converges “by completeness”, as an abbreviation of “by the monotonic convergence theorem”. This is justified because the monotonic convergence theorem is equivalent to the axiom of completeness; in fact, each half of the monotonic convergence theorem is individually equivalent to the axiom of completeness.

A set R is complete if any of the following hold:

1. Least upper bound property: for every subset $S \subseteq R$, if S is non-empty and has an upper bound, then S has a supremum in R.
2. Greatest lower bound property: for every subset $S \subseteq R$, if S is non-empty and has a lower bound, then S has an infimum in R.
3. Monotonic convergence theorem: every bounded above increasing sequence of elements of R is convergent.

4. Monotonic convergence theorem: every bounded below decreasing sequence of elements of R is convergent.
5. Bolzano-Weierstrass theorem: every bounded sequence of elements of R has a convergent subsequence.
6. Cauchy criterion: every Cauchy sequence of elements of R is convergent.
7. Infinite decimal sequences: Every infinite decimal sequence is convergent.

"Equivalence" here means that we can prove any of these results assuming only one of them, so each individual result as above can be used as an alternative formulation of the completeness axiom. Note, we have not yet shown these equivalences, mostly only having proved the forward direction from the least upper bound property.

34.4 Series

One useful type of sequence is a *series* – a sequence of sums. We already saw some of these in the previous section when we explored decimals, but in this section, we will develop some more theory around these series.

Definition 34.4.1. Let (a_n) be a sequence. The series

$$\sum_{n=1}^{\infty} a_i$$

has *partial sums* (s_n) given by

$$s_n = \sum_{i=0}^{n} a_n$$

That is, the nth partial sum is the sum of the first n terms of (a_n).

We say that the series

- *converges* if (s_n) converges, and if $(s_n) \to S$, we say that S is the *sum* or *limit* of the series;
- *diverges* if (s_n) fails to converge;
- *diverges to infinity* if (s_n) diverges to infinity;
- *diverges to minus infinity* if (s_n) diverges to minus infinity.

Note that there are two sequences associated with each series: the sequence being summed, (a_n), and the sequence of partial sums, (s_n). Also note that n is a bound variable in the first sum above, while n is free in the second.

We sometimes omit the bounds of the summation when the series is clear.

34.4.1 Properties of Convergent Series

Theorem (Sum Rule for Series)**.** *Suppose $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$ are convergent series. Then, for all $c,d \in \mathbb{R}$, $\sum_{n=1}^{\infty}(ca_n + db_n)$ is a convergent series, and,*

$$\sum_{n=1}^{\infty}(ca_n + db_n) = c\sum_{n=1}^{\infty} a_n + d\sum_{n=1}^{\infty} b_n$$

Proof.

$$\begin{aligned}\sum_{n=1}^{n}(ca_n + db_n) &= c\sum_{n=1}^{n} a_n + d\sum_{n=1}^{n} b_n \\ &\to c\sum_{n=1}^{\infty} a_n + d\sum_{n=1}^{\infty} b_n\end{aligned}$$

■

Theorem (Shift Rule for Series)**.** *Let $N \in \mathbb{N}$. Then, the series $\sum_{n=1}^{\infty} a_n$ converges if and only if $\sum_{n=1}^{\infty} a_{N+n}$ converges.*

Proof.

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{N} a_n + \sum_{n=1}^{\infty} a_{n+N}$$

$\sum_{n=1}^{N} a_n$ is a finite sum, so, if $\sum_{n=1}^{\infty} a_{n+N}$ converges to a finite value, we have $\sum_{n=1}^{\infty} a_n$ equal to the sum of two finite numbers, which is another finite number, so $\sum_{n=1}^{\infty} a_n$ converges if $\sum_{n=1}^{\infty} a_{n+N}$ converges.

Similarly,

$$\sum_{n=1}^{\infty} a_{n+N} = \sum_{n=1}^{\infty} a_n + \left(-\sum_{n=1}^{N} a_n\right)$$

$\sum_{n=1}^{N} a_n$ is a finite sum, so, if $\sum_{n=1}^{\infty} a_n$ converges to a finite value, we have $\sum_{n=1}^{\infty} a_{n+N}$ equal to the sum of two finite numbers, which is another finite number, so $\sum_{n=1}^{\infty} a_{n+N}$ converges if $\sum_{n=1}^{N} a_n$ converges.

As the implication has been shown in both directions, $\sum_{n=1}^{\infty} a_n$ converges if and only if $\sum_{n=1}^{\infty} a_{n+N}$ converges. ■

Usually, it is extremely difficult to find an explicit formula for the sum of a series. For instance, consider series of the form,

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

So, $\zeta(1)$ would be the harmonic series, for example.

It is known that $\zeta(n)$ converges for all positive integer $n \geq 2$, but there is no known explicit formula for the value of ζ evaluated at any odd integer.

The problem becomes even more difficult with series that can also contain negative terms.

So, we generally settle for proving whether a sequence converges or not, and we have a variety of tests that we can try. For now, we will mainly focus on series with only non-negative terms.

34.4.2 Boundedness Condition

Intuitively, a sequence should only converge if the sequence of its partial terms is bounded. However, if the sequence being summed consists of only non-negative terms, then this is not only necessary, but sufficient to prove convergence.

Theorem (Boundedness Theorem)**.** *Suppose $a_n \geq 0$ for all n. Then, $\sum_{n=1}^{\infty} a_n$ converges if and only if the sequence of partial sums $(s_n) = (\sum_{i=1}^{n} a_i)$ is bounded.*

Proof. $a_n \geq 0$, so (s_n) is increasing. If (s_n) is bounded, it is convergent by the monotonic convergence theorem, so $\sum_{n=1}^{\infty} a_n$ converges.

$\sum_{n=1}^{\infty} a_n = \lim_{k\to\infty} \sum_{n=1}^{k} a_n = \lim_{k\to\infty} s_k$. Suppose that s_k is not bounded. As (s_k) is increasing and unbounded, $(s_k) \to \infty$, so $\sum_{n=1}^{\infty} a_n$ diverges. By contraposition, if $\sum_{n=1}^{\infty} a_n$ converges, then (s_n) is bounded.

As the implication has been shown in both directions, $\sum_{n=1}^{\infty} a_n$ converges if and only if (s_n) is bounded. ■

34.4.3 Null Sequence Test

Theorem. *If (a_n) is not null, then $\sum_{n=1}^{\infty} a_n$ diverges.*

Proof. If $\sum_{n=1}^{\infty} a_n$ converges to some value, s, then $(s_n) \to s$, and $(s_{n+1}) \to s$ by shift rule. $a_{n+1} = s_{n+1} - s_n$, so $\lim_{n\to\infty} a_{n+1} = s - s = 0$, so (a_{n+1}) is a null sequence. By shift rule, (a_n) is also a null sequence.

By contraposition, if (a_n) is not a null sequence, then $\sum_{n=1}^{\infty} a_n$ does not converge. ■

The null sequence test does not require the series to have only non-negative terms. However, it is a test for divergence only, because the converse isn't true – the series generated from a null sequence does not always converge, as we will soon see. But first, another test:

34.4.4 Comparison Test

The comparison test allows us to test the convergence of a series by comparing it term by term against a series with known behaviour.

Theorem (Comparison Test)**.** *Suppose $0 \leq a_n \leq b_n$ for all n. If $\sum_{n=1}^{\infty} b_n$ converges, then $\sum_{n=1}^{\infty} a_n$ also converges, and $\sum_{n=1}^{\infty} a_n \leq \sum_{n=1}^{\infty} b_n$.*

Proof. Let $s_n = \sum_{i=1}^{n} a_n$ and $r_n = \sum_{i=1}^{n} b_n$. $0 \leq a_n \leq b_n$, so (s_n) and (r_n) are increasing, and,

$$\begin{aligned} a_1 &\leq b_1 \\ a_1 + a_2 &\leq b_1 + b_2 \\ a_1 + a_2 + \cdots + a_k &\leq b_1 + b_2 + \cdots + b_k \\ \sum_{n=1}^{k} a_n &\leq \sum_{n=1}^{k} b_n \\ s_n &\leq r_n \end{aligned}$$

If $\sum_{n=1}^{\infty} b_n$ converges to some value, r, then $(r_n) \to r$, so

$$\begin{aligned} 0 \leq s_n & \qquad \leq r_n \\ \lim_{n\to\infty} 0 \leq \lim_{n\to\infty} s_n &\leq \lim_{n\to\infty} r_n \\ 0 \leq \lim_{n\to\infty} s_n &\leq r \end{aligned}$$

So (s_n) is bounded. As (s_n) is also increasing, it is convergent by the completeness axiom. ■

So, if we can show that every term of a sequence is at most equal to the corresponding term in a second sequence, and the series generated by summing the second sequence converges, then the series generated by summing our original sequence also converges and does so to a smaller value. The contrapositive of the theorem also gives a second test for divergence:

Corollary (Comparison Test)**.** *Suppose $0 \leq a_n \leq b_n$ for all n. If $\sum_{n=1}^{\infty} a_n$ diverges, then $\sum_{n=1}^{\infty} b_n$ also diverges.*

Now we have enough machinery to see the counterexample to converse of the null sequence test.

34.4.5 Harmonic Series

The *harmonic series* is the series given by $a_n = \frac{1}{n}$, or,

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$$

It is named after the overtones and harmonics in music – the wavelengths of the overtones in a vibrating medium are terms of the harmonic series multiplied by the medium's fundamental wavelength. Each term of the harmonic series is also the harmonic mean (§34.1.2) of its preceding and following term, so the terms form a harmonic progression. The partial sums are also called *harmonic numbers*, often denoted with H_n. The harmonic series has many applications, particularly in discrete mathematics.

While (a_n) is a null sequence, the harmonic series is actually divergent:

Theorem (Divergence of the Harmonic Series)**.** *The harmonic series,*

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$$

diverges.

Proof. Replace each denominator of the harmonic series with the next largest power of two. Because the denominators are being replaced by larger numbers, this makes the series smaller.

$$\begin{aligned}\sum_{n=1}^{\infty} \frac{1}{n} &= 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \frac{1}{9} + \cdots \\ &\geq 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{4} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{16} + \cdots \\ &= 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right) + \left(\frac{1}{16} + \cdots\right. \\ &= 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots\end{aligned}$$

This second series clearly diverges to infinity. By the comparison test, the harmonic series diverges. ∎

The above proof also shows that

$$\sum_{n=1}^{2^k} \frac{1}{n} \geq 1 + \frac{k}{2}$$

for all positive integers k.

The harmonic series diverges extremely slowly – later, we will show that the harmonic series has a logarithmic growth rate.

34.4.6 Geometric Series

A geometric series is the sum of a geometric progression – a sequence of numbers where the ratio of each pair of consecutive terms is constant.

Lemma (Geometric Progression)**.** *If $r \neq 1$, then,*

$$\sum_{i=1}^{n} ar^{n-1} = a\left(\frac{r^n - 1}{r - 1}\right)$$

Proof.

$$\begin{aligned}
\sum_{i=1}^{n} ar^{n-1} &= a + ar + ar^2 + \cdots + ar^{n-1} \\
(1-r)\sum_{i=1}^{n} ar^{n-1} &= (1-r)(a + ar + ar^2 + \cdots + ar^{n-1}) \\
(1-r)\sum_{i=1}^{n} ar^{n-1} &= a + ar + ar^2 + \cdots + ar^{n-1} - ar - ar^2 - ar^3 - \cdots - ar^{n-1} - ar^n \\
(1-r)\sum_{i=1}^{n} ar^{n-1} &= a - ar^n \\
\sum_{i=1}^{n} ar^{n-1} &= \frac{a - ar^n}{1 - r} \\
\sum_{i=1}^{n} ar^{n-1} &= a\left(\frac{r^n - 1}{r - 1}\right)
\end{aligned}$$

■

Theorem (Geometric Series)**.** *The series $\sum_{n=0}^{\infty} x^n$ is convergent if and only if $|x| < 1$, and the sum is $\frac{1}{1-x}$. The series is divergent otherwise.*

Proof. Using the previous lemma, we have,

$$\begin{aligned}
s_n &= \sum_{n=0}^{\infty} x^n \\
&= \lim_{n\to\infty} \sum_{i=0}^{n} x^i \\
&= \lim_{n\to\infty} \frac{1 - x^{n+1}}{1 - x} \\
&= \lim_{n\to\infty} \frac{1}{1-x} - \frac{x^{n+1}}{1-x}
\end{aligned}$$

If $x > 1$, then (x^{n+1}) diverges, so (s_n) diverges by sum rule. If $x = 1$, then s_n evaluates to $1 + 1 + \cdots$, which diverges to infinity. If $|x| < 1$, then $(x^{n+1}) \to 0$, so $(s_n) \to \frac{1}{1-x}$. If $x \leq -1$, then (x^{n+1}) diverges, so (s_n) diverges.[*]

So, the series converges to $\frac{1}{1-x}$ if and only if $|x| < 1$, and diverges otherwise. ■

Geometric series are incredibly useful for the comparison test, because we know exactly when a geometric series converges.

[*] The proofs for the (x^{n+1}) sequences converging or diverging were set as exercises at the end of §34.2.6.

34.4.7 Ratio Test

All the previous tests rely on comparing two different series. Choosing a geometric series to compare against gives another simple way to frame the comparison test.

Theorem (Ratio Test)**.** *Suppose $a_n > 0$ for all $n \geq 1$, and $\left(\frac{a_{n+1}}{a_n}\right) \to l$. Then, $\sum_{n=1}^{\infty} a_n$ converges if $0 \leq l < 1$, and diverges if $l > 1$.*

Proof. Suppose $\left(\frac{a_{n+1}}{a_n}\right) \to l < 1$. Let $l < r < 1$, so $a_{n+1} < ra_n$ for any sufficiently large n, say, for all $n > N$. So, $a_{n+i} < r^i a_n$ for all $n > N$ and any integer $i > 0$, and,

$$\sum_{i=N=1}^{\infty} a_n = \sum_{i=1}^{\infty} a_{N+i} < \sum_{i=1}^{\infty} r^i a_N = a_N \sum_{i=1}^{\infty} r^i = a_N \frac{r}{1-r} < \infty$$

so the series converges.

Conversely, if $l > 1$, then $a_{n+1} > a_n$ for all sufficiently large n, so (a_n) is eventually increasing. Because $a_n > 0$ for all $n \geq 1$, and is eventually increasing, it cannot be a null sequence, so $\sum_{n=1}^{\infty} a_n$ diverges by the null sequence test. ■

Note that, if $l = 1$, then the ratio test is inconclusive – we do not gain any information from this test because there exist both convergent and divergent series which give $l = 1$.

34.4.8 Integral Test

As mentioned earlier, it is very difficult to find an explicit formula for the sum of a series, so we often settle for approximations or bounds on the limit. We can compare sums of functions with integrals, allowing us to use integration techniques to obtain very useful approximations.

First, we note that,

$$\sum_{k=2}^{n} f(k) = \int_1^n f(\lceil x \rceil)\, dx$$

The ceiling turns the function into a series of steps with width 1, so integrating the new stepped function is the same as summing the areas of the rectangles with heights determined by integer values of the function.

We can rewrite this as,

$$\sum_{k=2}^{n} f(k) = \int_1^n f(x)\, dx + \int_1^n f(\lceil x \rceil) - f(x)\, dx$$

If f is additionally decreasing, then $x \leq \lceil x \rceil$, so the second integral reduces the value of the whole expression, so,

$$\sum_{k=1}^{n} f(k) \leq \int_0^n f(x)\, dx$$

Through similar arguments, we also have,

$$\int_1^{n+1} f(x)\, dx \leq \sum_{k=1}^{n} f(k)$$

And we don't have to start at 1 either. Combining the two and generalising, we get,

$$\int_m^{n+1} f(x)\, dx \leq \sum_{k=m}^{n} f(k) \leq \int_{m-1}^{n} f(x)\, dx$$

giving us very close bounds to the series, as long as f is decreasing.

This gives us two very useful tests for convergence:

Theorem (Integral Test for Convergence)**.** *Suppose the function $f : [1,\infty) \to \mathbb{R}$ is non-negative and decreasing. Then $\sum_{n=1}^{\infty} f(n)$ converges if and only if the increasing sequence $(\int_1^n f(x)dx)$ is bounded, or equivalently, the sequence $(\int_1^n f(x)dx)$ is convergent.*

Theorem (Integral Test for Divergence)**.** *Suppose the function $f : [1,\infty) \to \mathbb{R}$ is non-negative and decreasing. Then $\sum_{n=1}^{\infty} f(n)$ diverges if and only if the increasing sequence $(\int_1^n f(x)dx)$ is unbounded, or equivalently, the sequence $(\int_1^n f(x)dx)$ is divergent.*

Proof. Integral test for convergence. If $\left(\int_1^n f(x)\,dx\right)$ is bounded, it is convergent by the completeness axiom as it is increasing. $0 \leq \sum_{k=2}^{n} f(k) \leq \int_1^n f(x)\,dx$, so $\sum_{k=2}^{n} f(k)$ converges by the comparison test. It follows that $\sum_{k=1}^{n} f(k)$ then also converges by the shift rule.

Integral test for divergence. Exercise. ■

Using the integral test, we can show that $\zeta(s)$ converges for $p > 1$, and diverges for $0 < p \leq 1$.

34.4.9 Alternating Series

Most of the tests we have covered so far only work with series that consist only of non-negative terms. However, for certain simpler series, we can extend some of these tests to handle some extra cases.

An *alternating series* is a series whose terms alternate from positive to negative. That is, series of the form,

$$\sum_{n=1}^{\infty}(-1)^{n+1}a_n$$

where (a_n) is a non-negative sequence.

For instance,

$$\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}$$

is the *alternating harmonic series*, and, unlike the regular harmonic series, the alternating harmonic series converges.

Theorem 34.4.1. *If (a_n) is decreasing and null, then the alternating series $\sum_{n=1}^{\infty}(-1)^{n+1}a_n$ is convergent.*

Proof. Suppose (a_n) is decreasing and null, and let s_n be the alternating series of (a_n). Then,

$$S_{2(k+1)+1} = S_{2k+1} - a_{2k+2} + a_{2k+3}$$

(a_n) is decreasing, so $a_{2k+3} \geq a_{2k+2}$, and,

$$\leq S_{2m+1}$$

so the subsequence of odd partial sums is decreasing.

Similarly,

$$\begin{aligned} S_{2(k+1)} &= S_{2k} + a_{2k+1} - a_{2k+2} \\ &\leq S_{2k+1} \end{aligned}$$

Because (a_n) is decreasing and null, we have $a_n > 0$ for all n, so,

$$\begin{aligned} S_{2k+1} &= S_{2k} + a_{2k+1} \\ S_{2k+1} &\geq S_{2k} \end{aligned}$$

Combining the inequalities, we have,

$$a_1 - a_2 = S_2 \leq S_{2m} \leq S_{2m+1} \leq S_1 = a_1$$

So both S_{2k} and S_{2k+1} are bounded. Because they are both also monotonic, they converge by completeness.

Then,

$$\lim_{k\to\infty} (S_{2k+1} - S_{2k}) = \lim_{k\to\infty} a_{2k+1} = 0$$

so they converge to the same limit. ■

The alternating series test requires that the sequence (a_n) is decreasing and null. If either condition is relaxed, then the alternating series may not converge, even if the terms of the sequence are still all non-negative.

For instance, $(a_n) = 1$ is a decreasing sequence, but is not a null sequence. The alternating series $s_n = \sum_{k=1}^{n} (-1)^{k+1} a_k$ has the subsequence of odd terms $s_{2k+1} = 1 + 1 + \cdots$ which diverges to infinity, while the subsequence of even terms $s_{2k} = -1 - 1 - \cdots$ diverges to minus infinity. Because (s_n) has subsequences which have different limits, (s_n) also diverges.

On the other hand, the sequence given by,

$$a_n = \begin{cases} \frac{2}{n} & n \text{ is even} \\ 0 & n \text{ is odd} \end{cases}$$

has alternating series $s_n = 0 + 1 + 0 + \frac{1}{2} + 0 + \frac{1}{3} + 0 + \cdots$, which is the harmonic series, which diverges.

Additionally, the alternating series test is sufficient, but not necessary for convergence.* For instance,

$$\sum_{n=2}^{\infty} \frac{(-1)^n}{n + (-1)^n}$$

is non-monotonic, but still converges.

34.4.10 General Series

The main reason why series with non-negative terms are easier to deal with is because the sequence of partial terms is then monotonic, allowing us just prove that the series is bounded above, and letting completeness prove convergence. However, in the general case, the sequence of partial sums is not guaranteed to be monotonic.

However, we can still use the Cauchy criterion, which does not require monotonicity.

A series $\sum_{n=1}^{\infty} a_n$ is *absolutely convergent* if $\sum_{n=1}^{\infty} |a_n|$ is convergent.

If $\sum_{n=1}^{\infty} a_n$ converges, but $\sum_{n=1}^{\infty} |a_n|$ diverges, then we say the series is *conditionally convergent.*

For instance, the alternating harmonic series converges, but the absolute value gives the harmonic series, which diverges, so the alternating harmonic series is conditionally convergent.

* Specifically, monotonicity is not necessary for convergence. We know from the null sequence test that (a_n) being a null sequence *is* necessary.

Theorem (Absolute Convergence)**.** *Every absolutely convergent series is convergent.*

Proof. Suppose $\sum_{n=1}^{\infty} a_n$ is absolutely convergent, so $\sum_{n=1}^{\infty} |a_n|$ is convergent and hence Cauchy. So, for any $\varepsilon > 0$, there exists N such that $|\sum_{i=m}^{n} |a_i|| = \sum_{i=m}^{n} |a_i| < \varepsilon$ whenever $n > m \geq N$.

By the triangle inequality,

$$\begin{aligned} \left|\sum_{i=m}^{n} a_i\right| &\leq \sum_{i=m}^{n} |a_i| \\ &< \varepsilon \end{aligned}$$

so $\sum_{i=m}^{n} a_i$ is Cauchy, and hence convergent. ■

Now, we can extend the ratio test to handle general series. In fact, the proof of this extended version is the exact same proof as before, just with absolute value bars instead of brackets.

Theorem (Ratio Test)**.** *Suppose* $\left|\frac{a_{n+1}}{a_n}\right| \to l$*. Then,* $\sum_{n=1}^{\infty} a_n$ *converges absolutely if* $0 \leq l < 1$*, and diverges if* $l > 1$*.*

Proof. Suppose $\left|\frac{a_{n+1}}{a_n}\right| \to L < 1$. Let $L < r < 1$, so $|a_{n+1}| < r|a_n|$ for any sufficiently large n, say, for all $n > N$. So $|a_{n+i}| < r^i|a_n|$ for all $n > N$ and any integer $i > 0$, so,

$$\sum_{i=N=1}^{\infty} |a_n| = \sum_{i=1}^{\infty} |a_{N+i}| < \sum_{i=1}^{\infty} r^i |a_N| = |a_N| \sum_{i=1}^{\infty} r^i = |a_N| \frac{r}{1-r} < \infty$$

so the series converges absolutely.

Conversely, if $L > 1$, then $|a_{n+1}| > |a_n|$ for all sufficiently large n, failing the null sequence test. ■

Note that, again, if $l = 1$, then this version of the ratio test is still inconclusive. There exist series which give $l = 1$ that converge absolutely, converge conditionally, and diverge.

For instance, consider the three series,

$$\sum_{n=1}^{\infty} 1,$$
$$\sum_{n=1}^{\infty} \frac{1}{n^2},$$
$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}$$

The first series is divergent, the second converges absolutely,[*] and the third converges conditionally.[†] The value of $\left|\frac{a_{n+1}}{a_n}\right|$ for each series is 1, $\frac{n^2}{(n+1)^2}$ and $\frac{n}{n+1}$, respectively, and in all three cases, $\lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| = 1$.

Corollary (Ratio Test Variant)**.** *If* $\left|\frac{a_{n+1}}{a_n}\right| \to \infty$*, then* $\sum_{n=1}^{\infty} a_n$ *diverges.*

[*] Finding an exact value for this series is known as the *Basel problem*, named after the hometown of the Bernoulli family who had attempted this problem. The problem was eventually solved by Euler, who proved that the sequence converges to $\frac{\pi^2}{6}$. Euler also generalised the problem to the reciprocals of other powers. Later, Riemann continued further with his zeta function (this series is $\zeta(2)$).

[†] This one converges to $\ln 2$, and is considerably easier to prove.

Proof. If $\left|\frac{a_{n+1}}{a_n}\right| \to \infty$, then there exists $N \in \mathbb{N}$ such that for all $n > N$, $\left|\frac{a_{n+1}}{a_n}\right| > C$ for any choice of $C > 0$. Let $C \geq 1$. It follows that, for all $n > N$, $|a_{n+1}| > C|a_n|$, so $|a_n|$ is strictly increasing for all $n > N$. $|a_n| \geq 0$, so $|a_n|$ cannot be null, so $\sum_{n=1}^{\infty} |a_n|$ diverges by the null sequence test. ■

34.5 Riemann's Rearrangement Theorem

Because addition is commutative and associative, for a finite set of numbers, we can add them up in any order, and obtain the same result. However, it turns out that this seemingly reasonable idea does not hold for infinite sums – rearranging the terms of certain types of series does not preserve their value.

A sequence (b_n) is a *rearrangement* of a sequence (a_n) if there exists a permutation (§12.4.7.1) $\sigma : \mathbb{N} \to \mathbb{N}$ such that $b_n = a_{\sigma(n)}$ for all n.

Theorem 34.5.1. *Suppose $\sum_{n=1}^{\infty} a_n$ is convergent and consists only of non-negative terms. Then, if (b_n) is a rearrangement of (a_n), then $\sum_{n=1}^{\infty} b_n$ converges, and $\sum_{n=1}^{\infty} b_n = \sum_{n=1}^{\infty} a_n$.*

Theorem 34.5.2. *Suppose $\sum_{n=1}^{\infty} a_n$ is absolutely convergent. Then, if (b_n) is a rearrangement of (a_n), then $\sum_{n=1}^{\infty} b_n$ is convergent and $\sum_{n=1}^{\infty} b_n = \sum_{n=1}^{\infty} a_n$.*

So far, nothing unusual. However, something interesting happens with conditionally convergent series. Because the positive terms of a conditionally convergent series converge to infinity, and the negative terms tend to minus infinity, we can rearrange the sequence so that we add up terms until we get past a certain point, L, then have some negative terms until we are back below L. Since the series is infinite, there's no problem with pulling more positive or more negative terms in from further back, since there's always infinitely many more to come. Repeating this procedure, the series jumps back and forth around L, and eventually, in the limit, converges to L.

Proof. Suppose $a_n \neq 0$ for all n. Now, consider the sequences given by,

$$a_n^+ = \frac{a_n + |a_n|}{2}, \qquad a_n^- = \frac{a_n - |a_n|}{2}$$

Then, ■

WIP

34.6 Functions

34.6.1 Terminology & Notation

Recall that a sequence is really just a function with domain $\mathbb{N}$, with a_n really being another way of writing $a(n)$ for some function $a : \mathbb{N} \to X$ (with the specific case of $X = \mathbb{R}$ in the previous sections). We can extend our previous results from sequences to more general functions, but first, we state some basic definitions.

Given two sets, A and B, a *function* from A to B assigns every element in A an element in B, and we write $f : A \to B$. If we are talking about mapping specific elements $a \in A$ to an element $b \in B$, we write $a \mapsto b$.

The set A is the *domain* of the function f, also written as $\text{dom}(f)$, and the set B is the *codomain* of f, also written as $\text{cod}(f)$ or $\text{cdm}(f)$.

If every element in A is mapped to a distinct element in B, the function is *injective*. This property can be written $f(a) = f(b) \implies a = b$ or $a \neq b \implies f(a) \neq f(b)$. If every element in B has an origin element, the function is *surjective*. This property can be written as $\forall b \in B, \exists a \in A$ such that $f(a) = b$. If a function is both injective and surjective, it is *bijective*.

The set $\operatorname{im}(f) = \{f(x) : x \in A\} \subseteq B$ is the *image* of the function, and is a subset of the codomain; surjectivity is also equivalent to $\operatorname{im}(f) = \operatorname{cdm}(f)$.

You may occasionally see the term “range” being used to refer to either the codomain, or the image of a function, but these notions are distinct. It is recommended that the term “range” is avoided in general – be specific and say “image” or “codomain” instead, as this is unambiguous.

An *interval* of the real line is a subset, I, of $\mathbb{R}$ with the property that if $x < y < z$, $x \in I$, and $z \in I$, then $y \in I$. That is, an interval contains all points between its endpoints. But, the endpoints may or may not be included themselves. This defines a few types of intervals of interest:

$$\begin{aligned} \{x : a \le x \le b\} &= [a,b] && \text{a \textit{closed} interval} \\ \{x : a < x < b\} &= (a,b) && \text{an \textit{open} interval} \\ \{x : a \le x < b\} &= [a,b) && \text{a \textit{half-open} interval} \\ \{x : a \le x\} &= [a,\infty) && \text{a \textit{half-infinite} interval} \end{aligned}$$

Note that we use an open interval bracket whenever we have ∞ as one of the endpoints, because ∞ is not a member of the standard real numbers.

Also note that sometimes “$]a,b[$” is written to mean (a,b), but we will not be using this notation.

Two functions, $f : A \to B$ and $g : X \to Y$ are *equal* if $A = X$, $B = Y$ and $f(x) = g(x)$ for all $x \in A$. That is, two functions are equal if their domains and codomains are equal, and they agree for all inputs.

34.6.2 Continuity

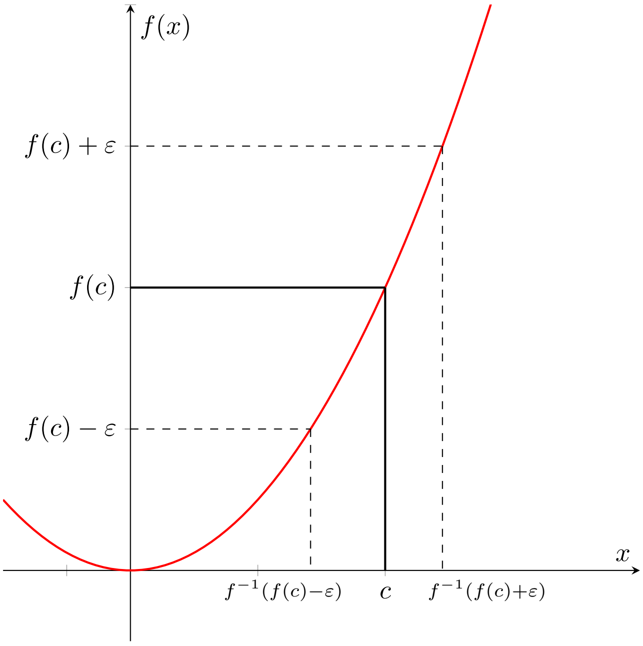
A function $f : I \to R$ is said to be *continuous at a point* $c \in I$ if for all $\varepsilon > 0$, there exists $\delta > 0$ such that if $x \in I$ and $|x - c| < \delta$, then $|f(x) - f(c)| < \varepsilon$.

Example. The function $x \mapsto x$ is continuous for all x. △

Proof. Let $\delta = \varepsilon$, so if $|x - c| < \delta$, then $|x - c| < \varepsilon$ and $|x - c| = |f(x) - f(c)| < \varepsilon$. ■

Example. The function $x \mapsto x^2$ is continuous for all x. △

Proof. We present a graphical method for working through these types of questions when the function is easy to invert.

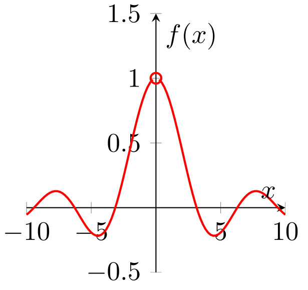


So clearly, if x is less than $\min\left(|c - f^{-1}\left(f(c) - \varepsilon|\right), |c - f^{-1}\left(f(c) + \varepsilon\right)|\right)$ away from c, then $|f(x) - f(c)| < \varepsilon$. Thus, if we let $\delta = \min\left(\left|c - \sqrt{c^2 - \varepsilon}\right|, \left|c - \sqrt{c^2 + \varepsilon}\right|\right)$, then $|f(x) - f(c)| < \varepsilon$ as required. ■

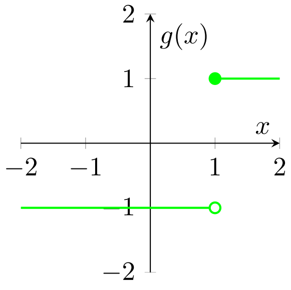
Let $L^+ = \lim_{x \to c^+} f(x)$, and $L^- = \lim_{x \to c^-} f(x)$. Using these quantities, we define three classes of discontinuities. We say that the function f has:

- a *removable discontinuity* at c if $L^+ = L^- = L$ and $f(c)$ exists, but $f(c) \neq L$. If $L^+ = L^-$ but $f(c)$ is undefined, then $f(c)$ is instead a removable *singularity.*
- a *jump discontinuity* at c if $L^+ \neq L^-$. In this case, $f(c)$ can take any value.
- an *essential discontinuity* at c if at least one of the limits L^+ and L^- do not exist.

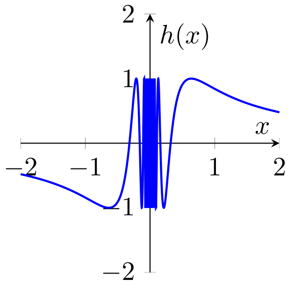
Example.



$$f(x) = \frac{\sin(x)}{x}$$



$$g(x) = \begin{cases} -1 & x < 1 \\ 1 & x \geq 1 \end{cases}$$



$$h(x) = \sin\left(\frac{1}{x}\right)$$

△

- $\lim_{x \to 0^+} f(x) = \lim_{x \to 0^-} f(x) = 1$, but $f(0)$ is undefined, so f has a removable singularity at 0.
- $\lim_{x \to 1^+} g(x) = 1 \neq 1 = \lim_{x \to 1^-} g(x)$, so g has a jump discontinuity at 1.
- $\lim_{x \to 0^+} h(x)$ and $\lim_{x \to 0^-} h(x)$ both fail to exist, so h has an essential discontinuity at 0.

Let $f : I \to \mathbb{R}$ and $c \in I$. Then, f is *sequentially continuous at a point c* if $f(x_n) \to f(c)$ for every sequence $(x_n) \subseteq I$ which converges to c.

Theorem 34.6.1. *A function is sequentially continuous if and only if it is continuous.*

Proof. Suppose $f : I \to \mathbb{R}$ is continuous, and let $(x_n) \subseteq I$ converge to c.

Because f is continuous, for any $\varepsilon > 0$, there exists $\delta > 0$ such that $|x - c| < \delta$ implies $|f(x) - f(c)| < \varepsilon$. Similarly, there exists $N > 0$ such that $|x_n - c| < \delta$ for all $n > N$. Then, if $n > N$, we have $|f(x_n) - f(c)| < \varepsilon$, so $f(x_n) \to f(c)$ and f is sequentially continuous.

Conversely, suppose f is not continuous at c, so there exists $\varepsilon > 0$ such that for all $\delta > 0$, there exists x such that $|x - c| < \delta$, but $|f(x) - f(c)| \geq \varepsilon$. Fix this ε, and for each $n \in \mathbb{N}$, let $\delta_n = \frac{1}{n}$, so there exists at least one choice of x_n such that $|x_n - c| < \delta_n$, but $|f(x) - f(c)| \geq \varepsilon$. Then, $(x_n) \to c$, but $f(x_n)$ does not converge to $f(c)$. ■

Theorem (Algebra of Continuous Functions)**.** *Let $f,g : I \to \mathbb{R}$ be functions continuous at a point $c \in I$. Then,*

- *$f + g$ is continuous at c;*
- *fg is continuous at c;*
- *$\frac{f}{g}$ and is continuous at c if $g(c) \neq 0$.*

Proof. Follows trivially from the algebra of sequences and the equivalence of continuity and sequential continuity. ■

Theorem (Continuity of Polynomials and Rational Functions)**.** *If p is a polynomial, then p is continuous over $\mathbb{R}$. If $r = \frac{p}{q}$ is the ratio of two polynomials, it is continuous wherever $q \neq 0$.*

Proof. $f(x_n) \to f(c)$ and $g(x_n) \to g(c)$, so applying algebra of sequences, we have $(f + g)(x_n) = f(x_n) + g(x_n) \to f(c) + g(c)$, so $(f + g)(x)$ is continuous by sequential continuity.

The proofs for the continuity of the product and ratio of continuous functions are similar. ■

Theorem (Composition of Continuous Functions)**.** *Let $f : I \to \mathbb{R}$ and $g : X \to I$. If g is continuous at c and f is continuous at $g(c)$, then the composition $f \circ g$ is continuous at c.*

Proof. Let (x_n) be a sequence in X converging to c. Then, $g(x_n) \to g(c) \in I$ and hence $f(g(x_n)) \to f(g(c))$. ■

34.7 The Intermediate Value Theorem

Intuitively, a continuous function is a function whose graph we can draw without lifting our pen off the paper, and it should therefore be obvious that a continuous function can't “skip” a value over an interval. However, the proof is non-trivial because we have only defined continuity at each point individually.

The strategy for the proof is then to find the point where the function is supposed to take the correct value and then use continuity at that point. When we are trying to demonstrate the existence of a particular real number, we also usually have to use a completeness axiom.

Theorem (Intermediate Value Theorem)**.** *Let $f : [a,b] \to \mathbb{R}$ be continuous and suppose $f(a) < k < f(b)$. Then, there exists $c \in [a,b]$ such that $f(c) = k$.*

Proof. Let $f(a) < u < f(b)$ and let $S = \{x \in [a,b] : f(x) \leq u\}$. S is non-empty as $a \in S$ and is bounded above by b. By completeness, the supremum, $c = \sup S$ exists. We will show that $f(c) = u$.

Let $\varepsilon > 0$. As f is continuous, there exists $\delta > 0$ such that $|f(x) - f(c)| < \varepsilon$ whenever $|x - c| < \delta$, so we have;

$$f(x) - \varepsilon < f(c) < f(x) + \varepsilon \tag{$\star$}$$

for all $x \in (c - \delta, c + \delta)$ by the interval property. By the properties of the supremum, there exists some $a^* \in (c - \delta, c]$ that is contained in S. By construction, a^* is within δ of c, so $|a^* - c| < \delta$ holds, and we may use the right side of ($\star$) to write:

$$f(c) < f(a^*) + \varepsilon \leq u + \varepsilon$$

Picking $a^{**} \in (c, c + \delta)$, we know $a^{**} \notin S$ since c is the supremum of S and $a^{**} > c$ by construction. Again, a^{**} is within δ of c, so,

$$f(c) > f(a^{**}) - \varepsilon > u - \varepsilon$$

and combining both inequalities, we have,

$$u - \varepsilon < f(c) < u + \varepsilon$$

and u is the only value of $f(c)$ such that the above inequality holds for all $\varepsilon > 0$, so $f(c) = u$. ■

Corollary (Continuous Image of Intervals)**.** *If $f : I \to \mathbb{R}$ is continuous over I, then the image of f is also an interval.*

Proof. If x and y are in the image of f, then by the IVT, every point between x and y is also in the image of f, so the image of f is an interval. ■

This corollary is actually equivalent to the IVT, and can be taken to be an alternative statement of the IVT.

The IVT has many applications, but one of the most obvious is that it guarantees the existence of solutions to equations:

Theorem (Bolzano)**.** *If a continuous function has values of opposite sign inside an interval, then it has a root in that interval.*

Example. There is a solution of the equation $x^3 + x - 1 = 0$ between 0 and 1.

Proof. Define $f(x) = x^3 + x - 1$. Then, $f(0) = -1 < 0$, while $f(1) = 1 > 0$. △

△

Theorem (Existence of Square Roots)**.** *Every positive real number has a unique positive square root.*

Proof. Let r be a positive real number and consider the function f defined by $x \mapsto x^2$. We have that $f(0) < r$ and $f(r + 1) = r^2 + 2r + 1 > r$, so there exists a number $c \in [0, r + 1]$ such that $f(c) = r$. It follows that $c^2 = R$, so c is a positive square root of r.

For uniqueness, suppose two distinct positive real numbers x and y exist such that $f(x) = r$ and $f(y) = r$. By trichotomy, and without loss of generality, suppose $0 < x < y$. As x and y are positive and real, we have $x^2 < y^2$, contradicting that $f(x) = r$ and $f(y) = r$. ■

Corollary 34.7.0.1. *For each positive real x and natural n, there is a unique positive nth root $x^{\frac{1}{n}}$, and the map $x \mapsto x^{\frac{1}{n}}$ is continuous.*

Proof. Exercise. ■

Theorem (Existence of Inverses)**.** *Let $f : [a,b] \to \mathbb{R}$ be continuous and strictly increasing. Then f has a continuous inverse, f^{-1} defined over its image.*

Proof. Let $f(a) = c$ and $f(b) = d$. Since f is increasing, the image of f lies between c and d. In fact, the image of f is exactly $[c,d]$.

For each y, there is a unique number x such that $f(x) = y$ as f is strictly increasing, so we let $f^{-1} = g : y \mapsto x$. By construction, g is increasing.

Let $\varepsilon > 0$, and suppose $f(x) = y \in (c,d)$, so $f(x - \varepsilon) < y < f(x + \varepsilon)$, so there exists δ such that,

$$f(x - \varepsilon) < y - \delta < y < y + \delta < f(x + \varepsilon)$$

and for any $z \in (y - \delta, y + \delta)$, we have

$$x - \varepsilon < g(z) < x + \varepsilon$$

so $|g(z) - g(y)| < \varepsilon$.

If instead $y = a$ or $y = b$, the argument is the same, but with $f(x - \varepsilon)$ and $f(x + \varepsilon)$ replaced by c or d. ■

34.8 The Extreme Value Theorem

A function $f : I \to \mathbb{R}$ is:

- *bounded above* if $\exists M \in \mathbb{R} : f(x) \leq M \,\forall\, x \in I$ – if there exists a real number M, called an *upper bound*, such that $f(x) \leq M$ for all $x \in I$;
- *bounded above* if $\exists m \in \mathbb{R} : f(x) \geq m \,\forall\, x \in I$ – if there exists a real number m, called a *lower bound*, such that $f(x) \geq m$ for all $x \in I$;
- *bounded* if it is bounded above *and* bounded below.

A continuous function on an open interval may take arbitrarily large values. For instance, $\frac{1}{x}$ is continuous on $(0,1)$, but $\lim_{x \to 0^+} \frac{1}{x} = \infty$. However, this cannot can't occur if a function f is continuous on a closed interval $[a,b]$, since the function is supposed to approach $f(a)$ as x approaches a. It turns out that indeed a continuous function on a closed interval must be bounded.

Theorem (Boundedness of Continuous Functions)**.** *If $f : [a,b] \to \mathbb{R}$ is continuous, then f is bounded.*

Proof. Suppose f is continuous, but unbounded. We construct a sequence, (x_n), where $|f(x_n)| \geq n$ for all n. By Bolzano-Weierstrass, we can construct a subsequence, (x_{n_i}) which converges to some value, say, x. Since the interval over which f is defined is closed, $x \in [a,b]$ by the closed interval rule. By sequential continuity, $f(x_{n_i}) \to f(x)$, but this is impossible as $f(x_{n_i})$ become arbitrarily large by the construction of (x_n). ■

As shown by the previous example, continuous functions on open intervals may not attain a maximum or minimum. But on closed intervals, this again cannot happen.

Theorem (Extreme Value Theorem)**.** *Let $f : [a,b] \to \mathbb{R}$ be continuous. Then, there exist numbers $c,d \in [a,b]$ such that $f(c) \leq f(x) \leq f(d)$ for all $x \in [a,b]$.*

Proof. Let M be the least upper bound of the set $S = \{f(x) : x \in [a,b]\}$. If there is no point, c, in the interval where $f(c) = M$, then, the function $g(x) = M - f(x)$ is strictly positive and continuous over $[a,b]$. By the algebra of limits, $\frac{1}{g(x)}$ is continuous, and therefore bounded. Let R be an upper bound of $\frac{1}{g(x)}$ over the interval $[a,b]$, noting that $R > 0$. Then, $\frac{1}{R} \leq g(x) = M - f(x)$, so $f(x) \leq M - \frac{1}{R}$, and $M - \frac{1}{R}$ is an upper bound of S. As R is positive, $M - \frac{1}{R} < M$, contradicting that M is the least upper bound of S.

It follows that the point c must exist, and f attains the maximum $f(c)$ over $[a,b]$. The proof for the minimum is similar. ■

34.9 Power Series

A *power series* is a series of the form $\sum_{n=0}^{\infty} a_n(x-c)^n$. We mostly focus on the specific case $c = 0$, with everything easily transferring over to the general case.

Theorem (Radius of Convergence I). *Let $\sum_{n=0}^{\infty} a_n x^n$ be a power series with $\sum_{n=0}^{\infty} a_n t^n$ convergent. Then, $\sum_{n=0}^{\infty} a_n x^n$ converges absolutely for all x such that $|x| < |t|$.*

Proof. Since $\sum_{n=0}^{\infty} a_n t^n$ converges, we know that $a_n t^n \to 0$ as $n \to \infty$ so the sequence of partial sums is bounded by some M such that $|a_n t^n| < M$ for all n. Now,

$$\begin{aligned}\sum_{n=0}^{N} |a_n x^n| &= \sum_{n=0}^{N} |a_n t^n| \left|\frac{x}{t}\right|^n \\ &\leq M \sum_{n=0}^{N} \left|\frac{x}{t}\right|^n \\ &\leq M \sum_{n=0}^{\infty} \left|\frac{x}{t}\right|^n \\ &= M \frac{1}{1 - \left|\frac{x}{t}\right|}\end{aligned}$$

which is finite, so the series converges absolutely. ■

Theorem (Radius of Convergence I). *Let $\sum_{n=0}^{\infty} a_n x^n$ be a power series. Then, exactly one of the following statements holds:*

- *There is a positive real R such that the series converges if $|x| < R$ and diverges if $|x| > R$.*
- *The series converges only if $x = 0$.*
- *The series converges for all real x.*

If R exists, it is called the *radius of convergence* of the series. In the second case, we say that the radius of convergence is 0, and in the last case, we say the radius of convergence is ∞, or that the series *converges everywhere*.

Theorem (Absolute Series). *Let $\sum_{n=0}^{\infty} a_n x^n$ be a power series with radius of convergence R. Then, $\sum_{n=0}^{\infty} |a_n| x^n$ also has radius of convergence R.*

Theorem (Geometric Series I). *The series $\sum_{n=0}^{\infty} x^n$ has radius of convergence $R = 1$.*

Theorem (Geometric Series II). *If p is real, the series $\sum_{n=0}^{\infty} p^n x^n$ has radius of convergence $R = \frac{1}{|p|}$.*

Proof. $\sum_{n=0}^{\infty} p^n x^n = \sum_{n=0}^{\infty} (px)^n$, which converges if $|px| \leq 1$, so $|x| \leq \frac{1}{|p|}$. ■

Theorem (Log Series)**.** *The series $\sum_{n=0}^{\infty} \frac{x^n}{n}$ has radius of convergence $R = 1$.*

Theorem (Continuity of Convergent Power Series)**.** *Let $\sum_{n=0}^{\infty} a_n x^n$ be a power series with radius of convergence R. Then, the function, $x \mapsto \sum_{n=0}^{\infty} a_n x^n$ is continuous over $(-R,R)$.*

Proof. Let $x \in (-R,R)$, and T such that $|x| < T < R$. It follows that $T \in (-R,R)$ so $\sum_{n=0}^{\infty} |a_n| T^n$ converges, so for each $\varepsilon > 0$ there exists N for which,

$$\sum_{n=N+1}^{\infty} |a_n| T^n < \frac{\varepsilon}{3}$$

Now, if $|y - x| < T - |x|$, $|y| < T$ and $x < T$, so,

$$\sum_{n=N+1}^{\infty} |a_n||x|^n < \frac{\varepsilon}{3} \text{ and } \sum_{n=N+1}^{\infty} |a_n||y|^n < \frac{\varepsilon}{3}$$

The partial sum $\sum_{n=0}^{N} |a_n| y^n$ is a polynomial in y, and polynomials are continuous, so there exists some δ_0 such that if $|y - x| < \delta_0$,

$$\left|\sum_{n=0}^{N} a_n y^n - \sum_{n=0}^{N} a_n x^n\right| < \varepsilon/3$$

So, letting $\delta = \min(\delta_0, T - |x|)$, if $|y - x| < \delta$, we have,

$$\begin{aligned} \left|\sum_{n=0}^{\infty} a_n y^n - \sum_{n=0}^{\infty} a_n x^n\right| &\leq \left|\sum_{n=N+1}^{\infty} a_n y^n\right| + \left|\sum_{n=0}^{N} a_n y^n - \sum_{n=0}^{N} a_n x^n\right| + \left|\sum_{n=N+1}^{\infty} a_n x^n\right| \\ &\leq \sum_{n=N+1}^{\infty} |a_n||y|^n + \left|\sum_{n=0}^{N} a_n y^n - \sum_{n=0}^{N} a_n x^n\right| + \sum_{n=N+1}^{\infty} |a_n||x|^n < \varepsilon \end{aligned}$$

■

34.9.1 The Exponential Function

The *exponential function* $\exp(x)$ is defined to be the series:

$$\exp(x) := \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

Theorem 34.9.1. *The exponential series converges everywhere.*

Proof.

$$\begin{aligned} \frac{x^{n+1}}{(n+1)!} \frac{n!}{x^n} &= \frac{x}{n+1} \\ &\to 0 \end{aligned}$$

so the series converges for all x by the ratio test. ■

We then also have that $x \mapsto \exp(x)$ is continuous by the continuity of convergent power series.

Theorem (Characteristic Property of the Exponential)**.** *The exponential function satisfies the following functional equation:*

$$\exp(x+y) = \exp(x)\exp(y)$$

with $\exp(1) = e$*, where* e *is Euler's constant.*

Proof. For a fixed $z \in \mathbb{R}$, consider the function $f(x) = \exp(x)\exp(z-x)$. Differentating with respect to x, we have $f'(x) = \exp(x)\exp(z-x) - \exp(x)\exp(z-x) = 0$, so $f(x)$ must be a constant function by the mean value theorem. At $x = 0$, $f(x) = \exp(z)$, but, as $f(x)$ is constant, we must have $f(x) = \exp(z)$ for all x, so $\exp(x)\exp(z-x) = \exp(z)$ for all x. Let $z = x+y$, and we have $\exp(x)\exp(y) = \exp(x+y)$. ∎

Theorem 34.9.2. $\exp(x) = e^x$ *for all real numbers* x*.*

Theorem (Inequalities for the Exponential)**.** *The following inequalities hold:*

- $1 + x \leq e^x$*;*
- $e^x \leq \frac{1}{1-x}$ *if* $x < 1$*.*

Proof. If $x \geq 0$ then $e^x = 1 + x + \frac{x^2}{2} + \cdots \geq 1 + x$, and if $0 \leq x < 1$, $e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots \leq 1 + x + x^2 + x^3 + \cdots = \frac{1}{1-x}$.

If $x < 0$, let $u = -x$ so $e^u \geq 1 + u$ implies $e^{-x} \geq 1 - x \Leftrightarrow \frac{1}{e^x} \geq 1 - x \Leftrightarrow \frac{1}{1-x} \geq e^x$, so the second inequality holds for all $x < 1$.

If $x \leq -1$, then $1 + x \leq 0$, but $e^x > 0$, so $e^x \geq 1 + x$ for $x \leq -1$. Now, if $-1 < x < 0$, then $0 < u < 1$ and so $e^u \leq \frac{1}{1-u}$, so $e^{-x} \leq \frac{1}{1+x} \Leftrightarrow \frac{1}{e^x} \leq \frac{1}{1+x} \Leftrightarrow \frac{1}{+}x \leq e^x$ ∎

Theorem 34.9.3. *The exponential function is strictly increasing, and its image is* (0,∞)*.*

Proof. Suppose $x < y$. Then, $e^y = e^{y-x}e^x \geq (1 + y - x)e^x > e^x$.

Since $e^x \geq 1 + x$, the exponential function takes arbitrarily large values for large choices of x, and since $e^{-x} = \frac{1}{e^x}$, the exponential function takes arbitrarily small values for very large negative choices of x. By the IVT, the exponential takes all positive values. ∎

34.9.2 The Logarithmic Function

The exponential function maps R to (0,∞) and is continuous and strictly increasing, so by the IVT, the exponential function has a continuous inverse defined over (0,∞). This inverse function is called the *natural logarithm*, written as ln or sometimes log (though log can also represent a logarithm with base 10, or rarely, 2).

The function ln : (0,∞) $\to \mathbb{R}$ satisfies $e^{\ln x} = x$ for all positive real x, and $\ln(e^y) = y$ for all real y. For all positive real a,b, we have $\ln(ab) = \ln a + \ln b$.

Using the natural logarithm, we can extend the definition of exponentiation to irrational exponents: if $x > 0$ and $p \in \mathbb{R}$, we define $x^p = \exp(p \log x)$.

Theorem (Tangent to the Logarithm)**.** *If* $x > 0$*, then* $\log x \leq x - 1$*.*

34.10 Limits

Let I be an open interval and f a real-valued function defined over I, except possibly at a point $c \in I$. We write

$$\lim_{x\to c} f(x) = L$$

if for every $\varepsilon > 0$, there exists $\delta > 0$ such that if $0 < |x-c| < \delta$, then $|f(x) - L| < \varepsilon$.

Theorem (Limits and Continuity)**.** *If $f : I \to \mathbb{R}$ is defined over the open interval I and $c \in I$, then f is continuous at c if and only if $\lim_{x\to c} f(x) = f(c)$.*

This gives an alterative characterisation of continuous functions: a function is continuous at a point if it is equal to its limit at that point.

Theorem (Continuous and Sequential Limits)**.** *If $f : I \setminus \{c\} \to \mathbb{R}$ is defined over the interval $I \setminus \{c\}$, then $\lim_{x\to c} f(x) = L$ if and only if for every sequence (x_n) of points in $I \setminus \{c\}$ which converges to c, we have $f(x_n) \to L$.*

Theorem (Algebra of Limits)**.** *If $f,g : I \setminus \{c\} \to \mathbb{R}$ are defined over the interval $I \setminus \{c\}$ and $\lim_{x\to c} f(x)$ and $\lim_{x\to c} g(x)$ exist, then,*

- $\lim_{x\to c}(f(x) + g(x)) = \lim_{x\to c} f(x) + \lim_{x\to c} g(x)$;
- $\lim_{x\to c} f(x)g(x) = \lim_{x\to c} f(x) \lim_{x\to c} g(x)$;
- *if* $\lim_{x\to c} g(x) \neq 0$, *then,*

$$\lim_{x\to c} \frac{f(x)}{g(x)} = \frac{\lim_{x\to c} f(x)}{\lim_{x\to c} g(x)}$$

Proof. Exercise. ■

Let $f : [a,b] \to \mathbb{R}$. We write

$$\lim_{x\to c^+} f(x) = L$$

for the *one-sided limit* of $f(x)$ to c from the right, if for every $\varepsilon > 0$, there exists $\delta > 0$ such that if $c < x < c + \delta$, then $|f(x) - L| < \varepsilon$. The one sided limit

$$\lim_{x\to c^-} f(x) = L$$

from the left is defined similarly.

Let I be an open interval and f a real-valued function defined over I, except possibly at a point $c \in I$. We write

$$\lim_{x\to c} f(x) = \infty$$

if for every $M > 0$ there is a $\delta > 0$ such that if $0 < |x - c| < \delta$ then $f(x) > M$. The limit

$$\lim_{x\to c} f(x) = -\infty$$

is defined similarly.

Let $f : \mathbb{R} \to \mathbb{R}$ be defined on all of $\mathbb{R}$ (or, on all sufficiently large inputs). We write

$$\lim_{x\to\infty} f(x) = L$$

if for every $\varepsilon > 0$ there is an N such that if $x > N$, then $|f(x) - L| < \varepsilon$.

Theorem (Uniqueness of Limits)**.** *If $f(x) \to M$ as $x \to c$, and $f(x) \to L$ as $x \to c$, then $M = L$.*

Proof. Similar to uniqueness of limits for sequences. ■

Theorem (Sandwich Theorem for Limits). *Let I be an interval containing the point a. Let g,f,h be functions defined over I, except possibly at a. If for every $x \in I$, we have $g(x) \leq f(x) \leq h(x)$ and $\lim_{x\to a} g(x) = \lim_{x\to a} h(x) = L$, then $\lim_{x\to a} f(x) = L$.*

Note that a does not have to lie within the interior of I, and can be an endpoint, with the limits above being evaluated as one-sided limits. Similarly, the statement holds for infinite intervals, where $x \to \pm\infty$.

34.11 The Derivative

Let $f : I \to \mathbb{R}$ and $c \in I$. The function f is *differentiable* at c if

$$\lim_{h\to 0} \frac{f(c+h) - f(c)}{h}$$

exists, and if so, we denote this limit by $f'(c)$.

Letting $x = c + h$, we can equivalently write the derivative as,

$$\lim_{x\to c} \frac{f(x) - f(c)}{x - c}$$

Theorem (Differentiability Implies Continuity). *If $f : I \to \mathbb{R}$ is differentiable at $c \in I$ then f is continuous at c.*

Proof. $f(x) - f(c) = \frac{f(x)-f(c)}{x-c} \cdot (x - c)$. $\frac{f(x)-f(c)}{x-c} \to f'(c)$ as $x \to c$ by the definition of a derivative, and $(x - c) \to 0$ as $x \to c$, so $\frac{f(x)-f(c)}{x-c} \cdot (x - c) \to f'(c) \cdot 0 = 0$, so $f(x) - f(c) \to 0$, and $f(x) \to f(c)$. ■

Note that the converse does not hold; continuous functions are not necessarily differentiable. For instance, $|x|$ is continuous but not differentiable at 0.

Theorem (Sum and Product Rule). *Suppose f,$g : I \to \mathbb{R}$ are differentiable at $c \in I$. Then, $f + g$ and fg are differentiable at c, and:*

- $(f + g)'(c) = f'(c) + g'(c)$;
- $(fg)'(c) = f(c)g'(c) + f'(c)g(c)$.

Proof. Exercise. ■

Theorem (Derivatives of Monomials). *If n is a natural number, then the derivative of $f(x) = x^n$ is $f'(x) = nx^{n-1}$.*

Proof. For $n = 1$, we have $f(x) = x$. For every c, and $h \neq 0$,

$$\begin{aligned}\lim_{h\to 0} \frac{f(c+h) - f(c)}{h} &= \lim_{h\to 0} \frac{c+h-c}{h} \\ &= 1\end{aligned}$$

Suppose the statement holds for arbitrary fixed $n \geq 1$. Then, $x^{n+1} = xf(x)$, so by the product rule, the derivative is $1 \cdot f(x) + xf'(x) = x^n + xnx^{n-1} = (n + 1)x^n$, which is the statement for $n + 1$, completing the inductive step. ■

We prove a lemma useful for proving the chain rule:

Lemma (Local Linearisation)**.** *Let $f : I \to \mathbb{R}$ be a function and let $c \in I$. Then, f is differentiable at c if and only if there is a number A and a function ε such that for all $x \in I$, we have,*

$$f(x) - f(c) = A(x - c) + \varepsilon(x)(x - c)$$

and $\varepsilon(x) \to 0$ as $x \to c$. If this happens, then $A = f'(c)$.

This lemma is sometimes called the Weierstrass–Caratheodory criterion. Essentially, the lemma states that if x is close to c, then $f(x)$ is approximately given by the linear function

$$x \mapsto f(c) + f'(c)(x - c)$$

which you may recognise as the first-degree Taylor approximation for f.

Proof. If the condition holds, then

$$\frac{f(x) - f(c)}{x - c} = A + \varepsilon$$

which approaches A as $x \to c$, and hence f is differentiable with derivative $f'(c) = A$.

Conversely, suppose f is differentiable at c. Define $A = f'(c)$ and define ε by

$$\varepsilon(x) = \begin{cases} \frac{f(x)-f(c)}{x-c} - A & x \neq c \\ 0 & c = x \end{cases}$$

If $x \neq c$, then

$$\begin{aligned} f(x) - f(c) &= \left(A + \frac{f(x) - f(c)}{x - c} - A\right)(x - c) \\ &= \big(A + \varepsilon(x)\big)(x - c) \\ &= A(x - c) + \varepsilon(x)(x - c) \end{aligned}$$

and if $x = c$, then

$$\begin{aligned} A(x - c) + \varepsilon(x)(x - c) &= A(0) + 0(0) \\ &= 0 \\ &= f(x) - f(x) \\ &= f(x) - f(c) \end{aligned}$$

as required. Then, we have

$$\begin{aligned} \lim_{x \to c} \varepsilon(x) &= \lim_{x \to c} \frac{f(x) - f(c)}{x - c} - A \\ &= \lim_{x \to c} \frac{f(x) - f(c)}{x - c} - f'(c) \\ &= f'(c) - f'(c) \\ &= 0 \end{aligned}$$

■

We are now ready to prove the chain rule.

Theorem (Chain Rule)**.** *Suppose $f : I \to \mathbb{R}$, $g : X \to I$, g is differentiable at $c \in X$ and f is differentiable at $g(c) \in I$. Then, the composition $f \circ g$ is differentiable at c, and,*

$$(f \circ g)'(c) = f'\big(g(c)\big) \cdot g'(c)$$

Proof. We have that f is differentiable at $g(c)$, so for all y, we have,

$$f(y) - f\big(g(c)\big) = f'\big(g(c)\big)\big(y - g(c)\big) + \varepsilon(y)\big(y - g(c)\big)$$

where $\varepsilon(y) \to 0$ as $y \to g(c)$, so,

$$f\big(g(x)\big) - f\big(g(c)\big) = f'\big(g(c)\big)\big(g(x) - g(c)\big) + \varepsilon\big(g(x)\big)\big(g(x) - g(c)\big)$$

and hence,

$$\frac{f\big(g(x)\big) - f\big(g(c)\big)}{x-c} = f'\big(g(c)\big)\frac{\big(g(x) - g(c)\big)}{x-c} + \varepsilon\big(g(x)\big)\frac{\big(g(x) - g(c)\big)}{x-c}$$

As $x \to c$, we have

$$\frac{g(x) - g(c)}{x-c} \to g'(c)$$

while $g(x) \to g(c)$, so $\varepsilon\big(g(x)\big) \to 0$ and hence,

$$\frac{f\big(g(x)\big) - f\big(g(c)\big)}{x-c} \to f'\big(g(c)\big)g'(c)$$

as required. ■

34.12 The Mean Value Theorem

Theorem (Rolle)**.** *Suppose $f : [a,b] \to \mathbb{R}$ is continuous over $[a,b]$ and differentiable over (a,b), and that $f(a) = f(b)$. Then, there is a point $c \in (a,b)$ such that $f'(c) = 0$.*

Proof. If f is constant over the interval, then f' is 0 everywhere over the interval. If not, it takes values distinct from $f(a) = f(b)$.

As f is continuous, there f attains a maximum and minimum within the interval $[a,b]$ by the extreme value theorem. Suppose f attains a maximum at $c \neq a,b$, so $c \in (a,b)$.

If $x \neq c$, then $f(x) \leq f(c)$, as c is a maximum, so

$$f(x) - f(c) \leq 0 \tag{$\star$}$$

Suppose $x > c$. Then, $x - c > 0$, so $\frac{1}{x-c} > 0$. Multiplying both sides of ($\star$) by $\frac{1}{x-c}$, we have,

$$\frac{f(x) - f(c)}{x-c} \leq 0$$

so $f'(c) \leq 0$

Now suppose $x < c$. Then, $x - c < 0$, so $\frac{1}{x-c} < 0$. Multiplying both sides of ($\star$) by $\frac{1}{x-c}$, we now need to swap the direction of inequality as we are multiplying by a negative value, so we have,

$$\frac{f(x) - f(c)}{x-c} \geq 0$$

so $f'(c) \geq 0$.

From the two above equations, we deduce that $f'(c) = 0$. ■

Theorem (Mean Value Theorem)**.** *Suppose $f : [a,b] \to \mathbb{R}$ is continuous over $[a,b]$ and differentiable over (a,b). Then, there is a point $c \in (a,b)$ such that,*

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

This theorem may be more useful in the form,

$$f(b) - f(a) = (b - a)f'(c)$$

Informally, the theorem states that, for any arc between two endpoints, a and b, there exists at least one point between the two endpoints such that the tangent to the arc is parallel to the line that passes through its endpoints.

Another statement of the theorem that more accurately reflects its name, is that there exists at least one point, c, between the two endpoints of the graph of $y = f(x)$ such that the slope at $f(c)$ is equal to the average slope of the graph between the endpoints.

Proof. Let

$$g(x) = f(x) - xr$$

where $r = \frac{f(b)-f(a)}{b-a}$, which is a constant. Then,

$$\begin{aligned} g(b) - g(a) &= (f(b) - br) - (f(a) - ar) \\ &= f(b) - f(a) - (b - a)r \\ &= f(b) - f(a) - (f(b) - f(a)) \\ &= 0 \end{aligned}$$

So, by Rolle's Theorem, there exists a point $c \in (a,b)$ such that $g'(c) = 0$. So, we have,

$$\begin{aligned} g'(x) &= f'(x) - r \\ g'(c) &= f'(c) - r \\ &= 0 \\ \implies f'(c) &= r \\ &= \frac{f(b) - f(a)}{b - a} \end{aligned}$$

■

Corollary (Functions with Positive Derivative)**.** *If $f : I \to \mathbb{R}$ is differentiable over the open interval I, and $f'(x) > 0$ for all $x \in I$, then f is strictly increasing over I.*

Proof. If there were two points a and b such that $a < b$ but $f(a) \geq f(b)$, then $b-a > 0$ and $f(a)-f(b) \geq 0$, so $f'(c) = \frac{f(b)-f(a)}{b-a} \leq 0$ for some point $c \in I$, contradicting that $f'(x) > 0$ for all $x \in I$. ■

Corollary (Functions with Zero Derivative)**.** *If $f : I \to \mathbb{R}$ is differentiable over the open interval I and $f'(x) = 0$ for all $x \in I$, then f is constant over I.*

Proof. By the MVT, there exists $c \in I$ such that $(x - a)f'(c) = f(x) - f(a)$. But, $f'(c) = 0$ for all $c \in I$, so $f(x) - f(a) = 0$ and $f(x) = f(a)$. As the choice of x and a were arbitrary, letting $f(a) = k$, we have $f(x) = k$ for all $x \in I$. ■

Theorem (Extrema and Derivatives)**.** *Suppose $f : [a,b] \to \mathbb{R}$ is continuous and is differentiable over (a,b). Then, f attains its maximum and minimum at points within the open interval where $f' = 0$, or at one of the endpoints, a or b.*

Example. Find the maximum of xe^{-x} over $\mathbb{R}$.

The derivative of $f(x) = xe^{-x}$ is $f'(x) = (1-x)e^{-x}$. e^{-x} is positive for all x, and $(1-x)$ is positive if $x < 1$ and negative if $x > 1$, so the whole expression is positive if $x < 1$ and negative if $x > 1$. By the MVT, the function increases until $x = 1$ and then decreases, so the maximum is attained at $x = 1$, where $f(1) = e^-1$. △

34.13 Inverses

Theorem (Derivatives of Inverses)**.** *Let $f : (a,b) \to \mathbb{R}$ be differentiable with positive derivative. Then, $g = f^{-1}$ is differentiable, and,*

$$g'(x) = \frac{1}{f'(g(x))}$$

Proof. Since f has positive derivative, it is continuous and strictly increasing and hence has a continuous inverse. Let (c,d) be the range of the image of f, and let $x \in (c,d)$ and $g(x) = y$. Then,

$$g'(x) = \lim_{u \to x} \frac{g(u) - g(x)}{u - x}$$

Let $v = g(u)$. As $u \to x$, we have $g(u) \to g(x)$ as g is continuous at x, so $v \to y$, and,

$$\begin{aligned} &= \lim_{v \to y} \frac{v - y}{f(v) - f(y)} \\ &= \frac{1}{f'(y)} \\ &= \frac{1}{f'(g(x))} \end{aligned}$$

■

34.14 Power Series II

Theorem (Differentiability of Power Series I)**.** *Let $\sum_{n=0}^{\infty} a_n x^n$ be a power series with radius of convergence R. Then, the series $\sum_{n=0}^{\infty} n a_n x^{n-1}$ has the same radius of convergence.*

Proof. The series $\sum_{n=0}^{\infty} |a_n| x^n$ has the same radius of convergence by absolute series theorem. Let $0 < x < y < R$, so $\sum_{n=0}^{\infty} |a_n| x^n$ and $\sum_{n=0}^{\infty} |a_n| y^n$ both converge. It follows that their difference,

$$\sum_{n=0}^{\infty} |a_n| y^n - \sum_{n=0}^{\infty} |a_n| x^n = \sum_{n=0}^{\infty} |a_n| (y^n - x^n)$$

also converges, and thus

$$\sum_{n=0}^{\infty} |a_n| \frac{(y^n - x^n)}{y - x}$$

also converges. But

$$\sum_{n=0}^{\infty} |a_n| \frac{(y^n - x^n)}{y - x} = \sum_{n=1}^{\infty} |a_n| (y^{n-1} + y^{n-2} x + \cdots + x^{n-1})$$

Noting that $y > x$, we have,

$$
\begin{aligned}
&\geq \sum_{n=1}^{\infty} |a_n|(x^{n-1} + x^{n-2}x + \cdots + x^{n-1}) \\
&\geq \sum_{n=1}^{\infty} |a_n|(\underbrace{x^{n-1} + x^{n-1} + \cdots + x^{n-1}}_{n}) \\
&\geq \sum_{n=1}^{\infty} |a_n| n x^{n-1}
\end{aligned}
$$

so $\sum_{n=0}^{\infty} n|a_n|x^{n-1}$ also converges, further implying that $\sum_{n=0}^{\infty} na_n x^{n-1}$ converges (absolutely). ■

Theorem (Differentiability of Power Series II)**.** *Let $\sum_{n=0}^{\infty} a_n x^n$ be a power series with radius of convergence R. Then, the function, $x \mapsto \sum_{n=0}^{\infty} a_n x^n$ is continuous and differentiable over $(-R,R)$.*

Proof. Let $x \in (-R,R)$, and T such that $|x| < T < R$. It follows that $T \in (-R,R)$ so, by the theorem above, $\sum_{n=0}^{\infty} n|a_n|T^{n-1}$ converges, so for each $\varepsilon > 0$ there exists N for which,

$$
\sum_{n=N+1}^{\infty} n|a_n|T^{n-1} < \frac{\varepsilon}{3}
$$

Now, if $|y - x| < T - |x|$, $|y| < T$ and $x < T$, so,

$$
\left| \sum_{n=N+1}^{\infty} na_n x^{n-1} \right| \leq \sum_{n=N+1}^{\infty} n|a_n||x|^{n-1} < \frac{\varepsilon}{3}
$$

and also

$$
\begin{aligned}
\left| \sum_{n=N+1}^{\infty} a_n \frac{y^n - x^n}{y - x} \right| &= \left| \sum_{n=N+1}^{\infty} a_n(y^{n-1} + y^{n-2}x + \cdots + x^{n-1}) \right| \\
&\leq \sum_{n=N+1}^{\infty} |a_n|(|y|^{n+1} + \cdots + |x|^{n-1}) \\
&\leq \sum_{n=N+1}^{\infty} n|a_n|T^{n-1} \\
&< \frac{\varepsilon}{3}
\end{aligned}
$$

The finite sum,

$$
\sum_{n=1}^{N} a_n(y^{n-1} + y^{n-2}x + \cdots + x^{n-1})
$$

is a polynomial in y equal to $\sum_{n=1}^{N} na_n x^{n-1}$ when $y = x$, so there exists a $delta_0 > 0$ such that if $0 < |y - x| < \delta_0$,

$$
\begin{aligned}
\left| \sum_{n=1}^{N} a_n \frac{y^n - x^n}{y - x} - \sum_{n=1}^{N} na_n x^{n-1} \right| &= \left| \sum_{n=1}^{N} a_n(y^{n-1} + y^{n-2}x) + \cdots + x^{n-1} - \sum_{n=1}^{N} na_n x^{n-1} \right| \\
&< \frac{\varepsilon}{3}
\end{aligned}
$$

So, letting $\delta = \min(\delta_0, T - |x|)$, if $|y - x| < \delta$, we have,

$$\left|\sum_{n=1}^{\infty} a_n \frac{y^n - x^n}{y-x} - \sum_{n=1}^{\infty} n a_n x^{n-1}\right| \leq$$
$$\left|\sum_{n=N+1}^{\infty} a_n \frac{y^n - x^n}{y-x}\right| + \left|\sum_{n=1}^{N} a_n \frac{y^n - x^n}{y-x} - \sum_{n=1}^{N} n a_n x^{n-1}\right| + \left|\sum_{n=N+1}^{\infty} n a_n x^{n-1}\right| < \varepsilon$$

■

Theorem (Derivative of the Exponential)**.** *The derivative of the exponential function is equal to the exponential function.*

Proof. Exercise. ■

34.15 The Trigonometric Functions

We define the (*circular*) *trigonmetric functions* as follows:

$$\begin{aligned}\cos(x) &:= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n} \\ &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots \\ \sin(x) &:= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} \\ &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots\end{aligned}$$

Theorem (Addition Formulae)**.** *For all $x,y \in \mathbb{R}$,*

$$\cos(x \pm y) = \cos(x)\cos(y) \mp \sin(x)\sin(y)$$
$$\sin(x \pm y) = \sin(x)\cos(y) \pm \cos(x)\sin(y)$$

Proof. Let $f(x) = \cos(x)\cos(z-x) - \sin(x)\sin(z-x)$. $f'(x) = 0$, so $f(x)$ is constant by the MVT. When $x = 0$, $f(0) = \cos(0)\cos(z) - \sin(0)\sin(z) = \cos(z)$, so $f(x) = \cos(z)$ for all x. Letting $z = x + y$, we have $\cos(x)\cos(y) - \sin(x)\sin(y) = f(x) = \cos(z) = \cos(x + y)$.

The proof for $\sin(x + y)$ is similar. ■

Theorem (Circular Property)**.** *For all $x \in \mathbb{R}$,*

$$\cos^2(x) + \sin^2(y) = 1$$

Proof. In the addition formula for $cos(x + y)$, let $y = -x$. By the even and odd properties of $\cos(x)$ and $\sin(x)$, we have, $1 = \cos(0) = \cos(x - x) = \cos(x)\cos(-x) - \sin(x)\sin(-x) = \cos^2(x) + \sin^2(x)$. ■

34.16 Taylor's Theorem

Theorem (Cauchy's Mean Value Theorem)**.** *If $f,g : [a,b] \to \mathbb{R}$ are continuous and differentiable over (a,b), and $g'(t) \neq 0$ for $t \in (a,b)$, then there exists a point c such that,*

$$\frac{f'(c)}{g'(c)} = \frac{f(b) - f(a)}{g(b) - g(a)}$$

Proof. Consider $h(x) = f(x)[g(b) - g(a)] - g(x)[f(b) - f(a)]$.

$$\begin{aligned} h(a) &= f(a)g(b) - f(a)g(a) - f(b)g(a) + f(a)g(a) \\ &= f(a)g(b) - f(b)g(a) \\ h(b) &= f(b)g(b) - f(b)g(a) - f(b)g(b) + f(a)g(b) = f(a)g(b) - f(b)g(a) \\ &= f(a)g(b) - f(b)g(a) \\ &= h(a) \end{aligned}$$

So $h(a) = h(b)$, and, by Rolle's Theorem, there exists a point $c \in (a,b)$ such that $h'(c) = 0$, and

$$\begin{aligned} h'(x) &= f'(x)[g(b) - g(a)] - g'(x)[f(b) - f(a)] \\ h'(c) &= f'(c)[g(b) - g(a)] - g'(c)[f(b) - f(a)] \\ &= 0 \end{aligned}$$

so $f'(c)[g(b) - g(a)] = g'(c)[f(b) - f(a)]$. If $g(a) = g(b)$, then a stationary point of g would exist in (a,b) by Rolle's Theorem, but g' is given to be non-zero over (a,b), so $g(a) \neq g(b)$. We can then divide both sides of the equation by $g'(c)[g(b) - g(a)]$, obtaining the result. ■

Theorem (L'Hôpital's Rule)**.** *If $f,g : I \to \mathbb{R}$ are differentiable on the open interval I and $f(c) = g(c) = 0$ at some point $c \in I$, then*

$$\lim_{x \to c} \frac{f(x)}{g(x)} = \lim_{x \to c} \frac{f'(x)}{g'(x)}$$

provided the second limit exists.

Proof. Suppose that

$$\lim_{x \to c} \frac{f'(x)}{g'(x)}$$

does exist. Then, it cannot be that $g'(x) = 0$ at a sequence of points converging to c, so there is some interval around c on which g' is non-zero (except perhaps at c itself). So, g' is non-zero on an interval on each side of c. This allows us to apply Cauchy's mean value theorem. *Because*$f(c) = g(c) = 0$,

$$\begin{aligned} \lim_{x \to c} \frac{f(x)}{g(x)} &= \lim_{x \to c} \frac{f(x) - 0}{g(x) - 0} \\ &= \lim_{x \to c} \frac{f(x) - f(c)}{g(x) - g(c)} \end{aligned}$$

As long as x is in the region around c where $g' \neq 0$, Cauchy's mean value theorem ensures that there is a point t (depending on x) between c and x where

$$\frac{f(x) - f(c)}{g(x) - g(c)} = \frac{f'(t)}{g'(t)}$$

As $x \to c$, we have $t \to c$, so

$$\frac{f(x)}{g(x)} = \frac{f(x) - f(c)}{g(x) - g(c)} \to \lim_{t \to c} \frac{f'(t)}{g'(t)}$$

■

Note that L'Hôpital's rule cannot be used on expressions like $\lim_{x \to 0} \frac{\sin(x)}{x}$, because L'Hôpital's rule relies on derivatives, but $\lim_{x \to 0} \frac{\sin(x)}{x}$ is used in to find the derivative of sin in the first place: applying it here would be circular.

Theorem (Generalised L'Hôpital's Rule)**.** *If $f,g : \mathbb{R} \to \mathbb{R}$ are differentiable, and either*

- $\lim_{x\to\infty} f(x) = 0$ *and* $\lim_{x\to\infty} g(x) = 0$*;*
- *or* $\lim_{x\to\infty} f(x) = \infty$ *and* $\lim_{x\to\infty} g(x) = \infty$*;*

then,

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = \lim_{x\to\infty} \frac{f'(x)}{g'(x)}$$

provided the second limit exists.

L'Hôpital's rule can be applied to some other indeterminate forms, given an appropriate transformation:

Indeterminate form	Conditions	Transformation to 0/0
$\frac{0}{0}$	$\lim\limits_{x\to c} f(x) = 0,\ \lim\limits_{x\to 0} g(x) = 0$	–
$\frac{\infty}{\infty}$	$\lim\limits_{x\to c} f(x) = \infty,\ \lim\limits_{x\to 0} g(x) = \infty$	$\lim\limits_{x\to c} \frac{f(x)}{g(x)} = \lim\limits_{x\to c} \frac{1/f(x)}{1/g(x)}$
$0 \cdot \infty$	$\lim\limits_{x\to c} f(x) = 0,\ \lim\limits_{x\to 0} g(x) = \infty$	$\lim\limits_{x\to c} f(x)g(x) = \lim\limits_{x\to c} \frac{f(x)}{1/g(x)}$
$\infty - \infty$	$\lim\limits_{x\to c} f(x) = \infty,\ \lim\limits_{x\to 0} g(x) = \infty$	$\lim\limits_{x\to c} \big(f(x) - g(x)\big) = \lim\limits_{x\to c} \frac{1/g(x) - 1/f(x)}{1/\big(f(x)g(x)\big)}$
0^0	$\lim\limits_{x\to c} f(x) = 0^+,\ \lim\limits_{x\to 0} g(x) = 0$	$\lim\limits_{x\to c} f(x)^{g(x)} = \exp \lim\limits_{x\to c} \frac{g(x)}{1/\ln\big(f(x)\big)}$
1^∞	$\lim\limits_{x\to c} f(x) = 1,\ \lim\limits_{x\to 0} g(x) = \infty$	$\lim\limits_{x\to c} f(x)^{g(x)} = \exp \lim\limits_{x\to c} \frac{\ln\big(f(x)\big)}{1/g(x)}$
∞^0	$\lim\limits_{x\to c} f(x) = \infty,\ \lim\limits_{x\to 0} g(x) = 0$	$\lim\limits_{x\to c} f(x)^{g(x)} = \exp \lim\limits_{x\to c} \frac{g(x)}{1/\ln\big(f(x)\big)}$

34.16.1 Taylor's Theorem with Remainders

Theorem (Taylor's Theorem with Lagrange Remainder)**.** *If $f : I \to \mathbb{R}$ is n times differentiable on the open interval I, and x,$a \in I$, then,*

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \cdots + \frac{f^{(n-1)}(a)}{(n-1)!}(x-a)^{n-1} + \frac{f^{(n)}(t)}{n!}(x-a)^n$$

for some $t \in (x{,}a)$.

Proof. The function,

$$g(x) = f(x) - \left(f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \cdots + \frac{f^{(n-1)}(a)}{(n-1)!}(x-a)^{n-1}\right)$$

satisfies $g(a) = 0$, $g'(a) = 0$, $\cdots$, $g^{(n-1)}(a) = 0$, and $g^{(n)}(x) = f^{(n)}(x)$, because the bracket on the RHS goes to 0 after n differentiations.

If we let,

$$h(x) = g(x) - g(b)\frac{(x-a)^n}{(b-a)^n}$$

then the first $(n-1)$ derivatives of h also vanish at $x = a$, but we also have $h(b) = 0$.

Now, we proceed inductively. Since $h(n) = h(a) = 0$, there exists a point $t_1 \in (a{,}b)$ such that $h'(t_1) = 0$ by Rolle's Theorem. Since $h'(t_1) = h(a) = 0$, we again apply Rolle's Theorem so a there exists a point

$t_2 \in (a,t_1)$ such that $h''(t_2) = 0$. Repeating this process, we eventually find a point $t = t_n$ where $h^{(n)}(t) = 0$.

$$\begin{aligned} h^{(n)}(t) &= g^{(n)}(t) - g(b)\frac{n!}{(b-a)^n} \\ &= 0 \end{aligned}$$

So,

$$\begin{aligned} g^{(n)}(t) &= g(b)\frac{n!}{(b-a)^n} \\ g(b) &= \frac{g^{(n)}(t)}{n!}(b-a)^n \end{aligned}$$

Recalling our definition of $g(x)$, we have,

$$= \frac{f^{(n)}(t)}{n!}(b-a)^n$$

■

Theorem (Taylor's Theorem with Cauchy Remainder). *If $f : I \to \mathbb{R}$ is n times differentiable on the open interval I, and $x,a \in I$, then,*

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \cdots + \frac{f^{(n-1)}(a)}{(n-1)!}(x-a)^{n-1} + \frac{f^{(n)}(t)}{(n-1)!}(x-t)^{n-1}(x-a)$$

for some $t \in (x,a)$.

Proof. Let G be continuous over $[a,x]$ and differentiable over (a,x) with $G' \neq 0$, and let

$$F(t) = f(t) + f'(t)(x-t) + \frac{f''(t)}{2!}(x-t)^2 + \cdots + \frac{f^{(n)}(t)}{n!}(x-t)^n$$

for $t \in [a,x]$. Note that $F(x) = f(x)$ due to every term after the first having $(x-t)$ as a factor. Then, by Cauchy's MVT, there exists some $c \in (a,x)$ such that,

$$\frac{F'(c)}{G'(c)} = \frac{F(x) - F(a)}{G(x) - G(a)} \tag{$\star$}$$

Note that $F(x) - F(a) = R_n(x)$ is exactly the remainder term of the Taylor polynomial of $f(x)$.

$$\begin{aligned} F'(t) =& f'(t) + [f''(t)(x-t) - f'(t)] + \left[\frac{f^{(3)}(t)}{2!}(x-t)^2 - \frac{f^{(2)}(t)}{1!}(x-t)\right] + \cdots \\ &+ \left[\frac{f^{(n+1)}(t)}{n!}(x-t)^n - \frac{f^{(n)}(t)}{(n-1)!}(x-t)^(n-1)\right] \\ =& \frac{f^{(n+1)}(t)}{n!}(x-t)^n \end{aligned}$$

Substituting the above into $(\star)$, we have,

$$R_n(x) = \frac{f^{(n+1)}(c)}{n!}(c-x)^n \frac{G(x) - G(a)}{G'(c)}$$

If we let $G(t) = (x-t)^{k+1}$, we get the Lagrange form of the remainder. If we let $G(t) = t - a$, we get the Cauchy form of the remainder. ■

34.17 Riemann Integration

Given a function $f : [a,b] \to \mathbb{R}$, we can interpret the Riemann integral as the signed area enclosed between the graph of f and the x-axis.

We formalise this notion with the use of Darboux sums.

34.17.1 Partitions

We begin by introducing some terminology for intervals and partitions.

An interval $[a,b]$ is *non-trivial* if $a < b$. Two intervals I and J are *almost-disjoint* if they have at most one common point – that is, $|I \cap J| = 1$.

Let $I = [a,b]$ be a non-trivial closed interval over $\mathbb{R}$. A *partition* of I is a collection $\{I_1, \dots, I_n\}$ of almost-disjoint non-trivial intervals called *subintervals* with union $\bigcup_i I_i = I$.

Note that, because a partition must be almost-disjoint, but union to the total interval, it is entirely determined by the set of points $\{x_i\}_{i=0}^n$ satisfying

$$a = x_0 < x_1 < \cdots < x_n = b$$

corresponding to the endpoints of the component intervals.

Given a partition of $P = \{I_1, \dots, I_n\}$ of an interval $I = [a,b]$, we define the quantities:

$$M := \sup_I f \qquad m := \inf_I f$$
$$M_k := \sup_{I_k} f \qquad m_k := \inf_{I_k} f$$

Note that if f is unbounded, then some of these quantities will be infinite.

Given a function $f : [a,b] \to \mathbb{R}$ and a partition $P = \{I_1, \dots, I_n\}$ of $[a,b]$, we define the *upper Darboux sum* of f with respect to P as:

$$U(f,P) := \sum_{k=1}^{n} M_k |I_k|$$

and similarly, the *lower Darboux sum* of f with respect to P as:

$$L(f,P) := \sum_{k=1}^{n} m_k |I_k|$$

Intuitively, the upper (lower) Darboux sum under-approximates (resp. over-approximates) the area bounded by f and the x-axis by approximating the area A under the function over each subinterval I_k as a rectangle with height $\inf_{x \in I_k} f(x)$ (resp. sup).

This gives, by construction,

$$m(b-a) \le L(f,P) \le A \le U(f,P) \le M(b-a)$$

where the outer terms are the Darboux sums using the whole interval as a partition. If A fails to exist, then the inequality is simply

$$m(b-a) \le L(f,P) \le U(f,P) \le M(b-a)$$

Denote by $\mathscr{P}$ the set of all partitions of $[a,b]$. Then we define the *upper Darboux integral* of f by:

$$U(f) := \inf_{p \in \mathscr{P}} U(f,P)$$

and similarly, the *lower Darboux integral*

$$L(f) \coloneqq \sup_{p \in \mathscr{P}} L(f,P)$$

We say that a bounded function $f : [a,b] \to \mathbb{R}$ is *Darboux integrable* or *Riemann integrable*[*] if $U(f) = L(f)$, and we define the Riemann integral $\int_a^b f(x)\,dx$ by

$$\int_a^b f(x)\,dx \coloneqq U(f) = L(f)$$

noting that unbounded functions are not Riemann integrable by this definition, as one of the sums will be infinite.

34.17.2 Refinements

A partition $Q = \{J_1, \dots, J_\ell\}$ of $[a,b]$ is a *refinement* of a partition $P = \{I_1, \dots, I_n\}$ if every subinterval $I_k \in P$ is the union of intervals $J_k \in Q$.

Using our alternative characterisation of partitions as collection of interval endpoints, $Q = \{y_0, \dots, y_\ell\}$ is a refinement of $P = \{x_0, \dots, x_n\}$ if and only if $P \subseteq Q$.

Note that this means that every partition is a refinement of itself. It is also possible for neither of two partitions to be refinements of each other.

Theorem 34.17.1. *Let $f : I \to \mathbb{R}$ be a bounded function, and P,Q be partitions of I, with Q a refinement of P. Then,*

$$L(f,P) \leq L(f,Q) \leq U(f,Q) \leq U(f,P)$$

That is, refining a partition gives a better approximation to the desired area.

Theorem 34.17.2. *Let $f : I \to \mathbb{R}$ be a bounded function, and P,Q be arbitrary partitions of I. Then,*

$$L(f,P) \leq U(f,Q)$$

Corollary 34.17.2.1. *Let $f : I \to \mathbb{R}$ be a bounded function. Then,*

$$L(f) \leq U(f)$$

Theorem 34.17.3. *Let $f : I \to \mathbb{R}$ be a bounded function. Then, f is Riemann integrable if and only if for every $\varepsilon > 0$, there exists a partition P of I such that*

$$U(f,P) - L(f,P) < \varepsilon$$

We give an alternative characterisation of Riemann integrability, through the use of sequences.

Theorem 34.17.4. *Let $f : I \to \mathbb{R}$ be a bounded function. Then, f is Riemann integrable if and only if there exists a sequence of partitions P_n such that*

$$\lim_{n\to\infty} U(f,P_n) - L(f,P_n) < \varepsilon$$

[*] The above sums are sometimes called "Riemann sums", but general Riemann sums take the height of the function at arbitrary points within each subinterval, often the leftmost and rightmost points, defining the *left* and *right* Riemann sums, while Darboux sums take the infimum and supremum instead.

Unlike upper and lower Darboux sums, left and right Riemann sums do not obey a nice inequality, but in the limit, the two notions agree, and indeed, a function is Darboux integrable if and only if it is Riemann integrable, and the values of the two integrals agree whenever they exist.

To mark the distinction, and for consistency with most other sources, "Darboux" is used above when describing these sums, but due to their equivalence in the limit, we will continue to use "Riemann" when describing these integrals.

34.17.3 Continuity & Integrability

We recall that a function $f : I \to \mathbb{R}$ is *continuous* at $x \in I$ if for every $\varepsilon > 0$, there exists a $\delta(x,\varepsilon) > 0$ such that for all $y \in I$,

$$|x - y| < \delta \to |f(y) - f(x)| < \varepsilon$$

noting that we may only talk about one-sided continuity for the endpoints of I. Then, we say that f is continuous on I if f is continuous at every $x \in I$, with the case of endpoints understood as one-sided continuity.

Note that, in this definition, δ is a function of both x and ε. If we restrict δ to be a function of ε, we obtain the definition of *uniform continuity*:

Given a function $f : I \to \mathbb{R}$, we say f is *uniformly continuous* if for every $\varepsilon > 0$, there exists $\delta(\varepsilon) > 0$ such that for all $x,y \in I$, we have,

$$|x - y| < \delta \to |f(y) - f(x)| < \varepsilon$$

The difference here is that, in uniform continuity there is a globally applicable δ that depends on only ε, while in (ordinary) continuity there is only a locally applicable δ that depends on both ε and x. Thus, continuity is a local property of a function – that is, whether a function f is continuous or not at a particular point x can be determined by looking only at the values of f in an arbitrarily small neighbourhood of x. Conversely, uniform continuity is a global property of a function.

Uniform continuity is a stronger continuity condition than continuity: that is, a function that is uniformly continuous is continuous, but a function that is continuous is not necessarily uniformly continuous.

In particular, functions that are unbounded on a bounded domain cannot be uniformly continuous. For instance, the function $f : (0,1) \to \mathbb{R}$ defined by $x \mapsto \frac{1}{x}$ approaches infinity at an increasing rate as x approaches the origin, so it is not possible to find a δ independent of x that satisfies the definition of continuity.

Functions that have gradients that become unbounded on an infinite domain also cannot be uniformly continuous. For instance, $f : \mathbb{R} \to \mathbb{R}$ defined by $x \mapsto e^x$ is continuous everywhere, but its gradient becomes arbitrarily large, so it is possible to find arbitrarily small intervals in which f varies by more than ε.

Theorem 34.17.5. *Let I be a compact subset of $\mathbb{R}$ (i.e. a closed interval), and suppose $f : I \to \mathbb{R}$ is continuous. Then, f is uniformly continuous.*

We now give some sufficient (but not necessary) conditions for Riemann integrability.

Theorem 34.17.6. *If $f : [a,b] \to \mathbb{R}$ is continuous, then it is Riemann integrable.*

Theorem 34.17.7. *If $f : [a,b] \to \mathbb{R}$ is monotonic, then it is Riemann integrable.*

34.17.4 Algebra of Integrals

Theorem 34.17.8. *Let $f,g : [a,b] \to \mathbb{R}$ be Riemann integrable functions, and let $c \in \mathbb{R}$. Then, $f + g$ and cf are Riemann integrable, and satisfy,*

$$\int_a^b cf = c\int_a^b f, \qquad \int_a^b (f + g) = \int_a^b f + \int_a^b g$$

Theorem 34.17.9. *Let $f,g : [a,b] \to \mathbb{R}$ be integrable functions such that $f(x) \leq g(x)$ for all $x \in [a,b]$. Then,*

$$\int_a^b f \leq \int_a^b g$$

Corollary 34.17.9.1. *If $f : [a,b] \to \mathbb{R}$ is integrable, then,*

$$m(b-a) \leq \int_a^b f \leq M(b-a)$$

Corollary 34.17.9.2. *If $f : [a,b] \to \mathbb{R}$ is continuous, then there exists $c \in [a,b]$ such that*

$$f(c) = \frac{1}{b-a} \int_a^b f$$

Theorem 34.17.10. *If $f : [a,b] \to \mathbb{R}$ is integrable, then $|f|$ is integrable, and,*

$$\left| \int_a^b f \right| \leq \int_a^b |f|$$

Theorem 34.17.11. *Let $f : [a,b] \to \mathbb{R}$ and $c \in (a,b)$. Then, f is integrable on $[a,b]$ if and only if it is integrable on $[a,c]$ and $[c,b]$, and moreover,*

$$\int_a^c f + \int_c^b f = \int_a^b f$$

Theorem 34.17.12. *If $f : [a,b] \to \mathbb{R}$ is integrable and $g : \mathbb{R} \to \mathbb{R}$ is continuous, then $g \circ f$ is integrable.*

Note that the composition of two integrable functions is not necessarily integrable.

Theorem 34.17.13. *If $f,g : [a,b] \to \mathbb{R}$ are integrable, then the product function fg is integrable, and, if additionally $\frac{1}{g}$ is bounded, then $\frac{f}{g}$ is integrable.*

34.17.5 Fundamental Theorem of Calculus

The fundamental theorem of calculus links the notions of differentiation and integration together as inverses.

Theorem 34.17.14 (FTC I)**.** *Let $f : [a,b] \to \mathbb{R}$ be continuous, and define $F : [a,b] \to \mathbb{R}$ by*

$$F(x) = \int_a^x f(t)\,dt$$

Then, F is uniformly continuous on $[a,b]$ and differentiable on (a,b), with $F'(x) = f(x)$ for all $x \in (a,b)$, and we say that F is an antiderivative of f.

Equivalently,

$$\frac{d}{dx} \int_a^x f(t)\,dt = f(x)$$

Proof. We compute the derivative of $F(x)$ from the definition:

$$\begin{aligned} F'(x) &= \lim_{h \to 0} \frac{F(x+h) - F(x)}{h} \\ &= \lim_{h \to 0} \frac{1}{h} \left[\int_a^{x+h} f(t)\,dt - \int_a^x f(t)\,dt \right] \\ &= \lim_{h \to 0} \frac{1}{h} \int_x^{x+h} f(t)\,dt \end{aligned}$$

By the mean value theorem for integrals, there exists $c \in [x,x+h]$ such that $f(c) \cdot h = \int_x^{x+h} f(t)\,dt$, so,

$$= \lim_{h \to 0} f(c)$$

$c \in [x,x+h]$, so by the sandwich theorem,

$$= f(x)$$

■

Corollary 34.17.14.1. *Let $f : [a,b] \to \mathbb{R}$ be continuous with antiderivative F on $[a,b]$. Then,*

$$\int_a^b f(t)\,dt = F(b) - F(a)$$

Theorem 34.17.15 (FTC II). *Let $f : [a,b] \to \mathbb{R}$ be integrable on $[a,b]$ with continuous antiderivative F on (a,b). Then,*

$$\int_a^b f(x)\,dx = F(b) - f(a)$$

Unlike in the corollary above, FTC II does not require continuity of f over $[a,b]$, and is thus a slightly stronger result.

Proof. We wish to show

$$L(f,P) \leq F(b) - F(a) \leq U(f,P)$$

for every partition P of $[a,b]$. By taking a supremum on the left, and infimum on the right, we obtain $L(f) \leq F(b) - F(a) \leq U(f)$, and since f is integrable, both sides reduce to equalities.

Now, consider any partition $P = \{a = x_0,x_1,\dots,x_{n-1},x_n = b\}$. On every interval $I_k = [x_{k-1},x_k]$, for every $c_k \in (x_{k-1},x_k)$ we have,

$$\inf_{I_k} f(x)(x_k - x_{k-1}) \leq f(c_k)(x_k - x_{k-1}) \leq \sup_{I_k} f(x)(x_k - x_{k-1})$$

As F is continuous on $[x_{k-1},x_k]$ and differentiable on (x_{k-1},x_k), by the mean value theorem there exists c_k such that $F(x_k) - F(x_{k-1}) = f(c_k)(x_k - x_{k-1})$, so we have,

$$\inf_{I_k} f(x)(x_k - x_{k-1}) \leq F(x_k) - F(x_{k-1}) \leq \sup_{I_k} f(x)(x_k - x_{k-1})$$

Summing over $k = 1$ to n, we have,

$$L(f,P) \leq \sum_{k=1}^{n} F(x_k) - F(x_{k-1}) \leq U(f,P)$$

This sum telescopes to,

$$\begin{aligned} L(f,P) &\leq F(x_0) - F(x_n) \leq U(f,P) \\ L(f,P) &\leq F(b) - F(a) \quad \leq U(f,P) \end{aligned}$$

thus proving the result. ■

Theorem 34.17.16. *If $f : [a,b] \to \mathbb{R}$ is integrable on $[a,b]$ and is continuous from the right at a, then,*

$$\lim_{h\to 0^+} \frac{1}{h}\int_a^{a+h} f(t)\,dt = f(a)$$

and similarly, if f is continuous from the left at b,

$$\lim_{h\to 0^+} \frac{1}{h}\int_{b-h}^{b} f(t)\,dt = f(b)$$

More generally, if (I_h) is a sequence of intervals such that $|I_h| \to 0$, $x \in I_h$ for all h, and f is continuous at x, then,

$$\lim_{h\to 0} \frac{1}{|I_h|}\int_{I_h} f(t)\,dt = f(x)$$

Integration by parts and u-substitution are both consequences of the fundamental theorem of calculus:

Theorem 34.17.17 (IBP)**.** *If $f,g : [a,b] \to \mathbb{R}$ are continuous on $[a,b]$ and differentiable on (a,b) such that f',g' are integrable on $[a,b]$, then,*

$$\int_a^b f(x)g'(x)\,dx = f(b)g(b) - f(a)g(a) - \int_a^b f'(x)g(x)\,dx$$

Theorem 34.17.18 (u-sub)**.** *Let $f : [a,b] \to \mathbb{R}$ be differentiable on $[a,b]$ (understood as one-sided differentiability at the endpoints) such that f' is integrable on $[a,b]$, and let g be continuous on $f\big([a,b]\big)$. Then,*

$$\int_a^b g\big(f(x)\big)f'(x)\,dx = \int_{f(a)}^{f(b)} g(u)\,du$$

34.17.6 Improper Integration

So far, we have only defined Riemann integrals for bounded functions over bounded intervals. Now, we extend this definition to include unbounded functions and/or unbounded intervals using limits. This extension is called an *improper Riemann integral.*

Let $f : (a,b] \to \mathbb{R}$ be Riemann integrable over every interval $[c,b] \subset (a,b]$. Then, the improper intergral of f on $[a,b]$ is defined as,

$$\int_a^b f(x)\,dx := \lim_{\varepsilon\to 0^+}\int_{a+\varepsilon}^{b} f(x)\,dx$$

If this limit is finite, then the improper integral *converges*, *diverging* otherwise.

Similarly, if $f : [a,b) \to \mathbb{R}$ is integrable over every interval $[a,c] \subset [a,b)$, then the improper intergral of f on $[a,b]$ is defined as,

$$\int_a^b f(x)\,dx := \lim_{\varepsilon\to 0^+}\int_{a}^{b-\varepsilon} f(x)\,dx$$

We can also define an inproper integral if the function is unbounded at an interior point c.

Let $f : [a,b] \setminus \{c\} \to \mathbb{R}$ be a function integrable on any closed interval not containing $c \in [a,b]$. That is, f is integrable on $[a,c-\varepsilon_1]$ and $[c+\varepsilon_2,b]$ for all sufficiently small $\varepsilon_1,\varepsilon_2 > 0$. Then,

$$\int_a^b f(x)\,dx := \lim_{\varepsilon_1\to 0^+}\int_{a}^{c-\varepsilon_1} f(x)\,dx + \lim_{\varepsilon_2\to 0^+}\int_{c+\varepsilon_2}^{b} f(x)\,dx$$

For unbounded domains of integration, we take a limit of ordinary integrals:

If $f : [a,\infty) \to \mathbb{R}$ is integrable for every interval $[a,y] \subset [a,\infty)$, then,

$$\int_a^\infty f(x)\,dx := \lim_{y\to\infty} \int_a^y f(x)\,dx$$

Similarly, if $f : (-\infty,b] \to \mathbb{R}$ is integrable for every interval $[y,b] \subset (-\infty,b]$, then,

$$\int_{-\infty}^b f(x)\,dx := \lim_{y\to-\infty} \int_y^b f(x)\,dx$$

and if $f : \mathbb{R} \to \mathbb{R}$ is integrable on every bounded interval $[a,b]$, then,

$$\int_{-\infty}^\infty f(x)\,dx := \lim_{a\to-\infty} \int_a^c f(x)\,dx + \lim_{b\to\infty} \int_c^b f(x)\,dx$$

for any $c \in \mathbb{R}$.

The space of functions that are improperly Riemann integrable forms a linear space: that is, if f and g are improperly integrable on the same domain, then $\alpha f + \beta g$ is also improperly integrable over the same domain for any $\alpha,\beta \in \mathbb{R}$.

Theorem 34.17.19 (Absolute Comparison Test)**.** *Let $f : [a,\infty) \to \mathbb{R}$ be integrable on $[a,b]$ for every $b > a$. If $\int_a^\infty |f| < \infty$, then $\int_a^\infty f$ converges, and we say that $\int_a^\infty f$ is absolutely convergent.*

Moreover, if $g : [a,\infty) \to [0,\infty)$ is a function such that $|f| \leq g$ and $\int_a^\infty g < \infty$, then $\int_a^\infty f$ is absolutely convergent.

34.18 Sequences and Series of Functions

34.18.1 Convergence

Let $(f_n)_{n=0}^\infty$ be a sequence of functions $f_n : \Omega \to \mathbb{R}$. We say that (f_n) *converges pointwise* to $f : \Omega \to \mathbb{R}$ if for every $x \in \Omega$,

$$\lim_{n\to\infty} f_n(x) = f(x)$$

and we denote this relation by $f_n \to f$.

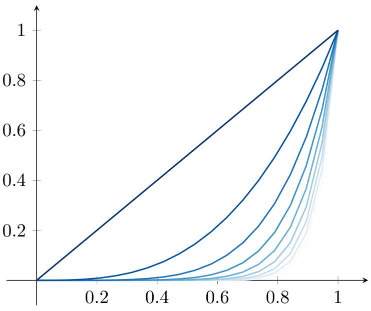
Intuitively, a sequence (f_n) of functions converges pointwise to a function f if, when we fix any choice of input value x, the resulting sequence of output terms $\big(f_n(x)\big)_{n=0}^\infty$ (which is just a sequence of real numbers) converges to the output value $f(x)$ in the usual sense.

Note that pointwise limit do not preserve continuity. That is, the pointwise limit of a sequence of continuous functions is not necessarily continuous.

For instance, consider the sequence of monomials

$$f_n(x) = x^n$$

defined over $[0,1]$.



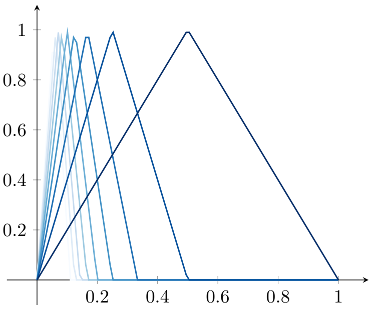
Being polynomials, each f_n is continuous, but the pointwise limit of the sequence is

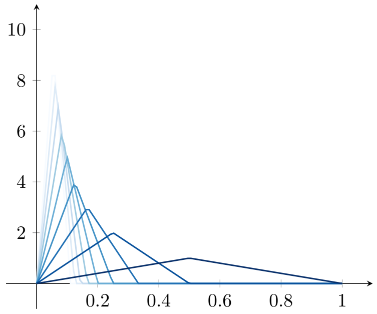
$$f(x) = \begin{cases} 0 & x \in [0,1) \\ 1 & x = 1 \end{cases}$$

which is discontinuous at $x = 1$.

Pointwise convergence is also very non-uniform in the sense that it is possible for $f_n(x) \to 0$ for all x, but $\sup_x |f_n(x) - f(x)| \to C > 0$, or even $\sup_x |f_n(x) - f(x)| \to \infty$ as $n \to \infty$, as in the next two examples:

$$g_n(x) = \begin{cases} 2nx & x \in [0,\frac{1}{2n}) \\ -2n\left(x - \frac{1}{n}\right) & x \in [\frac{1}{2n},\frac{1}{n}) \\ 0 & x \in [\frac{1}{n},1] \end{cases} \qquad h_n(x) = \begin{cases} 2n^2x & x \in [0,\frac{1}{2n}) \\ -2n^2\left(x - \frac{1}{n}\right) & x \in [\frac{1}{2n},\frac{1}{n}) \\ 0 & x \in [\frac{1}{n},1] \end{cases}$$





This latter sequence (h_n) is also called the *witch's hat*, and is often useful for providing counterexamples.

Every g_n and h_n are continuous piecewise linear functions, and both converge to the zero function $f = 0$. However, for every n, we have $g_n\left(\frac{1}{2n}\right) = 1$, so

$$\sup_{x \in [0,1]} |g_n(x) - 0| = 1$$

Similarly, for every n, we have $h_n\left(\frac{1}{2n}\right) = n$, so

$$\sup_{x \in [0,1]} |h_n(x) - 0| = \infty$$

Pointwise limits and integrals also do not interact well: even if the pointwise limit of a sequence of functions is integrable, we do not necessarily have $\lim \int f_n = \int \lim f_n$.

For instance, consider $f_n(x) = \chi_{[n,n+1)}(x)$, where χ_I is the indicator function of the set I. Clearly, f_n converges pointwise to $f = 0$, but

$$1 = \int_{-\infty}^{\infty} f_n \neq \int_{-\infty}^{\infty} f = 0$$

In a sense, the mass of the function “escapes to infinity” along the x-axis.

Similarly, $g_n(x) = n\chi_{(0,1/n)}(x)$ also converges pointwise to the zero function, but $\int_{-\infty}^{\infty} g_n = 1$ for all n, this time the mass escaping along the y-axis. Integrating the witch’s hat also provides a continuous example of this case.

Let $(f_n)_{n=0}^{\infty}$ be a sequence of functions $f_n : \Omega \to \mathbb{R}$. We say that (f_n) *converges uniformly* to $f : \Omega \to \mathbb{R}$ if for any $\varepsilon > 0$, there exists $N(\varepsilon)$ such that $|f_n(x) - f(x)| < \varepsilon$ for every $x \in \Omega$ and every $n > N(\varepsilon)$, and we denote this relation by $f_n \rightrightarrows f$.

Uniform convergence is to pointwise convergence what uniform continuity is to ordinary continuity: in uniform convergence, N depends only on ε, and not on x, while in pointwise continuity, we began by fixing a value of x.

To simplify notation, we define the ℓ^∞, *supremum* or *Chebyshev* norm by:

$$\|f\|_\infty \coloneqq \sup_{x \in \Omega} |f(x)|$$

Using this, we can simplify the definition of uniform convergence to:

$$f_n \rightrightarrows f \coloneqq \forall\varepsilon > 0, \exists N(\varepsilon), \forall n > N(e) : \|f_n - f\|_\infty < \varepsilon.$$

Theorem 34.18.1. *Uniform convergence implies pointwise convergence, but not the converse.*

A sequence (f_n) of functions in Ω is *uniformly Cauchy* if for every $\varepsilon > 0$, there exists $N(\varepsilon)$ such that $\|f_n - f_m\|_\infty < \varepsilon$ for all $n,m > N(\varepsilon)$.

Theorem 34.18.2. *A sequence (f_n) of functions is uniformly convergent if and only if it is uniformly Cauchy.*

Theorem 34.18.3. *If a sequence of continuous functions (f_n) in Ω converges uniformly to a function $f : \Omega \to \mathbb{R}$, then f is continuous.*

The space of bounded continuous functions on a space Ω is denoted $C_b(\Omega)$.

Theorem 34.18.4. *$(C_b(\Omega), \|\cdot\|_\infty)$ is a complete space: that is, every Cauchy sequence converges to a continuous bounded function, etc.*

Theorem 34.18.5. *Let (f_n) be a sequence of Riemann integrable functions $f_n : [a,b] \to \mathbb{R}$ that converges uniformly to a function $f : [a,b] \to \mathbb{R}$. Then, f is Riemann integrable and $\int f_n \to \int f$.*

Uniform convergence and differentiation do not interact as nicely. There are examples of sequences of differentiable functions (f_n), with (f_n) converging uniformly to f, but (f_n') does not converge to f' (or f' may fail to exist). This also does not hold even if the sequence is of infinitely differentiable functions.

34.18.2 Multivariate Continuity

We now introduce definitions of (uniform) continuity of functions defined over subsets of $\mathbb{R}^2$.

We write $C^k(\Omega)$ to denote the space of functions that are k times continuously differentiable over Ω, and $C^\infty(\Omega)$ for the space of functions infinitely differentiable over Ω, also called functions that are *smooth* over Ω.

A function $f : \Omega \subseteq \mathbb{R}^2 \to \mathbb{R}$ is continuous at $x \in \Omega$ if for every $\varepsilon > 0$, there exists $\delta(x,\varepsilon) > 0$ such that for all $y \in \Omega$,

$$\|x - y\| < \delta \to |f(y) - f(x)| < \varepsilon$$

Similarly, a function $f : (\Omega \subseteq \mathbb{R}^2) \to \mathbb{R}$ is uniformly continuous if for every $\varepsilon > 0$, there exists $\delta(\varepsilon) > 0$ such that for all $x,y \in \Omega$,

$$\|x - y\| < \delta \to |f(y) - f(x)| < \varepsilon$$

and again, the difference here is that δ is independent of x.

Theorem 34.18.6. *Let $\Omega \subset \mathbb{R}^2$ be closed and bounded. Then, any continuous function $f : \Omega \to \mathbb{R}$ is furthermore uniformly continuous.*

Theorem 34.18.7. *Let $f : [a,b] \times [c,d] \to \mathbb{R}$ be continuous. Define $I : [c,d] \to \mathbb{R}$ by*

$$I(t) := \int_a^b f(x,t)\,dx$$

Then, I is continuous.

Theorem 34.18.8 (Leibniz Integral Rule)**.** *Let $f, \frac{\partial f}{\partial t}$ be continuous functions on $[a,b] \times [c,d]$. Then, for any $t \in (c,d)$,*

$$\frac{d}{dt}\int_a^b f(x,t)\,dx = \int_a^b \frac{\partial f}{\partial t}(x,t)\,dx$$

Theorem 34.18.9 (Fubini's Theorem for Continuous Functions)**.** *Let $f : [a,b] \times [c,d] \to \mathbb{R}$ be continuous. Then,*

$$\int_a^b \int_c^d f(x,y)\,dy\,dx = \int_c^d \int_a^b f(x,y)\,dx\,dy$$

Theorem 34.18.10. *Let (f_n) be a sequence of C^1 functions on $[a,b]$, and suppose $f_n \to f$ (pointwise), and $f' \rightrightarrows g$ (uniformly). Then, f is C^1 and $g = f'$ (that is, $f_n' \rightrightarrows f'$).*

34.18.3 Series

We now define the notions of pointwise and uniform convergence for series of functions.

Let (f_k) be a sequence of functions $f_k : \Omega \to \mathbb{R}$, and let (S_n) be the sequence of partial sums of (f_k), with $S_n : \Omega \to \mathbb{R}$ defined by

$$S_n(x) = \sum_{k=1}^{n} f_k(x)$$

Then, the series

$$\sum_{k=1}^{\infty} f_k(x)$$

is said to converge pointwise to $S : \Omega \to \mathbb{R}$ in Ω if $S_n \to S$ pointwise in Ω, and to converge uniformly to S in Ω if $S_n \rightrightarrows S$ uniformly on Ω.

Theorem 34.18.11. *If (f_k) is a series of integrable functions $f_k : [a,b] \to \mathbb{R}$, and S_n converges uniformly, then $\sum_{k=1}^{\infty} f_k$ is Riemann integrable, and,*

$$\int \sum_{k=1}^{\infty} f_k = \sum_{k=1}^{\infty} \int f_k$$

Theorem 34.18.12. *Let (f_k) be a sequence of C^1 functions $f_k : [a,b] \to \mathbb{R}$ such that S_n converges pointwise, and suppose that $\sum_{k=1}^{n} f_k'$ converges uniformly. Then,*

$$\left(\sum_{k=1}^{\infty} f_k\right)' = \sum_{k=1}^{\infty} f_k'$$

That is, the series is differentiable and can be differentiated term-by-term.

Theorem 34.18.13 (Weierstrass M-test)**.** *Let (f_k) be a sequence of functions $f_k : \Omega \to \mathbb{R}$, and suppose that exists a sequence (M_k) of non-negative reals such that*

- *$|f_k(x)| \leq M_k$ for all $k \in \mathbb{N}$ and all $x \in \Omega$;*
- *$\sum_{k=1}^{\infty} M_k$ converges.*

Then, the series $\sum_{k=1}^{\infty} f_n(x)$ converges absolutely and uniformly on Ω.

Proof. We show that the partial sums $S_n = \sum_{k=1}^{n} f_k(x)$ is uniformly Cauchy. Now, since $\sum_{k=1}^{\infty} M_k$ converges, given $\varepsilon > 0$, there exists N such that

$$\sum_{k=m+1}^{n} M_k < \varepsilon$$

for all $m,n > N$. Now,

$$\begin{aligned}
|S_n(x) - S_m(x)| &= \left|\sum_{k=1}^{n} f_k(x) - \sum_{k=1}^{m} f_k(x)\right| \\
&= \left|\sum_{k=m+1}^{n} f_k(x)\right| \\
&\leq \sum_{k=m+1}^{n} |f_k(x)| \\
&\leq \sum_{k=m+1}^{n} M_k \\
&< \varepsilon
\end{aligned}$$

■

34.19 Complex Analysis

We quickly revisit some basic properties of the complex numbers.

The set of complex numbers $\mathbb{C}$ is given by

$$\mathbb{C} = \{x + iy : x,y \in \mathbb{R}\}$$

where i is the imaginary unit, satisfying $i^2 = -1$.

For a complex number $z = x + iy$, we denote

- the real component of z by $\Re(z) = x$;
- the complex component of z by $\Im(z) = y$;
- the modulus or norm of z by $|z| = \sqrt{x^2 + y^2}$;

- the complex conjugate of z by $\bar{z} = x - iy$.

Theorem 34.19.1. *The following statements hold for any complex numbers $z,w \in \mathbb{C}$.*

- $\bar{\bar{z}} = z$;
- $\overline{z+w} = \bar{z} + \bar{w}$;
- $\overline{zw} = \bar{z}\bar{w}$;
- $|\bar{z}| = |z|$;
- $|z|^2 = z\bar{z}$

A sequence $(z_n)_{n=1}^{\infty} \subset \mathbb{C}$ converges to a complex number $z \in \mathbb{C}$ if $\lim_{n\to\infty} |z_n - z| = 0$. That is, if for every $\varepsilon > 0$, there exists $N > 0$ such that $|z_n - z| < \varepsilon$ for all $n > N$.

A set $\Omega \subseteq \mathbb{C}$ is *open* if for every $x \in \Omega$, there exists $r > 0$ such that $\mathbb{B}_r(x) \subset \Omega$, and a set $\Omega \subseteq \mathbb{C}$ is closed if $\Omega^c = \mathbb{C} \setminus \Omega$ is open.

A set $K \subset \mathbb{C}$ is *sequentially compact* if every sequence $(x_j)_{j=1}^{\infty} \subset K$ has a convergent subsequence $(x_{j(\ell)})_{\ell=1}^{\infty}$ whose limit is in K.

A function $f : (\Omega \subseteq \mathbb{C}) \to \mathbb{C}$ is continuous at $z_0 \in \Omega$ if for every $\varepsilon > 0$, there exists a $\delta(x,\varepsilon) > 0$ such that for all $z \in \Omega$,

$$|z - z_0| < \delta \to |f(z) - f(z_0)| < \varepsilon$$

This definition is identical to that of continuity for real functions, but with $|\cdot|$ now being a norm on $\mathbb{C}$ rather than $\mathbb{R}$, and in fact, coincides with the definition of continuity for functions $\mathbb{R}^2 \to \mathbb{R}^2$.

34.19.1 Complex Differentiability

Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is differentiable at a point p if the limit

$$\lim_{h\to 0} \frac{f(p+h) - f(p)}{h}$$

exists, and this limit is the value of the derivative.

In contrast, a function $f : \mathbb{R}^n \to \mathbb{R}^k$ is differentiable at a point p if there exists a linear map $Df \in L(\mathbb{R}^n; \mathbb{R}^k)$ such that

$$\lim_{h\to 0} \frac{|f(p+h) - f(p) - Df(p)h|}{|h|} = 0$$

and this linear map Df is the value of the derivative.

We use this definition because for $k > 1$, there is no well-defined notion of division of vectors.

However, unlike in $\mathbb{R}^2$, $\mathbb{C}$ does have a notion of division we can use, so we can return to the original definition of differentiability, and so, differentiability for functions $\mathbb{C} \to \mathbb{C}$ is distinct from (and in many ways, more well-behaved than) functions $\mathbb{R}^2 \to \mathbb{R}^2$.

Let $\Omega \subseteq \mathbb{C}$ be an open set. A function $f : \mathbb{C} \to \mathbb{C}$ is *complex differentiable* at a point $z_0 \in \Omega$ if the limit

$$\lim_{h\to 0} \frac{f(z_0+h) - f(z_0)}{h}$$

exists, and this limit is the value of the derivative.

However, here, h is a complex number, so there are many ways we could send h to 0. If this limit exists, then its value should be independent of the path taken. We will write $f(x,y)$ as $u(x,y) + iv(x,y)$ to separate out the real and imaginary components.

Now, consider approaching along the real axis. We have,

$$\lim_{\substack{h\to 0\\ h\in\mathbb{R}}} \frac{f(z_0+h)-f(z_0)}{h} = \frac{\partial f}{\partial x}(z_0)$$

while approaching along the imaginary axis gives,

$$\lim_{\substack{h\to 0\\ h\in\mathbb{C}}} \frac{f(z_0+h)-f(z_0)}{h} = \frac{1}{i}\frac{\partial f}{\partial y}(z_0)$$

These values should be equal, and so,

$$\begin{aligned} i\frac{\partial f}{\partial x} &= \frac{\partial f}{\partial y} \\ i\left(\frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x}\right) &= \frac{\partial u}{\partial y} + i\frac{\partial v}{\partial y} \\ -\frac{\partial v}{\partial x} + i\frac{\partial u}{\partial x} &= \frac{\partial u}{\partial y} + i\frac{\partial v}{\partial y} \end{aligned}$$

Equating the real and imaginary components, we have,

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$$

or more compactly,

$$u_x = v_y, \qquad u_y = -v_x$$

These are the *Cauchy-Riemann equations.* For a complex derivative to exist, these equations must be satisfied.

Moreover, if $f : \mathbb{C} \to \mathbb{C}$ is a function that is differentiable when regarded as a function $f : \mathbb{R}^2 \to \mathbb{R}$, then f is complex differentiable if and only if the Cauchy-Riemann equations hold.

This means that if the components u and v are real-differentiable functions of two real variables, then $u + iv$ is a complex-valued real-differentiable function, and is furthermore complex-differentiable if and only if the Cauchy-Riemann equations hold. We can also replace the requirement that u and v are differentiable with the requirement that all partial derivatives of u and v are continuous (as this implies that u and v are real-differentiable).

Example. Consider the function $f : \mathbb{C} \to \mathbb{C}$ defined by $z \mapsto z^2$. u and v are clearly continuous, so f is real-differentiable.

$$\begin{aligned} f(z) &= (x+iy)^2 \\ &= x^2 - y^2 + 2ixy \end{aligned}$$

so,

$$u(x,y) = x^2 - y^2, \quad v(x,y) = 2xy$$

with partial derivatives

$$u_x = 2x, \quad u_y = -2y, \quad v_x = 2y, \quad v_y = 2x$$

satisfying the Cauchy-Riemann equations, so f is also complex-differentiable. △

A function $f : \Omega \to \mathbb{C}$ is *holomorphic* at $z_0 \in \Omega$ if there exists a neighbourhood $U \subseteq \Omega$ of z_0 such that f is complex-differentiable at all $z \in U$.

f is holomorphic in Ω if f is holomorphic at all $z \in \Omega$, and we say that f is *entire* if it is holomorphic on the whole of $\mathbb{C}$.

A general function $f : A \to B$ is *analytic* at a point if it is given locally by a convergent power series at that point. That is, f is analytic at x_0 if the Taylor series centred at x_0 converges pointwise to $f(x)$ for every x in a neighbourhood $U \subseteq B$. Note that a function may be complex-differentiable at a point, but not necessarily analytic.

Earlier, we mentioned that complex functions are sometimes more well-behaved than real functions; it turns out that a complex-valued function is analytic if and only if it is holomorphic, so the terms are sometimes used interchangably in the context of complex analysis.

Theorem 34.19.2 (Algebra of Complex Derivatives)**.** *Let $f,g : \Omega \subseteq \mathbb{C} \to \mathbb{C}$ be complex-differentiable functions. Then,*

$$(f+g)' = f' + g', \qquad (fg)' = f'g + fg', \qquad \left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}, \qquad (f(g))' = f'(g)g'$$

(assuming that $g \neq 0$ in the third expression, and that the domains and codomains are appropriate in the fourth).

Theorem 34.19.3. *The function $f : \mathbb{C} \to \mathbb{C}$ defined by $z \mapsto z^n$ is entire for all $n \in \mathbb{N}$, and $f'(z) = nz^{n-1}$.*

34.19.2 Power Series

We define the notions of convergence of series in $\mathbb{C}$ similarly to that of series in $\mathbb{R}$.

Let $(a_n)_{n=0}^{\infty}$ be a sequence of complex numbers $a_n \in \mathbb{C}$. The series $\sum_{n=0}^{\infty} a_n$ is *convergent* if the sequence of partial sums $S_k = \sum_{n=0}^{k} a_n$ is convergent in C, and is *absolutely convergent* if the series $\sum_{n=0}^{\infty} |a_n|$ is convergent in C.

The geometric series $\sum_{n=0}^{\infty} a_n$ is absolutely convergent if and only if $|z| < 1$ with limit

$$\sum_{n=0}^{\infty} z^n = \frac{1}{1-z}$$

and partial sums

$$S_k = \frac{1 - z^{k+1}}{1-z}$$

Theorem 34.19.4 (Ratio Test)**.** *Let $(a_n)_{n=0}^{\infty}$ be a sequence of complex numbers $a_n \in \mathbb{C}$ with $a_n \neq 0$ for all sufficiently large n, and define the quantity,*

$$L := \lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right|$$

Then,

- *if $L < 1$, then the series $\sum_{n=0}^{\infty} a_n$ converges absolutely;*
- *if $L > 1$, then the series $\sum_{n=0}^{\infty} a_n$ diverges;*
- *if $L = 1$ or the limit fails to exist, then the test is inconclusive.*

Using suprema and infima, we can strengthen this test: define the quantities,

$$R := \limsup \left|\frac{a_{n+1}}{a_n}\right|, \qquad r := \liminf \left|\frac{a_{n+1}}{a_n}\right|$$

Then,

- *if $R < 1$, then the series $\sum_{n=0}^{\infty} a_n$ converges absolutely;*
- *if $r > 1$, then the series $\sum_{n=0}^{\infty} a_n$ diverges;*
- *if $\left|\frac{a_{n+1}}{a_n}\right| > 1$ for all sufficiently large n, then the series $\sum_{n=0}^{\infty} a_n$ also diverges;*
- *otherwise, the test is inconclusive.*

Theorem 34.19.5. *Consider $\sum_{n=0}^{\infty} a_n$ and define the quantity,*

$$r := \limsup \sqrt[n]{|a_n|}$$

- *If $r < 1$, then the series converges;*
- *If $r > 1$, then the series diverges;*
- *If $r = 1$, then the test is inconclusive.*

The root test is stronger than the ratio test: whenever the ratio test determines the convergence or divergence of an infinite series, the root test does too, but not the converse.

Theorem 34.19.6. *Given any sequence $(a_n)_{n=0}^{\infty}$, there exists $R \in [0,\infty]$ such that*

$$\sum_{n=0}^{\infty} a_n z^n$$

converges for all $|z| < R$ and diverges for $|z| > R$.

More specifically, this value is given by,

$$R = \frac{1}{\limsup \sqrt[n]{a_n}}$$

Theorem 34.19.7. *Let $a_n \neq 0$ for all $n \geq N$ and suppose that $\lim_{n\to 0} \frac{|a_{n+1}|}{|a_n|}$ exists. Then, $\sum_{n=0}^{\infty} a_n z^n$ has radius of convergence,*

$$R = \lim_{n\to\infty} \frac{|a_n|}{|a_{n+1}|}$$

Theorem 34.19.8. *Suppose a series $\sum_{n=0}^{\infty} a_n z^n$ has radius of convergence R. Then, for all $|z| < R$, the function $f(z) = \sum_{n=0}^{\infty} a_n z^n$ is differentiable and,*

$$f'(z) = \sum_{n=1}^{\infty} n a_n z^{n-1}$$

That is, the derivative may be computed term by term.

Corollary 34.19.8.1. *Let $\sum_{n=0}^{\infty} a_n z^n$ be a power series with radius of convergence $R > 0$. Then, the function $f(z) = \sum_{n=0}^{\infty} a_n z^n$ is smooth (infinitely differentiable), and moreover,*

$$\frac{f^{(n)}(0)}{n!} = a_n$$

for all $n \in \mathbb{N}_0$.

Theorem 34.19.9. *Let $\sum_{n=0}^{\infty} a_n z^n$ be a power series with radius of convergence $R > 0$. Then, for every $r < R$, the sequence of functions,*

$$f_k := \sum_{n=0}^{k} a_n z^n$$

converges uniformly in $|z| \leq r$.

34.19.3 The Complex Exponential

In this section, we will write $\exp(z)$ instead of e^z to emphasise that these power series are definitions and not theorems, unlike the case for the real exponential.

We define the following power series for $z \in \mathbb{C}$.

$$\begin{aligned}
\exp(z) &:= \sum_{n=0}^{\infty} \frac{1}{n!} z^n \\
&= 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots \\
\cos(z) &:= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} z^{2n} \\
&= 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \frac{z^6}{6!} + \cdots \\
\cosh(z) &:= \sum_{n=0}^{\infty} \frac{1}{(2n)!} z^{2n} \\
&= 1 + \frac{z^2}{2!} + \frac{z^4}{4!} + \frac{z^6}{6!} + \cdots \\
\sin(z) &:= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} \\
&= z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \cdots \\
\sinh(z) &:= \sum_{n=0}^{\infty} \frac{1}{(2n+1)!} z^{2n+1} \\
&= z + \frac{z^3}{3!} + \frac{z^5}{5!} + \frac{z^7}{7!} + \cdots
\end{aligned}$$

These functions are entire, converging for any $z \in \mathbb{C}$.

Theorem 34.19.10. *The following identities hold for all $z \in \mathbb{C}$:*

$$\begin{aligned}
\cos(z) &= \frac{\exp(iz) + \exp(-iz)}{2}, & \cosh(z) &= \frac{\exp(z) + \exp(-z)}{2}, \\
\sin(z) &= \frac{\exp(iz) - \exp(-iz)}{2i}, & \sinh(z) &= \frac{\exp(z) - \exp(-z)}{2}
\end{aligned}$$

$$\cos(iz) = \cosh(z), \qquad \cosh(iz) = \cos(z), \qquad \sin(iz) = i \sinh(z), \qquad \sinh(iz) = i \sin(z)$$

Theorem 34.19.11. *The complex exponential function $\exp(z)$ satisfies the following:*

- *(Characteristic Property of the Exponential)* $\exp(z + w) = \exp(z)\exp(w)$ *for all $z,w \in \mathbb{C}$,* $\exp(1) = e$*;*

- $\exp(z) \neq 0$ *for all* $z \in \mathbb{C}$*;*
- $\exp(z) = 1$ *if and only if* $z = 2k\pi i$ *with* $k \in \mathbb{Z}$*;*
- $\exp(z) = -1$ *if and only if* $z = (2k+1)\pi i$ *with* $k \in \mathbb{Z}$*.*

The third property implies that $\exp(z+w) = \exp(z)$ if and only if $w = 2k\pi i$, $k \in \mathbb{Z}$, so the exponential function is periodic along the imaginary axis with period 2π.

34.19.4 The Complex Logarithm

Every complex number $z = x + iy \in \mathbb{C} \setminus \{0\}$ can be written as $re^{i\theta}$, where r is the *modulus* of z, $|z|$, and θ is the *phase* of z – the angle that the vector rooted at the origin pointing to z makes with the positive real axis, measured counterclockwise. Note that for $z = 0$, this angle is undefined, and in any other case, is unique only up to factors of 2π.

We define the multivalued *argument* function $\arg : \mathbb{C} \setminus \{0\} \to \mathcal{P}(\mathbb{R})$ by

$$\arg(z) = \{\theta \in \mathbb{R} : z = |z|e^{iz}\}$$

The argument function is not a function in the usual sense as the image of each input is not uniquely defined: in particular, if $\theta \in \arg(z)$, then $\theta + 2k\pi \in \arg(z)$ for all $k \in \mathbb{Z}$.

Theorem 34.19.12. *The argument function* $\arg(z)$ *satisfies the following:*

- $\arg(\alpha z) = \arg(z)$ *for all real* $\alpha > 0$*;*
- $\arg(\alpha z) = \arg(z) + \pi = \{\theta + \pi : \theta \in \arg(z)\}$ *for all real* $\alpha < 0$*;*
- $\arg(\bar{z}) = -\arg(z) = \{-\theta : \theta \in \arg(z)\}$*;*
- $\arg\left(\frac{1}{z}\right) = -\arg(z)$*;*
- $\arg(zw) = \arg(z) + \arg(w) = \{\theta + \phi : \theta \in \arg(z), \phi \in \arg(w)\}$*.*

We define the *principle value argument* function $\mathrm{Arg} : \mathbb{C} \setminus \{0\} \to (-\pi,\pi]$ by taking the angle in $\arg(z)$ that lies in the interval $(-\pi,\pi]$. Then, we have $\arg(z) = \{\mathrm{Arg}(z) + 2k\pi : k \in \mathbb{Z}\}$.

Note that the Arg function is not continuous in the entire complex plane. In particular, approaching the negative real axis from the clockwise direction yields $-\pi$, while approaching from the counterclockwise direction yields π. Making any other choice for the image of Arg leads to a similar issue along the half-line where we define the ends of the image, where the arguments will differ by 2π when approaching from different directions.

We wish to define an extension of the logarithm to the complex numbers, and to mark the distinction, we will write ln to denote the ordinary real logarithm in $\mathbb{R}_{\geq 0}$, and log to denote our complex extension. One defining characteristic of the real logarithm is that $x = \ln(y)$ if and only if $e^x = y$ – that is, the real logarithm is the inverse of the real exponential.

Since $e^z = e^{z+2k\pi i}$ for any $k \in \mathbb{Z}$, then if $w = \log(z)$, then so is $w + 2k\pi i$, so the complex logarithm is also multivalued.

Let $z,w \in \mathbb{C}$ such that $w = \log(z) = u + iv$. Then, we have,

$$\begin{aligned} z &= e^{\ln(z)} \\ &= e^w \\ &= e^{u+iv} \\ &= (e^u)e^{iv} \end{aligned}$$

But, $z = |z|e^{i\arg(z)}$, so, equating the modulus and argument, we have $e^u = |z|$, and $v = \arg(z)$, with the modulus equation in particular implying that $u = \ln|z|$.

We define the multivalued complex logarithm $\log : \mathbb{C} \setminus \{0\} \to \mathcal{P}(\mathbb{R})$ by

$$\log(z) := \ln|z| + i\arg(z)$$

again noting that $\log(z)$ is undefined for $z = 0$, as $\ln|z| = \ln(0)$ is undefined.

Theorem 34.19.13. *The complex logarithm function* $\log(z)$ *satisfies the following:*

- $\log(zw) = \log(z) + \log(w) \pmod{2\pi i}$*;*
- $\log\left(\frac{z}{w}\right) = \log(z) - \log(w) \pmod{2\pi i}$*;*
- $\exp(\log(z)) = z$*;*
- $\log(\exp(z)) = z \pmod{2\pi i}$*.*

We define the *principle branch logarithm* $\operatorname{Log} : \mathbb{C} \setminus \{0\} \to \mathbb{R}_{\geq 0}$ by

$$\operatorname{Log}(z) := \ln|z| + i\operatorname{Arg}(z)$$

Because the Arg function is discontinuous along the half-line $x \leq 0$, the Log function is also discontinuous along the same line: if we consider points $z = x + i\varepsilon$ for $x < 0$ and sufficiently small $\varepsilon > 0$, we have,

$$\lim_{\varepsilon \to 0} \operatorname{Log}(x \pm i\varepsilon) = \operatorname{Log}|x| \pm i\pi$$

so the function cannot be extended continuously along $\{x \leq 0\}$. This half-line is called a *branch cut*, and any definition of the principle value argument function results in such a half-line.

From the identity,

$$e^{\operatorname{Log}(z)} = z$$

we have,

$$e^{\operatorname{Log}(z)}(\operatorname{Log}(z))' = 1$$

and hence,

$$(\operatorname{Log}(z))' = \frac{1}{z}$$

With the complex extension of the natural logarithm, we can now define complex powers of complex numbers. Given $\alpha, z \in \mathbb{C}$ with $z \neq 0$, we define,

$$\begin{aligned} z^\alpha &:= e^{\operatorname{Log}(z^\alpha)} \\ &= e^{\alpha \operatorname{Log}(z)} \\ &= e^{\alpha \ln|z| + \alpha i \arg(z)} \\ &= e^{\alpha \ln|z| + \alpha i \operatorname{Arg}(z) + \alpha 2ki\pi} \\ &= e^{\alpha \ln|z| + \alpha i \operatorname{Arg}(z)} e^{\alpha 2ki\pi} \end{aligned}$$

where $k \in \mathbb{Z}$, and we see that complex powers can be multivalued. Specfically, if α is an integer, then $e^{\alpha 2ki\pi} = 1$, so there is only one value of z^α. If $\alpha = \frac{p}{q}$ is rational with $p \in \mathbb{Z}, q \in \mathbb{N}$ coprime, then

$$e^{\alpha 2ki\pi} = e^{\alpha 2(k+q)i\pi}$$

and z^α will assume q distinct values. If α is irrational, then z^α will take infinitely many values.

34.19.5 Complex Integration

For a function $f : [a,b] \to \mathbb{C}$, we define,

$$\int_a^b f(t)\,dt := \int_a^b \Re\big(f(t)\big)\,dt + i\int_a^b \Im\big(f(t)\big)\,dt$$

So, integrating a complex-valued function reduces to integrating two real-valued functions.

Theorem 34.19.14. *For every $f,g : [a,b] \to \mathbb{C}$ and every $\alpha,\beta \in \mathbb{C}$, we have,*

$$\int_a^b \alpha f(t) + \beta g(t)\,dt = \alpha\int_a^b f(t)\,dt + \beta\int_a^b g(t)\,dt$$

Theorem 34.19.15. *For any function $f : [a,b] \to \mathbb{C}$,*

$$\overline{\int_a^b f(t)\,dt} = \int_a^b \overline{f(t)}\,dt$$

$$\left|\int_a^b f(t)\,dt\right| \le \int_a^b \big|f(t)\big|\,dt$$

34.19.6 Contour Integrals

The previous definition of an integral is a natural extension of real integration for integrating functions $\mathbb{R} \to \mathbb{C}$, but what would it mean to integrate a function $\mathbb{C} \to \mathbb{C}$? Single integrals only make sense when evaluated along 1 dimensional curves, so there is no natural extension in this case.

Because of this, we will only consider integrals of complex-valued functions *along curves* in the complex plane called *contours*:

$$\int_\Gamma f\,dz$$

where Γ is a contour in $\mathbb{C}$. To evaluate such an integral, we begin by parametrising Γ by a function $\gamma : [a,b] \to \mathbb{C}$ given by $\gamma(t) = x(t) + iy(t)$. We will also require that γ is C^1, as we will require a well-defined tangent at every point of the curve.

Given a function $f : \Omega \subseteq \mathbb{C} \to \mathbb{C}$ and a contour $\Gamma \subset \Omega \subseteq \mathbb{C}$ parametrised by $\gamma : [a,b] \to \mathbb{C}$, the *contour integral* of f over Γ is given by:

$$\begin{aligned}\int_\Gamma f\,dz &:= \int_a^b f\big(\gamma(t)\big)\gamma'(t)\,dt \\ &= \int_a^b \Re\Big(f\big(\gamma(t)\big)\gamma'(t)\Big)\,dt + \int_a^b \Im\Big(f\big(\gamma(t)\big)\gamma'(t)\Big)\,dt\end{aligned}$$

If Γ is only piecewise C^1, then we define,

$$\int_\Gamma f\,dz := \sum_{i=1}^n \int_{\Gamma_i} f\,dz$$

where $(\Gamma_i)_{i=1}^n$ are the C^1 components of Γ.

Theorem 34.19.16. *Let $f : \Omega \subseteq \mathbb{C} \to \mathbb{C}$ and $\Gamma \subset \Omega$ such that $f\big|_\Gamma$ is continuous, and parametrise Γ by $\gamma^+ : [a,b] \to \mathbb{C}$. Then,*

- *If γ^- represents the parametrisation of Γ in the opposite direction from γ^+, then,*

$$\int_{\gamma^-} f\,dz = -\int_{\gamma^+} f\,dz$$

If Γ has an attached notion of direction or orientation , we call it a directed curve or directed contour. In this case, we denote by $-\Gamma$ the same curve swept in the opposite direction, so we may reformulate the above result without reference to any particular parametrisation by:

$$\int_{-\Gamma} f\,dz = -\int_{\Gamma} f\,dz$$

- *If $\tilde{\gamma} : [\tilde{a},\tilde{b}] \to \mathbb{C}$ is another parametrisation of Γ that preserves orientation, then,*

$$\int_{\tilde{\gamma}} f\,dz = \int_{\gamma} f\,dz$$

This property is called reparametrisation invariance.

Given a function $f : \mathbb{C} \to \mathbb{C}$ and a curve parametrised by $\gamma : [a,b] \to \mathbb{C}$, we define,

$$\int_{\gamma} f\,d\bar{z} := \int_a^b f\big(\gamma(t)\big)\overline{\gamma'(t)}\,dt$$

Note that, unlike for functions $f : [a,b] \to \mathbb{C}$, in general, for contour integrals,

$$\overline{\int_{\gamma} f(z)\,dz} \neq \int_{\gamma} \overline{f(z)}\,dz$$

We instead have,

$$\overline{\int_{\gamma} f(z)\,dz} = \int_{\gamma} \overline{f(z)}\,d\bar{z}$$

Given a function $f : \mathbb{C} \to \mathbb{C}$ and a curve parametrised by $\gamma : [a,b] \to \mathbb{C}$, we define,

$$\int_{\gamma} |f||dz| := \int_a^b \Big|f\big(\gamma(t)\big)\Big|\big|\gamma'(t)\big|\,dt$$

Note that $\int_{\gamma} |f||dz| \geq 0$, so we have,

$$\left|\int_{\gamma} f\,dz\right| \leq \int_{\gamma} |f||dz|$$

If $f(z) = 1$, then we also have,

$$\int_{\gamma} |dz| = L(\gamma)$$

where $L(\gamma)$ is the length of γ.

Theorem 34.19.17. *Suppose that Ω is an open set, and $F : \Omega \subseteq \mathbb{C} \to \mathbb{C}$ is holomorphic, such that $f(z) := \frac{dF}{dz}$ is continous. Let $\gamma : [a,b] \to \Omega$ be a C^1 curve. Then,*

$$\int_{\gamma} f\,dz = F\big(\gamma(b)\big) - F\big(\gamma(a)\big)$$

A set $\Omega \subseteq \mathbb{C}$ is *connected* if it cannot be expressed as the union of non-empty open sets Ω_1 and Ω_2 such that $\Omega_1 \cap \Omega_2 = \varnothing$.

An open connected set $\Omega \subseteq \mathbb{C}$ is *simply connected* if every closed curve in Ω can be continuously deformed to a point (more precisely, every closed curve is homotopic to a constant function,§38.9.2).

An example of a set that is not simply connected is an annulus (a 2D torus; the region bounded by two circles of differing radii centred on the same point): any closed curve that encircles the hole cannot be continuous deformed into a point as it must always encircle the hole.

Theorem (Cauchy). *Let Ω be non-empty, open, and simply connected, and let $\gamma \subset \Omega$ be a continuous closed curve. If $f : \Omega \to \mathbb{C}$ is holomorphic, then,*

$$\int_\gamma f(z)\,dz = 0$$

This theorem says that the integral vanishes around regions where the integrand is holomorphic.

Theorem (Deformation of Contours). *Let $\Omega \subseteq \mathbb{C}$ be non-empty, open, and simply connected. Let $f : \Omega \to \mathbb{C}$ be holomorphic, and let $\gamma_1, \gamma_2 : [a,b] \to \mathbb{C}$ be simple regular piecewise C^1 paths in Ω with $\gamma_1(a) = \gamma_2(a)$ and $\gamma_1(b) = \gamma_2(b)$. Then,*

$$\int_{\gamma_1} f(z)\,dz = \int_{\gamma_2} f(z)\,dz$$

Informally, this means that contours may be deformed over regions where the integrand is holomorphic, and this does not change the value of the integral along the contour.

A parametrisation of a simple closed curve is *positively oriented* if, when following the direction of parametrisation, the interior is to our left, and is *negatively oriented* otherwise. For example, the counterclockwise parametrisation of the unit circle given by $\gamma(t) = (\cos(t), \sin(t))$ is positively oriented.

However, take an annulus, for example. This region has two boundary curves; an *exterior* and *interior* boundary. The exterior boundary is positively oriented if it has a counterclockwise parametrisation, but the interior boundary is positively oriented if it has a clockwise parametrisation.

Corollary 34.19.17.1. *Let $\Omega \subseteq \mathbb{C}$ be a region bounded by two simple curves, γ_1 exterior and γ_2 interior, both oriented positively, and let f be a function holomorphic over $\Omega \cup \gamma_1 \cup \gamma_2$. Then,*

$$\int_{\gamma_1} f(z)\,dz + \int_{\gamma_2} f(z)\,dz = 0$$

If we denote by γ_2^- the counterclockwise parametrisation of γ_2, then,

$$\int_{\gamma_1} f(z)\,dz = \int_{\gamma_2^-} f(z)\,dz$$

That is, the integral is the same along both curves when both are parametrised in the same direction.

Given a simple closed C^1 curve γ, we denote by $I(\gamma)$ the region interior to γ, and by $E(\gamma)$ the region exterior to γ.

Theorem (Cauchy's Integral Formula). *Let $\gamma : [a,b] \to \mathbb{C}$ be a positively oriented simple closed C^1 curve, and suppose g is a function holomorphic over $\gamma \cup I(\gamma)$. Then, for all $z_0 \in I(\gamma)$,*

$$g(z_0) = \frac{1}{2\pi i} \int_\gamma \frac{g(z)}{z - z_0}\,dz$$

Proof. Fix $z_0 \in I(\gamma)$ and choose r sufficiently small such that $B_r(z_0) \subset I(\gamma)$. By deformation of contours, we have,

$$\frac{1}{2\pi i}\int_\gamma \frac{g(z)}{z-z_0}\,dw = \frac{1}{2\pi i}\int_{\partial B_r(z_0)} \frac{g(z_0)}{z-z_0}\,dz$$

reducing the problem to considering γ as $\partial B_r(z_0)$. Notice that this integral is the same for every r sufficiently small. For now, we have,

$$\begin{aligned}\frac{1}{2\pi i}\int_{\partial B_r(z_0)} \frac{g(z)}{z-z_0}\,dz &= \frac{1}{2\pi i}\int_{\partial B_r(z_0)} \frac{g(z_0)+g(z)-g(z_0)}{z-z_0}\,dz \\ &= \frac{1}{2\pi i}\int_{\partial B_r(z_0)} \frac{g(z_0)}{z-z_0}\,dz + \frac{1}{2\pi i}\int_{\partial B_r(z_0)} \frac{f(z)-f(z_0)}{z-z_0}\,dz\end{aligned}$$

Denote the first integral by I, and the second integral by J. Note that $I = g(z_0)$ since

$$\frac{1}{2\pi i}\int_{\partial B_r(z_0)} \frac{g(z_0)}{z-z_0}\,dw = g(z_0)\frac{1}{2\pi i}\int_{\partial B_r(z_0)} \frac{1}{z-z_0}\,dz = g(z_0)$$

It remains to show that $J = 0$. Since g is holomorphic in $I(\gamma)$, given any $\varepsilon > 0$, there exists r sufficiently small such that

$$\left|g(z) - g(z_0)\right| \leq \varepsilon$$

for all $z \in \partial B_r(z_0)$. Parametrise $\partial B_r(z_0)$ counterclockwise by $\gamma(t) = z_0 + re^{it}$ for $t \in [0,2\pi)$. Then, we have $\gamma'(t) = ire^{it}$ and therefore

$$\begin{aligned}|J| &= \left|\frac{1}{2\pi i}\int_{\partial B_r(z_0)} \frac{g(z)-g(z_0)}{z-z_0}\,dz\right| \\ &\leq \left|\frac{1}{2\pi i}\int_0^{2\pi} \frac{g(z_0+re^{it})-g(z_0)}{re^{it}} ire^{it}\,dt\right| \\ &\leq \frac{1}{2\pi}\int_0^{2\pi} \left|g(z_0+re^{it})-g(z_0)\right|\,dt \\ &\leq \varepsilon\end{aligned}$$

Since ε is arbitrary, we obtain the result. ■

This theorem says that we can recover the value of g at any point z by integrating along a closed curve around that point, provided the curve is sufficiently regular, positively oriented, and contained in $I(\gamma)$. This is a very significant difference when compared to smooth functions in $\mathbb{R}^2$, for example.

Note that since the curve γ is a compact set, for any point $z_0 \in I(\gamma)$, the expression $z - z_0$ found in the denominator is bounded away from zero, indicating that we can differentiate the formula with respect to z_0 to obtain

$$g'(z_0) = \frac{1}{2\pi i}\int_\gamma \frac{g(z)}{(z-z_0)^2}\,dz$$

Of course, this requires justifying moving the derivative inside the integral. We have assumed that g is holomorphic, so $g'(z_0)$ exists. This expression would then give a way of computing this derivative. Notably, the right side can be differentiated arbitrarily many times, even though we have only assumed that g is differentiable once: this formula implies that every holomorphic function is smooth. We will show that this formula indeed holds:

Theorem 34.19.18. *Let $\gamma : [a,b] \to \mathbb{C}$ be a positively oriented simple closed C^1 curve, and suppose g is a function holomorphic over $\gamma \cup I(\gamma)$. Then, for all $z_0 \in I(\gamma)$, g is smooth (infinitely differentiable), and the nth derivative is given by*

$$g^{(n)}(z) = \frac{n!}{2\pi i}\int_\gamma \frac{g(z)}{(z-z_0)^{n+1}}\,dz$$

Proof. Note that the Cauchy's integral formula corresponds to the case $n = 0$. To prove the result for $n = 1$, we use Cauchy's integral formula on the difference quotient

$$\frac{g(z_0+h)-g(z_0)}{h} = \frac{1}{h}\left(\frac{1}{2\pi i}\int_\gamma \frac{g(z)}{z-z_0-h}\,dz - \frac{1}{2\pi i}\int_\gamma \frac{g(z)}{z-z_0}\,dz\right)$$

By deformation of contours, we can choose γ as $\partial B_{2r}(z_0)$ with $B_{2r}(z_0) \subset I(\gamma)$. We have, operating on the right side,

$$\begin{aligned}\frac{g(z_0+h)-g(z_0)}{h} &= \frac{1}{2\pi i}\int_{\partial B_{2r}(z_0)} \frac{g(z)}{(z-z_0-h)(z-z_0)}\,dz \\ &= \frac{1}{2\pi i}\int_{\partial B_{2r}(z_0)} \frac{g(z)}{(z-z_0)^2}\,dz + \frac{1}{2\pi i}\int_{\partial B_{2r}(z_0)} g(z)\left(\frac{1}{(z-z_0-h)(z-z_0)} - \frac{1}{(z-z_0)^2}\right)dz \\ &= \frac{1}{2\pi i}\int_{\partial B_{2r}(z_0)} \frac{g(z)}{(z-z_0)^2}\,dz + \frac{1}{2\pi i}\int_{\partial B_{2r}(z_0)} \frac{hg(z)}{(z-z_0-h)(z-z_0)^2}\,dz\end{aligned}$$

Denote the integral on the right by I. The task is now to prove that $I \to 0$ as $h \to 0$.

Recall that we may choose r arbitrarily small without affecting the values of the integrals above. Choose $|h| < r$ such that for all $z \in B_{2r}(z_0)$, we have

$$\begin{aligned}|z-z_0-h| &\geq |z-z_0| - |h| \\ &> r\end{aligned}$$

where the first inequality is from the reverse triangle inequality, and the second is from the fact that $|z - z_0| = 2r$ for any point $z \in \partial B_{2r}(z_0)$. Parametrising $\partial B_{2r}(z_0)$ with $\gamma(t) = z_0 + 2re^{it}$ for $t \in [0,2\pi)$, we have $\gamma'(t) = 2rie^{it}$, and therefore $\left|\gamma'(t)\right| \leq 2r$. Since g is holomorphic, in particular, it is continuous, and therefore there exists $M > 0$ such that $\left|g(z)\right| \leq M$ for all $z \in \partial B_{2r}(z_0)$. From this, we have

$$\begin{aligned}\left|\int_{\partial B_{2r}(z_0)} \frac{hg(z)}{(z-z_0-h)(z-z_0)^2}\,dz\right| &\leq \int_0^{2\pi} \frac{hM}{r(2r)^2}2r\,dt \\ &= \frac{\pi M}{r^2}h\end{aligned}$$

which tends to 0 as $h \to 0$, proving the result for $n = 1$. Now, we induct on n. Assume the result holds for all values less than some fixed arbitary n. Then, writing the incremential quotient for $(n-1)$, we have

$$\frac{g^{(n-1)}(z_0+h)-g^{(n-1)}(z_0)}{h} = \frac{1}{h}\left(\frac{(n-1)!}{2\pi i}\int_\gamma \frac{g(z)}{(z-z_0-h)^n}\,dz - \frac{(n-1)!}{2\pi i}\int_\gamma \frac{g(z)}{(z-z_0)^n}\,dz\right)$$

By deformation of contours, we can choose γ as $\partial B_{2r}(z_0)$ with $B_{2r}(z_0) \subset I(\gamma)$. We have, operating on the right side,

$$\begin{aligned}\frac{g^{(n-1)}(z_0+h)-g^{(n-1)}(z_0)}{h} &= \frac{(n-1)!}{2\pi ih}\int_{\partial B_{2r}(z_0)} \frac{g(z)\big((z-z_0)^n-(z-z_0-h)^n\big)}{(z-z_0-h)^n(z-z_0)^n}\,dz \\ &= \frac{n!}{2\pi i}\int_{\partial B_{2r}(z_0)} \frac{g(z)}{(z-z_0)^{n+1}}\,dz \\ &\quad + \frac{(n-1)!}{2\pi i}\int_{\partial B_{2r}(z_0)} g(z)\left(\frac{\big((z-z_0)^n-(z-z_0-h)^n\big)}{h(z-z_0-h)^n(z-z_0)^n} - \frac{n}{(z-z_0)^2}\right)dz \\ &= \frac{n!}{2\pi i}\int_{\partial B_{2r}(z_0)} \frac{g(z)}{(z-z_0)^{n+1}}\,dz\end{aligned}$$

$$+\frac{(n-1)!}{2\pi i}\int_{\partial B_{2r}(z_0)} g(z)\frac{(z-z_0)^{n+1}-(z-z_0-h)^n(z-z_0)-nh(z-z_0-h)^n}{h(z-z_0-h)^n(z-z_0)^{n+1}}\,dz$$

Denote the integral on the right by I. As before, the task is to prove that $I \to 0$ as $h \to 0$. Choose h and the parametrisation as above.

If we can show that

$$\left|\frac{(z-z_0)^{n+1}-(z-z_0-h)^n(z-z_0)-nh(z-z_0-h)^n}{h}\right| \leq C|h| \qquad (*)$$

where C is a constant possibly depending on r, then the result will follow, since, as before $|g| \leq M$ and $|z-z_0-h| \geq |z-z_0| - |h| > r$ implies

$$\left|\frac{1}{(z-z_0-h)^n(z-z_0)^n}\right| \leq \frac{1}{(2r)^n r^n}$$

So, to prove $(*)$, note that the binomial theorem gives

$$(z-z_0-h)^n = \sum_{j=0}^{n}\binom{n}{j}(z-z_0)^{n-j}(-h)^j$$

and therefore

$$\begin{aligned}(z-z_0)^{n+1}-(z-z_0-h)^n(z-z_0)-nh(z-z_0-h)^n &= -\sum_{j=2}^{n}\binom{n}{j}(z-z_0)^{n+1-j}(-h)^j \\ &\quad - nh\sum_{j=1}^{n}(z-z_0)^{n-j}(-h)^n\end{aligned}$$

which is of order h^2, proving the result. ■

These theorems are often helpful in reverse for evaluating contour integrals.

Theorem (Taylor Series Expansion). *Let f be holomorphic on $\mathbb{B}_r(a)$ for $a \in \mathbb{C}$, $r > 0$. Then, there exist unique constants c_n, $n \in \mathbb{N}$ such that, fo all $z \in \mathbb{B}_r(a)$*

$$f(z) = \sum_{n=0}^{\infty} c_n(z-a)^n$$

That is, a holomorphic function is analytic.

Moreover, the coefficients c_n are given by

$$\begin{aligned}a_n &= \frac{1}{2\pi i}\int_\gamma \frac{f(w)}{(w-a)^{n+1}}\,dw \\ &= \frac{f^{(n)}(a)}{n!}\end{aligned}$$

where γ is any positively oriented parametrisation of a simple closed curve $\Gamma \subset \mathbb{B}_r(a)$ that is piecewise C^1 with $a \in I(\gamma)$.

A function f defined on a subset of $\mathbb{C}$ is said to have a *pole* of order $m \in \mathbb{N}$ at $a \in \mathbb{C}$ if there is a neighbourhood U of a such that for any $z \in U$,

$$f(z) = \frac{c_{-m}}{(z-a)^m} + \frac{c_{m-1}}{(z-a)^{m-1}} + \cdots + \frac{c_{-2}}{(z-a)^2} + \frac{c_{-1}}{(z-a)} + \phi(z)$$

where ϕ is analytic in U, $(c_{-k})_{k=1}^m \subset \mathbb{C}$, and $c_{-m} \neq 0$. The coefficient c_{-1} is called the *residue* of f at a, also denoted $\operatorname{Res}\bigl(f(a)\bigr)$. This expansion is also called a *Laurent polynomial.*

A function that is holomorphic at all points of an open subset $\Omega \subseteq \mathbb{C}$ apart from some poles is said to be *meromorphic* on Ω.

Theorem (Cauchy's Residue Theorem)**.** *Let $\gamma \subset \mathbb{C}$ be a simple closed positively oriented piecewise C^1 curve, and let f be meromorphic on $I(\gamma)$ with poles $(z_k)_{k=1}^n \subset I(\gamma)$. Then,*

$$\int_\gamma f = 2\pi i \sum_{k=1}^{n} \operatorname{Res}\bigl(f(z_k)\bigr)$$

Lemma 34.19.19. *Let $f,g : U \to \mathbb{C}$ be holomorphic on an open neighbourhood U of $a \in \mathbb{C}$, and suppose $g(a) = 0$, but $g'(a) \neq 0$. Then, provided $f(a) \neq 0$, the function $\frac{f}{g}$ has a pole of order 1 at a, and,*

$$\operatorname{Res}\left(\frac{f}{g}(a)\right) = \frac{f(a)}{g'(a)}$$

34.19.7 Examples of Contour Integration

Example (Integrating around one pole)*.* Let

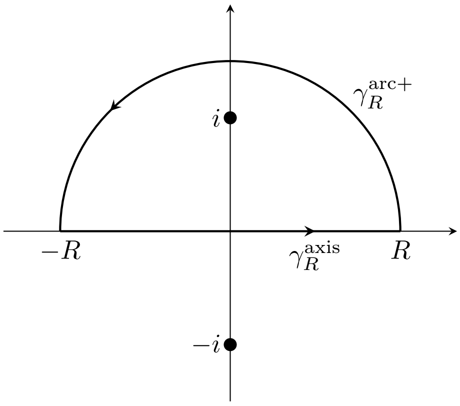
$$I = \int_{-\infty}^{\infty} \frac{1}{x^2+1}\, dz$$

Denote the integrand by $f(z) = \frac{1}{z^2+1}$, which factorises to

$$f(z) = \frac{1}{(z-i)(z+i)}$$

and we can see that f has poles at i and $-i$, and is analytic elsewhere.

For $R > 0$, we construct a positively oriented contour γ_R consisting of a line segment along the real axis from $-R$ to R, then closing the contour with a semicircular arc in the complex plane:



This contour can be decomposed into the line segment and the arc and parametrised separately by paths $\gamma_R^{\text{axis}} : [-R,R] \to \mathbb{C}$ and $\gamma_R^{\text{arc}+} : [0,\pi] \to \mathbb{C}$ given by

$$\gamma_R^{\text{axis}}(t) = t$$
$$\gamma_R^{\text{arc}+}(t) = Re^{it}$$

Since I is absolutely convergent, we have

$$\begin{aligned} I &= \lim_{R\to\infty} \int_{\gamma_R^{\text{axis}}} f(z)\,dz \\ &= \lim_{R\to\infty} \left(\oint_{\gamma_R} f(z)\,dz - \int_{\gamma_R^{\text{arc}+}} f(z)\,dz \right) \end{aligned}$$

We have

$$\oint_{\gamma_R} f(z)\,dz = \int_{\gamma_R^{\text{axis}}} f(z)\,dz + \int_{\gamma_R^{\text{arc}+}} f(z)\,dz$$

By deformation of contours,

$$\int_{\gamma_R} f(z)\,dz = \int_{\partial B_1(i)} f(z)\,dz$$

for all $R > 0$, and Cauchy's integral formula with $g(z) = \frac{1}{z+i}$ yields

$$\begin{aligned} \oint_{\partial B_1(i)} f(z)\,dz &= \oint_{\partial B_1(i)} \frac{1}{(z-i)(z+i)}\,dz \\ &= \oint_{\partial B_1(i)} \frac{\frac{1}{z+i}}{z-i}\,dz \\ &= 2\pi i \cdot \left.\frac{1}{z+i}\right|_{z=i} \\ &= \pi \end{aligned}$$

Or alternatively, the residue at i is given by

$$\begin{aligned} f(z) &= \frac{1/(z+i)}{(z-i)} \\ \operatorname{Res}[f(z)]_{z=i} &= \frac{1}{z+i} \\ &= \frac{1}{2i} \end{aligned}$$

so the residue theorem gives

$$\begin{aligned} \oint_{\partial B_1(i)} f(z)\,dz &= 2\pi i \operatorname{Res}[f(z)]_{z=i} \\ &= 2\pi i \frac{1}{2i} \\ &= \pi \end{aligned}$$

Next,

$$\begin{aligned} \int_{\gamma_R^{\text{arc}+}} \frac{1}{z^2+1}\,dz &= \int_0^\pi \frac{1}{(Re^{i\theta})^2+1} \frac{dz}{d\theta}\,d\theta \\ &= \int_0^\pi \frac{1}{R^2e^{i2\theta}+1} iRe^{i\theta}\,d\theta \\ &= \int_0^\pi \frac{iRe^{i\theta}}{R^2e^{i2\theta}+1}\,d\theta \end{aligned}$$

$$\begin{aligned}\left|\int_{\gamma_R^{\text{arc}+}} \frac{1}{z^2+1}\,dz\right| &\leq \int_0^\pi \left|\frac{iRe^{i\theta}}{R^2e^{i2\theta}+1}\right| d\theta \\ &= \int_0^\pi \frac{R}{|Re^{i2\theta}+1|}\,d\theta \\ &\leq \int_0^\pi \frac{R}{|R^2-1|}\,d\theta \\ &= \pi\frac{R}{|R^2-1|}\end{aligned}$$

so the integral along the arc vanishes as $R \to \infty$.

So, we have

$$\begin{aligned} I &= \lim_{R\to\infty} \int_{\gamma_R^{\text{axis}}} f(z)\,dz \\ &= \lim_{R\to\infty}\left(\oint_{\gamma_R} f(z)\,dz - \int_{\gamma_R^{\text{arc}+}} f(z)\,dz\right) \\ &= \pi \end{aligned}$$

△

In this example, it would not have mattered if we chose to close the contour in the lower half of the plane rather than the upper half. That is, if we instead defined $\gamma_R^{\text{arc}-} : [0,\pi] \to \mathbb{C}$ by $\gamma_R^{\text{arc}-}(t) = Re^{-it}$, then almost identical reasoning yields the same answer.

This is not always the case.

Example (Choosing the correct contour). Let

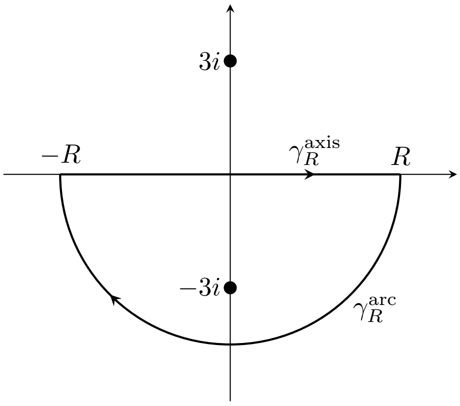
$$I = \int_{-\infty}^{\infty} \frac{e^{ix}}{x^2+9}\,dz$$

Denote the integrand by $f(z) = \frac{e^{iz}}{z^2+9}$, which factorises to

$$f(z) = \frac{e^{iz}}{(z-3i)(z+3i)}$$

so f has poles at $3i$ and $-3i$, and is analytic elsewhere.

Construct a negatively oriented contour γ_R consisting of a line segment along the real axis from $-R$ to R, then closing the contour with a semicircular arc in lower half-plane:



We have $\lim_{R\to\infty}\int_{\gamma_R^{\text{arc}-}} f(z)\,dz = 0$ since

$$\begin{aligned}
\lim_{R\to\infty}\left|\int_{\gamma_R^{\text{arc}-}} f(z)\,dz\right| &\le \lim_{R\to\infty}\int_0^\pi \left|\frac{e^{iRe^{-it}}}{(Re^{-it})^2+9}\right|\cdot|-iRe^{-it}|\,dt \\
&= \lim_{R\to\infty}\int_0^\pi \left|\frac{e^{iR\left(\cos(-t)+i\sin(-t)\right)}}{(Re^{-it})^2+9}\right|\cdot|-iRe^{-it}|\,dt \\
&\le \lim_{R\to\infty}\int_0^\pi \frac{e^{-R\sin(t)}}{|R^2-9|}\cdot R\,dt \\
&\le \lim_{r\to\infty}\frac{\pi R}{|R^2-9|} \\
&= 0
\end{aligned}$$

where the second last line follows from $0 \le e^{-R\sin(t)} \le 1$ for all $t\in[0,\pi]$ and all $R>0$. So, by the absolute convergence of I, we have,

$$\begin{aligned}
I &= \lim_{R=\to\infty}\left(\int_{\gamma_R^{\text{axis}}} f(z)\,dz + \int_{\gamma_R^{\text{arc}-}} f(z)\,dz\right) \\
&= \lim_{R\to\infty}\oint_{\gamma_R} f(z)\,dz
\end{aligned}$$

Deformation of contours and Cauchy's integral formula then give

$$\begin{aligned}
I &= \lim_{R\to\infty}\oint_{\gamma_R} f(z)\,dz \\
&= -2\pi i\cdot\left.\frac{e^{iz}}{z-3i}\right|_{z=-3i} \\
&= \frac{\pi}{3e^3}
\end{aligned}$$

Note that we have a minus sign on the second line because γ_R is negatively oriented. △

In the above, it is important that the contour is closed over the lower half-plane, or otherwise the integral along the arc does not vanish:

$$\begin{aligned}
\lim_{R\to\infty}\left|\int_{\gamma_R^{\text{arc}+}} f(z)\,dz\right| &\le \lim_{R\to\infty}\int_0^\pi \left|\frac{e^{-iRe^{it}}}{(Re^{it})^2+9}\right|\cdot|iRe^{it}|\,dt \\
&= \lim_{R\to\infty}\int_0^\pi \left|\frac{e^{-iR\left(\cos(t)+i\sin(t)\right)}}{(Re^{it})^2+9}\right|\cdot|iRe^{it}|\,dt \\
&\le \lim_{R\to\infty}\int_0^\pi \frac{e^{R\sin(t)}}{|R^2-9|}\cdot R\,dt \\
&= \infty
\end{aligned}$$

We can similarly show that $\int_{-\infty}^{\infty}\frac{e^{ix}}{x^2+9}\,dx = \frac{\pi}{3e^3}$ by closing the contour in the upper half-plane.

Now, recall that for all $z\in\mathbb{C}$, we have

$$\begin{aligned}
\cos(z) &= \frac{1}{2}(e^{iz}+e^{-iz}) & \sin(z) &= \frac{1}{2i}(e^{iz}-e^{-iz}) \\
&= \Re(e^{iz}) & &= \Im(e^{iz})
\end{aligned}$$

So any real trigonometric integral can be converted into a complex exponential integral. For instance,

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{\cos(x)}{x^2+9}\,dz &= \int_{\infty}^{\infty} \frac{\frac{1}{2}(e^{ix}+e^{-ix})}{x^2+9}\,dx \\ &= \frac{1}{2}\left(\int_{\infty}^{\infty} \frac{e^{ix}}{x^2+9}\,dx + \int_{\infty}^{\infty} \frac{e^{-ix}}{x^2+9}\,dx\right) \\ &= \frac{\pi}{3e^3}\end{aligned}$$

Example (Complexifying the integrand). Let

$$I = \int_0^{2\pi} e^{\cos(t)} \cos(\sin(t))\,dt$$

Note that

$$\begin{aligned}I &= \Re\left(\int_0^{2\pi} e^{e^{i}t}\,dt\right) \\ &= \Re\left(\frac{1}{i}\int_0^{2\pi} \frac{e^{e^{it}}}{e^{it}} \cdot ie^{it}\,dt\right) \\ &= \Re\left(\frac{1}{i}\oint_{S^1} \frac{e^z}{z}\,dz\right)\end{aligned}$$

Cauchy's integral formula gives

$$\int_{S^1} \frac{e^z}{z}\,dz = 2\pi i$$

so

$$\begin{aligned}I &= \Re\left(\frac{1}{i}\oint_{S^1} \frac{e^z}{z}\,dz\right) \\ &= \Re\left(2\pi\right) \\ &= 2\pi\end{aligned}$$

△

Example (Pole at the origin). Let

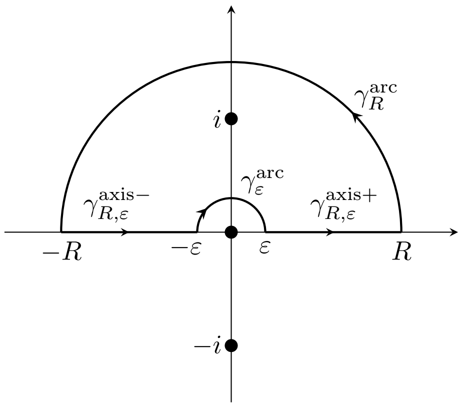
$$I = \int_{-\infty}^{\infty} \frac{\sin(x)}{x(x^2+1)}\,dx$$

We can replace this with an exponential:

$$I = \Im\left(J\right), \qquad J = \int_{-\infty}^{\infty} \frac{e^{ix}}{x(x^2+1)}\,dx$$

Denote the integrand by $f(z) = \frac{e^{iz}}{z(z^2+1)}$, which has poles at 0, i, and $-i$. This time, a semicircular contour does not work, since we cannot integrate over the pole at the origin.

We create a new contour γ by modifying the previous contour, adding another semicircle around the origin:


γ_R^{arc}
i
$\gamma_\varepsilon^{\text{arc}}$
$\gamma_{R,\varepsilon}^{\text{axis}-}$
$\gamma_{R,\varepsilon}^{\text{axis}+}$
$-R$
$-\varepsilon$
ε
R
$-i$

which decomposes into four paths

$$\begin{aligned}\gamma_{R,\varepsilon}^{\text{axis}+} &= t, && t \in [\varepsilon,R]\\ \gamma_R^{\text{arc}} &= Re^{it}, && t \in [0,\pi]\\ \gamma_{R,\varepsilon}^{\text{axis}-} &= t, && t \in [-R,-\varepsilon]\\ \gamma_\varepsilon^{\text{arc}} &= \varepsilon e^{it}, && t \in [-\pi,0]\end{aligned}$$

Then,

$$J = \lim_{R\to\infty}\lim_{\epsilon\to 0}\left(\int_{\gamma_{R,\varepsilon}^{\text{axis}+}} f(z)\,dz + \int_{\gamma_{R,\varepsilon}^{\text{axis}-}} f(z)\,dz\right)$$

as the integral converges absolutely.

It is routine to check that $\lim_{R\to\infty}\int_{\gamma_R^{\text{arc}}} f(z)\,dz = 0$. For the other arc, we have

$$\begin{aligned}\lim_{\varepsilon\to 0}\int_{\gamma_\varepsilon^{\text{arc}}} f(z)\,dz &= \lim_{\varepsilon\to 0}\int_{-\pi}^{0} \frac{e^{i\varepsilon e^{-it}}}{\varepsilon e^{i-t}\big((\varepsilon e^{it})^2+1\big)}\cdot(-i\varepsilon e^{-it})\,dt\\ &= -i\lim_{\varepsilon\to 0}\int_0^\pi \frac{e^{i\varepsilon e^{it}}}{(\varepsilon e^{it})^2+1}\,dt\\ &= -i\int_0^\pi \lim_{\varepsilon\to 0}\frac{e^{i\varepsilon e^{it}}}{(\varepsilon e^{it})^2+1}\,dt\\ &= -i\int_0^\pi 1\,dt\\ &= -\pi i\end{aligned}$$

where uniform convergence of the integrand allowed the limit to commute in the second last line.

Deformation of contours and Cauchy's integral formula then give

$$\begin{aligned}J + \lim_{\varepsilon\to 0}\int_{\gamma_\varepsilon^{\text{arc}}} f(z)\,dz &= 2\pi i\cdot\left.\frac{e^{iz}}{z(z+i)}\right|_{z=i}\\ &= -\frac{\pi}{e}i\end{aligned}$$

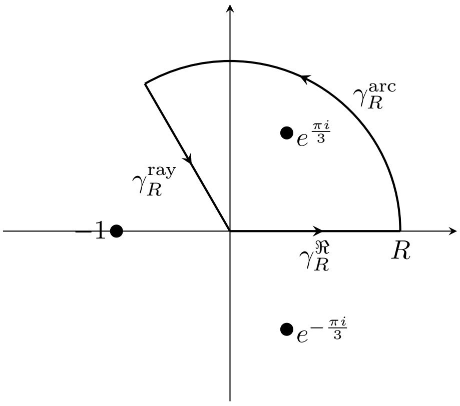
so $J = (\pi - \frac{\pi}{e})i$, so $I = \pi - \frac{\pi}{e}$. △

Example (Pie contour). Let

$$I = \int_0^\infty \frac{1}{x^3+1}\,dx$$

Denote the integrand by $f(z) = \frac{1}{z^3+1}$.

Extending the half-line to a semicircle contour does not work because of the pole at -1. Instead, for $R > 0$, we take a line segment along the real-axis from 0 to R, and another line segment off the real-axis, before joining them with an arc:



which decomposes into three paths

$$\begin{aligned}
\gamma_R^{\Re} &= t, & t &\in [0,R] \\
\gamma_R^{\text{arc}} &= Re^{it}, & t &\in [0,\tfrac{2\pi}{3}] \\
\gamma_R^{\text{ray}} &= -e^{\frac{2\pi i}{3}} t, & t &\in [-R,0]
\end{aligned}$$

Note

$$I = \lim_{R\to\infty} \int_{\gamma_R^{\Re}} f(z)\, dz$$

It is routine to check that $\lim_{R\to\infty} \int_{\gamma_R^{\text{arc}}} f(z)\, dz = 0$. Furthermore,

$$\begin{aligned}
\lim_{R\to\infty} \int_{\gamma_R^{\text{ray}}} f(z)\, dz &= \lim_{R\to\infty} \int_{-R}^{0} \frac{1}{(-e^{\frac{2\pi i}{3}})^3 + 1} \cdot (-e^{\frac{2\pi i}{3}})\, dt \\
&= -e^{\frac{2\pi i}{3}} \lim_{R\to\infty} \int_0^R \frac{1}{t^3+1}\, dt \\
&= -e^{\frac{2\pi i}{3}} I
\end{aligned}$$

Deformation of contours and Cauchy's integral formula then give

$$\begin{aligned}
I - e^{\frac{2\pi i}{3}} I &= \lim_{R\to\infty} \left(\int_{\gamma_R^{\Re}} f(z)\, dz + \int_{\gamma_R^{\text{arc}}} f(z)\, dz + \int_{\gamma_R^{\text{ray}}} f(z)\, dz \right) \\
&= 2\pi i \cdot \left. \frac{1}{(z+1)(z - e^{-\frac{2\pi i}{3}})} \right|_{z = e^{\frac{\pi i}{3}}} \\
&= \pi \left(\frac{1}{\sqrt{3}} - \frac{1}{3} i \right)
\end{aligned}$$

so $I = \frac{2\pi}{3\sqrt{3}}$ △

Example (Keyhole contour). Let

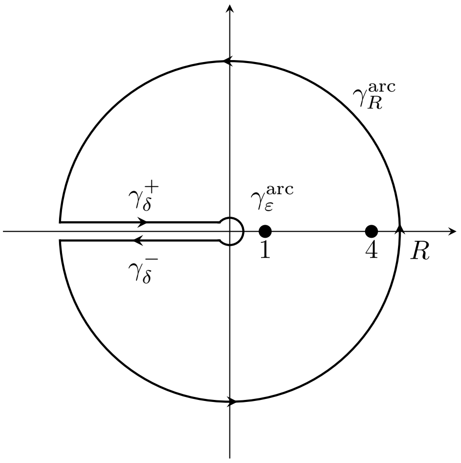
$$I = \int_{-\infty}^{0} \frac{z^{\frac{1}{2}}}{z^2 - 5z + 4}\, dz$$

This does not make sense as a real integral, as we are taking the square root of a negative number in the numerator. However, in the complex plane, this is simply a contour integral along the negative real axis with integrand $f(z) = \frac{z^{\frac{1}{2}}}{z^2-5z+4}$ which factorises into

$$f(z) = \frac{z^{\frac{1}{2}}}{(z-1)(z-4)}$$

so f has poles at 1 and 4. Since $z^{\frac{1}{2}} = e^{\frac{1}{2}\log(z)}$, f requires a branch cut to be defined continuously.

As is standard, we take the branch cut along the half-line $(-\infty,0]$. We cannot integrate f along this branch cut, so we use a contour γ made up of an arc of radius $0 < \varepsilon < 1$, an arc of radius $4 < R$, and two line segments displaced from the negative real-axis by $0 < \delta \leq \varepsilon$.



which decomposes into

$$\begin{aligned}
\gamma_R^{\text{arc}} &= Re^{it}, \quad t \in [-\pi + \arcsin(\tfrac{\delta}{R}),\pi - \arcsin(\tfrac{\delta}{R})] \\
\gamma_\varepsilon^{\text{arc}} &= \varepsilon e^{-it}, \quad t \in [-\pi + \arcsin(\tfrac{\delta}{\varepsilon}),\pi - \arcsin(\tfrac{\delta}{\varepsilon})] \\
\gamma_\delta^+ &= t + i\delta, \quad t \in [-\sqrt{R^2-\delta^2}, -\sqrt{\varepsilon^2-\delta^2}] \\
\gamma_\delta^- &= t - i\delta, \quad t \in [\sqrt{\varepsilon^2-\delta^2},\sqrt{R^2-\delta^2}]
\end{aligned}$$

We have

$$\oint_\gamma f(z)\,dz = \int_{\gamma_R^{\text{arc}}} f(z)\,dz + \int_{\gamma_\delta^+} f(z)\,dz + \int_{\gamma_\varepsilon^{\text{arc}}} f(z)\,dz + \int_{\gamma_\delta^-} f(z)\,dz$$

It is routine to check that the integrals along γ_R^{arc} and $\gamma_\varepsilon^{\text{arc}}$ vanish as $R \to \infty$ and $\varepsilon \to 0$, respectively. For the line segments, we have

$$\begin{aligned}
\lim_{R\to\infty}\lim_{\varepsilon\to 0}\lim_{\delta\to 0}\int_{\gamma_\delta^+} f(z)\,dz &= \lim_{R\to\infty}\lim_{\varepsilon\to 0}\lim_{\delta\to 0}\int_{-\sqrt{R^2-\delta^2}}^{-\sqrt{\varepsilon^2-\delta^2}} \frac{(t+i\delta)^{\frac{1}{2}}}{(t+i\delta)^2-5(t+i\delta)+4}\,dt \\
&= \int_{-\infty}^{0} \frac{t^{\frac{1}{2}}}{t^2-5t+4}\,dt \\
&= I
\end{aligned}$$

and

$$\lim_{R\to\infty}\lim_{\varepsilon\to 0}\lim_{\delta\to 0}\int_{\gamma_\delta^-} f(z)\,dz = \lim_{R\to\infty}\lim_{\delta\to 0}\lim_{\delta\to 0}\int_{\sqrt{\varepsilon^2-\delta^2}}^{\sqrt{R^2-\delta^2}} \frac{(-t-i\delta)^{\frac{1}{2}}}{(-t-i\delta)^2-5(-t-i\delta)+4}\cdot(-1)\,dt$$

$$
\begin{aligned}
&= \int_0^{\infty} \frac{(-t)^{\frac{1}{2}}}{(-t)^2 - 5(-t) + 4}\, dz \\
&= e^{it} \int_0^{-\infty} \frac{t^{\frac{1}{2}}}{t^2 - 5t + 4} \\
&= \int_{-\infty}^{0} \frac{t^{\frac{1}{2}}}{t^2 - 5t + 4} \\
&= I
\end{aligned}
$$

So,

$$\int_\gamma f(z)\, dz = 2I$$

The contour encircles two poles, so we have,

$$
\begin{aligned}
I &= \frac{1}{2} \oint_{\gamma_R} f(z)\, dz \\
&= \frac{1}{2} \cdot 2\pi i \left(\left. \frac{z^{\frac{1}{2}}}{z-1} \right|_{z=4} + \left. \frac{z^{\frac{1}{2}}}{z-4} \right|_{z=1} \right) \\
&= \frac{\pi i}{3}
\end{aligned}
$$

△

So far, every integrand we have seen has only had poles of order 1. However, we can still handle higher order poles with the generalised Cauchy's integral formula.

Example (Higher order poles). Let

$$I = \int_{-\infty}^{\infty} \frac{\cos(x)}{(x^2+1)^4}\, dx$$

Let $f(z) = \frac{e^{iz}}{(z^2+1)^4}$ which factorises as

$$f(z) = \frac{e^{iz}}{(z-i)^4 (z+i)^4}$$

so f has poles at i and $-i$. Then,

$$I = \Re\left(\int_{-\infty}^{\infty} f(x)\, dx \right)$$

Taking the usual semicircular contour γ_R of radius $R > 0$ in the upper half-plane, it is routine to check that

$$\int_{-\infty}^{\infty} f(x)\, dx = \lim_{R \to \infty} \oint_{\gamma_R} f(z)\, dz$$

(i.e. the integral along the arc vanishes as $R \to \infty$). Cauchy's integral formula then gives

$$
\begin{aligned}
I &= \Re\left(\lim_{R \to \infty} \oint_{\gamma_R} f(z)\, dz \right) \\
&= \frac{2\pi i}{3!} \cdot g^{(3)}(i) \qquad \text{where } g(z) = \frac{e^{iz}}{(z+i)^4} \\
&= \frac{37\pi}{48e}
\end{aligned}
$$

△

Example (Contour contains no poles). Let

$$I = \int_{-\infty}^{\infty} \frac{1}{z^2 - 3iz - 2}\, dz$$

Denote the integrand by $f(z) = \frac{1}{z^2 - 3iz - 2}$ which factorises as

$$f(z) = \frac{1}{(z - i)(z - 2i)}$$

so f has poles at i and $2i$, and is analytic elsewhere. Both of these poles lie in the upper half-plane.

Taking the semicircular contour γ_R of radius $R > 0$ in the *lower* half-plane, it is routine to check that

$$I = \lim_{R \to \infty} \int_{\gamma_R} f(z)\, dz$$

Cauchy's theorem then gives $I = 0$. △

34.19.8 Liouville's Theorem

Theorem (Liouville). *Let $f : \mathbb{C} \to \mathbb{C}$ be entire (analytic over $\mathbb{C}$) and bounded. Then, f is constant.*

Proof. Given two points x and y, consider the open balls $\mathbb{B}_r(x)$ and $\mathbb{B}_r(y)$, where $r > |x - y|$. For sufficiently large r, the two balls coincide except for an arbitrarily small proportion of their volume. Since f is bounded and entire functions are harmonic, by the mean value property, the averages of f over the two balls are arbitarily close so f takes the same value at x and y. Since x and y were arbitrary, f is constant. ∎

Theorem (Fundamental Theorem of Algebra). *Every non-constant polynomial $p \in \mathbb{C}[x]$ has a root in $\mathbb{C}$ – that is, there exists $\alpha \in \mathbb{C}$ such that $p(\alpha) = 0$.*

Proof. Suppose for a contradiction that $|p(z)| \neq 0$ for all $z \in \mathbb{C}$. Define $f : \mathbb{C} \to \mathbb{C}$ by $f(z) = \frac{1}{p(z)}$. Since p does not vanish, f is holomorphic on all of $\mathbb{C}$ since it is the composition of two holomorphic functions ($\frac{1}{z}$ is holomorphic outside of the origin).

If $p(z) = \sum_{k=0}^{n} c_k z^k$ with $c_n \neq 0$ for $n > 0$, then at infinity, the polynomial behaves like $c_n z^n$, as that is the highest power. T hus, $|p(z)| \to \infty$ as $z \to \infty$, and satisfies $|p(z)| > 1$ for all $|z| > R$ for some $R > 0$.

So, f is less than 1 for all $|z| > R$, and f is bounded on the compact set $|z| \leq R$ since it is continuous. Then, by Liouville's theorem, f is constant, and hence p is constant, which is a contradiction. ∎

Theorem 34.19.20. *Let Ω be open, and let $f_n : \Omega \to \mathbb{C}$ be a sequence of analytic functions. If f_n converges uniformly to f, then f is analytic.*

Proof. Fix a point $z_0 \in \Omega$. Choose $r > 0$ sufficiently small such that $B_r(z_0) \subset \Omega$. Since f_n is analytic on Ω, Cauchy's formula yields

$$f_n(z_0) = \frac{1}{2\pi i} \int_{\partial B_r(z_0)} \frac{f_n(z)}{z - z_0}\, dz$$

Taking limits as n tends to infinity, we have

$$f(z_0) = \lim_{n \to \infty} \frac{1}{2\pi i} \int_{\partial B_r(z_0)} \frac{f_n(z)}{z - z_0}\, dz$$

If we could commute the limit and the integral, we would obtain

$$f(z_0) = \frac{1}{2\pi i} \int_{\partial B_r(z_0)} \frac{f(z)}{z - z_0}\, dz$$

which implies that f is differentiable (in fact, smooth), and an expression for its derivatives is given by Theorem 34.19.18.

(In the proof of that theorem, it was assumed that f is analytic. However, if the above integral expression is already given, analyticity of the left side can be proved using only that f is bounded in $\partial B_r(z_0)$: $f\big|_{\partial B_r(z_0)}$ is continuous as the uniform limit of a sequence of analytic, and hence continuous, functions, so f is bounded as $\partial B_r(z)$ is closed and bounded.)

To show that the limit may be moved inside the integral, parametrise $\partial B_r(z_0)$ by $\gamma(t) = z_0 + re^{it}$ for $t \in [0{,}2\pi)$. Then, $\gamma'(t) = ire^{it}$, so

$$\begin{aligned}\int_{\partial B_r(z_0)} \frac{f_n(z)}{z - z_0}\, dz &= \int_0^{2\pi} \frac{f_n(z_0 + re^{it})}{re^{it}} ire^{it}\, dt \\ &= i \int_0^{2\pi} f_n(z_0 + re^{it})\, dt \qquad (1)\end{aligned}$$

As a function of t, $f_n(z_0 + re^{it})$ converges uniformly to $f(z_0 + re^{it})$, so the integral of the sequence also converges uniformly to the integral of the limit, so we have

$$\begin{aligned}\lim_{n\to\infty} \int_{\partial B_r(z_0)} \frac{f_n(z)}{z - z_0}\, dz &= \lim_{n\to\infty} i \int_0^{2\pi} f_n(z_0 + re^{it}) \\ &= i \int_0^{2\pi} f(z_0 + re^{it})\end{aligned}$$

Replacing f_n by f in (1), and reading the chain of equalities in reverse, we have

$$i \int_0^{2\pi} f(z_0 + re^{it})\, dt = \int_{\partial B_r(z)} \frac{f(z)}{z - z_0}$$

obtaining the result. ■

Chapter 35

Asymptotics

"Hold infinity in the palm of your hand."

— William Blake, *Auguries of Innocence*

Chapter 36

Variational Principles

"We are not randomly suboptimal in our decisions. We are systematically suboptimal."

— Leland Wilkinson, *The Grammar of Graphics, Statistics, and Computing*

Chapter 37

Point-Set Topology

> *"A linguist would be shocked to learn that if a set is not closed that does not mean that it is open, or again that 'E is dense in E' does not mean the same thing as 'E is dense in itself'."*
>
> — John Edensor Littlewood, *A mathematicians' Miscellany*

Topology is the branch of mathematics concerned with continuity and connectedness, and properties of spaces that are invariant under continuous deformations.

For instance, the *hairy ball theorem* of algebraic topology states that there is no nonvanishing continuous tangent vector field on the sphere (or "you can't comb the hair on a coconut flat"). The object in question being a sphere is not actually important to the theorem, and it holds on any smooth blob that can be continuously deformed into a sphere. Note that this excludes say, a torus, which has a hole and thus cannot be continuously deformed into a sphere. Topology makes precise this distinction between a sphere and a torus ("homotopy classes"), and also formalises what it means to continously deform an object ("homeomorphisms").

In this chapter, we investigate the topologies of sets, in what is known as *point-set* or *set-theoretic* topology, which has more of an analytic flavour. We begin by generalising the notion of lengths and distances with normed and metric spaces, before investigating topological properties of those spaces and dropping the precise measure of distance altogether in topological spaces.

In later chapters, we will investigate topological spaces using algebraic invariants and techniques.

37.1 Normed Spaces

A *norm* on a real or complex vector space X is a map $\| \cdot \| : X \to \mathbb{R}_{\geq 0}$ such that,

(*i*) $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$ (point separating or positive-definiteness);

(*ii*) $\|\lambda \mathbf{x}\| = \lambda \|\mathbf{x}\|$ for all $\lambda \in \mathbb{R}$ or $\mathbb{C}$ and all $\mathbf{x} \in X$ (absolute homogeneity);

(*iii*) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (triangle inequality).

Note that these axioms imply that $\|\mathbf{x}\| \geq 0$ for all $\mathbf{x} \in X$. The pair $(X, \| \cdot \|)$ is then called a *normed space.*

As an example, the absolute value function $|\cdot| : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a norm on the one-dimensional vector spaces $\mathbb{R}$ and $\mathbb{C}$.

For a vector $\mathbf{x} = (x_1,x_2,\dots,x_n)$ in the vector space $\mathbb{R}^n$, we define the *Euclidean* or *standard* norm as,

$$\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{n} |x_i|^2}$$

We also have the *taxicab* or *Manhattan* norm,

$$\|\mathbf{x}\|_{\ell^1} = \sum_{i=1}^{n} |x_i|$$

and the *uniform* or *maximum* norm,

$$\|\mathbf{x}\|_{\ell^\infty} = \max_{1\leq i\leq n} \|x_i\|$$

The *closed unit ball* denoted $\overline{\mathbb{B}}$ or $\mathfrak{B}$ in the normed space $(X,\|\cdot\|)$ is the set,

$$\mathfrak{B}_X = \{\mathbf{x} \in X : \|\mathbf{x}\| \leq 1\}$$

The *open unit ball* denoted $\mathbb{B}$ or B in the normed space $(X,\|\cdot\|)$ is the set,

$$B_X = \{\mathbf{x} \in X : \|\mathbf{x}\| < 1\}$$

Let X be a vector space. A subset $K \subseteq X$ is *convex* if, whenever $\mathbf{x},\mathbf{y} \in K$, then $\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in K$ for $0 \leq \lambda \leq 1$. Informally, a set is convex if the straight line segment connecting any two points in the set is contained within the set.

Lemma (Convexity of Balls)**.** *In any normed space $(X,\|\cdot\|)$, the open and closed unit balls are convex.*

Proof. We show the case for the closed ball. The proof for the open ball is analogous.

Let $\mathbf{x},\mathbf{y} \in \mathfrak{B}_X$, then $\|\mathbf{x}\| \leq 1$ and $\|\mathbf{y}\| \leq 1$. Then, for $0 \leq \lambda \leq 1$,

$$\begin{aligned} \|\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\| &\leq |\lambda|\|\mathbf{x}\| + |1-\lambda|\|\mathbf{y}\| && \text{[Triangle Inequality]} \\ &\leq \lambda + (1-\lambda) \\ &= 1 \end{aligned}$$

so $\|\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\| \leq 1$, giving $\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in \mathfrak{B}_X$, as required. ■

Lemma (Equivalence of Convexity and Triangle Inequality)**.** *Suppose a function $N : X \to \mathbb{R}_{\geq 0}$ satisfies requirements (i) and (ii) of a norm, and in addition, that the set $\mathfrak{B} := \{\mathbf{x} \in X : N(\mathbf{x}) \leq 1\}$ is convex. Then, N satisfies the triangle inequality,*

$$N(\mathbf{x}+\mathbf{y}) \leq N(\mathbf{x}) + N(\mathbf{y})$$

and thus defines a norm on X.

Proof. If $N(\mathbf{x}) = 0$, then $\mathbf{x} = \mathbf{0}$ and

$$N(\mathbf{x}+\mathbf{y}) = N(\mathbf{y}) = N(\mathbf{x}) + N(\mathbf{y})$$

so we can assume $N(\mathbf{x}),N(\mathbf{y}) > 0$.

In this case, $\mathbf{x}/N(\mathbf{x}) \in \mathfrak{B}$ and $\mathbf{y}/N\mathbf{y} \in \mathfrak{B}$, so by convexity of $\mathfrak{B}$,

$$\frac{N(\mathbf{x})}{N(\mathbf{x})+N(\mathbf{y})}\left(\frac{\mathbf{x}}{N(\mathbf{x})}\right) + \frac{N(\mathbf{y})}{N(\mathbf{x})+N(\mathbf{y})}\left(\frac{\mathbf{y}}{N(\mathbf{y})}\right) \in \mathfrak{B}$$

So,

$$\frac{\mathbf{x}+\mathbf{y}}{N(\mathbf{x})+N(\mathbf{y})} \in \mathfrak{B}$$

By homogeneity,

$$N\left(\frac{\mathbf{x}+\mathbf{y}}{N(\mathbf{x})+N(\mathbf{y})}\right) = \frac{N(\mathbf{x}+\mathbf{y})}{N(\mathbf{x})+N(\mathbf{y})} \leq 1$$

and multiplying through by $N(\mathbf{x})+N(\mathbf{y})$ gives the result. ■

Because verifying the triangle inequality can be quite difficult, this lemma provides a simple way to check if a function defines a norm based on verifying convexity of the closed unit ball the function would generate.

For $p \in [1,\infty]$, the ℓ^p norms on $\mathbb{R}^n$ are defined by,

$$\|\mathbf{x}\|_{\ell^p} \coloneqq \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$$

The standard norm corresponds to the choice $p = 2$, the taxicab norm to $p = 1$, and the max norm to $p = \infty$.

Theorem (Minkowski's Inequality in $\mathbb{R}^n$)**.** *For all* $1 \leq p \leq \infty$*, if* $\mathbf{x}$,$\mathbf{y} \in \mathbb{R}^n$*, then,*

$$\|\mathbf{x}+\mathbf{y}\|_{\ell^p} \leq \|\mathbf{x}\|_{\ell^p} + \|\mathbf{y}\|_{\ell^p}$$

Proof. If $p = \infty$, this is straightforward. For $p \in [1,\infty)$, the function $t \mapsto |t|^p$ is convex, so if $\mathbf{x}$,$\mathbf{y} \in \mathfrak{B}$, then,

$$\begin{aligned}\left\|\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\right\|_{\ell^p}^p &= \sum_{i=1}^{n} \left|\lambda x_i + (1-\lambda)y + i\right|^p \\ &\leq \sum_{i=1}^{n} \lambda|x_i|^p + (1-\lambda)|y_i|^p \\ &\leq 1\end{aligned}$$

and so $\lambda\mathbf{x} + (1-\lambda)y \in \mathfrak{B}$ and $\mathfrak{B}$ is convex. The result then follows from §37.1. ■

Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on X are *equivalent* if there exists constants $0 < c_1 \leq c_2$ such that,

$$c_1\|\mathbf{x}\|_1 \leq \|\mathbf{x}\|_2 \leq c_2\|\mathbf{x}\|_1$$

for every $\mathbf{x} \in X$, or equivalently, if,

$$c_1\mathfrak{B}_{(X,\|\cdot\|_1)} \subseteq \mathfrak{B}_{(X,\|\cdot\|_2)} \subseteq c_2\mathfrak{B}_{(X,\|\cdot\|_1)}$$

This notion of equivalence forms an equivalence relation on the space of norms of X.

Theorem (Equivalence of Finite Norms)**.** *All norms on a finite-dimensional vector space are equivalent.*

The *sequence space* ℓ^p, $1 \leq p \leq \infty$, consists of all sequences $(\mathbf{x}_i)_{i=1}^{\infty}$ such that,

$$\sum_{i=1}^{\infty} |\mathbf{x}_i|^p < \infty$$

(in the case of $p = \infty$, ℓ^∞ is the space of bounded sequences) equipped with the corresponding ℓ^p norm.

The ℓ^p spaces are infinite-dimensional, with the standard basis being given by $(\mathbf{e}_i)_{i=1}^{\infty} Minkowski's$, where,

$$\mathbf{e}_i = (0, 0, \underbrace{\ldots, 1}_{i\text{th place}}, 0, \ldots)$$

Note that for any $1 \leq q < p \leq \infty$, there are elements of ℓ^p that are not elements of ℓ^q – for instance, the sequence $(i^{-\frac{1}{q}})_{i=1}^{\infty}$ – so the ℓ^p spaces are nested within each other, with ℓ^∞ being the largest, and ℓ^1 the smallest.

Theorem (Minkowski's Inequality in ℓ^p)**.** *For all $1 \leq p \leq \infty$, if $x, y \in \ell^p$, then $x + y \in \ell^p$ and,*

$$\|x + y\|_{\ell^p} \leq \|x\|_{\ell^p} + \|y\|_{\ell^p}$$

Proof. The case $p = \infty$ is again straightforward. For $p \in [1,\infty)$, given $x, y \in \ell^p$, we can use Minkowski's inequality in $\mathbb{R}^n$ to guarantee that

$$\begin{aligned} \left(\sum_{i=1}^{n} |x_i + y_i|^p\right)^{\frac{1}{p}} &\leq \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{n} |y_i|^p\right)^{\frac{1}{p}} \\ &\leq \|\mathbf{x}\|_{\ell^p} + \|y\|_{\ell^p} \end{aligned}$$

Taking the limit as $n \to \infty$, we deduce the result. ■

37.1.1 Normed Subspaces

If $(X, \|\cdot\|)$ is a normed space, and Y is a subspace of X, then $(Y, \|\cdot\|)$ is another normed space. Strictly speaking, the norm $\|\cdot\|_Y$ defined on Y is the restriction of the norm $\|\cdot\|$ on X to Y, but we denote them by the same symbol as the implied meaning is clear.

For example, c_0, the space of all null sequences, is a subspace of ℓ^∞. The space c_{00}, the space of all sequences with only a finite number of non-zero terms, is a subspace of ℓ^p for all $p \in [1,\infty]$.

37.1.2 Spaces of Continuous Functions

We denote by $C\big([a,b]\big)$ the space of (real-valued) continuous functions defined on the interval $[a,b]$. The usual norm to use on this space is the supremum norm,

$$\|f\|_\infty := \sup_{x \in [a,b]} |f(x)|$$

but by the extreme value theorem, this is equivalent to

$$= \max_{x \in [a,b]} |f(x)|$$

The L^p norms are defined on this space analogously to how the ℓ^p norms are defined on $\mathbb{R}^n$:

$$\|f\|_{L^p} := \left(\int_a^b |f(x)|^p\right)^{\frac{1}{p}}$$

37.2 Metric Spaces

In many situations, we care less about a notion of length than about a generalised notion of distance. This generalisation is given in the form of a metric.

Let X be any set. A *metric* d on X is a map $d : X \times X \to \mathbb{R}_{\geq 0}$ such that, (*i*) $d(x,y) = 0$ if and only if $x = y$ (point separating or positive-definiteness);

(*ii*) $d(x,y) = d(y,x)$ for all $x,y \in X$ (symmetry);

(*iii*) $d(a,b) \leq d(a,x) + d(x,b)$ for every $a,b,x \in X$ (triangle inequality).

Note that these axioms imply that $d(x,y) \geq 0$ for all $x,y \in X$. The pair (X,d) is then called a *metric space.*

Theorem (Induced Metric)**.** *If* $(X,\|\cdot\|)$ *is a normed space, then* $d(\mathbf{x},\mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$ *defines a metric on* X*.*

Proof. We verify the metric axioms:

(*i*) If $\mathbf{x} = \mathbf{y}$, then $d(\mathbf{x},\mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = \|\mathbf{0}\| = 0$; if $d(\mathbf{x},\mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = 0$, then $\mathbf{x} = \mathbf{y}$;

(*ii*) $d(\mathbf{x},\mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = \|(-1)(\mathbf{y} - \mathbf{x})\| = |-1|\|\mathbf{y} - \mathbf{x}\| = \|\mathbf{y} - \mathbf{x}\| = d(\mathbf{y},\mathbf{x})$;

(*iii*) $d(\mathbf{a},\mathbf{b}) = \|\mathbf{a} - \mathbf{b}\| \leq \|\mathbf{a} - \mathbf{x}\| + \|\mathbf{x} - \mathbf{b}\| = d(\mathbf{a},\mathbf{x}) + d(\mathbf{x},\mathbf{b})$. ■

The *Euclidean* or *standard* metric on $\mathbb{R}^n$ is the metric induced by the Euclidean ℓ^2 norm:

$$\begin{aligned} d_2(\mathbf{x},\mathbf{y}) &= \|\mathbf{x} - \mathbf{y}\|_{\ell^2} \\ &= \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \end{aligned}$$

The *discrete metric* on any non-empty set X is defined as,

$$d(x,y) := \begin{cases} 0 & x = y \\ 1 & x \neq y \end{cases}$$

Every point in a discrete metric space is equally distanced from every other distinct point. The discrete metric is useful for counterexamples, as it is very different from metrics that arise from norms.

Let L be the set of words of length n constructed from a finite alphabet Σ of characters. The *Hamming distance* on L is defined as the number of places in the strings which disagree. For example, the strings $abcdef$ and $aacdef$ have a Hamming distance of 1. This metric has important applications in (en)coding and information theory, as it measures, among other things, the error or noise in a signal.

Let G be a graph. The *graph metric* defined on the vertex set of G is the number of edges in a shortest path connecting two vertices.

The *jungle river metric* on $\mathbb{R}^2$ is defined as,

$$d\big((x_1,y_1)(x_2,y_2)\big) = \begin{cases} |y_1 - y_2| & x_1 = x_2 \\ |y_1| + |x_1 - x_2| + |y_2| & x_1 \neq x_2 \end{cases}$$

37.2.1 Metric Subspaces and Product Spaces

If (X,d) is a metric space, and A is a subset of X, then the restriction $d|_A$, also denoted d_A, of d to A is also a metric on A. Then, we say that (A,d_A) is a (*metric*) *subspace* of (X,d).

For instance, any set $A \subseteq \mathbb{R}$ equipped with the usual Euclidean metric with the appropriate restriction is a metric subspace of $\mathbb{R}$.

If (X,d_1) and (Y,d_2) are metric spaces, we can define a metric on the Cartesian product of their underlying sets. In fact, there are many ways to do so:

Lemma 37.2.1. *Let (X,d_1) and (Y,d_2) be metric spaces. Then, for any $1 \le p \le \infty$,*

$$\varrho_p\big((x_1,y_1),(x_2,y_2)\big) \coloneqq \begin{cases} \Big(\big(d_1(x_1,x_2)\big)^p + \big(d_2(y_1,y_2)\big)^p\Big)^{\frac{1}{p}} & 1 \le p < \infty \\ \max\big(d(x_1,x_2),d(y_1,y_2)\big) & p = \infty \end{cases}$$

defines a metric on $X \times Y$.

That is, we perform a pairwise computation analogous to the ℓ^p metrics on the components of the points. This similarly extends to any finite product of metric spaces.

Theorem 37.2.2. *Given a finite collection $\big((X_i,d_i)\big)_{i=1}^n$ of metric spaces,*

$$\varrho_p\big(\mathbf{a},\mathbf{b}\big) = \begin{cases} \left(\displaystyle\sum_{i=1}^n \big(d_i(a_i,b_i)\big)^p\right)^{\frac{1}{p}} & 1 \le p < \infty \\ \displaystyle\max_{1\le i\le n} \left(d_i(a_i,b_i)\right) & p = \infty \end{cases}$$

defines a metric on $\prod_{i=1}^n X_i$.

37.2.2 Open and Closed Sets

Let (X,d) be a metric space. The *open ball* centred at $a \in X$ of radius r is the set,

$$B(a,r) = \{x \in X : d(x,a) < r\}$$

also denoted by $\mathbb{B}(a,r)$ or $\mathbb{B}_r(a)$

Similarly, the *closed ball* centred at $a \in X$ of radius r is the set,

$$\overline{B}(a,r) = \{x \in X : d(x,a) \le r\}$$

also denoted by $\overline{\mathbb{B}}(a,r)$ or $\overline{\mathbb{B}}_r(a)$.

Example. The open ball of radius 1 centred at 0 in $\mathbb{R}$ under the Euclidean metric is the interval $\mathbb{B}(0,1) = (-1,1)$. In the subspace $[0,2] \subset \mathbb{R}$, the same ball is instead given by $\mathbb{B}(0,1) = [0,1)$, so balls depend on the ambient space containing them. △

Let (X,d) be a metric space. A subset $S \subseteq X$ is *bounded* if there exists $a \in X$ and $r > 0$ such that $S \subset \mathbb{B}(a,r)$.

A subset $U \subseteq X$ is *open in X* if for every $x \in U$ there exists $\varepsilon > 0$ such that $B(x,\varepsilon) \subset U$. A subset $F \subseteq X$ is *closed in X* if its complement $X \setminus F$ is open.

Note that the definition of a closed set here is different from the one given in §45.3, where closed sets are defined to be sets closed under sequential limits, but these definitions are equivalent in metric spaces. However, in more general topological spaces, these definitions are *not* equivalent.

Example.

- In any metric space (X,d), X and $\varnothing$ are both simultaneously open and closed (or *clopen*).
- In $\mathbb{R}$, open intervals are open and closed intervals are closed. Half-open intervals are neither open nor closed.
- In a discrete metric space, every singleton set $\{x\} \subseteq X$ is open (take any $\varepsilon < 1$).

△

Sets can be open, closed, both (clopen), or neither, so the adjectives “open” and “closed” do not have all of their usual intuitive connotations when used in a mathematical context.

Lemma (Open Balls)**.** *Open balls are open sets.*

Proof. Let (X,d) be a metric space, and let $a \in X$ and $r > 0$. Let $y \in \mathbb{B}(a,r)$ so $d(y,a) < r$, and take $\varepsilon := r - d(y,a) > 0$. Then, $\mathbb{B}(y,\varepsilon) \subset \mathbb{B}(a,r)$, since, if $d(z,y) < \varepsilon$, we have,

$$d(z,a) \leq d(z,y) + d(y,a) < \varepsilon + d(y,a) = r$$

■

Corollary (Closed Balls)**.** *Closed balls are closed sets.*

Lemma (Open Finite Intersection)**.** *If $(U_i)_{i=1}^n$ is a finite collection of sets open in (X,d), then $\bigcap_{i=1}^n U_i$ is open in (X,d).*

Proof. Take $x \in \bigcap_{i=1}^n U_i$. Then, for each i, $x \in U_i$, so there exists $\varepsilon_i > 0$ such that $\mathbb{B}(x,\varepsilon_i) \subset U_i$. If $\varepsilon := \min(\varepsilon_1, \ldots ,\varepsilon_n)$, then,

$$\mathbb{B}(x,\varepsilon) \subseteq \mathbb{B}(x,\varepsilon) \subset U_i$$

for all i, and hence $\mathbb{B}(x,\varepsilon) \subset \bigcap_{i=1}^n U_i$. ■

However, the countable intersection of open sets is not necessarily open. For example, $\left(-\frac{1}{n},\frac{1}{n}\right)$ is open in $\mathbb{R}$ for all n, but,

$$\bigcap_{n=1}^{\infty}\left(-\frac{1}{n},\frac{1}{n}\right) = \{0\}$$

which is not open in $\mathbb{R}$.

Corollary (Closed Finite Union)**.** *If $(F_i)_{i=1}^n$ is a finite collection of sets closed in (X,d), then $\bigcup_{i=1}^n F_i$ is closed in (X,d).*

Proof. By De Morgan’s laws,

$$X \setminus \bigcup_{i=1}^{n} F_i = \bigcap_{i=1}^{n}(X \setminus F_i)$$

As F_i is closed, $X \setminus F_i$ is open, so $\bigcap_{i=1}^n (X \setminus F_i)$ is the finite intersection of open sets, and hence $X \setminus \bigcup_{i=1}^n F_i$ is open. It follows that $\bigcup_{i=1}^n F_i$ is closed. ■

Again, the countable union of closed sets is not necessarily closed. For example, $\left[-1 + \frac{1}{n},1 - \frac{1}{n}\right]$ is closed in $\mathbb{R}$ for all n, but,

$$\bigcup_{n=1}^{\infty}\left[-1 + \frac{1}{n},1 - \frac{1}{n}\right] = (-1,1)$$

which is not closed in $\mathbb{R}$.

Lemma (Open Arbitrary Union)**.** *If $(U_i)_{i\in\mathcal{I}}$ is an arbitrary collection of sets open in (X,d), then $\bigcup_{i\in\mathcal{I}} U_i$ is open in (X,d).*

Proof. If $x \in \bigcup_{i\in\mathcal{I}} U_i$, then $x \in U_i$ for some $i \in \mathcal{I}$. Since U_i is open, there exists $\varepsilon > 0$ such that $B(x,\varepsilon) \subset U_i \subseteq \bigcup_{i\in\mathcal{I}} U_i$, so $\bigcup_{i\in\mathcal{I}} U_i$ is open. ■

Corollary (Closed Arbitrary Intersection)**.** *If $(F_i)_{i\in\mathcal{I}}$ is an arbitrary collection of sets closed in (X,d), then $\bigcup_{i\in\mathcal{I}} F_i$ is closed in (X,d).*

Proof.

$$X \setminus \bigcap_{i\in\mathcal{I}} F_i = \bigcup_{i\in\mathcal{I}} (X \setminus F_i)$$

and apply the preceding lemma. ■

37.2.3 Convergence of Sequences

We will now rephrase the ε-δ notion of convergence of sequences from analysis in terms of open sets in metric spaces.

A sequence $(x_n)_{n=1}^{\infty}$ in a metric space (X,d) converges to $x \in X$ if,

$$\lim_{n\to\infty} d(x_n,x) = 0$$

or equivalently, in terms of open balls, for every $\varepsilon > 0$, there exists $N \geq 1$ such that,

$$x_n \in \mathbb{B}(x,\varepsilon)$$

for all $n \geq N$.

Lemma 37.2.3. *A sequence in a metric space can have at most one limit.*

Proof. Suppose that $(x_n) \to a$ and $(x_n) \to b$ so,

$$\lim_{n\to\infty} d(x_n,a) = \lim_{n\to\infty} d(x_n,b) = 0$$

Then,

$$0 \leq d(a,b) \leq d(a,x_n) + d(x_n,b) \to 0$$

so $d(a,b) = 0$ and hence $a = b$. ■

This may be rather familiar from analysis, but it turns out that this result may not hold in more general spaces.

We can now characterise convergence purely in terms of open sets, without directly invoking the metric:

Lemma (Open Set Convergence)**.** *Let $(x_n)_{n=1}^{\infty} \subset (X,d)$ be a sequence. Then, $(x_n) \to x \in X$ if and only if for every open set U containing x, there exists $N \geq 1$ such that $x_n \in U$ for all $n \geq N$.*

Proof. If $(x_n) \to x$ and $U \ni x$ is open, then $\mathbb{B}(x,\varepsilon) \subset U$ for some $\varepsilon > 0$. There exists $N \geq 1$ such that $d(x_n,x) < \varepsilon$ for all $n \geq N$. That is, $x_n \in \mathbb{B}(x,\varepsilon) \subseteq U$ for all $n \geq N$.

Conversely, suppose that for every open set U containing x there is an $N \geq 1$ such that $x_n \in U$ for all $n \geq N$. Then, given $\varepsilon > 0$, the set $\mathbb{B}(x,\varepsilon)$ is an open set containing x, so there exists $N \geq 1$ such that $x_n \in \mathbb{B}(x,\varepsilon)$ for all $n \geq N$. That is, $d(x_n,x) < \varepsilon$ for all $n \geq N$, so $(x_n) \to x$. ■

Lemma (Sequential Closure)**.** *A subset F of a metric space is closed if and only if whenever a sequence $(x_n)_{n=1}^{\infty} \subset F$ converges to some $x \in X$, then $x \in F$.*

Proof. For the forward implication, suppose F is closed and $(x_n)_{n=1}^{\infty} \subseteq F$ converges to x. Suppose that $x \notin F$. Since $X \setminus F$ is open, then by the previous lemma, there exists $N \geq 1$ such that $x_n \in X \setminus F$ for all $n \geq N$. But this contradicts that $x_n \in F$, so $x \in F$.

For the reverse implication, suppose otherwise that F is not closed. That is, $X \setminus F$ is not open. Then there exists some $x \in X \setminus F$ such that there is no ε such that $\mathbb{B}(x,\varepsilon) \subseteq X \setminus F$. Then, for each $k \in \mathbb{N}$, there exists $x_k \in \mathbb{B}\left(x,\frac{1}{k}\right)$ such that $x_k \notin X \setminus F$, i.e., such that $x_k \in F$. Then, $(x_k) \to x$ but $x \notin F$. ■

37.3 Continuity

37.3.1 Metric Continuity

Let $f : (X,d_X) \to (Y,d_Y)$ be a function between two metric spaces. For each point $p \in X$, we write,

$$\lim_{x \to p} f(x) = y \in Y$$

if, for every $\varepsilon > 0$, there exists $\delta > 0$ such that whenever $0 < d_X(x,p) < \delta$, we have $d_Y(f(x),y) < \varepsilon$.

Then, f is *continuous at a point* $p \in X$ if $\lim_{x \to p} f(x) = f(p)$. That is, if for every $\varepsilon > 0$, there exists a $\delta > 0$ such that $d_X(x,p) < \delta \to d_Y(f(x),f(p)) < \varepsilon$. We also say that f is *continuous on a (sub)set* $S \subseteq X$ if it is continuous at every point $p \in S$.

A function $f : X \to Y$ is *Lipschitz continuous* if there exists a constant $M \geq 0$ such that

$$d_Y(f(x),f(y)) \leq M d_Y(x,y)$$

for every $x,y \in X$, and we say that M is a *Lipschitz constant* or *modulus of (uniform) continuity* for f. Lipschitz continuity implies continuity, as we can take $\delta = \frac{\varepsilon}{M}$.

Let $A \subset X$ be non-empty. We define the distance of a point $x \in X$ from the set A to be,

$$d(x,A) = \inf_{a \in A} d(x,a)$$

Lemma 37.3.1. *If $A \subset X$ is non-empty, then the function $x \mapsto d(x,A)$ is Lipschitz with modulus* 1.

Proof. Let $x,y \in X$. Then, for every $a \in A$, we have,

$$d(x,A) \leq d(x,a) \leq d(x,y) + d(y,a)$$

Taking the infimum, we have,

$$\begin{aligned} d(x,A) &\leq d(x,y) + d(y,A) \\ d(x,A) - d(y,A) &\leq d(x,y) \end{aligned}$$

The situation is symmetric with respect to y, so we also have,

$$d(y,A) - d(x,A) \leq d(x,y)$$

giving,

$$|d(x,A) - d(y,A)| \leq d(x,y)$$

■

Lemma (Sequential Continuity). *Let (X,d_X) and (Y,d_Y) be metric spaces, and let $(x_n)_{i=n}^{\infty} \subset X$ be a sequence such that $(x_n) \to x \in X$. Then, a function $f : X \to Y$ is continuous at x if and only if $f(x_n) \to f(x)$.*

Lemma (Algebra of Continuous Functions)**.** *Let (X,d) be a metric space. Then,*

- *If $f,g : X \to \mathbb{R}$ are continuous, then $f + g$ and fg are continuous, and f/g is continuous at all points x where $g(x) \neq 0$;*
- *If $(Y,\| \cdot \|)$ is a normed vector space, and $f,g : X \to Y$ are continuous, then $f + g$ is coninuous.*

Continuity and open sets are closely related, but perhaps not in the intuitive way one might expect. In particular, the image of an open set under a continuous function need not be open (or closed). For instance, $\sin : \mathbb{R} \to \mathbb{R}$ sends the open set $(-\pi,\pi)$ to the closed set $[-1,1]$, and the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = \frac{1}{1+x^2}$ sends the open set $\mathbb{R}$ to the set $(0,1]$, which is neither open nor closed.

Instead, the *preimage* of any open set under a continuous function is open. If $f : X \to Y$ is a function and $A \subseteq Y$, then we write the preimage of A under f as,

$$f^{-1}[A] = \{x \in X : f(x) \in A\}$$

Note that this does not require that f is invertible.

We now characterise continuity in terms of open sets:

Theorem (Characterisation of Continuity)**.** *For any function $f : X \to Y$ between metric spaces (X,d_X) and (Y,d_Y), the following statements are equivalent:*

1. *f is continuous at all points of X;*
2. *$f^{-1}[U]$ is open whenever $U \subseteq Y$ is open;*
3. *$f^{-1}[\mathcal{F}]$ is closed whenever $\mathcal{F} \subseteq Y$ is closed.*

Proof. $(1 \to 2)$: Suppose f is continuous. Take any open set $U \subseteq Y$ and some point $x \in f^{-1}[U]$.. Then $f(x) \in U$, which is open, so there exists $\varepsilon > 0$ such that $\mathbb{B}_Y\big(f(x),\varepsilon\big) \in U$, i.e., $d_Y\big(f(x),y\big) < \varepsilon$ implies that $y \in U$. Since f is continuous, there exists $\delta > 0$ such that $d_X(x',x) < \delta$ implies that $d_Y\big(f(x'),f(x)\big) < \varepsilon$, so if $x' \in \mathbb{B}_X(x,\delta)$, we have $f(x') \in \mathbb{B}_y\big(f(x),\varepsilon\big) \subseteq U$, i.e., $\mathbb{B}_X(x,\delta) \in f^{-1}[U]$. Hence $f^{-1}[U]$ is open.

$(2 \to 1)$: Suppose that $f^{-1}[U]$ is open whenever $U \subseteq Y$ is open, and take $x \in X$ and $\varepsilon > 0$. $\mathbb{B}_Y\big(f(x),\varepsilon\big)$ is open in Y, so $f^{-1}\big[\mathbb{B}_y\big(f(x),\varepsilon\big)\big]$ is open in X. Since this set contains x, $\mathbb{B}_X(x,\delta) \subseteq f^{-1}\big[\mathbb{B}_y\big(f(x),\varepsilon\big)\big]$ for some $\delta > 0$. But this inclusion says precisely that $d_Y\big(f(x),f(x')\big) < \varepsilon$ whenever $d_X(x',x) < \delta$, so f is continuous at x. Since $x \in X$ was arbitrary, f is continuous.

$(1 \leftrightarrow 3)$: Similar to previous. ■

Note that this does not imply that the image of an open (closed) set under a continuous function is open (resp. closed): only inverse images preserve the topology of a set.

Lemma (Continuity of Compositions)**.** *Suppose that (X,d_X), (Y,d_Y), and (Z,d_Z) are metric spaces, and $f : X \to Y$ and $g : Y \to Z$ are continuous functions. Then, the composition $g \circ f : X \to Z$ is continuous.*

A direct ε-δ proof is long and tedious, but using the inverse image characterisation of continuity simplifies the proof considerably:

Proof. If $U \subseteq Z$ is open, then $g^{-1}[U]$ is open in Y, and hence $f^{-1}\big[g^{-1}[U]\big] = (g \circ f)^{-1}[U]$ is open in X. ■

37.3.2 Topologically Equivalent Metrics

Suppose we have two metrics d_1 and d_2 defined on a set X. We have characterised continuity in terms of open sets, so if the open sets (X,d_1) are the same as the open sets in (X,d_2), then any function $f : X \to Y$ that is continuous on (X,d_1) should be continuous on (X,d_2).

More formally, we note that $f = f \circ \mathrm{id}_X$, so we should require that the identity function $\mathrm{id}_X : X \to X$ is also continuous. But id_X is continuous from (X,d_1) to (X,d_2) if and only if every set that is open in (X,d_2) is also open in (X,d_1).

Lemma 37.3.2. *Suppose d_1 and d_2 are metrics on X. Then, the following statements are equivalent:*

(i) *Every set that is open in (X,d_2) is open in (X,d_1);*

(ii) *For any metric space (Y,d_Y), if $g : X \to Y$ is continuous as a function $(X,d_2) \to (Y,d_Y)$, then g is continuous as a function $(X,d_1) \to (Y,d_Y)$;*

(iii) *For any metric space (Y,d_Y), if $f : Y \to X$ is continuous as a function $(Y,d_Y) \to (X,d_1)$, then f is continuous as a function $(Y,d_Y) \to (X,d_2)$.*

Proof. We only show $(i) \leftrightarrow (ii)$, as the proof of $(i) \leftrightarrow (iii)$ is similar.

It follows from (i) that the identity map $\mathrm{id}_X : (X,d_1) \to (X,d_2)$ is continuous. So, if $g : (X,d_2) \to Y$ is continuous, then the composition $g \circ \mathrm{id}_X : (X,d_1) \to Y$ is continuous.

For the reverse implication, take $(Y,d_Y) = (X,d_2)$ and $g = \mathrm{id}_X : (X,d_2) \to (X,d_1)$. Since g is continuous from (X,d_2) to (X,d_2), it is continuous from (X,d_1) to (X,d_2), so for every open set U in (X,d_2), $g^{-1}[U] = U$ is open in (X,d_1). ■

Also applying this lemma in reverse, we obtain,

Theorem 37.3.3. *Suppose d_1 and d_2 are metrics on X. Then, the following statements are equivalent:*

(i) *The open sets in (X,d_2) and (X,d_1) coincide;*

(ii) *For any metric space (Y,d_Y), $g : X \to Y$ is continuous as a function $(X,d_1) \to (Y,d_Y)$ if and only if it is continuous as a function $(X,d_2) \to (Y,d_Y)$;*

(iii) *For any metric space (Y,d_Y), $f : Y \to X$ is continuous as a function $(Y,d_Y) \to (X,d_1)$ if and only if it is continuous as a function $(Y,d_Y) \to (X,d_2)$.*

In this case, we say that d_1 and d_2 are *topologically equivalent.*

Two metrics d_1 and d_2 on X are *Lipschitz equivalent* if there exist constants $0 < m \leq M < \infty$ such that,

$$md_1(x,y) \leq d_2(x,y) \leq Md_1(x,y)$$

for all $x,y \in X$.

Lemma 37.3.4. *If d_1 and d_2 are Lipschitz equivalent on X, then d_1 and d_2 are topologically equivalent.*

Recall that two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on a vector space X are equivalent if there exist constants $0 < m \leq M < \infty$ such that,

$$m\|\mathbf{x}\| \leq \|\mathbf{x}\| \leq M\|\mathbf{x}\|$$

for all $\mathbf{x} \in X$.

Each norm induces a metric $d_i(x,y) = \|x - y\|_i$ on X. The following corollary is immeditae from the previous lemma:

Corollary 37.3.4.1. *Metrices induced by equivalent norms are topologically equivalent.*

Example. The metrics induced by the ℓ^p norms on $\mathbb{R}^n$, $1 \le p \le \infty$ are topologically equivalent to each other (since the norms are equivalent). △

Example. The metrics $d(x,y)$ and $d_1(x,y) := \min\bigl(d(x,y),1\bigr)$ are topologically equivalent (they have the same open sets). △

As illustrated by the previous example, topologically equivalent metrics are not necessarily Lipschitz equivalent. However, for normed spaces, we have:

Lemma 37.3.5. *If X is a vector space and two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on X induce topologically equivalent metrics, then the norms are equivalent.*

Proof. Since the metrics are topologically equivalent, the identity map $\mathrm{id}_X : (X,d_1) \to (X,d_2)$ is continuous; this is the same as considering the identity map between the two normed spaces $(X,\|\cdot\|_1)$ and $(X,\|\cdot\|_2)$. In particular, the identity map is continuous at 0, so there exists $\delta > 0$ such that $\|x\|_2 < 1$ whenever $\|x\|_1 < \delta$.

For $y \in X$, take $x = \frac{\delta y}{2\|y\|_1}$, so that $\|x\|_1 = \frac{\delta}{2} < \delta$. It follows that

$$\left\|\frac{\delta y}{2\|y\|_1}\right\|_2 < 1$$

That is, $\|y\|_2 < \frac{2}{\delta}\|y\|_1$. Similarly, the identity map is continuous from $(X,\|\cdot\|_2)$ into $(X,\|\cdot\|_1)$, so $\|y\|_1 \le \frac{2}{\delta'}\|y\|_2$. ■

Example. The norms $\|\cdot\|_{L^1}$ and $\|\cdot\|_{L^2}$ on $C[0,1]$ are not topologically equivalent. △

37.3.3 Isometries and Homeomorphisms

Suppose $f : X \to Y$ is a bijection such that,

$$d_X(x,y) = d_Y(f(x),f(y))$$

for all $x,y \in X$. That is, f *preserves distances*. Then, f is an *isometry*, and X and Y are *isometric*. Isometric spaces are essentially the same metric spaces, just with different labelling of points, and in fact, isometries are exactly the isomorphisms of metric spaces.

If f and f^{-1} are both additionally continuous (f is *bicontinuous*), then f is a *homeomorphism*, and X and Y are *homeomorphic*. If two spaces are homeomorphic, their open sets coincide, and the spaces are essentially the same *topological* spaces, just with different labelling of points, and in fact, homeomorphisms are exactly the isomorphisms of topological spaces.

Example.

- Every metric space is homeomorphic to itself under the identity map.
- Any two open intervals (a,b) and (α,β) are homeomorphic under

$$f(x) = \alpha + \frac{\beta-\alpha}{b-a}(x-a)$$

- $(-1,1)$ is homeomorphic to $\mathbb{R}$ under

$$f(x) = \tan\left(\frac{\pi x}{2}\right) \qquad \text{or} \qquad f(x) = \frac{x}{1-|x|}$$

- Any open interval is homeomorphic to $\mathbb{R}$ by composing the previous two examples.

- The square is homeomorphic to the circle under a radial projection mapping.

△

Two metrics d_1 and d_2 on X are topologically equivalent if and only if the identity map $\mathrm{id}_X : (X,d_1) \to (X,d_2)$ is a homeomorphism.

37.3.4 Topological Properties

If a property P of a metric space is preserved under homeomorphism, then P is a *topological property.* Informally, topological properties are generally those properties that are set-theoretic in nature, and do not care about the exact notion of distance imposed on the space.

Example. Topological properties on a space X:

- X is open in X;
- X is closed in X;
- X is finite; countably infinite; uncountable;
- X has a point x such that $\{x\}$ is open in X (an *isolated point*);
- X has no isolated points;
- Every subset of X is open;
- Every continuous real-valued function on X is bounded.

△

Example. Non-topological properties (they intrinsically depend on the metric in some way):

- X is bounded;
- For each $r > 0$ there exists a finite set F such that every ball of radius r intersects F (X is *totally bounded*);

△

37.4 Topological Spaces

In light of this, it seems that many properties of a space do not depend on our exact choice of measure of distance, and we have already characterised convergence and continuity in metric spaces entirely in terms of open sets. We then may be prompted to dispense with the metric entirely, and define a new kind of space entirely in terms of open sets.

A *topology* on a set T is a collection of subsets $\mathcal{T} \subseteq \mathcal{P}(T)$, which we will call the "*open sets*" of T, such that,

(T1) T and $\varnothing$ are open;

(T2) The intersection of finitely many open sets is open;

(T3) Arbitrary unions of open sets are open.

The pair $(T,\mathcal{T})$ is then a *topological space.*

Example.

- In any metric space (X,d), the induced collection of open sets forms a topology on the underlying set X;

- The *discrete topology* – every set is open (induced by the discrete metric);
- The *indiscrete* or *trivial topology* – only T and $\varnothing$ are open;
- The *cofinite topology* – a set is open if it is T, $\varnothing$, or if its relative complement in T is finite;
- The *cocountable topology* – a set is open if it is T, $\varnothing$, or its relative complement in T is countable;
- The *Zariski topology* on $\mathbb{R}^n$ – a set is open if it is $\mathbb{R}^n$, $\varnothing$, or its complement is the set of zeros of some polynomial $p \in \mathbb{R}[x]$.

△

Not every topology is induced by a metric. That is, there does not necessarily always exist a metric on T that induces the same collection of open sets as $\mathcal{T}$. More formally, there does not always exist a metric d such that the metric space (T,d) is homeomorphic to the topological space $(T,\mathcal{T})$. If such a metric exists, then $(T,\mathcal{T})$ is *metrisable.*

Theorem 37.4.1. *The indiscrete topology is not metrisable on any set with more than one point.*

Proof. Suppose the indiscrete topology is on T is induced by some metric d on T. Let $x,y \in T$ be distinct, so $d(x,y) = \varepsilon > 0$. The open ball $\mathbb{B}(x,\varepsilon/2)$ is open in (T,d). This ball contains x, so it is not the empty set, and it does not contain y, so it is not T. But, $\varnothing$ and T are the only open sets in the indiscrete topology. ■

It is sometimes possible to compare two topologies on the same space if they are subsets of one another. If $\mathcal{T}_1$ and $\mathcal{T}_2$ are topologies on T, such that $\mathcal{T}_1 \subseteq \mathcal{T}_2$, then we say that $\mathcal{T}_1$ is *coarser* than $\mathcal{T}_2$, or that $\mathcal{T}_2$ is *finer* than $\mathcal{T}_1$. This defines a partial order on the set of topologies on T.

Sometimes we will say that a topology $\mathcal{T}$ is the coarsest or smallest topology that satisfies a given property P. This means that if $\mathcal{T}'$ also satisfies P, then $\mathcal{T}' \subseteq T$.

The discrete topology is the finest possible topology, and the indiscrete topology is the coarsest possible topology.

The *closed* sets in a topological space are the complements of open sets. By De Morgan's laws, the collection $\mathcal{F}$ of closed sets satisfies:

(F1) T and $\varnothing$ are closed;

(F2) The union of finitely many closed sets is closed;

(F3) Arbitrary intersections of closed sets are closed.

Because the closed sets completely determine the open sets, we can equivalently define a topology in terms of its open sets. In some cases, this is easier then specifying the open sets; for instance, the cofinite topology can be more naturally specified as the topology with finite closed sets.

37.4.1 Bases

A *base* or *basis* for a topology $\mathcal{T}$ on T is a collection $\mathcal{B} \subseteq \mathcal{T}$ such that every set in $\mathcal{T}$ is the union of sets in $\mathcal{B}$.

Example. A set U is open in a metric space (X,d) if for every $x \in U$, there exists $\varepsilon_x > 0$ such that $\mathbb{B}(x,\varepsilon_x) \subset U$, so,

$$U = \bigcup_{x \in U} \mathbb{B}(x,\varepsilon_x)$$

so in any metric space, the collection of all open balls forms a basis for the induced topology. △

A topology may have several distinct bases, but each basis generates a unique topology.

Theorem (Uniqueness of Topology for Basis)**.** *If $\mathcal{B}$ is a basis for two topologies $\mathcal{T}$ and $\mathcal{T}'$, then $\mathcal{T} = \mathcal{T}'$.*

Proof. Every set in $\mathcal{T}'$ is a union of sets in $\mathcal{B} \subset \mathcal{T}$, so $\mathcal{T}' \subseteq \mathcal{T}$. The situation is symmetric, so $\mathcal{T} \subseteq \mathcal{T}'$ and $\mathcal{T} = \mathcal{T}'$. ■

Theorem (Synthetic Bases)**.** *If $\mathcal{B}$ is a basis for $\mathcal{T}$ on T, then,*

(B1) T is the union of some sets from $\mathcal{B}$;

(B2) If $B_1, B_2 \in \mathcal{B}$, then $B_1 \cap B_2$ is the union of some sets from $\mathcal{B}$.

Conversely, let T be a set and let $\mathcal{B} \subseteq \mathcal{P}(T)$ satisfy (B1) and (B2). Then there is a unique topology $\mathcal{T}$ on T whose basis is $\mathcal{B}$; that is, the open sets are exactly those formed from union of sets from $\mathcal{B}$.

$\mathcal{T}$ is then also the smallest topology that contains $\mathcal{B}$.

Proof. We verify that the set $\mathcal{T}$ generated by the basis is a topology:

(T1) T is the union of sets from $\mathcal{B}$ by (B1);

(T2) If $U, V \in \mathcal{T}$, then $U = \bigcup_{i\in\mathcal{I}} B_i$ and $V = \bigcup_{j\in\mathcal{J}} C_j$ with $B_i, C_i \in \mathcal{B}$. Then,

$$U \cap V = \bigcup_{i\in\mathcal{I}, j\in\mathcal{J}} B_i \cap C_j$$

which is a union of sets in $\mathcal{B}$ by (B2), and is hence an element of $\mathcal{T}$;

(T3) Any union of union of sets from $\mathcal{B}$ is a union of sets from $\mathcal{B}$.

So, $\mathcal{T}$ is a topology on T. ■

A *sub-basis* for a topology $\mathcal{T}$ on a set T is a collection $\mathcal{B} \subseteq T$ such that every set in T is a union *of finite intersections* of sets in $\mathcal{B}$.

Example. One sub-basis of $\mathbb{R}$ with the standard topology is given by,

$$\mathcal{B} = \big\{(a,\infty), (-\infty, b) : a, b \in \mathbb{R}\big\}$$

as intersections give the open intervals (a,b) which form a normal basis. △

Theorem 37.4.2. *If $\mathcal{B}$ is any collection of subsets of a set T whose union is T, then there is a unique topology $\mathcal{T}$ on T with sub-basis $\mathcal{B}$, formed precisely from the collection of all unions of finite intersections of sets from $\mathcal{B}$.*

Proof. If $\mathcal{B}$ is a sub-basis for a topology $\mathcal{T}$, then this topology has $\mathcal{D}$, the collection of finite intersections of elements of $\mathcal{B}$ as a basis. But $\mathcal{D}$ satisfies (B1) and (B2), so there is a unique topology $\mathcal{T}$ with basis $\mathcal{D}$ by the previous lemma, which is also the unique topology with sub-basis $\mathcal{B}$. ■

$\mathcal{T}$ is the smallest topology on T that contains $\mathcal{B}$.

37.4.2 Topological Subspaces and Finite Product Spaces

If $(T,\mathcal{T})$ is a topological space, and $S \subseteq T$, then the *subspace topology* on S is,

$$\mathcal{T}_S = \{U \cap S : U \in \mathcal{T}\}$$

and we call $(S,\mathcal{T}_S)$ a *(topological) subspace* of $(T,\mathcal{T})$.

Lemma (Induced Subspaces)**.** *Suppose (X,d) is a metric space with induced topology $\mathcal{T}$. If $S \subseteq X$, then the subspace topology $\mathcal{T}_S$ on S corresponds to the topology on S induced by the metric subspace $(S,d|_S)$.*

If $(T_1,\mathcal{T}_1)$ and $(T_2,\mathcal{T}_2)$ are topological spaces, then the *product topology* on $T_1 \times T_2$ is the topology $\mathcal{T}$ with basis,

$$\mathcal{B} = \{U_1 \times U_2 : (U_1,U_2) \in T_1 \times T_2,\}$$

and we call $(T_1 \times T_2,\mathcal{T})$ the *(topological) product* of T_1 and T_2.[*]

Intuitively, the topological product is the smallest topology for which the left and right projections $\pi_1 : T_1 \times T_2 \to T_1$ and $\pi_2 : T_1 \times T_2 \to T_2$ are continuous.

For finite n, the product topology on $\mathbb{R}^n$ agrees with the topologies induced by any of the ϱ_p metrics.

37.4.3 Closures, Interiors, and Boundaries

Let $(T,\mathcal{T})$ be a topological space. A *neighbourhood* of a point $x \in T$ is a set $N \subseteq T$ that contains an open set $U \in \mathcal{T}$ such that $x \in U \subseteq N$. An *open neighbourhood* of $x \in T$ is an open set $U \in \mathcal{T}$ that contains x. General neighbourhoods are not used as often, so the unqualified term alone sometimes refers to an open neighbourhood.

The *closure* $\overline{A}$ of a set $A \subseteq T$ is the intersection of all closed sets that contain A:

$$\overline{A} = \bigcap_{\substack{A\subseteq\mathcal{F} \\ \mathcal{F}\text{ closed}}} V$$

Note that if A is non-empty, then $\overline{A}$ is non-empty; $\overline{A}$ is also always closed, since it is the intersection of closed sets; the closure of A is therefore the minimal closed set that contains A. It follows that A is closed if and only if $A = \overline{A}$.

For any sets A and B,

- $A \subseteq B \to \overline{A} \subseteq \overline{B}$;
- $\overline{A \cup B} = \overline{A} \cup \overline{B}$;
- in general, $\overline{A \cap B} \neq \overline{A} \cap \overline{B}$.

We give an alternative characterisation of closures:

Theorem (Characterisation of Closures)**.** *Given $A \subseteq T$, the closure $\overline{A}$ is the set,*

$$\begin{aligned}\overline{A} &= \{x \in T : U \cap A \neq \varnothing \textit{ for all open neighbourhoods } U \in \mathcal{T} \textit{ of } x\} \\ &= \{x \in T : \textit{every open neighbourhood of } x \textit{ intersects } A\}\end{aligned}$$

[*] More correctly, this is the *box topology*, but for finitely many products, this topology agrees with the true product topology discussed in §37.4.7. For infinite products, the product topology is more well-behaved, as it is a categorical product.

Proof. Let $x \in \overline{A}$. Suppose there is an open set U such that $x \in U$ and $U \cap A = \varnothing$. Then, $T \setminus U \supseteq A$. Since $T \setminus U$ is closed, $\overline{A} \subseteq T \setminus U$. However, this gives a contradiction, since $x \in \overline{A} \cap U$, and so $\overline{A} \cap U \neq \varnothing$. Therefore,

$$\overline{A} \subseteq \{x \in T : U \cap A \neq \varnothing \text{ for all open neighbourhoods } U \in \mathcal{T} \text{ of } x\}$$

Now suppose $x \in T$ is such that $U \cap A \neq \varnothing$ for every open set that contains x, but $x \notin \overline{A}$. Then, $x \notin F$ for some closed set that contains A. So, we have an open set $T \setminus F$ which contains x and satisfies $(T \setminus F) \cap A = \varnothing$, a contradiction. Therefore,

$$\overline{A} \supseteq \{x \in T : U \cap A \neq \varnothing \text{ for all open neighbourhoods } U \in \mathcal{T} \text{ of } x\}$$

■

It follows easily from this lemma that in $\mathbb{R}$, we have $\overline{\mathbb{Q}} = \overline{\mathbb{R} \setminus \mathbb{Q}} = \mathbb{R}$. This shows that in general,

$$\overline{H \cap K} \neq \overline{H} \cap \overline{K}$$

e.g., take $H = \mathbb{Q}$ and $K = \mathbb{R} \setminus \mathbb{Q}$. Then, $\overline{H} = \overline{K} = \mathbb{R}$ but $H \cap K = \varnothing = \overline{H \cap K}$.

In a metric spacce, we have another simple characterisation of the closure:

Theorem (Closure in Metric Spaces). *If (X,d) is a metric space, and $A \subseteq X$, then,*

$$\overline{A} = \{\textit{limits of convergent sequences in } A\}$$

Proof. If $(x_n)_{n=1}^{\infty}$ is a sequence in A, then it is also a sequence $\overline{A}$. If the sequence converges to $x \in X$, then, since $\overline{A}$ is closed, $x \in \overline{A}$ by sequential closure.

Conversely, if $x \in \overline{A}$, then for every $n \geq 1$, we hae $\mathbb{B}\left(x,\frac{1}{n}\right) \cap A \neq \varnothing$, so there exists $x_n \in A$ with $d(x_n,a) < \frac{1}{n}$. Clearly, $(x_n) \to a$ as $n \to \infty$. ■

The *interior* A° of a set $A \subseteq T$ is the union of all open subsets of A:

$$A^\circ = \bigcup_{\substack{U \subseteq A \\ U \text{ open}}} U$$

Since A° is the union of open sets, it is open, and it is contained in A. It is the maximal open subset of A. So, A is open if and only if $A = A^\circ$.

For any sets A and B,

- $A \subseteq B \to A^\circ \subseteq B^\circ$;
- $(A \cap B)^\circ = A^\circ \cap B^\circ$;
- in general, $(A \cup B)^\circ \neq A^\circ \cup B^\circ$.

We give an alternative characterisation of interiors:

Theorem (Characterisation of Interiors). *Given $A \subseteq T$, the interior A° is the set of all points for which A is an open neighbourhood:*

$$A^\circ \coloneqq \{x \in T : x \in U \subseteq A, U \in \mathcal{T}\}$$

Proof. If $x \in A^\circ$, then $x \in U$ for some open set $U \subseteq A$, so A is a neighbourhood of x. Conversely, if A is a neighbourhood of x, then there is an open subset $U \subseteq A$ such that $x \in U$, so $x \in A^\circ$. ■

Theorem 37.4.3. *If $A \subseteq T$, then,*

1. $A^\circ = T \setminus \overline{T \setminus A}$
2. $\overline{A} = T \setminus (T \setminus A)^\circ$

In this way, the interior operator is dual to the closure operator.

Proof. (1): If $x \in A^\circ$, then A is a neighbourhood of X that does not intersect $T \setminus A$, so $x \notin \overline{T \setminus A}$,and $x \in T \setminus \overline{T \setminus A}$. If $x \in T \setminus \overline{T \setminus A}$, then $x \notin \overline{T \setminus A}$, so there is an open set containing x that does not meet $T \setminus A$. So this open set is a subset of A, so $x \in A^\circ$, and hence $A^\circ = T \setminus \overline{T \setminus A}$.

(2): Similar to (dual of) previous. ■

The *boundary* ∂A of a set A is the set of all points x such that every neighbourhood of x intersects both A and and its complement:

$$\partial A = \{x \in T : \text{if } U \ni x \text{ is open, then } U \cap A \neq \varnothing \text{ and } U \cap (T \setminus A) \neq \varnothing\}$$

It is immediate from the definition that

$$\partial A = \overline{A} \cap \overline{T \setminus A}$$

so ∂A is always closed. By the previous theorem, we also have,

$$\begin{aligned}\partial A &= \overline{A} \cap (T \setminus A^\circ)\\ &= \overline{A} \setminus A^\circ\end{aligned}$$

Example. In $\mathbb{R}$, $\partial(a,b) = \partial[a,b] = \{a,b\}$; $\partial\mathbb{Q} = \mathbb{R}$. △

Let $S \subseteq T$. A point $x \in T$ is a *limit point* of S if every neighbourhood of x intersects $S \setminus \{x\}$. Note that a limit point of S need not lie within S. Intuitively, a limit point is "nearby" other points in S, in that, if we remove x from S and look in some neighbourhood around x, then we still see some other points contained in S. In contrast, a point in S that is in S that is not a limit point of S is an *isolated point.*

Example.

- If $S = (0,1) \subset \mathbb{R}$, then every point in $[0,1]$ is a limit point of S;
- If $S = [0,1] \cup \{2\}$, then 2 is not a limit point of S, as we can find a neighbourhood containing 2 that does not intersect S, say, $(1.5,2.5)$, so 2 is an isolated point of S.

△

Note that if S is closed, then it contains all its limit points, or else we would have a limit point $x \in T \setminus S$, which would be an open set containing x that does not intersect $S = S \setminus \{x\}$, so,

$$\overline{S} = S \cup \{x : x \text{ is a limit point of } S\}$$

A subset $A \subseteq T$ is,

- *dense* in T if $\overline{A} = T$;
- *nowhere dense* in T if $(\overline{A})^\circ = \varnothing$;
- *meagre* in T, or *of the first category* in T, if it is a union of countably many nowhere dense sets.

Example. $\mathbb{Q}$ is dense in $\mathbb{R}$, as is $\mathbb{R} \setminus \mathbb{Q}$. In $\mathbb{R}$, singleton sets are nowhere dense; so $\mathbb{Q}$ is meagre in $\mathbb{R}$. However, $\mathbb{R} \setminus \overline{\mathbb{Q}} = \varnothing$, so $\mathbb{Q}$ is not nowhere dense. △

Equivalently, a set $A \subseteq T$ is nowhere dense if $T \setminus \overline{A}$ is dense in T since,

$$(\overline{A})^\circ = T \setminus \overline{T \setminus \overline{A}}$$

If A is closed, this reduces to if $T \setminus A$ is dense.

37.4.4 The Cantor Set

We construct a set that is pathological in many ways and serves as a useful counterexample to many propositions.

The (*middle third*) *Cantor set* is constructed as follows:

(0) Let $C_0 = [0,1] \subset \mathbb{R}$.

(1) Remove the open middle third of this set, leaving,

$$C_1 = \left[0,\frac{1}{3}\right] \cup \left[\frac{2}{3},1\right]$$

(n) From each of the 2^{n-1} closed intervals from C_{n-1}, remove the open middle third to give a new set that consists of 2^n closed intervals.

Note that C_n consists of 2^n closed intervals, each of length 3^{-n}, so their total length is $\left(\frac{2}{3}\right)^n \to 0$ as $n \to \infty$.

Then, the set,

$$C = \bigcap_{n=0}^{\infty} C_n$$

is the (middle third) Cantor set.

Since each C_n is closed, C is closed as it is the intersection of closed sets. C is also non-empty as it contains the endpoints of every open interval removed, but the interior of C is empty, as the set would otherwise have non-zero length. Since C is closed, it is nowhere dense in [0,1].

We have $\partial C = C$, since $\overline{C} = C$, and $C^\circ = \varnothing$. C also contains no isolated points; for any $\varepsilon > 0$, any point in C was at some point in an interval of length less than $\varepsilon/2$, and the two endpoints of this interval are both in C.

37.4.5 The Hausdorff Property

Recall that a topological space $(T,\mathcal{T})$ is said to be metrisable if there is a metric d on T such that $\mathcal{T}$ consists of the open sets in (T,d).

Not all topological spaces are metrisable, as we have seen with the indiscrete topology. But, there are other more natural topologies that cannot be derived from a metric. One way to show that a topology is not metrisable is to find a property that all metrisable spaces must satisfy and show that it fails to hold. We will give such a necessary condition for a topological space to be metrisable in terms of convergence of sequences.

A sequence $(x_n)_{n=1}^{\infty} \subseteq T$ in a topological space $(T,\mathcal{T})$ converges to $x \in T$ if for every open neighbourhood U of x, there exists $N \geq 1$ such that $x_n \in U$ for all $n \geq N$.

Note that this is the same definition of convergence for metric spaces we found earlier. However, in topological spaces, this can lead to some unusual behaviours. Take T to have the indiscrete topology where only T and $\varnothing$ are open. Then, any sequence $(x_n) \subset T$ converges to any point x in T; the only open neighbourhood of x is $U = T$, and $x_n \in T$ for all $n \geq 1$, so $(x_n) \to x$ for all $x \in T$. The problem here is that in the indiscrete topology, distinct points cannot be separated into distinct open sets. This motivates the next definition:

A topological space T is *Hausdorff* if for any distinct $x,y \in T$, there exist disjoint open sets U,V such that $x \in U$ and $y \in V$.

Theorem 37.4.4. *Metric spaces are Hausdorff.*

Proof. Let (X,d) be a metric space. Take distinct $x,y \in X$, and let $\varepsilon = d(x,y) > 0$. Then, $x \in \mathbb{B}(x,\frac{\varepsilon}{2})$ and $y \in \mathbb{B}(y,\frac{\varepsilon}{2})$, but $\mathbb{B}(x,\frac{\varepsilon}{2}) \cap \mathbb{B}(y,\frac{\varepsilon}{2}) = \varnothing$. ■

Example. The indiscrete topology on at least two points is not Hausdorff, since for any pair of points x,y, the only open set containing x is T, which also contains y. So, the indiscrete topology on at least two points is not metrisable. △

Example. The co-finite topology on any infinite set is not Hausdorff: since any two open sets have finite complements, they must intersect. So, the co-finite topology is not metrisable △

Theorem 37.4.5. *In a Hausdorff space T, any sequence has at most one limit.*

Proof. Suppse $(x_n) \to x$ and $(x_n) \to y$ with $x \neq y$. Then, there exist disjoint open sets U and V such that $x \in U$ and $y \in V$. By the definition of convergence, there exist $N_1 \geq 1$ and $N_2 \geq 1$ such that $x_n \in U$ for all $n \geq N_1$ and $x_n \in V$ for all $n \geq N_2$.

If $n \geq \max(N_1,N_2)$, we must simultaneously have $x_n \in U$ and $x_n \in V$, but $U \cap V = \varnothing$. ■

Note that the converse of this theorem does not hold: there exist non-Hausdorff topologies in which convergent sequences have unique limits.

37.4.6 Topological Continuity

We previously characterised continuity between metric spaces in terms of open sets. We now reverse this to *define* continuity between topological spaces to be in terms of open sets.

A map $f : T_1 \to T_2$ between two topological spaces $(T_1,\mathcal{T}_2)$ and $(T_2,\mathcal{T}_2)$ is *continuous* if $f^{-1}[U] \subseteq T_1$ is open in T_1 whenever $U \subseteq T_2$ is open in T_2.

Example.

- Any constant map that sends every $x \in T_1$ to some fixed $c \in T_2$ is continuous as $f^{-1}[U] = T_1$ if $c \in U$ and $f^{-1}[U] = \varnothing$ if $c \notin U$.
- The identity map $f : T_1 \to T_1$ is continuous if the domain and codomain have the same topology.
- Continuous maps between metric spaces are also continuous maps between the induced topological spaces.
- Any map with a discrete domain is continuous since every set in the domain is open.

△

To check that a map is continuous, it suffices to check that it is continuous on a sub-basis. (Since every basis is also a sub-basis, we may also check for a basis.)

Lemma 37.4.6. *Suppose that $f : T_1 \to T_2$ is a map between two topological spaces $(T_1,\mathcal{T}_2)$ and $(T_2,\mathcal{T}_2)$, and that $\mathcal{B}$ is a sub-basis for the topology $\mathcal{T}_2$. Then, f is continuous if and only if $f^{-1}[B]$ is open T_1 for every $B \in \mathcal{B}$.*

Proof. The reverse direction is clear since every element of the sub-basis is an element of $\mathcal{T}_2$.

Now, any element U of $\mathcal{T}_2$ can be written as $U = \bigcup_i D_i$ where each set D_i is a finite intersection of some elements $\{B_j\}_{j=1}^{n} \subseteq \mathcal{B}$. So,

$$f^{-1}[U] = f^{-1}\left(\bigcup_i D_i\right)$$

$$\begin{aligned} &= \bigcup_i f^{-1}[D_i] \\ &= \bigcup_i f^{-1}\left[\bigcap_{j=1}^n B_j\right] \\ &= \bigcup_i \bigcap_{j=1}^n f^{-1}[B_j] \end{aligned}$$

The innermost intersection is open by assumption, so $f^{-1}[U]$ is thus the union of open sets and is hence open. ■

Lemma (Continuity of Compositions)**.** *Suppose that $(X,\mathcal{T}_X)$, $(Y,\mathcal{T}_Y)$, and $(Z,\mathcal{T}_Z)$ are topological spaces, and $f : X \to Y$ and $g : Y \to Z$ are continuous functions. Then, the composition $g \circ f : X \to Z$ is continuous.*

Proof. If $U \subseteq Z$ is open, then $g^{-1}[U]$ is open in Y, and hence $f^{-1}\big(g^{-1}[U]\big) = (g \circ f)^{-1}[U]$ is open in X. ■

We now discuss continuity in product spaces. Suppose that $(X,\mathcal{T}_X)$ and $(Y,\mathcal{T}_Y)$ are two topological spaces. For the product $X \times Y$, we define the left and right projections,

$$\pi_1 : X \times Y \to X, \qquad \pi_2 : X \times Y \to Y$$

by

$$\pi_1(x,y) = x, \qquad \pi_2(x,y) = y$$

Lemma 37.4.7. *The projection mappings are continuous in the product topology.*

Proof. If $U \subseteq X$ is open, then $\pi_1^{-1}[U] = U \times Y$, which is open. A similar argument shows π_2 is continuous. ■

Theorem (Componentwise Continuity)**.** *A function $f : T \to X \times Y$ with components $f = (f_1,f_2)$ is continuous if and only if its components are continuous.*

Proof. Since π_i is continuous, so is $f_i = \pi_i \circ f$.

For the reverse implication, the open sets $U \times V$, with U open in X and V open in Y form a basis of $X \times Y$, and we have,

$$f^{-1}[U \times V] = f_1^{-1}[U] \cap f_2^{-1}[V]$$

is open in T. ■

If we consider maps from T into $\mathbb{R}$ (or more generally, into any field K), then we can consider sums, products, and quotients of these maps.

Lemma (Algebra of Continuous Functions)**.** *If $f,g : T \to \mathbb{R}$ are continuous, then $f + g$ and fg are continous, and f/g is continous at all points x where $g(x) \neq 0$.*

Proof. We give the argument for $f+g$. The intervals (a,∞) and $(-\infty,b)$ for every $a,b \in \mathbb{R}$ are a sub-basis for the topology of $\mathbb{R}$, so it is sufficient to show that $(f+g)^{-1}\big[(a,\infty)\big]$ and $(f+g)^{-1}\big[(-\infty,b)\big]$ are open in T.

$$\begin{aligned} f(x)+g(x) > a \quad &\leftrightarrow \quad f(x) > a - g(x) \\ &\leftrightarrow \quad f(x) > r \quad \text{and} \quad r > a - g(x) \quad \text{for some } r \\ &\leftrightarrow \quad f(x) > r \quad \text{and} \quad g(a) > a - r \quad \text{for some } r \end{aligned}$$

It follows that

$$\big\{x : f(x)+g(x) > a\big\} = \bigcup_{r \in \mathbb{R}} \big\{x : f(x) > r\big\} \cap \big\{x : g(x) > a - r\big\}$$

which is open. Similarly,

$$f^{-1}\big[(-\infty,b)\big] = \big\{x : f(x)+g(x) < b\big\}$$

is open. A similar argument works for products and quotients. ■

Alternative proof. The function $\sigma : \mathbb{R}^2 \to \mathbb{R}$ given by $\sigma(x,y) = x+y$ is continuous from $\mathbb{R}^2$ to $\mathbb{R}$. In fact, it is Lipschitz continuous from $(\mathbb{R}^2,\varrho_1)$ into $\mathbb{R}$ since

$$\begin{aligned} d_{\mathbb{R}}(x_1+y_1,x_2+y_2) &= \big|(x_1+y_1)-(x_2+y_2)\big| \\ &\le |x_1-x_2| + |y_1-y_2| \\ &\le \varrho_1\big((x_1,y_1),(x_2,y_2)\big) \end{aligned}$$

The map $x \mapsto \big(f(x),g(x)\big)$ is continous from T into $\mathbb{R}^2$ by componentwise continuity, so the composition $\big(\sigma \circ (f,g)\big)(x) = f(x)+g(x)$ is continuous from T into $\mathbb{R}$. A similar argument works for products and quotients. ■

Example. The map from $\mathbb{R}^2$ to $\mathbb{R}^2$ given by

$$(x,y) \mapsto \big(x+y, \sin(x^2y^3)\big)$$

is continuous because both components are the sum, product, and composition of the continuous functions π_1, π_2, and sin. △

37.4.7 The Projective Topology

Consider a set T (without a topology), a collection of topological spaces $(T_i,\mathcal{T}_i)_{i\in\mathcal{I}}$, and a collection of maps $f_i : T \to T_i$. This data defines a topology on T:

The *projective topology* on T is the coarsest topology for which all the maps $f_i : T \to T_i$ are continuous.

Recall that the coarsest topology is the one with the smallest collection of open sets. In order for f_i to be continuous, we must hae that $f^{-1}[U]$ is open whenever $U \in \mathcal{T}_i$, so the projective topology contains

$$\mathcal{B} \coloneqq \bigcup_{i\in\mathcal{I}} \big\{f_i^{-1}[U] : U \in \mathcal{T}_i\big\}$$

This is not a basis for a topology, since if $B_1,B_2 \in \mathcal{B}$, then $B_1 \cap B_2$ is not necessarily a union of sets in $\mathcal{B}$. However, the union of sets in $\mathcal{B}$ is equal to T, since for any i, $f_i^{-1}[T_i] = T$, so we can apply Theorem 37.4.2 to get that the projective topology is the unique topology with $\mathcal{B}$ as a sub-basis.

An example of this topology is given by the product topology. Let $(T_1,\mathcal{T}_1)$ and $(T_2,\mathcal{T}_2)$ be topological spaces. Recall that the product topology $\mathcal{T}$ on $T_1 \times T_2$ is the topology with basis

$$\mathcal{B} = \{U_1 \times U_2 : U_1 \in \mathcal{T}_1, U_2 \in \mathcal{T}_2\}$$

The product topology $\mathcal{T}$ may also be characterised as the coarsest topology, $\mathcal{T}'$, for which the two projection maps are continuous.

First note that for any $U_1 \in \mathcal{T}_1$, $\mathcal{T}'$ must contain $\pi_1^{-1}[U_1] = U_1 \times T_2$, and similarly, for any $U_2 \in \mathcal{T}_2$, $\mathcal{T}'$ must contain $\pi_2^{-1}[U_2] = T_1 \times U_2$. So, $\mathcal{T}'$ must also contain the intersection of such sets, i.e., $U_1 \times U_2$. That is, $\mathcal{T}' \supseteq \mathcal{B}$, and therefore $\mathcal{T}' \supseteq \mathcal{T}$. Conversely, by definition, for $U_1 \in \mathcal{T}_1$, $\mathcal{T}$ contains $U_1 \times T_2 = \pi_1^{-1}[U_1]$, and for any $U_2 \in \mathcal{T}_2$, $\mathcal{T}$ contains $T_1 \times U_2 = \pi_2^{-1}[U_2]$. So, $\mathcal{T}$ is a topology that makes π_1 and π_2 continuous. Since $\mathcal{T}'$ is the coarsest such topology, $\mathcal{T} \supseteq T'$. Thus $\mathcal{T} = \mathcal{T}'$ as required.

We now use this approach to define the product topology for an arbitrary product.

Let $(T_i,\mathcal{T}_i)_{i\in\mathcal{I}}$ be an arbitrary collection of topological spaces. Their product $T = \prod_{i\in\mathcal{I}} T_i$ is the set of all functions $x : \mathcal{I} \to \bigcup_{i\in\mathcal{I}} T_i$ such that $x(i) \in T_i$. (We will soon discuss this definition in more detail.) The product topology $\mathcal{T}$ on T is the coarsest topology for which all the projections

$$\pi_i : T \to T_i : x \mapsto x(i)$$

are continuous. We then call the topological space $(T,\mathcal{T})$ the *topological product* of the spaces $(T_i,\mathcal{T}_i)_{i\in\mathcal{I}}$, and a sub-basis for the product topology consists of all sets of the form

$$\prod_{i\in\mathcal{I}} U_i$$

where $U_i \in \mathcal{T}_i$, with $U_i = T_i$ for all but finitely many i.

Now, why do we define T to be a set of functions? Let us consider the ordinary binary product $X_1 \times X_2$ of two sets X_1 and X_2. We are familiar with this being defined as a set of ordered pairs:

$$X_1 \times X_2 = \big\{(x_1,x_2) : x_1 \in X_1, x_2 \in X_2\big\}$$

Similarly, for any collection $(X_i)_{i=1}^n$, we may define the product to be a set of ordered n-tuples:

$$\prod_{i=1}^n X_i = \big\{(x_1,\dots,x_n) : \forall i \in \{1,\dots,n\}, x_i \in X_i\big\}$$

Note that this definition agrees with induction on the binary case, up to a natural isomorphism of sets to account for associativity.

Now, suppose that we have a collection $(X_i)_{i\in\mathbb{N}}$ of sets indexed by the natural numbers. Then, we can define the product of these sets to be a set of sequences:

$$\begin{aligned}\prod_{i\in\mathbb{N}} X_i &= \big\{(x_1,x_2,x_3,\dots) : \forall i \in \mathbb{N}, x_i \in X_i\big\} \\ &= \big\{(x_i)_{i=1}^\infty : \forall i \in \mathbb{N}, x_i \in X_i\big\}\end{aligned}$$

More generally, if we have a collection of sets indexed by a countable set, we can simply biject to (an initial segment of) $\mathbb{N}$ and again define the product to be a set of sequences.

Now suppose we have a collection $(X_i)_{i\in\Lambda}$, where Λ is uncountable. How might we define

$$\prod_{i\in\Lambda} X_i$$

We might be tempted to write the elements formally as $(x_i)_{i\in\Lambda}$, but what does this notation really mean?

We go back to the case where the indexing set is $\mathbb{N}$, and the elements in our product are sequences. Now, recall that a sequence $(x_i)_{i\in\mathbb{N}} \subseteq X$ is really just a function $x : \mathbb{N} \to X$, and the notation $x_i = x(i)$ is just syntactic sugar.

In the product of $(X_i)_{i\in\mathbb{N}}$, we have $x_i \in X_i$ for all i, so the codomain of our sequence is just the union $\bigcup_{i\in\mathbb{N}} X_i$, and the elements (sequences) $(x_i)_{i\in\mathbb{N}}$ in the product are precisely the functions

$$x : \mathbb{N} \to \bigcup_{i\in\mathbb{N}} X_i$$

satisfying $x(i) \in X_i$.

Note that nothing here is actually specific to $\mathbb{N}$ – only the cardinality of the indexing set is really relevant – so this holds more generally.

For instance, in a binary product, an element (x_1,x_2) is isomorphic to a function $x : \{1,2\} \to X_1 \cup X_2$, where $x(1) = x_1$ and $x(2) = x_2$ (again $\{1,2\}$ is arbitrary – we just need any two-element indexing set, but the arguments matching up with the indices makes this a sensible choice). For an n-ary product, we have functions

$$x : \{1,\dots,n\} \to \bigcup_{i\in\{1,\dots,n\}} X_i$$

where again, $x(i) = x_i$.

More concretely, the isomorphism

$$\prod_{i\in\mathcal{I}} X_i \cong \left\{ x : \mathcal{I} \to \bigcup_{i\in\mathcal{I}} X_i : \forall i \in \mathcal{I}, x(i) \in X_i \right\}$$

is given by $(x_i)_{i\in\mathcal{I}} \mapsto (x : i \mapsto x_i)$, with inverse $x \mapsto \big(x(i)\big)_{i\in\mathcal{I}}$.

Given this, it is natural to *define* the product set

$$\prod_{i\in\Lambda} X_i$$

to be the set of functions

$$x : \Lambda \to \bigcup_{i\in\Lambda} X_i : \forall i \in \Lambda, x(i) \in X_i$$

so when we write $(x_i)_{i\in\Lambda}$ for an element of this product, we mean a function $x : \Lambda \to \bigcup_{i\in\Lambda} X_i$ defined by $x(i) = x_i$ for all $i \in \Lambda$.

37.4.8 Homeomorphisms

Recall that a homeomorphism between metric spaces is a bijective and bicontinuous map. The notion of a homeomorphism between topological spaces is essentially the same.

Let $(T_1,\mathcal{T}_1)$ and $(T_2,\mathcal{T}_2)$ be topological spaces. A function $f : T_1 \to T_2$ is a homeomorphism if it is bijective and any of the following equivalent conditions hold:

(*i*) both f and f^{-1} are continuous;

(*ii*) $U \subseteq T_2$ is open in T_2 if and only if $f^{-1}[U] \subseteq T_1$ is open in T_1;

(*iii*) $V \subseteq T_1$ is open in T_1 if and only if $f(V) \subseteq T_2$ is open in T_2.

If a homeomorphism between $(T_1,\mathcal{T}_1)$ and $(T_2,\mathcal{T}_2)$ exists, then we say that $(T_1,\mathcal{T}_1)$ and $(T_2,\mathcal{T}_2)$ are *homeomorphic*.

A property of topological spaces is a *topological invariant* or *topological property* if it is preserved by homeomorphisms.

Example.

- T is finite;
- T is Hausdorff;
- T is metrisable;
- every continuous real-valued function on T is bounded.

△

To show that two topological spaces are not homeomorphic, we can show that one space has a topological invariant that the other does not. For instance, every continuous real-valued function on $[0,1]$, but not $\mathbb{R}$, is bounded, so $[0,1]$ and $\mathbb{R}$ are not homeomorphic.

37.5 Compactness

A *cover* of a set A is a collection $\mathcal{U}$ of sets whose union contains A. That is,

$$A \subseteq \bigcup_{U \in \mathcal{U}} U$$

and we say that the elements of $\mathcal{U}$ *cover* A. A *subcover* of a cover $\mathcal{U}$ is a subset of $\mathcal{U}$ whose elements still cover A. A cover is *open* if every element of the cover is open.

Example.

- $\mathcal{U} = \{(n-2,n+2) : n \in \mathbb{Z}\}$ is an (open) cover of $\mathbb{R}$, with one possible subcover given by $S = \{(n-2,n+2) : n \in 2\mathbb{Z}\}$;
- $\mathcal{U} = \{(n,n+1) : n \in \mathbb{Z}\}$ is not a cover of $\mathbb{R}$ since it does not cover the integers.

△

Note that a subcover is a subset of a cover – we do *not* modify (the size of) sets within the cover. That is, while $S = \{(n-1,n+1) : n \in \mathbb{Z}\}$ covers $\mathbb{R}$, it is *not* considered a subcover of $\mathcal{U} = \{(n-2,n+2) : n \in \mathbb{Z}\}$ because $S \not\subseteq \mathcal{U}$.

A topological space T is *compact* if every open cover of T has a finite subcover.

Example.

- $(0,1)$ is not compact because $\mathcal{U} = \{(0,a) : a \in (0,1)\}$ is an open cover with no finite subcover;
- $\mathbb{R}$ is not compact because $\mathcal{U} = \{(-\infty,a), a \in \mathbb{R}\}$ has no finite subcover.

△

A subset S of T is compact if every open cover of S by subsets of T has a finite subcover. This is equivalent to S being compact with respect to the subspace topology.

Lemma 37.5.1. *If $(T,\mathcal{T})$ is a topological space and $S \subseteq T$, then the two notions of compactness above are equivalent.*

Theorem (Heine–Borel)**.** *Any closed interval $[a,b]$ is a compact subset of $\mathbb{R}$ with the standard topology.*

Proof. Let $\mathcal{U}$ be a cover of $[a,b]$ by open subsets of $\mathbb{R}$, and let A denote the set of all points $p \in [a,b]$ such that $[a,p]$ can be covered by a finite subcover S of $\mathcal{U}$. We note that A is non-empty as $[a,a] = \{a\}$ can certainly be covered.

A is bounded above by b, so we can define $c \coloneqq \sup A \leq b$. Since $a \leq c \leq b$, we must have $c \in U$ for some open set $U \in \mathcal{U}$. Since U is open, there exists $\delta > 0$ such that $(c-\delta,c+\delta) \subseteq U$.

Since $c = \sup A$, there exists at least one point $x \in A$ such that $x \in (c-\delta,c]$. Since $[a,x]$ can be covered by S, and $(c-\delta,c+\delta) \subset U \in \mathcal{C}$, it follows that

$$[a,c+\delta) = [a,x] \cup (c-\delta,c+\delta)$$

also be covered by a finite collection of sets from $\mathcal{U}$ – namely $S \cup \{U\}$.

If $c < b$, then this yields a finite subcover of

$$\big[a, \min(c - \tfrac{\delta}{2},b)\big]$$

which contradicts that $c = \sup A$, so $c = b$, and hence a finite subcover of $\mathcal{U}$ covers $[a,b+\delta) \supset [a,b]$, so $[a,b]$ is compact. ■

Lemma (Closed in Compact is Compact)**.** *Any closed subset S of a compact space T is compact.*

Proof. Let $\mathcal{U}$ be any cover of S by open subsets of T. Because S is closed, $T \setminus S$ is open, so $\mathcal{U} \cup \{T \setminus S\}$ is an open cover of T.

By the compactness of T, there exists an finite open subcover of this cover. This subcover (with extraneous elements like $T \setminus S$ removed) also covers S, so S is compact. ■

Lemma (Compact in Hausdorff is Closed)**.** *Any compact subspace S of a Hausdorff space T is closed.*

Proof. Let $a \in T \setminus S$. For each $x \in S$, there exist disjoint open sets $U(x)$ and $V(x)$ containing a and x, respectively. The open sets $\{U(x) : x \in S\}$ form an open cover of S, so there is a finite subcover $\big\{U(x_i)\big\}_{i=1}^{n}$ of S. Then,

$$V_a = \bigcap_{i=1}^{n} V(x_i)$$

is open as it is the finite intersection of open sets; contains a as $V(x)$ contains a for all x by construction; and is disjoint from S by the Hausdorff property, and hence $V_a \subseteq T \setminus S$ for any a. Then,

$$T \setminus S = \bigcup_{a \in T \setminus S} V_a$$

so $T \setminus S$ is the union of open sets and is hence open, so S is closed in T. ■

Lemma (Compact in Metric is Bounded)**.** *Any compact subspace S of a metric space (X,d) is bounded.*

Proof. Fix any $a \in X$. For any $x \in S$, $x \in \mathbb{B}(a,r)$ for all $r > d(a,x)$, so S is covered by the collection of open balls $\mathcal{U} = \{\mathbb{B}(a,r) : r > 0\}$. By compactness of S, there is a finite subcover $S = \{\mathbb{B}(a,r_i)\}_{i=1}^{n}$, so

$$S \subseteq \bigcup_{i=1}^{n} \mathbb{B}(a,r_i) = \mathbb{B}(a, \max_i r_i)$$

and K is bounded. ■

Corollary (Compact in $\mathbb{R}$ iff Closed, Bounded)**.** *A subset S of $\mathbb{R}$ with the standard topology is compact if and only if it is closed and bounded.*

Proof. Since $\mathbb{R}$ is a metric space, any compact subset is bounded; since $\mathbb{R}$ is Hausdorff, any compact subset is closed.

For the converse, if $S \subset \mathbb{R}$ is bounded, then there exists $r > 0$ such that $S \subseteq [-r,r]$, which is compact in $\mathbb{R}$. Then, S is a closed subset of a compact set, so S is compact. ■

Theorem 37.5.2. *Let $F_1 \supseteq F_2 \supseteq F_3 \supseteq \cdots$ be a chain of non-empty closed subsets of a compact space T. Then,*

$$\bigcap_{i=1}^{\infty} F_i \neq \varnothing$$

37.5.1 Compact Products and Compact Subsets of $\mathbb{R}^n$

Theorem (Tychonov)**.** *The product of any collection of compact spaces is compact with the product topology.*

Theorem (Heine–Borel in $\mathbb{R}^n$)**.** *A subset of $\mathbb{R}^n$ is compact if and only if it is closed and bounded.*

Proof. Let S be a compact subset of $\mathbb{R}^n$. It follows from a previous lemma that S is bounded. Metric spaces are also Hausdorff, so S is closed.

For the converse, suppose that S is bounded, so there exists $r > 0$ such that $S \subseteq [-r,r]^n$. Since $[-r,r]$ is compact (by Heine–Borel in $\mathbb{R}$), it follows that $[-r,r]^n$ is compact (by Tychonov). If S is closed, then it is a closed subset of a compact space, and is hence compact. ■

Note that this result does not hold in general metric spaces. For instance, (0,1) is bounded and is closed in itself, but is not compact.

37.5.2 Continuous Functions on Compact Sets

Theorem (Continuous Image of Compact is Compact)**.** *Let $f : T \to S$ be a continuous function between topological spaces. If T is compact, then $f(T) \subseteq S$ is compact.*

Proof. Suppose $\mathcal{U}$ is an open cover of $f(T)$. Then, $f^{-1}[U]$ is open for all $U \in \mathcal{U}$, and the collection $\left\{f^{-1}[U] : U \in \mathcal{U}\right\}$ of these sets covers T. Because T is compact, it has a finite subcover $\left\{f^{-1}[U_i]\right\}_{i=1}^{n}$, and hence $\{U_i\}_{i=1}^{n}$ is a finite subcover of $f(T)$. ■

This corollary shows that compactness is a topological property.

Theorem 37.5.3. *Let $f : T \to S$ be a continuous bijection. If T is compact and S is Hausdorff, then f is a homeomorphism.*

Proof. Let $K \in T$ be closed, and hence compact. Then, $f(K)$ is compact, and, since S is Hausdorff, $f(K)$ is closed. The inverse image of K under f^{-1} (that is, under $(f^{-1})^{-1} = f$) is then closed, so f^{-1} is continuous. ■

Corollary 37.5.3.1. *If T is non-empty and compact, then a continuous function $f : T \to \mathbb{R}$ is bounded and attains its bounds.*

37.5.3 Lebesgue Numbers and Uniform Continuity

Let $\mathcal{U}$ be an open cover of a metric space (X,d). A number $\delta > 0$ is called a *Lebesgue number* for $\mathcal{U}$ if for every $x \in X$, there exists an open set $U \in \mathcal{U}$ such that $\mathbb{B}(x,\delta) \subseteq U$.

In general, open covers do not have a Lebesgue number. For instance, $\mathcal{U} = \left\{(\frac{x}{2},x) : x \in (0,1)\right\}$ form an open cover of (0,1), but the covering sets become arbitrary small as $x \to 0$, so no Lebesgue number exists.

Lemma (Lebesgue's Number Lemma)**.** *Every open cover $\mathcal{U}$ of a compact metric space (X,d) has a Lebesgue number.*

A map $f : (X,d_X) \to (Y,d_Y)$ between metric spaces is *uniformly continuous* if for every $\varepsilon > 0$ there exists $\delta > 0$ such that $d_Y(f(x),f(y)) < \varepsilon$ whenever $d_X(x,y) < \delta$ for all $x,y \in X$.

As usual in uniform definitions, δ may depend only on ε and not on x nor y.

Theorem (Compact Continuous is Uniform). *A continuous map from a compact metric space into a metric space is uniformly continuous.*

37.5.4 Sequential Compactness

A subset K of a metric space (X,d) is *sequentially compact* if every sequence in K has a convergent subsequence whose limit lies in K.

Lemma 37.5.4. *If K us a sequentially compact subset of a metric space, then any open cover of K has a Lebesgue number.*

Proof. Suppose $\mathcal{U}$ is an open cover of K that does not have a Lebesgue number. Then, for every $\varepsilon > 0$, there exists $x \in K$ such that $\mathbb{B}(x,\varepsilon)$ is not contained in any element of $\mathcal{U}$.

Let $(x_n)_{i=1}^{\infty}$ be a sequence such that $\mathbb{B}(x_n,\frac{1}{n})$ is not contained in any element of $\mathcal{U}$ as above. By sequential compactness, (x_n) has a convergent subsequence $(x_{n_i}) \to x \in K$, and since $\mathcal{U}$ covers K, $x \in U$ for some open set $U \in \mathcal{U}$.

Since $\mathcal{U}$ is open, there exists $\varepsilon > 0$ such that $\mathbb{B}(x,\varepsilon) \subseteq U$. Now, take suficiently large i such that $d(x_{n_i},x) < \frac{\varepsilon}{2}$ and $\frac{1}{n_i} < \frac{\varepsilon}{2}$. Then, $\mathbb{B}(x_{n_i},\frac{1}{n_i}) \subseteq \mathbb{B}(x,\varepsilon) \subseteq U$, contradicting the construction of $(x_n) \supseteq (x_{n_i})$. ■

Theorem (Sequentially Compact is Compact). *A metric subspace is sequentially compact if and only if it is compact.*

The equivalence of compactness and sequential compactness in metric spaces show that they are also equivalent in normed spaces.

However, there exist closed bounded subsets in general normed spaces that are not compact. For instance, the closed unit ball in ℓ^p is not compact for any $1 \leq p \leq \infty$: consider the sequence of basis vectors $(\mathbf{e}_i)_{i=1}^{\infty}$. This sequence has no convergence subsequence, as any such subsequence would necessarily be Cauchy, but,

$$\|\mathbf{e}_i - \mathbf{e}_j\|_{\ell^p} = \begin{cases} 2^{\frac{1}{p}} & 1 \leq p < \infty \\ 1 & p = \infty \end{cases}$$

In fact, the compactness of the closed unit ball is necessary and sufficient for a normed space to be finite-dimensional:

Theorem 37.5.5. *A normed space is finite-dimensional if and only if its closed unit ball is compact.*

37.6 Connectedness

A pair of sets (A,B) is a *partition* of a topological space T if $T = A \cup B$ and $A \cap B = \varnothing$, and we say that A and B *partition* T.

It is clear from the definition that if A and B partition T, they are open if and only if they are closed.

A topological space T is *connected* if the only partitions of T into open (closed) sets are $(T,\varnothing)$ and $(\varnothing,T)$, and is *disconnected* otherwise.

Lemma (Characterisation of Disconnected Spaces). *The following statements are all equivalent:*

(*i*) *T is disconnected;*

(*ii*) *T has a partition into two non-empty open sets;*

(*iii*) *T has a partition into two non-empty closed sets;*

(*iv*) *T has a clopen subset equal to neither $\varnothing$ nor T;*

(*v*) *there is a continuous function from T to the two-point set $\{0,1\}$ with the discrete topology.*

Proof.

(*i*) ↔ (*ii*): Follows from the definition of disconnected.

(*ii*) ↔ (*iii*): If A and B are open and $T = A \cup B$, then $A = T \setminus B$ and $B = T \setminus A$ are closed in T.

(*ii*) ↔ (*iv*): As above, if A,B satisfy the hypotheses of (*ii*), they are clopen, so (*iv*) holds. Now, suppose (*iv*) holds, and let $A \subset T$ be non-empty and clopen. Then, $B = T \setminus A$ is open and A,B partition T.

(*ii*) ↔ (*v*): Let χ_A be the indicator function of A. Then, $\chi_A^{-1}\big[\{1\}\big] = A$ and $\chi_A^{-1}\big[\{1\}\big] = B$ and $\chi_A^{-1}\big[\{0,1\}\big] = T$, which are all open, so f is continuous. For the converse, suppose $f : T \to \{0,1\}$ is a continuous surjection. Define $A := f^{-1}\big[\{0\}\big]$ and $B := f^{-1}\big[\{1\}\big]$. Both A and B are open as f is continuous; are non-empty as f is surjective; and $A \cup B = T$ and $A \cap B = \varnothing$, so A and B partition T. ■

Statement (*v*) allows us to show that a space is connected by showing that any continuous function $T \to \{0,1\}$ must be a constant function. Equivalently, we can show that a space is disconnected by exhibiting a non-constant (or equivalently, surjective) function $T \to \{0,1\}$.

Statement (*iv*) shows that if a subset of T is clopen, then it is either empty or is all of T.

A subset S of T is connected (disconnected) if $(S,\mathcal{T}_S)$ is connected (disconnected). That is, S is connected (disconnected) as a topological space under the subspace topology.

A subset $S \subseteq T$ is *separated* by sets $U,V \in \mathcal{T}$ if,

- $S \subseteq U \cup V$;
- $S \cap U \neq \varnothing$;
- $S \cap V \neq \varnothing$;
- $S \cap U \cap V = \varnothing$.

That is, S is at least (partially) contained within both U and V individually, contained entirely within U and V together, but is not contained in any overlap between U and V (if any such overlap exists).

Theorem 37.6.1. *A subspace $(S,\mathcal{T}_S)$ of a space $(T,\mathcal{T})$ is disconnected if and only if it is separated by some sets $U,V \in \mathcal{T}$.*

Proof. If S is disconnected, then there are non-empty $A,B \in \mathcal{T}_S$ such that $S = A \cup B$ and $A \cap B = \varnothing$. By the definition of the subspace topology, there exist $U,V \in \mathcal{T}$ such that $A = U \cap S$ and $B = V \cap S$. Then, U and V separate S. Conversely, if U and V separate S, then $U \cap S$ and $V \cap S$ partition S and S is disconnected. ■

37.6.1 Connected Subsets of $\mathbb{R}^n$

An *interval* of the real line is any set of the form,

- $[a,b] = \{x \in \mathbb{R} : a \leq x \leq b\}$;
- $[a,b) = \{x \in \mathbb{R} : a \leq x < b\}$;

- $(a,b] = \{x \in \mathbb{R} : a < x \leq b\}$;
- $(a,b) = \{x \in \mathbb{R} : a < x < b\}$.

where $a,b \in \mathbb{R} \cup \{-\infty, +\infty\}$, with infinite values allowed only with strict inequalities.

Lemma 37.6.2. *A set $I \subseteq \mathbb{R}$ is an interval if and only if whenever $x,z \in I$ and $x < y < z$, then $z \in I$.*

That is, an interval contains all points between any pair of points in the interval.

Proof. The intervals above are all defined to have this property. For the converse, suppose $I \subseteq \mathbb{R}$ satisfies this property, and let $a = \inf I$ and $b = \sup I$. Certainly, $(a,b) \subseteq I$, for if $z \in (a,b)$, then there exists $\alpha,\beta \in I$ with $\alpha < z < \beta$ by the definition of a and b, which implies that $z \in I$. Now,

$$(a,b) \subseteq I \subseteq (a,b) \cup \{a,b\}$$

■

Theorem (Intervals are Connected)**.** *A subset of $\mathbb{R}$ is connected if and only if it is an interval.*

Proof. If an interval I is not connected, then there is a continuous surjective map $f : I \to \{0,1\}$. Note that when considered as a function $f : I \to \mathbb{R}$, then this is also continuous, since given any open subset U of $\mathbb{R}$, we have

$$f^{-1}[U] = \begin{cases} f^{-1}\big[\{0\}\big] & 0 \in U, 1 \notin U \\ f^{-1}\big[\{1\}\big] & 0 \notin U, 1 \in U \\ f^{-1}\big[\{0\}\big] \cup f^{-1}\big[\{1\}\big] & 0,1 \notin U \\ \varnothing & 0,1 \notin U \end{cases}$$

and all these sets are open. But, if $f(x) = 0$ and $f(y) = 1$, then f takes all values in between by the intermediate value function, which is a contradiction.

For the reverse implication, suppose $I \subseteq \mathbb{R}$ is not an interval, so there exist x,y,z such that $x < z < y$, and $x,y \in I$, but $z \notin I$. Let $A = (-\infty,z) \cap I$ and $B = (z,\infty) \cap I$. Then, A and B are disjoint, open in I (by definition of the subspace topology), and non-empty, since $x \in A$ and $y \in B$. We have $I = A \cup B$, since $z \notin I$. So, I is not connected. ■

37.6.2 Operations on Connected Sets

Theorem (Union of Overlapping Connected Sets)**.** *If $(C_i)_{i \in \mathcal{I}}$ are connected subsets of T and $C_i \cap C_j \neq \varnothing$ for all $i,j \in \mathcal{I}$, then,*

$$K = \bigcup_{i \in \mathcal{I}} C_i$$

is connected.

Proof. Suppose $f : K \to \{0,1\}$ is continous. Since each C_i is connected, $f(C_i) = \{\delta_i\}$ where δ_i is either 0 or 1 for each i. Since $C_i \cap C_j$ is non-empty for all $i,j \in \mathcal{I}$, it follows that $f(C_i)$ takes the same value for every $i \in I$, so f must be a constant function and hence K is connected. ■

Lemma 37.6.3. *Suppose C and D are connected subsets of T and $\overline{C} \cap D \neq \varnothing$. Then, $C \cup D$ is connected.*

Proof. Let $K = C \cup D$, and suppose $f : K \to \{0,1\}$ is continous (and $\{\}$). Suppose $f(C) = \{0\}$ and $f(D) = \{1\}$ (we just require that C and D have disjoint images in $\{0,1\}$).

$\{1\}$ is open in $\{0,1\}$, so $f^{-1}\big[\{1\}\big]$ is open in K, and is hence $f^{-1}\big[\{1\}\big] = U \cap K$ for some open set $U \in T$ by the definition of the subspace topology.

Since $\overline{C} \cap D \neq \varnothing$, there exists $x \in D$ such that every open neighbourhood of x in T intersects C. $U \ni x$ is such a set, so $U \cap C \neq \varnothing$. But, $C \subseteq K$, so this is the same as,

$$\begin{aligned} \varnothing &\neq U \cap C \\ &= U \cap (K \cap C) \\ &= f^{-1}[\{1\}] \cap C \end{aligned}$$

but $f(C) = \{0\}$, so this implies $\{1\} \cap \{0\} \neq \varnothing$. It follows that C and D cannot have disjoint images in $\{0,1\}$, so f must be constant on K, so K is connected. ■

Theorem (Union of Connected Subspaces). *Suppose C and $(C_i)_{i \in \mathcal{I}}$ are connected subspaces of T and $\overline{C} \cap C_i \neq \varnothing$ for all $i \in \mathcal{I}$. Then,*

$$C \cup \bigcup_{i \in \mathcal{I}} C_i$$

is connected.

Proof. Define $C_i' := C \cup C_i$ for $i \in \mathcal{I}$. Then, each C_i' is connected by the previous lemma. We also have $C_i' \cap C_j' = (C \cup C_i) \cap (C \cup C_j) = C \cup (C_i \cap C_j)$ so $C_i' \cap C_j' \neq \varnothing$ for all $i,j \in \mathcal{I}$, and $\bigcup_{i \in \mathcal{I}} C_i' = \bigcup_{i \in \mathcal{I}} C \cup C_i = C \cup \bigcup_{i \in \mathcal{I}} C_i = K$, so the C_i' and K satisfy the hypotheses of the previous theorem, and hence K is connected. ■

Corollary (Subsets of Closure). *If $C \subseteq T$ is connected, then so is any set K satisfying $C \subseteq K \subseteq \overline{C}$.*

Proof. $K = C \cup \bigcup_{x \in K} \{x\}$, and $\{x\} \cap \overline{C} \neq \varnothing$ for all $x \in K$. ■

Theorem (Continuous Image of Connected is Connected). *Let $f : T \to S$ be a continuous function between topological spaces. If T is connected, then $f(T) \subseteq S$ is connected.*

Proof. Suppose T is connected. If $f(T)$ is disconnected, then there exists a non-constant $g : f(T) \to \{0,1\}$. But then, $g \circ f$ is a continuous non-constant function $T \to \{0,1\}$, contradicting that T is connected. It follows that no such g exists, so $f(T)$ is connected. ■

This corollary shows that connectedness is a topological property, i.e., if T is connected and $T \cong S$, then S is connected.

Theorem (Connected Product). *If topological spaces T and S are connected, then the topological product $T \times S$ is connected.*

Proof. Let $t \in T$, $s \in S$ and define $C := T \times \{s_0\}$ and $C_t := \{t\} \times S$. Then, C is homeomorphic to T and C_t is homeomorphic to S, so both are connected. $C \cap C_t$ is non-empty as $(t,s) \in C, C_t$, so $C \cap \overline{C_t} \supset C \cap C_t$ is non-empty. We also have,

$$T \times S = C \cup \bigcup_{t \in T} C_t$$

so $T \times S$ is connected as a union of connected subspaces. ■

To show that a set is connected, we can show that it can be constructed from the continuous images of known connected sets (often intervals), via products, and via unions.

Example.

- $\mathbb{R}^2$ is connected as $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ and $\mathbb{R} = (-\infty,\infty)$ is connected as it is an interval.

- The circle S^1 is connected, as it is the continuous image of $[0,2\pi]$ by $t \mapsto (\sin t, \cos t)$.
- The unit square is connected, as it is the image of S^1 under a radial projection mapping, which is continuous.
- $\mathbb{R}^2 \setminus \{(0,0)\}$ is connected, since $\mathbb{R}^n \setminus \{0\} = \bigcup_{r\in\mathbb{R}} rS^1$ is the union of circles.
- The *topologist's sine curve*

$$\mathscr{S} := \left\{ \left(x, \sin\frac{1}{x} \right) : x \in \mathbb{R}, x \neq 0 \right\} \cup \big\{(0,0)\big\}$$

  is a connected subspace of $\mathbb{R}^2$: let $S_- = \big\{ \big(x, \sin(1/x)\big) : x < 0 \big\}$ and $S_+ = \big\{ \big(x, \sin(1/x)\big) : x > 0 \big\}$, and $O = \big\{(0,0)\big\}$. S_- and S_+ are connected as the images of the intervals $(-\infty,0)$ and $(0,\infty)$, respectively, under the continuous map $x \mapsto \big(x, \sin(1/x)\big)$, and O is a point, so it is connected.

  Note that $O \subseteq \overline{S_-}$ and $O \subseteq \overline{S_+}$, so both $S_- \cup O$ and $S_+ \cup$ are connected by Lemma 37.6.3. Since $(S_- \cup O) \cap (S_+ \cup O) = O \neq \varnothing$, it follows that

$$\mathscr{S} = S_- \cup O \cup S_+$$

  is connected as it is the union of overlapping connected sets.
- The *harmonic comb*

$$\mathscr{H} := \big\{ (x,y) : y = 0, x \in (0,1] \big\} \cup \left\{ \left(\frac{1}{n}, y \right) : n \in \mathbb{N}, y \in [0,1] \right\} \cup \big\{(0,1)\big\}$$

  is connect as the union of vertical lines, all of which intersect the horizontal line, plus $(0,1)$, which is contained in the closure of the vertical lines.

△

We have shown that connectedness is a topological invariant. More useful, however, is that the property "$T\setminus\{x\}$ is connected for every $x \in T$" is a topological property. That is, if $f : T \to S$ is a homeomorphism, then for any $y \in S$, the set $S \setminus \{y\}$ is the continuous image of $T \setminus \{x\}$ for some $x \in T$. This can be used to show that certain sets are not homeomorphic by finding a point that disconnect one set when removed, while proving no such points exist for the other.

Example.

- $\mathbb{R}^2$ is not homeomorphic to $\mathbb{R}$: $\mathbb{R} \setminus \{0\}$ is disconnected, but $\mathbb{R}^2 \setminus \big\{(0,0)\big\}$ is still connected.
- $[0,1]$ is not homeomorphic to S^1: removing an interior point from $[0,1]$ disconnects the set, but removing a point from the circle leaves it connected.
- Similarly, $[0,1]$ is not homeomorphic to the unit square.

△

37.6.3 Connected Components

We can define an equivalence relation on a topological space T by having $x \sim y$ if and only if $x,y \in C$ for some connected $C \subseteq T$. This relation is clearly reflexive and symmetric. For transitivity, suppose $x \sim y$ and $y \sim z$, so $x,y \in C_1$ and $y,z \in C_2$. $C_1 \cap C_2$, is non-empty as it contains y, so $C_1 \cup C_2$ is connected and $x,z \in C_1 \cup C_2$, so $x \sim z$.

The equivalence classes of $\sim$ are called the *connected components* of T.

- The connected component C containing x is the union of all connected subsets of T that contain x:

$$C = \bigcup_{x \in S \subseteq T} S$$

- Connected components are connected;
- Connected components are closed;
- Connected components are maximal connected subsets of T.

Example. The connected components of $T = (0{,}1) \cup (1{,}2)$ are $(0{,}1)$ and $(1{,}2)$; the connected components of $\mathbb{Q}$ are the singleton sets $\{p\}$, where $p \in \mathbb{Q}$; the connected components of $\mathbb{R} \setminus \mathbb{Q}$ are the singleton sets $\{q\}$, where $q \in \mathbb{R} \setminus \mathbb{Q}$. △

Since the continuous image of a connected space is connected, the number of connected components is a topological invariant.

37.6.4 Path-Connected Spaces

Given two points s and t in a topological space T, a *path* from s to t, or a *s-t-path*, is a continuous map $\varphi : [0{,}1] \to T$ such that $\varphi(0) = s$ and $\varphi(1) = t$.

A space T is *path-connected* if every pair of points T can be joined by a path in T.

Theorem (Path-Connected is Connected)**.** *A path-connected space is connected.*

Proof. Fix $s \in T$, and let $t \in T$. The path $C_v = \varphi\big([0{,}1]\big)$ is then connected as it is the continuous image of a connected space. Then, $T = \{u\} \cup \bigcup_{v \in T} C_v$, and each C_v contains u, so T is connected. ■

In general, the converse of this theorem does not hold, so path-connectedness is a stronger notion of connectedness. However, there are some specific cases where the two are equivalent.

Theorem 37.6.4. *Connected open subsets of $\mathbb{R}^n$ are path-connected.*

Proof. Let $U \subseteq \mathbb{R}^n$ be connected and open. Let $u \in U$, and let A be the set of all points in U that can be reached from u by a path contained in U. Let $B = U \setminus A$. We will show that B is empty by proving that if it is not, then A and B form a partition of U.

Let $a \in A$. Since U is open, there exists $\varepsilon > 0$ such that $\mathbb{B}(a,\varepsilon) \subseteq U$, so there is a path joining a to any $x \in \mathbb{B}(a,\varepsilon)$. Concatenating the path from u to a to this path from a to x yields a path from u to x, so $\mathbb{B}(a,\varepsilon) \subseteq \mathbb{A}$, so A is open.

For any $b \in B$, we have $\mathbb{B}(b,\varepsilon) \subseteq U$, so if there is a path from u to $z \in \mathbb{B}(b,\varepsilon)$, there would be a path from u to x, so $\mathbb{B}(b,\varepsilon) \subseteq B$, so B is also open.

Now, if B is non-empty, we have $U = A \cup B$, $A \cap B \cap U = \varnothing$, $A \cap U \neq \varnothing$, and $B \cap U \neq \varnothing$. But U is connected. ■

Theorem 37.6.5. *Connected components of open subsets of $\mathbb{R}^n$ are open.*

Proof. Let $U \subseteq \mathbb{R}^n$ be open and let C be one of its connected components. If $x \in C \subseteq U$, then there exists $\varepsilon > 0$ such that $\mathbb{B}(x,\varepsilon) \subseteq U$ as U is open. But C is the union of all connected subsets of U that contain x, so $\mathbb{B}(x,\varepsilon) \subseteq C$ and C is open. ■

Theorem 37.6.6. *A subset U of $\mathbb{R}$ is open if and only if it is the union of countably many disjoint open intervals:*

$$U = \bigcup_{i\in\mathcal{I}} (a_i,b_i), \qquad (a_i,b_i) \cap (a_j,b_j) = \varnothing \text{ for all } i \neq j$$

Proof. Any union of open sets is open. For the converse, let $U \subseteq \mathbb{R}$ be open, and let $(C_i)_{i\in\mathcal{I}}$ be the collection of its connected components, which are mutually disjoint. These components are open by the previous theorem, and since they are open and connected, they are open intervals. Then, for each C_i, we can pick a rational q_i in C_i, so we can index the connected components by $\mathbb{Q}$, which is countable. ■

37.7 Completeness in Metric Spaces

Recall that a sequence $(x_n)_{n=1}^{\infty}$ converges in a metric space (X,d) if and only if it is Cauchy. That is, if for every $\varepsilon > 0$, there exists N such that, for all $n,m \geq N$,

$$d(x_n,x_m) < \varepsilon$$

A metric space (X,d) is *complete* if every Cauchy sequence in X converges, and is *incomplete* otherwise.

It is implicit in the definition that the limit must lie in X. So, for example, $\mathbb{R}$ and $\mathbb{C}$ are complete, but $(0,1)$ is not complete, as $(\frac{1}{n})_{n=1}^{\infty}$ is Cauchy, but $\frac{1}{n} \to 0 \notin (0,1)$. Since $\mathbb{R}$ and $(0,1)$ are homeomorphic (given by $x \mapsto \frac{1}{1+e^{-x}}$, or any other sigmoid curve), this shows that completeness is *not* a topological property. A topological space X that is metrisable with at least one metric d on X such that (X,d) is a complete metric space is called *completely metrisable.*

Theorem 37.7.1 (Complete Subset is Closed)**.** *Let (X,d) be a metric space, and let $S \subseteq X$. If $(S,d|_S)$ is complete, then S is closed in X.*

Proof. Suppose $(x_n)_{n=1}^{\infty} \subseteq S$ with $(x_n) \to x$. Then, (x_n) is Cauchy in S, and thus converges to some $y \in S$. Since $x_n,y \in S$, it follows that $d|_S(x_n,y) = d(x_n,y)$, so $(x_n) \to y$ in X, i.e., $x = y$, and S is closed. ■

Theorem 37.7.2 (Closed in Complete is Complete)**.** *Let (X,d) be a metric space, and let $S \subseteq X$. If (X,d) is complete and S is closed, then $(S,d|_S)$ is complete.*

Proof. If $(x_n)_{n=1}^{\infty} \subseteq S$ is Cauchy in S, then (x_n) is also Cauchy in X, and thus converges to some $x \in X$. Since S is closed, $x \in S$, so $(x_n) \to x$ in S. ■

Theorem (Compact Metric is Complete)**.** *Any compact metric space (X,d) is complete.*

Proof. If $(x_n)_{n=1}^{\infty}$ is a Cauchy sequence in X, then it has a convergent subsequence (x_{n_i}) (since compact implies sequentially compact in a metric space) with $(x_{n_i}) \to x \in X$. But if a Cauchy sequence has a convergent subsequence, then the whole sequence converges to x: given any $\varepsilon > 0$, find N such that

$$d(x_n,x_n) < \frac{\varepsilon}{2}$$

for all $n,m \geq N$, and I such that $n_I \geq N$ and $d(x_{n_i},x) < \frac{\varepsilon}{2}$ for all $i \geq I$. Then, for all $k \geq n_I$, we have

$$\begin{aligned} d(x_k,x) &\leq d(x_k,x_{n_I}) + d(x_{n_I},x) \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\ &= \varepsilon \end{aligned}$$

■

37.7.1 Examples of Complete Spaces

Note that all our examples will be normed spaces: a normed space is *complete* if it is complete as a metric space, i.e., a Cauchy sequence is $(x_n)_{i=1}^{\infty}$ such that for every $\varepsilon > 0$, there exists N such that

$$\|x_n - x_m\| < \varepsilon$$

for all $n,m \geq N$, and any such sequence should converge to some $x \in X$, i.e., $\|x_n - x\| \to 0$ as $n \to \infty$.

A complete normed space is also called a *Banach* space.

Theorem 37.7.3. *$\mathbb{R}^d$ is Banach for all $d \in \mathbb{N}$.*

Proof. Let $(\mathbf{x}^{(i)})_{i=1}^{\infty}$ (the indices are written as superscripts as we need to access the subscripts to index components) be a Cauchy sequence in $\mathbb{R}^d$. Then, for every $\varepsilon > 0$, there exists $N(\varepsilon)$ such that

$$\left\|\mathbf{x}^{(n)} - \mathbf{x}^{(m)}\right\| = \sqrt{\sum_{i=1}^{d} \left|x_i^{(n)} - x_i^{(m)}\right|^2} < \varepsilon$$

for all $m,n \geq N(\varepsilon)$. In particular, for each $i = 1,\ldots,d$ we have

$$\left|x_i^{(n)} - x_i^{(m)}\right| < \varepsilon$$

for all $m,n \geq N(\varepsilon)$, so $\left(x_i^{(n)}\right)_{i=1}^{\infty}$ is a Cauchy sequence. Since Cauchy sequences of real numbers converge, $\left(x_i^{(n)}\right) \to x_i$ for some $x_i \in \mathbb{R}$. Now, set $x = (x_1,\ldots,x_d)$. Then,

$$\lim_{n\to\infty} \left\|\mathbf{x}^{(n)} - \mathbf{x}\right\| = \lim_{n\to\infty} \sqrt{\sum_{i=1}^{d} \left|x_i^{(n)} - x_i\right|^2} = 0.$$

so $\left(\mathbf{x}^{(n)}\right) \to \mathbf{x}$. ■

We proved this theorem using the standard Euclidean norm, but since all norms are equivalent on $\mathbb{R}^n$, $\mathbb{R}^n$ is complete in any norm.

Theorem 37.7.4. *ℓ^p is complete for all $1 \leq p \leq \infty$*

Theorem 37.7.5. *For any non-empty set X, the space $B(X)$ of bounded real-valued functions defined on X, $f : X \to \mathbb{R}$ under the supremum norm,*

$$\|f\|_\infty := \sup_{x\in X} |f(x)|$$

is complete.

Proof. Let $(f_n)_{n=1}^{\infty}$ be a Cauchy sequence in $B(X)$. Then, for every $\varepsilon > 0$ there exists $N(\varepsilon)$ such that

$$\|f_n - f_m\|_\infty = \sup_{x\in X}\left|f_n(x) - f_m(x)\right| < \varepsilon$$

for all $n,m \geq N(\varepsilon)$. In particular, for each $x \in X$, we have

$$\left|f_n(x) - f_m(x)\right| < \varepsilon$$

for all $n,m \geq N(\varepsilon)$, so $\left(f_n(x)\right)_n^{\infty}$ is a Cauchy sequence in $\mathbb{R}$. Since $\mathbb{R}$ is complete, $f_n(x)$ converges for each $x \in X$. Now, define $f : X \to \mathbb{R}$ by setting

$$f(x) = \lim_{n\to\infty} f_n(x)$$

for each $x \in X$. For any $\varepsilon > 0$, we have

$$\big|f_n(x) - f(x)\big| \leq \varepsilon$$

for all $n \geq N(\varepsilon)$, by letting $m \to \infty$ in the previous equation.

Since $N(\varepsilon)$ does not depend on x, this implies that

$$\big|f_{N(1)}(x) - f(x)\big| \leq 1$$

for every $x \in X$, so f is bounded, i.e., an element of $B(X)$, and that,

$$\|f_n - f\|_\infty \leq \varepsilon$$

for all $n \geq N(\varepsilon)$, i.e., that $(f_n) \to f$ in the supremum norm. ■

Theorem 37.7.6. *For any non-empty topological space $(T,\mathcal{T})$, the space $C_b(T)$ of bounded and continuous real-valued functions defined on X, $f : X \to \mathbb{R}$ under the supremum norm is complete.*

Proof. Suppose $f \in \overline{C_b(T)}$, where the closure is taken in $B(T)$. Then, for any $\varepsilon > 0$, there exists $f_\varepsilon \in C_b(T)$ such that $\|f - f_\varepsilon\|_\infty < \varepsilon$. Next, we show that for any $a \in \mathbb{R}$, we have

$$\big\{x : f(x) > a\big\} = \bigcup_{\varepsilon>0} \big\{x : f\varepsilon(x) > a + \varepsilon\big\}$$

Indeed, if $f(x) > a$, then we can take $\varepsilon = \frac{f(x)-a}{2}$ and then

$$\begin{aligned} f_\varepsilon(x) &= f(x) - \big(f(x) - f_\varepsilon(x)\big) \\ &> f(x) - \varepsilon \\ &= a + \varepsilon \end{aligned}$$

while if $f_\varepsilon(x) > a + \varepsilon$, then

$$\begin{aligned} f(x) &= f_\varepsilon(x) - \big(f_\varepsilon(x) - f(x)\big) \\ &> (a + \varepsilon) - \varepsilon \\ &= a \end{aligned}$$

which proves the equality. Now, since each f_ε is continuous, each set in the union is open, and so $f^{-1}\big[(a,\infty)\big]$ is open. A similar argument works for $\big\{x : f(x) < a\big\}$, and f is continuous since $\big\{(a,\infty),(-\infty,a) : a \in \mathbb{R}\big\}$ forms a sub-basis for the open sets of $\mathbb{R}$. ■

Theorem 37.7.7. *For any non-empty compact topological space T, the space $C(T)$ of continuous real-valued functions defined on X is complete under the maximum norm*

$$\|f\|_\infty = \max_{x\in T}\big|f(x)\big|$$

Proof. If $f \in C(T)$ and T is compact, then f is bounded, so $C(T) = C_b(T)$, and f attains its bounds, so

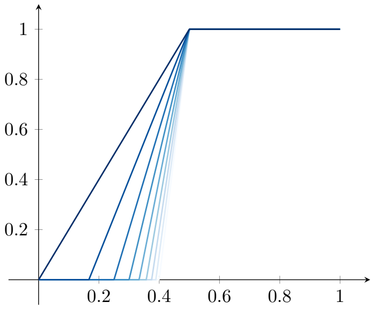
$$\sup_{x\in T}\big|f(x)\big| = \max_{x\in T}\big|f(x)\big|$$

■

37.7.2 Completions

Consider the space $C[0,1]$ of continuous functions defined on $[0,1]$ under the L^1 norm. This space is not complete as there exist Cauchy sequences that do not converge to a function in $C[0,1]$. For example,

$$f_n(x) = \begin{cases} 0 & 0 \le x < \frac{1}{2} - \frac{1}{n} \\ 1 - n(\frac{1}{2} - x) & \frac{1}{2} - \frac{1}{n} \le x < \frac{1}{2} \\ 1 & \frac{1}{2} < x \le 1 \end{cases}$$



$f_n(x)$ for $n = 2,3,\ldots$

For $n > m$,

$$\begin{aligned} \|f_n - f_m\|_{L^1} &= \left(\int_0^1 |f_n(x) - f_m(x)|^1 \, dx \right)^{\frac{1}{1}} \\ &= \int_0^1 |f_n(x) - f_m(x)| \, dx \\ &\le \frac{1}{m} \end{aligned}$$

so this sequence is Cauchy. f converges in the L^1 norm to the function

$$f(x) = \begin{cases} 0 & 0 \le x < \frac{1}{2} \\ 1 & \frac{1}{2} \le x \le 1 \end{cases}$$

since

$$\begin{aligned} \|f_n - f\|_{L^1} &= \int_0^1 |f_n(x) - f(x)| \, dx \\ &= \int_{\frac{1}{2}-\frac{1}{n}}^{\frac{1}{2}} |f_n(x)| \\ &\le \frac{1}{n} \end{aligned}$$

Clearly, $f \notin C[0,1]$, so $C[0,1]$ with the L^1 norm is incomplete.

An incomplete space can be *completed* by adding in the missing limit points, resulting in the *completion* of that space. There are two methods of completing an incomplete space A:

(i) Find a complete metric space $X \supset A$ such that A is dense in X. That is, $\overline{A} = X$.

(ii) Find a complete metric space X and an isometry $\mathfrak{i} : A \to Y$ with $Y \subseteq X$ and $\overline{Y} = X$.

Example.

(*i*) $\mathbb{Q}$ is dense in $\mathbb{R}$ and $\mathbb{R}$ is complete, so $\mathbb{R}$ is the completion of $\mathbb{Q}$.

(*ii*) $\mathbb{R}$ is the completion of $\mathbb{Q}$ with isometry $\mathfrak{i}(x) = x$ or $\mathfrak{i}(x) = -x$.

△

The second method is more flexible in that we do not have to find a complete space that contains exactly A, but only a space isometric to A.

Theorem 37.7.8. *Every metric space* (X,d) *can be isometrically embedded into the complete space metric space* $B(X)$*.*

Proof. Given (X,d), define $\mathfrak{i} : X \to B(X)$ by choosing some $a \in X$ and then setting

$$[\mathfrak{i}(x)](z) = d(z,x) - d(z,a)$$

Note that for every $z \in X$ we have

$$\begin{aligned}\big|[\mathfrak{i}(x)(z)]\big| &= \big|d(z,x) - d(z,a)\big| \\ &\leq d(x,a)\end{aligned}$$

so $\mathfrak{i}(x) \in B(X)$. Since

$$\begin{aligned}\big|[\mathfrak{i}(x)](z) - [\mathfrak{i}(y)](z)\big| &= \big|d(z,x) - d(z,a)\big| \\ &\leq d(x,y)\end{aligned}$$

and we have equality when $z = x$ or $z = y$, it follows that

$$\big\|\mathfrak{i}(x) - \mathfrak{i}(y)\big\|_{\infty} = d(x,y)$$

so the map $\mathfrak{i}$ is an isometry of (X,d) onto a subset of $B(X)$. ■

Corollary 37.7.8.1. *Every metric space has a completion.*

Proof. Embed (X,d) into $B(X)$ via the previous theorem. Then $\overline{\mathfrak{i}(X)}$ (with the closure taken in $B(X)$) is a closed subset of a complete space and is thus complete, and clearly, $\mathfrak{i}(X)$ is dense in $\overline{\mathfrak{i}(X)}$. ■

One can also complete any normed space to find a complete normed space, but the construction is significantly more involved.

37.8 The Contraction Mapping Theorem

A map $f : X \to X$ is a *contraction* if

$$d\big(f(x),f(y)\big) \leq \kappa d(x,y)$$

for all $x,y \in X$ and some $\kappa \in [0,1)$. The smallest such value of κ is called the *Lipschitz constant* of f.

Any contraction is continuous, so if $(x_n) \to x$, we have $\big(f(x_n)\big) \to f(x)$.

Theorem (Contraction Mapping). *Let* (X,d) *be a non-empty complete metric space, and* $f : X \to X$ *be a contraction. Then,* f *has a unique fixed point in* X*.*

This theorem is also known as the Banach Fixed Point theorem.

Proof. Let $x_0 \in X$, and set $x_{n+1} = f(x_n)$. Then, for any $j \in \mathbb{N}$,

$$\begin{aligned} d(x_{j+1},x_j) &\leq \kappa d(x_j,x_{j-1}) \\ &\leq \kappa^2 d(x_{j-1},x_{j-2}) \\ &\leq \kappa^3 d(x_{j-2},x_{j-3}) \\ &\vdots \\ &\leq \kappa^j d(x_1,x_0) \end{aligned}$$

so if $k > i$,

$$\begin{aligned} d(x_k,x_i) &\leq \sum_{i=j}^{k-1} d(x_{i+1},x_i) \\ &\leq \sum_{i=j}^{k} \kappa^i d(x_1,x_0) \\ &\leq \frac{\kappa^i}{1-\kappa} d(x_1,x_0) \end{aligned}$$

It follows that (x_n) is a Cauchy sequence in X. Since X is complete, $(x_n) \to x$ for some $x \in X$, and since f is continuous, $\big(f(x_n)\big) \to f(x)$, so,

$$\begin{aligned} x &= \lim_{n\to\infty} x_{n+1} \\ &= \lim_{n\to\infty} f(x_n) \\ &= f(x) \end{aligned}$$

Any such x must also be unique, since if $f(x) = x$ and $f(y) = y$, then

$$\begin{aligned} d(x,y) &= d\big(f(x),f(y)\big) \\ &\leq \kappa d(x,y) \end{aligned}$$

so $(1-\kappa)d(x,y) = 0$ and $x = y$. ■

We can use this theorem to prove the local existence and uniqueness of solutions of ordinary differential equations:

Theorem (Picard–Lindelöf)**.** *Suppose $f : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous:*

$$\big|f(x) - f(y)\big| \leq L|x-y|$$

for all $x,y \in \mathbb{R}^n$. Then, for any $x_0 \in \mathbb{R}^n$, the differential equation

$$\dot{x} = f(x), \qquad x(0) = x_0$$

has a unique solution on $[-T,T]$ for any $T < 1/L$.

Proof. Rewrite the equation as

$$x(t) = x_0 + \int_0^t f\big(x(s)\big)\,ds$$

so $x : [-T,T] \to \mathbb{R}^n$ solves the equation if it is a fixed point of the map

$$\mathcal{F} : C\big([-T,T]\big) \to C\big([-T,T]\big)$$

given by

$$\big[\mathcal{F}(x)\big](t) \coloneqq x_0 + \int_0^t f\big(x(s)\big)\, ds$$

We use the contraction mapping theorem in the space $X :- C\big([-T,T]\big)$ with the supremum metric. This map $\mathcal{F}$ is a contraction on X if $LT < 1$ since

$$\begin{aligned}\Big|\big[\mathcal{F}(x)\big](t) - \big[\mathcal{F}(y)\big](t)\Big| &= \left|\int_0^t f\big(x(s)\big) - f\big(y(s)\big)\, ds\right| \\ &\le \int_0^t \Big|f\big(x(s)\big) - f\big(y(s)\big)\Big|\, ds \\ &\le \int_0^t L\big|x(s) - y(s)\big|\, ds \\ &\le LT\|x-y\|_\infty\end{aligned}$$

so

$$\big\|\mathcal{F}(x) - \mathcal{F}(y)\big\|_\infty \le LT\|x-y\|_\infty$$

■

37.9 The Arzelà-Ascoli Theorem

Let (X,d_X) and (Y,d_Y) be metric spaces. A family F of continuous functions $X \to Y$ is,

- *equicontinuous at* $x_0 \in X$ if for every $\varepsilon > 0$ there exists $\delta > 0$ such that $d_Y\big(f(x_0) - f(x)\big) < \varepsilon$ for every $f \in F$ and $x \in X$ such that $d_X(x_0,x) < \delta$;
- (*pointwise*) *equicontinuous* if it is equicontinuous at every $x \in X$;
- *uniformly equicontinuous* if for every $\varepsilon > 0$ there exists $\delta > 0$ such that $d_Y\big(f(x),f(y)\big) < \varepsilon$ for every $f \in F$ and $x,y \in X$ such that $d_X(x,y) < \delta$;

For comparison, the statement "all functions f in F are continuous" means that for every $\varepsilon > 0$, every $f \in F$, and every $x_0 \in X$, there exists a $\delta > 0$ such that $d_Y\big(f(x_0),f(x)\big) < \varepsilon$ for all $x \in X$ such that $d_X(x_0,x) < \delta$. Then,

- for *continuity*, δ may depend on ε, f, and x_0;
- for *uniform continuity*, δ may depend on ε and f;
- for *pointwise equicontinuity*, δ may depend on ε, x_0;
- and for *uniform equicontinuity*, δ may depend only on ε.

Lemma 37.9.1. *If X is compact, then $A \subseteq C(X)$ is pointwise equicontinuous if and only if it is uniformly equicontinuous.*

A sequence $(f_n)_{n=1}^\infty$ of functions is *uniformly bounded* if there exists $M \in \mathbb{R}$ such that $\|f_n\|_\infty \le M$ for all n.

Lemma (Diagonal Subsequence)**.** *Let $(f_n)_{n=1}^\infty$ be a uniformly bounded sequence of functions $X \to \mathbb{R}$. Let $D = \{x_k\}_{k=1}^\infty \subseteq X$ be a countable subset of X. Then, (f_n) has a subsequence f_{n_i} such that the sequence of real numbers $f_{n_i}(x_k)$ converges as $i \to \infty$ for each $x_k \in D$.*

Proof. Since (f_n) is uniformly bounded, the sequence of real numbers $\big(f_n(x)\big)$ is bounded for every $x \in X$.

Since $\big(f_n(x_1)\big)$ is bounded, by Bolzano-Weierstrass, (f_n) has a subssequence $(f_{n_{1,i}})$ such that $\big(f_{n_{1,i}}(x_1)\big)$ converges. Let S_1 be the set of these indices:

$$S_1 = \{n_{1,i}\}_{i\in\mathbb{N}} \subseteq \mathbb{N}$$

Since $\big(f_n(x_2)\big)$ is bounded (here $n_{1,i} \in S_1$), by Bolzano-Weierstrass, $(f_{n_{1,i}})$ has a subssequence $(f_{n_{2,i}})$ such that $\big(f_{n_{2,i}}(x_1)\big)$ converges. Let S_2 be the set of these indices:

$$S_2 = \{n_{2,i}\}_{i\in\mathbb{N}} \subseteq S_1$$

We continue this way.

Suppose $k-1$ steps have been completed, and we already have a sequence $f_{n_{k-1,i}}$ with the set $S_{k-1} = \{n_{k-1,i}\}_{i\in\mathbb{N}}$. Since $\big(f_{n_{k-1,i}}(x_k)\big)$ is bounded, by Bolzano-Weierstrass, $(f_{n_{k-1,i}})$ has a subsequence $f_{n_{k,i}}$ such that $\big(f_{n_{k,i}}(x_k)\big)$ converges. Let S_k be the set of these indices:

$$S_k = \{n_{k,i}\}_{i\in\mathbb{N}} \subseteq S_{k-1}$$

Then,

$$S_k \subseteq S_{k-1} \subseteq \cdots \subseteq S_1 \subseteq \mathbb{N}$$

This process can be continued forever for every $k \in \mathbb{N}$. We now select the "diagonal subsequence". For each positive integer i, let $r_i = n_{i,i}$, i.e., the ith smallest number of S_i. Note that for each k, at most the first $k-1$ terns of the sequence $(f_{r_i})_{i=1}^{\infty}$ are not included in the sequence $(f_{n_{k,i}})_{i=1}^{\infty}$ since

$$r_i = n_{i,i \in S_i \subseteq S_k}$$

if $i \geq k$. Therefore, as $\big(f_{n_{k,i}}(x_k)\big)$ converges, $\big(f_{r_i}(x_k)\big)$ converges too, for every $k \geq 1$. ■

Lemma 37.9.2. *Every compact metric space contains a countable dense set $D = \{x_k\}_{k\in\mathbb{N}}$.*

Proof. For each positive integer n, the open balls of of radius $\frac{1}{n}$ form an open cover $\big\{\mathbb{B}\left(x,\frac{1}{n}\right) : x \in X\big\}$ of X. It has a finite subcover consisting of M_n balls

$$\mathbb{B}\left(x_{n,1},\frac{1}{n}\right),\ldots,\mathbb{B}\left(x_{n,M_n},\frac{1}{n}\right)$$

Let D be the countable set that contains all these points $x_{n,i}$, $n \geq 1$, $i \in [1,M_n]$.

To show that D is dense, it is enough to prove that D intersects every open ball $\mathbb{B}(y,r)$ where $y \in X$ and $r > 0$. Let n be such that $r > \frac{1}{n}$. The point y must be an element of one of the balls $\mathbb{B}(x_{n,i},\frac{1}{n})$, so $d(y,x_{n,i}) < \frac{1}{n} < r$. Then, $x_{n,i} \in \mathbb{B}(y,r) \cap D$, so $\mathbb{B}(y,r) \cap D \neq \varnothing$. ■

Lemma 37.9.3. *Let (X,d) be a compact metric space, and let D be dense subset of X. Let (f_n) be a uniformly equicontinuous sequence in $C(X)$ such that $f_n(x)$ converges for every $x \in D$. Then, (f_n) converges in the maximum norm.*

Proof. Let $\varepsilon > 0$. By unform equicontinuity, there exists a $\delta > 0$ such that $d(x,y) < \delta$ implies $\big|f_n(x) - f_n(y)\big| < \varepsilon$ for all n.

The collection of open balls of radius $\delta/2$, $\big\{\mathbb{B}(y,\delta/2) : y \in X\big\}$ is an open cover of X. As X is compact, there is a finite subcover $\{\mathbb{B}(y_i,\delta/2)\}_{i=1}^{M}$ of X.

Since D is dense in X, there are points $x_i \in D \cap \mathbb{B}(y_i,\delta/2)$ for $1 \leq i \leq M$. Since $x_i \in D$, $\lim_{n\to\infty} f_n(x_i)$ exists, so there is an integer N_i such that

$$\big|f_m(x_i) - f_n(x_i)\big| < \varepsilon$$

for all $n,m \geq N_i$.

Let $N = \max_{1\leq i\leq M} N_i$, and let $x \in X$. Then, $x \in \mathbb{B}(y_i,\delta/2)$ for some i and

$$\begin{aligned} d(x,x_i) &\leq d(x,y_i) + d(y_i,x_i) \\ &< \delta \end{aligned}$$

Then, if $m,n \geq N$, we have

$$\begin{aligned} \left|f_m(x) - f_n(x)\right| &= \left|f_m(x) - f_m(x_i) + f_m(x_i) - f_n(x_i) + f_n(x_i) - f_n(x)\right| \\ &\leq \left|f_m(x) - f_m(x_i)\right| + \left|f_m(x_i) - f_n(x_i)\right| + \left|f_n(x_i) - f_n(x)\right| \\ &< \varepsilon + \varepsilon + \varepsilon \\ &= 3\varepsilon \end{aligned}$$

where we used uniform continuity and the choice of δ for the first and third summand, and the choice of N for the second. Taking the maximum over all $x \in X$, we obtain

$$\begin{aligned} \|f_m - f_n\|_\infty &= \max_{x\in X}\left|f_m(x) - f_n(x)\right| \\ &\leq 3\varepsilon \end{aligned}$$

for all $m,n \geq N$. This means that (f_n) is a Cauchy sequence in the maximum norm, as $\varepsilon > 0$ was arbitrary, and N depends on ε. Since $C(X)$ is complete, (f_n) converges in the maximum norm. ■

Theorem (Arzelà-Ascoli)**.** *Let X be a compact metric space. Suppose that the sequence $(f_n)_{n=1}^\infty$ in $C(X)$ is uniformly bounded and uniformly equicontinuous. Then, (f_n) has a subsequence that converges in the maximum norm to a function $f \in C(X)$.*

Proof. Since X is compact, there is a countable dense set $D \subseteq X$ by Lemma 37.9.2. Since (f_n) is uniformly bounded, we can apply the diagonal subsequence lemma to obtain $(f_{r_i})_{i=1}^\infty$ such that $f_{r_i}(x)$ converges for every $x \in D$. Since (f_n), and thus (f_{r_i}), are uniformly equicontinuous, (f_{r_i}) converges in the maximum norm by Lemma 37.9.3. ■

Here is one application of the Arzelà-Ascoli theorem:

Corollary (Peano)**.** *Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ is continuous. Then, there exists $T > 0$ such that the differential equation*

$$\dot{x} = f\big(t,x(t)\big), \qquad x(0) = x_0$$

has at least one solution for $t \in (-T,T)$.

Proof sketch. Assume, for simplicity, that $x_0 = 0$. First we construct "approximate solutions".

For each positive integer n, let $x_n : [0,\infty) \to \mathbb{R}$ be the unique continuous function that is linear on each of the intervals $[\frac{i}{n},\frac{i+1}{n})$ such that the (right) derivatives satisfy

$$\dot{x}_n(t) = f\left(\frac{i}{n},x_n\left(\frac{i}{n}\right)\right)$$

if $t \in [\frac{i}{n},\frac{i+1}{n})$, for every integer $i \geq 0$; and $x(0) = x_0 = 0$.

Assume that $\left|f(t,x)\right| \leq M$ if $|t| \leq 1$ and $|x| \leq 1$. Set $T = \min(1,1/M)$. Then, the approximate solutions x_n on $[0,T]$ are uniformly bounded (by 1) and uniformly equicontinuous (for every $\varepsilon > 0,\delta = \varepsilon/M$ works).

By the Arzelà-Ascoli theorem, (x_n) in $C\big([0,T]\big)$ has a subsequence that converges in the maximum norm. One then shows that the limit is a solution of the differential equation for $t \in [0,T)$. ■

37.9.1 Completeness in Compact Metric Spaces

A metric space (X,d) is *totally bounded* or *precompact* if for every $\varepsilon > 0$, there is a finite ε*-net* in X, i.e., X can be covered by a finite collection of balls of radius ε:

$$X \subseteq \bigcup_{i=1}^{n} \mathbb{B}(x_i,\varepsilon)$$

Note that any totally bounded set is bounded, as a 1-net gives $X \subseteq \bigcup_{i=1}^{n} \mathbb{B}(x_i,1)$, so for every $x \in X$, $d(x,x_1) < r := 1 + \max_{1\leq i\leq n} d(x_1,x_i)$. The converse is not true, however, as shown by any infinite set equipped with the discrete metric.

Lemma 37.9.4. *A subspace Y of a metric space (X,d) is totally bounded if and only if for every $\varepsilon > 0$, there is a finite collection of points $\{x_i\}_{i=1}^{n} \subseteq X$ such that*

$$Y \subseteq \bigcup_{i=1}^{n} \mathbb{B}(x_i,\varepsilon)$$

The forward direction is clear: if $x_i \in Y$, then $x_i \in X$.

Let $\varepsilon > 0$ and find a collection $\{x_i\}_{i=1}^{n}$ such that

$$Y \subseteq \bigcup_{i=1}^{n} \mathbb{B}\left(x_i,\frac{\varepsilon}{2}\right)$$

We can assume that $Y \cap \mathbb{B}(x_i,\frac{\varepsilon}{2}) \neq \varnothing$ for each i; otherwise we can just remove the ball centred at x_i from the cover for each such ball.

Now for each i, choose a point $y_i \in Y \cap \mathbb{B}(x_i,\frac{\varepsilon}{2})$. Then,

$$\mathbb{B}\left(x_i,\frac{\varepsilon}{2}\right) \subseteq \mathbb{B}(y_i,\varepsilon)$$

and so,

$$Y \subseteq \bigcup_{i=1}^{n} \mathbb{B}(y_i,\varepsilon)$$

as required.

Lemma 37.9.5. *A subspace Y of a totally bounded metric space X is totally bounded.*

Proof. As X is totally bounded, for any $\varepsilon > 0$, there is a finite collection of points $\{x_i\}_{i=1}^{n} \subseteq X$ such that

$$Y \subseteq X \subseteq \bigcup_{i=1}^{n} \mathbb{B}(x_i,\varepsilon)$$

so Y is totally bounded by the previous lemma. ■

Lemma 37.9.6. *If a subspace Y of a metric space X is totally bounded, then so is $\overline{Y}$.*

Proof. Given $\varepsilon > 0$, let $\{x_i\}_{i=1}^{n}$ be an $\varepsilon/2$-net for Y. Then, this is an $\varepsilon - net$ for $\overline{Y}$ since given any $y \in \overline{Y}$, there exists $x \in Y$ with $d(x,y) < \varepsilon/2$ and x_i such that $d(x,x_i) < \varepsilon/2$, so $d(y,x_i) < \varepsilon$. ■

Theorem 37.9.7. *Any sequence in a totally bounded metric space (X,d) has a Cauchy subsequence.*

Proof. Take a sequence $(x_n)_{n=1}^{\infty} \subseteq X$.

Since X is totally bounded, it has a finite $\frac{1}{2}$-net, so there is at least one ball $\mathbb{B}(y_1,\frac{1}{2})$ containing infinitely many elements of (x_n). All elements of this subsequence are within distance 1 of each other.

Choose n_1 such that $x_{n_1} \in \mathbb{B}(y_1,\frac{1}{2})$ and let

$$X_1 = \left\{x_i : i > n_1, x_i \in \mathbb{B}(y_1,\tfrac{1}{2})\right\}$$

Since X has a finite $\frac{1}{4}$-net, there is at least one ball $\mathbb{B}(y_2,\frac{1}{4})$ containing infinitely many elements of X_1, and all these points are within distance $\frac{1}{2}$ of each other. Choose n_2 such that $x_{n_2} \in \mathbb{B}(y_2,\frac{1}{4})$ and let

$$X_2 = \left\{x_i : i > n_2, x_i \in \mathbb{B}(y_2,\tfrac{1}{4})\right\}$$

Continuing in this way, we obtain a sequence (x_{n_i}) of (x_n) that is Cauchy since $x_{n_i} \in \mathbb{B}(y_j, 2^{-j})$ for all $i \geq j$. ∎

Theorem 37.9.8. *A subspace Y of a complete metric space (X,d) is compact if and only if it is closed and totally bounded.*

Proof. If Y is compact, then it is closed as it is compact in a Hausdorff space X, and totally bounded since the open cover $\left\{\mathbb{B}(x,\varepsilon) : x \in Y\right\}$ has a finite subcover which functions as an ε-net.

Conversely, if Y is totally bounded, then any sequence in Y has a Cauchy subsequence. Since X is complete, this subsequence converges, and since Y is closed, the limit of this sequence lies in Y, so Y is sequentially compact. Since (Y,d) is a metric space, Y is compact. ∎

Theorem 37.9.9. *A subspace Y of a complete metric space is bounded if and only if its closure is compact.*

Proof. If Y is totally bounded, then $\overline{Y}$ is totally bounded and so compact by the previous theorem.

If $\overline{Y}$ is compact, then it is totally bounded by the previous theorem and hence Y is bounded. ∎

37.9.2 The Generalised Arzelà-Ascoli Theorem

The Arzelà-Ascoli theorem also gives a characterisation of the compact subsets of $C(X)$ when X is a compact metric space.

Theorem 37.9.10. *Let X be a compact metric space. A subset A of $C(X)$ is totally bounded if and only if it is bounded and equicontinuous.*

Proof. If A is totally bounded, then it is bounded. Since A is totally bounded, for any $\varepsilon > 0$ there exist $f_1, \ldots, f_n$ such that for every $f \in A$, there is an i with

$$\|f - f_i\|_\infty < \varepsilon$$

Since each of the f_i are uniformly continuous, there exists $\delta > 0$ such that for $i = 1, \ldots, n$, $|f_i(x) - f_i(y)| < \varepsilon/3$ whenever $d(x,y) < \delta$. Then, for any $f \in A$, choose j such that $\|f_j - f\|_\infty < \varepsilon/3$; it follows that if $d(x,y) < \delta$, we have

$$\begin{aligned}
\left|f(x) - f(y)\right| &= \left|f(x) - f_j(x) + f_j(x) - f_j(y) + f_j(y) - f(y)\right| \\
&\leq \left|f(x) - f_j(x)\right| + \left|f_j(x) - f_j(y)\right| + \left|f_j(y) - f(y)\right| \\
&\leq \|f - f_j\|_\infty + \left|f_j(x) + f_j(y)\right| + \|f_j - f\|_\infty \\
&< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3}
\end{aligned}$$

$$= \varepsilon$$

so A is uniformly equicontinuous (and hence equicontinuous).

Now, suppose A is bounded and equicontinuous, and let $\varepsilon > 0$. For every $x \in X$, by equicontinuity of A, we find $\delta(x) > 0$ such that for all $y \in \mathbb{B}\big(x,\delta(x)\big)$ and every $f \in A$,

$$\big|f(y) - f(x)\big| < \frac{\varepsilon}{3}$$

Since X is compact, there is a finite set $\{x_i\}_{i=1}^n \subseteq X$ such that

$$X \subseteq \bigcup_{i=1}^n \mathbb{B}\big(x_i,\delta(x_i)\big)$$

We now make a collection F of elements of A that form a finite ε-net. For any $\{q_i\}_{i=1}^n$ with $q_i \in \mathbb{Z}$ for which there exists a $g \in A$ with

$$g(x_i) \in \left[\frac{q_i\varepsilon}{3},\frac{(q_i+1)\varepsilon}{3}\right]$$

we choose one such g and add it to F. Since A is bounded, there are only finitely many such choices of $\{q_i\}_{i=1}^n$ and so there are only finitely many functions in F.

Now, given any $f \in A$, for each i there are q_i such that

$$f(x_i) \in \left[\frac{q_i\varepsilon}{3},\frac{(q_i+1)\varepsilon}{3}\right]$$

and so there is a $g \in F$ such that

$$g(x_i) \in \left[\frac{q_i\varepsilon}{3},\frac{(q_i+1)\varepsilon}{3}\right]$$

which implies that $\big|f(x_i) - g(x_i)\big| < \frac{\varepsilon}{3}$ for each $i \in [1,n]$.

Now, for each $x \in X$, we can find j such that $x \in \mathbb{B}\big(x_j,\delta(x_j)\big)$, and then

$$\begin{aligned}\big|f(x) - g(x)\big| &= \big|f(x) - f(x_j) + f(x_j) - g(x_j) + g(x_j) - g(x)\big| \\ &\leq \big|f(x) - f(x_j)\big| + \big|f(x_j) - g(x_j)\big| + \big|g(x_j) - g(x)\big| \\ &\leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} \\ &= \varepsilon\end{aligned}$$

from which it follows that $\|f - g\|_\infty < \varepsilon$, i.e., A is totally bounded. ■

Corollary (Generalised Arzelà-Ascoli theorem)**.** *Let X be a compact metric space. A subset A of $C(X)$ is compact if and only if it is closed, bounded and equicontinuous.*

One application is as follows:

Theorem 37.9.11. *Suppose $f : \mathbb{R} \to \mathbb{R}$ is continuous. Then there exists $\delta > 0$ such that the differential equation*

$$\dot{x} = f(x), \qquad x(0) = x_0$$

has at least one solution for $t \in (-\delta,\delta)$.

37.10 The Baire Category Theorem

If S is a non-empty subset of a metric space (X,d), we define

$$\operatorname{diam}(S) = \sup_{x,y\in S} d(x,y)$$

Note that S is bounded if and only if $\operatorname{diam}(S) < \infty$.

Theorem (Cantor)**.** *If (X,d) is a complete metric space and (F_n) is a decreasing sequence of non-empty closed subsets of X such that* $\operatorname{diam}(F_n) \to 0$*, then*

$$\bigcap_{n=1}^{\infty} F_n \neq \varnothing$$

Proof. For each $n \in \mathbb{N}$, choose some $x_n \in F_n$. Then, for al $i \geq n$, we have $x_n \in F_n$. So, if $i,j \geq n$, we have $x_i,x_j \in F_n$, so $d(x_i,x_j) \leq \operatorname{diam}(F_n)$. It follows that (x_n) is Cauchy, and so $(x_n) \to x$ for some $x \in X$.

Since F_n is closed and $x_i \in F_n$ for all $i \geq n$, it follows that $x \in F_n$ for each n. So, $x \in \bigcap_{n=1}^{\infty} F_n$, i.e., the intersection is non-empty. ■

Before we move on to the next theorem, we discuss various notions of "local compactness".

Let X be a topological space. Then, it may satisfy a variety of generally non-equivalent conditions:

(1) Every point of X has a compact neighbourhood;

(2) Every point of X has a closed compact neighbourhood;

(2') Every point of X has a neighbourhood whose closure is compact (is *relatively compact* or *precompact*);

(2") Every point of X has a local base of relatively compact neighbourhoods;

(3) Every point of X has a local base of compact neighbourhoods;

(4) Every point of X has a local base of closed compact neighbourhoods;

We have the following logical relations between these conditions:

- Each condition implies (1);
- Conditions (2), $(2')$, and $(2'')$ are equivalent;
- Conditions (2) and (3) do not imply each other;
- Condition (4) implies (2) and (3);
- Compactness implies (1) and (2), but not (3) or (4);
- These are all equivalent if X is Hausdorff.

Spaces that satisfy (1) are called *weakly locally compact*, as they satisfy the weakest of these conditions. Spaces that satisfy any of (2), $(2')$, and $(2'')$, are called *locally relatively compact* or sometimes *strongly locally compact* in contrast to weakly compact spaces. Spaces satisfying (4) are called *locally compact regular*.

A *Baire space* is a topological space X that satisfies any of the following equivalent conditions:

1. Every countable intersection of dense open sets is dense;
2. Every countable union of closed sets with empty interior has empty interior;

3. Every meagre set has empty interior;
4. Every non-empty open set is non-meagre;
5. Every comeagre set (a set with meagre complement) is dense;
6. Whenever a countable union of closed sets has an interior point, at least one of the closed sets has an interior point.

A *pseudometric* is a generalisation of a metric that is not necessarily point separating. That is, the distance between two distinct points may be zero under a pseudometric. Every metric space is a pseudometric space.

Theorem (Baire Category Theorem I)**.** *Every complete pseudometric space is a Baire space. In particular, every completely metrisable topological space is a Baire space.*

Theorem (Baire Category Theorem II)**.** *Every locally compact regular space is a Baire space. In particular, every locally compact Hausdorff space is a Baire space.*

Neither of these theorems imply the other, since there are complete metric spaces that are not locally compact (e.g. any infinite-dimensional Banach space), and there are locally compact Hausdorff spaces that are not metrisable (e.g. any uncountable product of non-trivial compact Hausdorff spaces).

We prove a variant of the first Baire category theorem, using the first characterisation of Baire spaces:

Theorem (Baire Category Theorem)**.** *Every complete metric space is a Baire space.*

That is, if $\{G_k\}_{k=1}^{\infty}$ is a countable collection of open dense subsets of a complete metric space (X,d), then

$$G \coloneqq \bigcap_{k=1}^{\infty} G_k$$

is dense in X.

A set is called *residual* if it contains a countable intersection of open dense sets (like G in the above theorem).

Proof. Take $x \in X$ and $r \geq 0$; we need to show that $\mathbb{B}(x,r) \cap G$ is non-empty. Since each G_n is open and dense, we can find $y \in G_n$ and $s > 0$ such that

$$\mathbb{B}(x,r) \cap G_n \supseteq \mathbb{B}(y,2s) \supseteq \overline{\mathbb{B}}(y,s)$$

First choose $x_1 \in X$ and $r_1 < 1/2$ such that

$$\overline{\mathbb{B}}(x_1,r_1) \subseteq \mathbb{B}(x,r) \cap G_1$$

then take $x_2 \in X$ and $r_2 < 2^{-2}$ such that

$$\overline{\mathbb{B}}(x_2,r_2) \subseteq \mathbb{B}(x_1,r_1) \cap G_2$$

and inductively, take $x_n \in X$ and $r_n < 2^{-n}$ such that

$$\overline{\mathbb{B}}(x_n,r_n) \subseteq \mathbb{B}(x_{n-1},r_{n-1}) \cap G_n$$

This yields a sequence of nested closed sets

$$\overline{\mathbb{B}}(x_1,r_1) \supseteq \overline{\mathbb{B}}(x_2,r_2) \supseteq \overline{\mathbb{B}}(x_3,r_3) \supseteq \cdots$$

Since (X,d) is complete, by Cantor's theorem, there exists $x_0 \in X$ such that

$$x_0 \in \bigcap_{i=1}^{\infty} \overline{\mathbb{B}}(x_i,r_i)$$

Now observe that $x_0 \in \overline{\mathbb{B}}(x_1,r_1) \subseteq \mathbb{B}(x,r)$, and that $x_0 \in \overline{\mathbb{B}}(x_n,r_n) \subseteq G_n$ for every $n \in \mathbb{N}$. It follows that $x_0 \in \mathbb{B}(x,r) \cap G$, and hence $\mathbb{B}(x,r) \cap G$ is non-empty, and G is dense in X. ■

An alternative formulation of this theorem says that you cannot make a complete metric spaces from the countable union of sets that are too "small". Recall that a subset W of (X,d) is *nowhere dense* if $\overline{W}^{\circ} = \varnothing$, or equivalently by Theorem 37.4.3, if $X \setminus \overline{W}$ is dense in T, as

$$\varnothing = \overline{W}^{\circ} = X \setminus \overline{X \setminus \overline{W}}$$

gives

$$X = \overline{X \setminus \overline{W}}$$

Corollary 37.10.0.1. *Let $\{F_i\}_{i=1}^{\infty}$ be a countable collection of nowhere dense subsets of a non-empty complete metric space (X,d). Then,*

$$\bigcup_{i=1}^{\infty} F_i \neq X$$

Or more concisely, a complete metric space is not meagre in itself.

Proof. The sets $X \setminus \overline{F_i}$ are a countable collection of open dense sets. It follows that

$$\bigcap_{i=1}^{\infty} X \setminus \overline{F_i} = X \setminus \bigcup_{i=1}^{\infty} \overline{F_i}$$

is dense, and in particular, non-empty. ■

Lemma 37.10.1. *The Cantor set C is uncountable.*

Proof. Since C is a closed subset of $\mathbb{R}$, it is complete as a metric space. For every $x \in C$, there are points C arbitrary close to x, so $C \setminus \{x\}$ is dense in C. Since $\{x\}$ is closed, this shows that $\{x\}$ is nowhere dense. Then, we cannot have $C = \bigcup_{i=1}^{\infty} x_i$, so C is uncountable. ■

Chapter 38

Algebraic Topology

> *"Time and space were, from Death's point of view, merely things that he'd heard described. When it came to Death, they ticked the box marked Not Applicable. It might help to think of the universe as a rubber sheet, or perhaps not."*
>
> — Terry Pratchett, *Hogfather*

38.1 Glossary

isomorphism, $\cong$ — A morphism with a two-sided inverse. Does not necessarily correspond with a bijective morphism in any given category.

homeomorphism, $\cong$ — A bicontinuous bijection of topological spaces; an isomorphism in **Top**.

quotient map — The function $q : X \to X/\sim$ canonically defined by $x \mapsto [x]_\sim$.

identification map — A continuous surjection that preserves openness of sets in both directions.

cover — A collection of (open) sets whose union is the covered space.

Lebesgue number — Given a cover $\mathcal{U}$ of a metric space (X,d), a number $\delta > 0$ is a Lebesgue number for $\mathcal{U}$ if for every $x \in X$ there exists an open set $U \in \mathcal{U}$ such that $\mathbb{B}(x,\delta) \subseteq U$. Or equivalently, $\delta > 0$ is a Lebesgue number for $\mathcal{U}$ if every subset $S \subseteq X$ with diameter at most $\operatorname{diam}(S) \leq \delta$ is contained within some member of the cover.

retract — A subset $A \subseteq X$ is a retract of X if there is a continuous map $r : X \to A$ such that $r\big|_A = \mathrm{id}_A$, called the retraction.

(strong) deformation retract

A subset $A \subseteq X$ is a (strong) deformation retract of X if there exists a one-parameter family of maps $f_t : X \to X$, $t \in I$ (or by uncurrying, a single map $F : X \times I \to X$), such that $f_0 = \mathrm{id}_X$; $f_1(X) = A$; and $f_t\big|_A = \mathrm{id}_A$ for all $t \in I$.

weak deformation retract

Same as the above, but the final condition is relaxed to only $t = 1$.

(free) homotopy, $f \simeq g$

A continuous map $F : X \times I \to Y$ is a homotopy between the maps f_1 and f_2.

relative homotopy, $f \overset{A}{\simeq} g$

A homotopy that additionally fixes some subspace $A \subseteq X$ for all $t \in I$.

homotopy relative to the boundary, $f \overset{\partial}{\simeq} g$

A relative homotopy where A is the pair of endpoints of the paths.

linear homotopy

The homotopy between f and g given by $x \mapsto (1-t)f(x) + tg(x)$.

homotopy equivalence, $X \simeq Y$

A relaxation of isomorphism that only requires that the composition is homotopic and not equal to the identity.

contractible

A space is contractible if it is homotopy equivalent to the one-point space.

locally contractible

A space X is locally contractible if for every $x \in X$ and every open neighbourhood $U \subseteq X$ of x, there exists an open neighbourhood $V \subseteq U$ of x that is contractible.

null-homotopic

A map is null-homotopic if it is homotopic to some constant map.

connected

A space is connected if it cannot be partitioned into two disjoint open sets.

path-connected

A space is path-connected if every pair of points may be connected by a path.

simply connected

A space is simply connected if it is path-connected and every pair of points has exactly one homotopy class of paths between them.

neighbourhood basis

A neighbourhood basis at a point x is a collection $\mathcal{B}$ of neighbourhoods of x such that for any neighbourhood V of x, there exists a neighbourhood $B \in \mathcal{B}$ such that $B \subseteq V$.

locally connected

A space is locally connected if every point admits a neighbourhood basis consisting of connected sets.

locally path-connected

A space is locally path-connected if every point admits a neighbourhood basis consisting of path-connected sets.

cover

A covering of a space X is a map $p : \tilde{X} \to X$ is a map $p : \tilde{X} \to X$ such that for every point $x \in X$, there exists an open neighbourhood $U_x \subseteq X$ of x whose preimage $p^{-1}[U_x] = \bigsqcup_{d \in D_x} V_d$ is a disjoint union of open sets $(V_i)_{i \in I_x}$, and the restriction $p\big|_{V_i} : V_i \to U_x$ is a homeomorphism for every $i \in I_x$. Such an open set U_x is said to be evenly covered by p, and the open sets V_i are called the sheets of the covering. The pair $(\tilde{X},p)$ is then a cover of X.

fibre

The preimage of a singleton set under a covering.

deck transformation

Given a covering $p : \tilde{X} \to X$, a deck transformation is a homeomorphism $\tau : \tilde{X} \to \tilde{X}$ such that $p \circ \tau = p$.

lift, $\tilde{f}$

Given a covering $p : \tilde{X} \to X$ and a map $f : Y \to X$, a lift of f is a map $\tilde{f} : Y \to \tilde{X}$ such that $p \circ \tilde{f} = f$.

induced homomorphism, f_*

Given a pointed map $f : (X,x_0) \to (Y,y_0)$, the induced homomorphism of fundamental groups is the homomorphism $f_* : \pi_1(X,x_0) \to \pi_1(Y,y_0)$ defined by $[\alpha] \mapsto [f \circ \alpha]$.

odd and even function

A map $f : X \to Y$ is odd if $f(-x) = -f(x)$; and even if $f(-x) = f(x)$; for all $x \in X$.

wedge sum, $\vee$

The one-point union of a collection of spaces; the disjoint union of a collection of pointed spaces with the basepoints identified. This identified point is a natural basepoint for the wedge sum, and picking this point makes the wedge sum associative and commutative (up to homeomorphism).

reduced word

A word in a free product is reduced if it does not contain any identities, and if every pair of consecutive letters is not from the same group.

38.2 Review of Point-Set Topology

38.2.1 Metric Spaces

Let X be any set. A *metric* d on X is a map $d : X \times X \to \mathbb{R}_{\geq 0}$ such that,

(*i*) $d(x,y) = 0$ if and only if $x = y$ (point separating or positive-definiteness);

(*ii*) $d(x,y) = d(y,x)$ for all $x,y \in X$ (symmetry);

(*iii*) $d(a,b) \leq d(a,x) + d(x,b)$ for every $a,b,x \in X$ (triangle inequality).

Note that these axioms imply that $d(x,y) \geq 0$ for all $x,y \in X$. The pair (X,d) is then called a *metric space.*

Let (X,d) be a metric space. The *open ball* centred at $a \in X$ of radius r is the set

$$\mathbb{B}(a,r) = \{x \in X : d(x,a) < r\}$$

also denoted by $B(a,r)$ or $\mathbb{B}_r(a)$. If $r = 1$, we say that the ball is a *unit ball*, and we omit r from the notation.

In a metric space (X,d), a set $U \subseteq X$ is said to be *open in* X if for every point $x \in U$, there exists some $\varepsilon > 0$ such that $\mathbb{B}(x,\varepsilon) \subset U$. A set $U \subseteq X$ is said to be *closed in* X if its complement is open in X. If the ambient set X is clear, then we omit the "in X" and just say that a set is open or closed.

Example.

- In any metric space (X,d), X and $\varnothing$ are both simultaneously open and closed (or *clopen*).
- In $\mathbb{R}$, open intervals are open and closed intervals are closed. Half-open intervals are neither open nor closed.
- In a discrete metric space, every singleton set $\{x\} \subseteq X$ is open (take any $\varepsilon < 1$).

△

Sets can be open, closed, both (clopen), or neither, so the adjectives "open" and "closed" do not have all of their usual intuitive connotations when used in a mathematical context.

Lemma (Open Finite Intersection)**.** *If $(U_i)_{i=1}^n$ is a finite collection of sets open in (X,d), then $\bigcap_{i=1}^n U_i$ is open in (X,d).*

Proof. Take $x \in \bigcap_{i=1}^n U_i$. Then, for each i, $x \in U_i$, so there exists $\varepsilon_i > 0$ such that $\mathbb{B}_{\varepsilon_i}(x) \subset U_i$. If $\varepsilon := \min(\varepsilon_1, \dots ,\varepsilon_n)$, then,

$$\mathbb{B}_\varepsilon(x) \subseteq \mathbb{B}_{\varepsilon_i}(x) \subset U_i$$

for all i, and hence $\mathbb{B}_\varepsilon(x) \subset \bigcap_{i=1}^n U_i$. ∎

Lemma (Open Arbitrary Union)**.** *If $(U_i)_{i\in\mathcal{I}}$ is an arbitrary collection of sets open in (X,d), then $\bigcup_{i\in\mathcal{I}} U_i$ is open in (X,d).*

Proof. If $x \in \bigcup_{i\in\mathcal{I}} U_i$, then $x \in U_i$ for some $i \in \mathcal{I}$. Since U_i is open, there exists $\varepsilon > 0$ such that $B(x,\varepsilon) \subset U_i \subseteq \bigcup_{i\in\mathcal{I}} U_i$, so $\bigcup_{i\in\mathcal{I}} U_i$ is open. ∎

By De Morgan's laws, we also have:

Corollary (Closed Finite Union)**.** *If $(F_i)_{i=1}^n$ is a finite collection of sets closed in (X,d), then $\bigcup_{i=1}^n F_i$ is closed in (X,d).*

Proof.

$$X \setminus \bigcup_{i=1}^{n} F_i = \bigcap_{i=1}^{n} (X \setminus F_i)$$

As F_i is closed, $X \setminus F_i$ is open, so $\bigcap_{i=1}^{n}(X \setminus F_i)$ is the finite intersection of open sets, and hence $X \setminus \bigcup_{i=1}^{n} F_i$ is open. It follows that $\bigcup_{i=1}^{n} F_i$ is closed. ■

Corollary (Closed Arbitrary Intersection). *If $(F_i)_{i\in\mathcal{I}}$ is an arbitrary collection of sets closed in (X,d), then $\bigcup_{i\in\mathcal{I}} F_i$ is closed in (X,d).*

Proof.

$$X \setminus \bigcap_{i\in\mathcal{I}} F_i = \bigcup_{i\in\mathcal{I}} (X \setminus F_i)$$

As F_i is closed, $X \setminus F_i$ is open, so $\bigcup_{i\in\mathcal{I}}(X \setminus F_i)$ is the intersection of open sets, and hence $X \setminus \bigcap_{i\in\mathcal{I}} F_i$ is open. It follows that $\bigcap_{i\in\mathcal{I}} F_i$ is closed. ■

38.2.2 Topological Spaces

Many properties of a metric space do not depend on our exact choice of metric, and many familiar notions such as convergence and continuity may be defined in terms of open sets, with no mention of a metric at all. This motivates the introduction of a more general kind of space defined entirely in terms of open sets.

A *topology* on a set X is a collection Ω of subsets of X, such that

(T1) X and $\varnothing$ are open;

(T2) If $(U_i)_{i\in\mathcal{I}} \subseteq \Omega$, then $\bigcup_{i\in\mathcal{I}} U_i \in \Omega$ (arbitrary unions of open sets are open);

(T3) If $U,V \in \Omega$, then $U \cap V \in \Omega$ (binary intersections of open sets are open).

The pair (X,Ω) is then a *topological space.* We call the sets in Ω “open”. Note that by induction, (T3) implies that the finite intersection of open sets is open.

These axioms mimic the ways open sets in metric spaces behave, but without reference to any kind of metric. Every metric space induces a topological space; and conversely, if a topology is induced by some metric, then the topology is said to be *metrisable.*

We often omit the topology from notation and speak about a set X as a topological space alone. Additionally, unless otherwise stated, when considering a metric space (X,d) as a topological space, the topology used will always be the topology induced from the metric.

The *closed* sets in a topological space are the complements of open sets. By De Morgan’s laws, the collection $\mathcal{F}$ of closed sets satisfies:

(T1’) T and $\varnothing$ are closed;

(T2’) Arbitrary intersections of closed sets are closed.

(T3’) The union of finitely many closed sets is closed;

Let (X,Ω) be a topological space. A set $\mathcal{B} \subseteq \Omega$ is a *basis* for the topology Ω, or that $\mathcal{B}$ *generates* the topology Ω, if every open set can be written as the union of sets in $\mathcal{B}$. That is, for each $U \in \Omega$, there exists a collection $\{B_i\}_{i\in\mathcal{I}} \subseteq \mathcal{B}$ such that $\bigcup_{i\in\mathcal{I}} B_i = U$.

Given $x \in X$, a set $\mathcal{B} \subseteq \Omega$ is a *neighbourhood basis* for x, if for every open set U containing x there is a set in the basis containing x that is a subset of U. That is, if for every $U \in \Omega$ with $x \in U$, there exists $B \in \mathcal{B}$ such that $x \in B \subseteq U$.

38.2.3 Maps and Topological Equivalence

Let X,Y be topological spaces. A function $f : X \to Y$ is *continuous* if for any open set $U \subseteq X$, the preimage $f^{-1}[U] = \big\{x \in X : f(x) \in U\big\}$ is open. Continuous functions are sometimes abbreviated to *maps.*

Lemma 38.2.1 (Pasting Lemma)**.** *Let $X = A \cup B$, with A,B both closed or both open in X, and let $f : X \to Y$ be a function such that the restrictions $f\big|_A$ and $f\big|_B$ are continuous. Then, f is continuous.*

Proof. We prove the case for open A,B.

Let $U \subseteq Y$ be open in Y. Then, $f^{-1}[U] = f\big|_A^{-1}[U] \cup f\big|_B^{-1}[U]$, and because $f\big|_A$ and $f\big|_B$ are continuous, $f\big|_A^{-1}[U]$ and $f\big|_B^{-1}[U]$ are open in A and B, respectively. Because A and B are open, $f\big|_A^{-1}[U]$ and $f\big|_B^{-1}[U]$ are also open in X, so $f^{-1}[U]$ is open in X as it is the union of open sets, and hence f is continuous.

Exchanging "open" with "closed" in the previous yields a completely analogous proof for closed A,B. ■

Given topological spaces X and Y, a continuous map $f : X \to Y$ is a (topological) isomorphism or a *homeomorphism* if there exists a continuous map $g : Y \to X$ such that

$$f \circ g = \mathrm{id}_Y\,, \qquad g \circ f = \mathrm{id}_X$$

If a homeomorphism between X and Y exists, then X and Y are isomorphic topological spaces, or are *homeomorphic*, and we denote this relation (as usual) as $X \cong Y$.

38.2.4 The Subspace Topology

Let (X,Ω) be a topological space, and $S \subseteq X$ be a subset. The *subspace topology* on S is the set

$$\Omega_S = \{U \cap S : U \in \Omega\} \tag{38.1}$$

and we call (S,Ω_S) a *subspace* of (X,Ω).

Example.

- The (unit) *n-sphere* $\mathbb{S}^n$ or S^n is a subspace of $\mathbb{R}^{n+1}$ defined by

$$S^n = \left\{x \in \mathbb{R}^{n+1} : \|x\|_2 = \sum_{i=1}^{n+1} x_i^2 = 1\right\}$$

  Note that the superscript denotes the dimension of the sphere, and not the ambient space it is contained within.

- The (closed, unit) *n-disc* $\mathbb{D}^n$ or D^n is a subspace of $\mathbb{R}^n$ defined by

$$D^n = \left\{x \in \mathbb{R}^n : \|x\|_2 = \sum_{i=1}^{n+1} x_i^2 \leq 1\right\}$$

  The unit disk is a special case of a closed ball centred at the origin with radius 1.

△

38.2.5 Product Spaces

Let X,Y be topological spaces. The *product topology** on $X \times Y$ is the topology generated by sets of the form $U \times V$ with U and V open in X and Y, respectively.

Example.

- The product topology on $\prod_{i=1}^{n} \mathbb{R}$ coincides with the Euclidean topology on $\mathbb{R}^n$ (the topology induced by the ℓ^2 metric).
- The *topological torus* $\mathbb{T}^n$ or T^n is defined as the n-fold product of 1-spheres:

$$T^n = \prod_{i=1}^{n} S^1$$

  Unless otherwise qualified, "torus" usually refers to T^2 – the surface of a doughnut.

△

38.2.6 Disjoint Unions

Given a family of sets $\{X_i\}_{i\in\mathcal{I}}$, the *disjoint union* of this family is the set

$$\bigsqcup_{i\in\mathcal{I}} X_i = \bigcup_{i\in\mathcal{I}} \{(x,i) : x \in X_i\}$$

Each set in the disjoint union is forced to be disjoint from every other via the use of the auxilliary index i, marking which set each element came from, so taking a disjoint union cannot lose information like a union. Intuitively, each of the sets X_i is canonically isomorphic to the set $\tilde{X_i} = X_i \times \{i\}$, so each set is equipped with a canonical embedding into the disjoint union, and furthermore, the images of these embeddings partition the disjoint union.

Given two topological spaces X and Y, we can endow the disjoint union $X \sqcup Y$ of the underlying sets with a topology generated by the basis consisting of sets of the form $U \times \{i\}$ for some $i \in \mathcal{I}$ and $U \subseteq X_i$ open.

Intuitively, in the disjoint union, the component spaces are now considered to be part of a single new space, but each space is completely detached and isolated from every other space, and retains its original local topology.

38.2.7 The Quotient Topology

Recall that an equivalence relation $\sim$ on a set X is a relation such that for all x,y,$z \in X$,

- $x \sim x$ (reflexivity);
- if $x \sim y$, then $y \sim x$ (symmetry);
- if $x \sim y$ and $y \sim z$, then $x \sim z$ (transitivity).

The *equivalence class* $[x]$ of an element $x \in X$ under an equivalence relation $\sim$ is the set of all elements of X equivalent to x. That is, the set

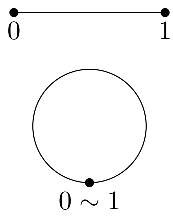
$$[x] = \{y \in X : x \sim y\}$$

The set of equivalence classes of an equivalence relation is denoted by $X/\sim$ and read as "the quotient of X by $\sim$", and the *quotient map* is the function $q : X \to X/\sim$ defined by $x \mapsto [x]$.

* More properly, this is the *box topology*, and not the true product topology, which is defined to be the coarsest topology such that the projections onto each component are all continuous. For finite product spaces, these topologies coincide, but for infinite products, the box topology is too fine and fails to satisfy a universal property.

If X is a topological space, then the *quotient topology* on the set $X/\sim$ is defined to have a set $U \subseteq X/\sim$ open if and only if $q^{-1}[U] = \{x \in X : q(x) = [x] \in U\}$ is open in X. Note that by definition, the quotient map is continuous.

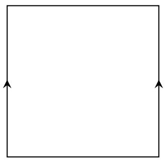
Example. Consider the unit interval $I = [0,1]$, and let $x \sim y$ if and only if $x = y$ and $0 \sim 1$ and $1 \sim 0$. The quotient set $I/\sim$, sometimes written as $I/0 \sim 1$ as only 0 and 1 are identified, then consists of the classes $[x] = \{x\}$ for $x \in (0,1)$ and $[0] = [1] = \{0,1\}$, so the endpoints of the interval have been "glued together" into a circle, and in fact, the resulting space with the quotient topology is homeomorphic to S^1.



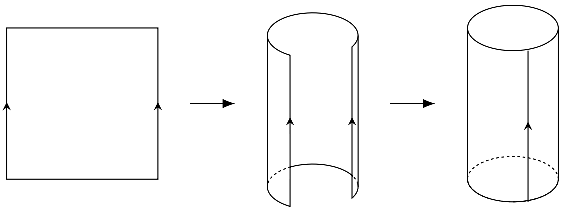
△

More generally, let $A \subseteq X$ be a subset of a topological space. This subset naturally induces the equivalence relation defined by $x \sim y$ if and only if $x = y$ or $\{x,y\} \subseteq A$, so every point of A is identified into a single equivalence class, while the points $x \in X \setminus A$ have singleton equivalence classes $[x] = \{x\}$. By an abuse of notation, we write X/A for the corresponding quotient space where all the points of A are identified into one point.

Example. Consider the square I^2. Define an equivalence relation by $(x,y) \sim (x',y')$ if and only if $(x,y) = (x',y')$ or $y = y'$ and $\{x,x'\} = \{0,1\}$. That is, we identify points on the left boundary with points on the right boundary with the same y-value. Visually, we represent this by marking an arrow on the square:



Then, we may identify marked edges together, with the arrows pointing in the same direction:



And we can see that the quotient space $I^2/\sim$ is homeomorphic to a cylinder (without the end faces). △

The quotient map is an example of an *identification map* – a continuous surjective function $f : X \to Y$ between topological spaces X and Y such that $U \subseteq Y$ is open if and only if $f^{-1}[U] \subseteq X$ is open.

The reverse direction follows from continuity, so a identification map may also be characterised as a surjective map which also preserves open sets under direct images. Or put another way, $f : X \to Y$ is an identification map if and only if Y has the finest topology such that f is continuous (the *final topology* with respect to f).

Theorem 38.2.2. *A surjective map $f : X \to Y$ is an identification map if and only if for every space Z and every function $g : Y \to Z$, $g \circ f$ is continuous if and only if g is continuous.*

38.3 Compactness

A *cover* of a set A is a collection $\mathcal{U}$ of sets whose union contains A. That is,

$$A \subseteq \bigcup_{U \in \mathcal{U}} U$$

and we say that the elements of $\mathcal{U}$ *cover* A. A *subcover* of a cover $\mathcal{U}$ is a subset of $\mathcal{U}$ whose elements still cover A. A cover is *open* if every element of the cover is open.

Example.

- $\mathcal{U} = \{(n-2,n+2) : n \in \mathbb{Z}\}$ is an (open) cover of $\mathbb{R}$, with one possible subcover given by $S = \{(n-2,n+2) : n \in 2\mathbb{Z}\}$;
- $\mathcal{U} = \{(n,n+1) : n \in \mathbb{Z}\}$ is not a cover of $\mathbb{R}$ since it does not cover the integers.

△

A topological space T is *compact* if every open cover of T has a finite subcover.

Example.

- (0,1) is not compact because $\mathcal{U} = \{(0,a) : a \in (0,1)\}$ is an open cover with no finite subcover;
- $\mathbb{R}$ is not compact because $\mathcal{U} = \{(-\infty,a),a \in \mathbb{R}\}$ has no finite subcover.

△

Note that, because compactness depends only on the open sets of a topological space, it is a topological invariant.

38.3.1 Lebesgue Numbers

Let $\mathcal{U}$ be an open cover of a metric space (X,d). A number $\delta > 0$ is called a *Lebesgue number* for $\mathcal{U}$ if for every $x \in X$, there exists an open set $U \in \mathcal{U}$ such that $\mathbb{B}(x,\delta) \subseteq U$.

In general, open covers do not have a Lebesgue number. For instance, $\mathcal{U} = \left\{(\frac{x}{2},x) : x \in (0,1)\right\}$ form an open cover of (0,1), but the covering sets become arbitrary small as $x \to 0$, so no Lebesgue number exists.

Lemma (Lebesgue's Number Lemma)**.** *Every open cover $\mathcal{U}$ of a compact metric space (X,d) has a Lebesgue number.*

38.4 Diagrams

The structure of a collection of objects and morphisms (sets and set functions, topological spaces and continuous maps, etc.) is often visually represented as a directed graph, called a *diagram.* We are already familiar with the notation $A \to B$ to denote a morphism from A to B, but we can also draw larger diagrams with more objects and morphisms to represent more structure at once. For instance, this diagram depicts 3 objects with morphisms between them:

$$\begin{tikzcd}
X \arrow[r, "f"] \arrow[dr, "h"'] & Y \arrow[d, "g"] \\
& Z
\end{tikzcd}$$

A diagram is *commutative* if for every pair of objects in the diagram, all routes between them are equal. For instance, the diagram above is commutative if and only if $h = g \circ f$. This also justifies the omission of identity morphisms in general diagrams; they don't meaningfully add any additional paths to the diagram.

38.4.1 Isomorphisms

Suppose we have objects A and B and morphisms $f : A \to B$ and $g : B \to A$ such that the following diagram is commutative:

$$\begin{tikzcd}
A \arrow[loop left, "\mathrm{id}_A"] \arrow[r, shift left, "f"] & B \arrow[l, shift left, "g"] \arrow[loop right, "\mathrm{id}_B"]
\end{tikzcd}$$

That is, $f \circ g = \mathrm{id}_B$ and $g \circ f = \mathrm{id}_A$, so f and g are *mutually inverse.* Then, we say that f and g are *isomorphisms*, and we alternatively label g by f^{-1}. If an isomorphism between a pair of objects A and B exists, we say that A and B are *isomorphic* and we write $A \cong B$.

Isomorphic objects are, as far as the ambient category is concerned, effectively identical – anything you can say about one object will apply just as well to any other isomorphic object.

38.5 The Fundamental Problem

The *fundamental problem* in topology is to classify topological spaces up to homeomorphism. That is, given two topological spaces X and Y, can we determine whether $X \cong Y$ or not?

To show that two spaces are homeomorphic, one only needs to provide a homeomorphism. To prove that they are *not* homeomorphicm is much more difficult. This involves finding a property that is invariant under homeomorphism that is satisfied by one space, but not the other.

Example. $\{*\} \ncong \mathbb{R}$ because $\{*\}$ is finite (bounded, countable, compact, etc.), but $\mathbb{R}$ is not. △

Example. $\mathbb{R} \ncong \mathbb{R}^2$ because $\mathbb{R}$ can be disconnected by removing one point, but $\mathbb{R}^2$ cannot. △

But what about $\mathbb{R}^2$ and $\mathbb{R}^3$? Or $\mathbb{R}^3$ and $\mathbb{R}^4$? Or more generally, $\mathbb{R}^n$ and $\mathbb{R}^m$?

Compactness and cut-point arguments don't work in the general case, and most other topological invariants we have seen are also not sufficiently powerful to distinguish these spaces. One might think these pairs of spaces are not homeomorphic, as one feels "bigger"; but set-theoretically, they all have the same cardinality (apart from $\mathbb{R}^0 \cong \{*\}$). It turns out that showing that two real vector spaces are not homeomorphic is non-trivial.

Theorem 38.5.1 (Invariance of Domain, Brouwer 1912)**.** $\mathbb{R}^n \cong \mathbb{R}^m$ *if and only if* $m = n$.

We will develop some tools that will allow us to prove a partial version of this theorem in low dimensions.

38.5.1 Retractions

A *pair* (X,A) consists of a topological space X and a subspace $A \subseteq X$. When $A = \{x\}$ is a single point, we instead write (X,x), and call the pair a *pointed space* (we sometimes call X alone a pointed space with *basepoint* x).

A *map of pairs* $f : (X,A) \to (Y,B)$ is a continuous function $f : X \to Y$ such that $f(A) \subseteq B$. If A and B are points, then f is a *pointed* or *based* map.

A subset $A \subseteq X$ is a *retract* of X if there is a map $r : X \to A$, called the *retraction*, such that

$$r\big|_A = \mathrm{id}_A$$

That is, r surjects X onto A while keeping all points of A fixed.

Example. For any pointed space (X,x_0), the unique constant map $r : X \to \{x_0\}$ is a retraction. △

Example. $\mathbb{R}^2 \setminus \{0\}$ retracts to S^1 via $r(x) = \frac{x}{\|x\|}$. △

Example. I does not retract to $\{0,1\}$, as the continuous image of a connected space must be connected. △

The following generalisation is non-trivial, and we will only be able to prove the $n = 2$ case later.

Theorem 38.5.2 (Brouwer)**.** *The disk D^n does not retract to S^{n-1}*

A subset $A \subseteq X$ is a (*strong*) *deformation retract* of X if there exists a one-parameter family of maps $f_t : X \to X$, $t \in I$ (or by uncurrying, a single map $F : X \times I \to X$), such that

- $f_0 = \mathrm{id}_X$;
- $f_1(X) = A$;
- $f_t\big|_A = \mathrm{id}_A$ for all $t \in I$.

Or, for all $x \in X$ and $a \in A$,

- $F(x,0) = x$;
- $F(X,1) = A$;
- $F(a,t) = a$ for all $t \in I$;

(A *weak deformation retract* relaxes the final condition for only $t = 1$. We will take the unqualified term "deformation retract" to always refer to the strong case.) Note that, by construction, f_1 is a retraction from X to A.

Example. $\mathbb{R}^n$ retracts to 0 via $F(x,t) = (1-t)x$. This is the *straight-line* or *linear homotopy.* △

Example. $\mathbb{R}^n \setminus \{0\}$ deformation retracts to S^{n-1} via $F(x,t) = (1-t)x + t\frac{x}{\|x\|}$ △

Intuitively, a deformation retract continuously shrinks a space onto a subspace; as the parameter t increases, the image of F continuously transitions from all of X to only all of A, with A being fixed throughout the entire process.

We can also view F as a kind of mapping between the retraction f_1 and the identity $f_0 = \mathrm{id}_X$ on X, smoothly transforming one map to the other – and in fact, this kind of parametrised deformation between two maps defines a construction called a *homotopy.*

38.5.2 Homotopy

Let X and Y be topological spaces. A (*free*) *homotopy* is a continuous map $F : X \times I \to Y$. If $f_t(x) = F(x,t)$, then we say that F is a homotopy from f_0 to f_1. Two maps $f,g : X \to Y$ are *homotopic* if there exists a homotopy $F : X \times I \to Y$ such that $f = f_0$ and $g = f_1$, and we write $f \simeq g$ to denote this relation.

Theorem 38.5.3. *Homotopy is an equivalence relation on the set of continuous maps between two given topological spaces. That is, if $f,g,h : X \to Y$ are continuous maps, then*

(*i*) $f \simeq f$;

(*ii*) *If* $f \simeq g$, *then* $g \simeq f$;

(*iii*) *If* $f \simeq g$ *and* $g \simeq h$, *then* $f \simeq h$.

Proof.

(*i*) The constant homotopy $F(x,t) = f(x)$ is a homotopy between f and f.

(*ii*) If F is a homotopy from f to g, then $F\big(-,(1-t)\big)$ is a homotopy from g to f.

(*iii*) If F is a homotopy from f to g and G is a homotopy from g to h, then

$$H(x,t) = \begin{cases} F(x,2t) & t \leq \frac{1}{2} \\ G(x,2t-1) & t > \frac{1}{2} \end{cases}$$

is a homotopy from f to h, with continuity given by the pasting lemma.

■

Recall that two spaces X and Y are homeomorphic if there exist a pair of maps between them with compositions equal to identities:

$$\begin{tikzcd} X \arrow[r, shift left, "f"] & Y \arrow[l, shift left, "g"] \end{tikzcd}$$

$$g \circ f = \mathrm{id}_X \qquad \text{and} \qquad f \circ g = \mathrm{id}_Y$$

If we relax these conditions and only require that these compositions are *homotopic* to identities, then we obtain a weaker notion of likeness called *homotopy equivalence*:

$$\begin{tikzcd} X \arrow[r, shift left, "f"] & Y \arrow[l, shift left, "g"] \end{tikzcd}$$

$$g \circ f \simeq \mathrm{id}_X \qquad \text{and} \qquad f \circ g \simeq \mathrm{id}_Y$$

We also say that f and g are *homotopy inverse* to one another.

Equality induces homotopy, but not the converse, so homeomorphic spaces are homotopy equivalent, but not the converse. More importantly, this means that two spaces that are not homotopy equivalent cannot be homeomorphic, allowing us another method to prove that two spaces are topologically distinguishable.

A space is always homotopy equivalent to any of its deformation retracts.

Example. $\mathbb{R}^n \setminus \{0\} \simeq S^{n-1}$, as S^{n-1} is a deformation retract of $\mathbb{R}^n$.

In more detail, the homotopy equivalence is witnessed by the inclusion mapping $f : S^{n-1} \hookrightarrow \mathbb{R}^n \setminus \{0\}$ and the retract $g : \mathbb{R}^n \setminus \{0\} \to S^{n-1}$ defined by $x \mapsto \frac{x}{\|x\|}$. Then, $g \circ f = \mathrm{id}_{S^{n-1}}$, and $f \circ g$ is homotopic to $\mathrm{id}_{\mathbb{R}^n}$ via the straight-line deformation retract $F(x,t) = (1-t)x + t\frac{x}{\|x\|}$ found earlier. △

Theorem 38.5.4. *Homotopy equivalence is an equivalence relation on the class of topological spaces.*

Proof. Symmetry and reflexivity are obvious. For transitivity, suppose $X \simeq Y$ and $Y \simeq Z$ witnessed by maps

$$\begin{tikzcd} X \arrow[r, shift left, "f_1"] & Y \arrow[l, shift left, "g_1"] \arrow[r, shift left, "f_2"] & Z \arrow[l, shift left, "g_2"] \end{tikzcd}$$

and homotopies F_1 from $f_1 \circ g_1$ to id_Y and F_2 from $f_2 \circ g_2$ to id_Z. (Note that this diagram does not necessarily commute.)

Then, $f = f_2 \circ f_1$ and $g = g_1 \circ g_2$ are homotopy equivalence maps, with the homotopy from $f \circ g$ to id_Z given by

$$F(z,t) = \begin{cases} f_2 \circ F_1(z,2t) \circ g_2 & t \leq \frac{1}{2} \\ F_2(z,2t-1) & t > \frac{1}{2} \end{cases}$$

■

A topological space X is *contractible* if $X \simeq \{*\}$. Or equivalently, if id_X is homotopic to a constant map (is *null-homotopic*).

Example. Euclidean space of any dimension is contractible: $\mathbb{R}^n \simeq \mathbb{R}^0$ for any $n \in \mathbb{N}$.

Consider the unique constant map $f : \mathbb{R}^n \to \mathbb{R}^0$ defined by $x \mapsto 0$ and the inclusion map $g : \mathbb{R}^0 \hookrightarrow \mathbb{R}^n$.

Then, we have $f \circ g = \mathrm{id}_{\mathbb{R}^0}$, and the composition $g \circ f : \mathbb{R}^n \to \mathbb{R}^n$ maps everything to zero, and is homotopic to the identity via the straight-line homotopy

$$F : \mathbb{R}^n \times I \to \mathbb{R}^n : (x,t) \mapsto tx$$

△

Example. Because $\{*\} \cong \mathbb{R}^0$, the above implies that $\mathbb{R}^n \simeq \mathbb{R}^m$ for all n,m. More generally, for any topological space X, $X \times \mathbb{R}^n \simeq X$. △

In general, it is much more difficult to show that a space is not contractible. For instance, the proof that S^n is not contractible for any $n \geq 1$ is non-trivial.

We can also compare the notions of contractibility with that of deformation retracts.

Theorem 38.5.5. *If X deformation retracts to a point $x_0 \in X$, then it is contractible.*

Proof. Consider the retraction given by the unique constant map $f : X \to \{x_0\}$, and the inclusion mapping $g : \{x_0\} \hookrightarrow X$. We have $f \circ g = \mathrm{id}_{\{x_0\}}$, and $g \circ f = f_1$ and $\mathrm{id}_X = f_0$, where f_t is the deformation retraction. The deformation retraction then gives the required homotopy. ■

Note that the converse does not hold, as the ordinary free homotopy demanded by a contractible space does not have to keep x_0 fixed throughout the homotopy.

Theorem 38.5.6. *The sphere S^n is not contractible for any $n \geq 0$.*

This theorem is highly non-trivial; we will only be able to prove the case $n = 1$ using the homotopy theory developed here.

38.5.3 Paths

Let X be a topological space, and $x,y \in X$ be two points. A *path* from x to y is a continuous map $f : I \to X$ with $f(0) = x$ and $f(1) = y$. We can view $f(s)$ as the position of a particle traveling along some curve in X as s varies from 0 to 1.

Note however, that a path is distinct from its image, and in particular, may not be injective. For instance, the image of the path $s \mapsto \exp(4\pi i s)$ in $S^1 \subset \mathbb{C}$ is the circle S^1, but the path itself travels around the circle twice and is distinct from, for example, the path $s \mapsto \exp(2\pi i s)$.

Given two paths $f,g : I \to X$ with $f(1) = g(0)$, the path $f * g : I \to X$ defined by

$$(f * g)(s) = \begin{cases} f(2s) & s \in [0,\frac{1}{2}] \\ g(2s-1) & s \in [\frac{1}{2},1] \end{cases}$$

is called the *concatenation* of f and g, with continuity given by the pasting lemma. Intuitively, the concatenation traverses the first path at double speed, then the second path at double speed.

Given a path $f : I \to X$ from x to y, the *reverse path* $\overline{f}$ defined by $\overline{f}(s) = f(1-s)$ is the path from y to x obtained by traversing f in the opposite direction.

We will often change the arguments to a map in concatenations and other similar operations, so it is helpful to be able to rescale any interval $[a,b]$ to $[0,1]$. This can be done via the affine map

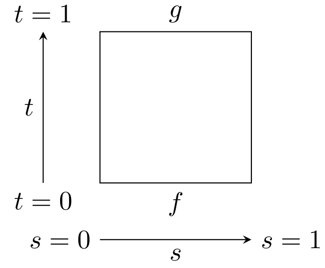
$$f(x) = \frac{x-a}{b-a}$$

i.e. translate down by a to reach zero, then rescale by the difference to reach 1.

Example. In the above concatenation, we have the intervals $[0,\frac{1}{2}]$ and $[\frac{1}{2},1]$. The rescaled argument of f is given by $\left(\frac{1}{2}-0\right)^{-1}(s-0) = 2s$, and of g by $\left(1-\frac{1}{2}\right)^{-1}\left(s-\frac{1}{2}\right) = 2s-1$. △

Because paths are maps between topological spaces, we can also consider homotopies of paths. Given two paths $f,g : I \to X$, a homotopy between them is given by a map $F : I \times I \to X$ satisfying $f_0 = f$ and $f_1 = g$.

Because the domain of such a homotopy is a square $I \times I$, it can be visualised as



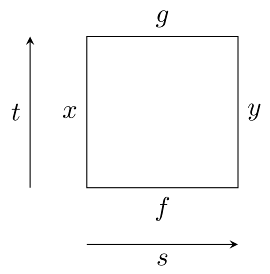
Each horizontal slice of the square at represents one of the functions f_t, with the bottom edge being f and the top edge being g, while each vertical slice represents the trajectory of a fixed argument s under the continuous deformation from f to g. This representation isn't very interesting yet, but will become helpful once we consider homotopies of concatenations.

The notion of free homotopy is, however, too weak to be very useful, since every path is homotopic to a constant path (i.e. deformation retract to the constant map at any point on the path), and we don't get much useful information from this. Instead, we can consider only paths that share endpoints, and define a more restricted notion of homotopy.

Let $x,y \in X$ and $f,g : I \to X$ be paths from x to y. In contrast to a free homotopy, a homotopy *relative to the boundary* or *endpoints* (sometimes abbreviated to "*rel boundary*") from f to g is a homotopy $F : I \times I \to X$ satisfying

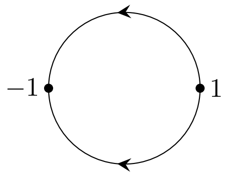
- $f_0 = f$;
- $f_1 = g$;
- $f_t(0) = x$ for all $t \in I$;
- $f_t(1) = y$ for all $t \in I$.

or as a diagram,



That is, a homotopy relative to boundaries is a continuous deformation of one path to another that keeps the endpoints of the paths fixed for all values of the parameter t. If there exists a homotopy relative to boundaries between f and g, then we write $f \overset{\partial}{\simeq} g$ to denote this relation.

Example. The paths $f,g : I \to S^1$ defined by $f(s) = \exp(\pi i s)$ and $g(s) = \exp(-\pi i s)$ traverse the upper and lower halves of the circle, respectively.



These paths are homotopic, as both can deformation retract to, for example, the point 1. However, they are not homotopic relative to boundaries. Intuitively, there is no way to continuously deform one to the other due to the hole in the circle that the two paths enclose. Proving this formally, however, is difficult. △

Lemma 38.5.7. *For any pair of points $x,y \in X$, relative homotopy is an equivalence relation on the set of paths from x to y.*

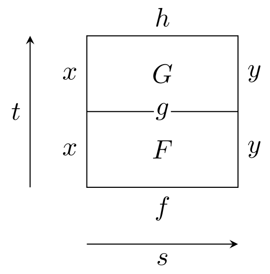
Proof. The proof is almost identical to that of free homotopy:

(i) The constant homotopy is a homotopy relative to boundaries from a path to itself.

(i) If F is a relative homotopy from f to g, then $F\big(-,(1-t)\big)$ is a relative homotopy from g to f.

(i) If F is a relative homotopy from f to g and G is a relative homotopy from g to h, then

$$H(s,t) = \begin{cases} F(s,2t) & t \leq \frac{1}{2} \\ G(s,2t-1) & t > \frac{1}{2} \end{cases}$$



is a relative homotopy from f to h.

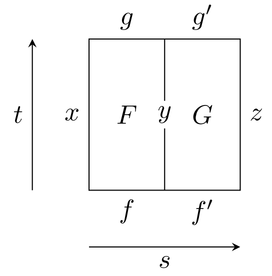
■

Lemma 38.5.8. *Let $f,g : I \to X$ be paths from x to y satisfying $f \overset{\partial}{\simeq} g$ and $f',g' : I \to X$ be paths from y to z satisfying $f \overset{\partial}{\simeq} g$. Then, $f * f' \simeq g * g'$.*

Proof. The proof is identical to that of transitivity in the previous lemma with the roles of s and t reversed.

If F is a relative homotopy from f to g and G is a relative homotopy from g to h, then

$$H(s,t) = \begin{cases} F(2s,t) & t \leq \frac{1}{2} \\ G(2s-1,t) & t > \frac{1}{2} \end{cases}$$



is a relative homotopy from $f * f'$ to $g * g'$. ■

More generally, a homotopy between maps $f,g : Z \to X$ may be relative to any subspace $A \subseteq X$. That is, the homotopy fixes the elements of the subspace A, and we write $f \overset{A}{\simeq} g$ if such a homotopy exists. A homotopy relative to boundaries is then the special case where the subspace consists of the two endpoints of the paths involved.

If we write $\iota : A \hookrightarrow X$ for the inclusion of A into X, then a deformation retract is just a special case of a retraction $r : X \to A$ such that $\iota \circ r$ is homotopic to id_X, relative to A.

38.5.4 Loops

A *loop* is a special case of a path where the two endpoints coincide. That is, a continuous map $f : I \to X$ with $f(0) = f(1) = x_0 \in X$, and we say that f is a loop *based at* x_0 or *with basepoint* x_0.

Because loops are a special case of paths, homotopy relative to boundaries is also an equivalence relation on the set of loops at some basepoint, so given a fixed point x_0, we can form equivalence classes of the form

$$[f] = \left\{(g : I \to X) : g(0) = g(1) = x_0, g \overset{\partial}{\simeq} f\right\}$$

A homotopy relative to boundaries between loops is also called a *based homotopy*, since the preserved subspace is a single point, as in a pointed space.

Given a pointed space (X,x_0), we denote the set of homotopy classes of loops based at x_0 as

$$\pi_1(X,x_0) = \left\{[f] : f(0) = f(1) = x_0\right\}$$

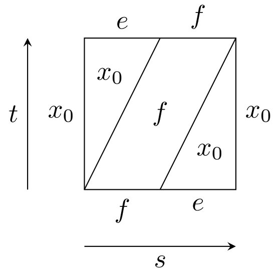
The concatenation of two loops based at x_0 is also a loop based at x_0, and we also have that if $f \simeq g$ and $f' \simeq g'$, then $f * f' \simeq g * g'$, so concatenation is compatible with homotopy. This allows us to define an operation

$$[f] \bullet [g] := [f * g]$$

For any pointed space (X,x_0), the set $\pi_1(X,x_0)$ equipped with this operation forms a group, called the *fundamental group* or *first homotopy group* of (X,x_0).

Theorem 38.5.9. *For any pointed space* (X,x_0)*, the* $\big(\pi_1(X,x_0),\bullet\big)$ *is a group, with unit* $[e]$*, where* e *is the constant loop, and the inverse* $[f]^{-1}$ *of the element* $[f]$ *is the class* $[\overline{f}]$*, where* $\overline{f}(t) = f(1-t)$ *is the inverse loop.*

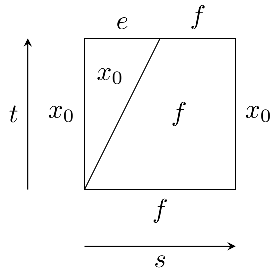
Proof. A homotopy showing $f * e \overset{\partial}{\simeq} e * f$ can be given as a diagram



To find the equation for this homotopy, we find the interval where f is applied to s; $[\frac{t}{2},\frac{1+t}{2}]$, as the homotopy will be constant outside of this interval; then find the affine function that varies from 0 to 1 as s varies from $\frac{t}{2}$ to $\frac{1+t}{2}$; $\left(\frac{1+t}{2} - \frac{t}{2}\right)^{-1}\left(s - \frac{t}{2}\right) = 2s - t$:

$$F(s,t) = \begin{cases} x_0 & s \in [0,\frac{t}{2}] \\ f(2s-t) & s \in [\frac{t}{2},\frac{1+t}{2}] \\ x_0 & s \in [\frac{1+t}{2},1] \end{cases}$$

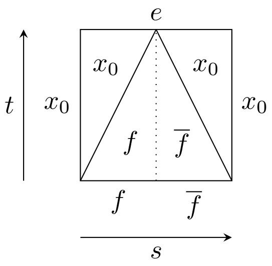
The homotopy $e * f \overset{\partial}{\simeq} f$ is then given by



We again find the argument, $\left(1 - \frac{t}{2}\right)^{-1}\left(s - \frac{t}{2}\right) = \frac{2s-t}{2-t}$, and then set the function to be constant past the linear bound:

$$F(s,t) = \begin{cases} x_0 & s \in [0,\frac{t}{2}] \\ f(\frac{2s-t}{2-t}) & s \in [\frac{t}{2},1] \end{cases}$$

For inverses, let $f : I \to X$ be a loop at x_0, and let $\overline{f} : I \to X$ be the loop defined by $\overline{f}(s) = f(1-s)$. Then, the homotopy $f * \overline{f} \overset{\partial}{\simeq} e$ is given by

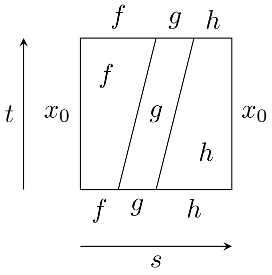


The concatenation $f * \overline{f}$ represents walking along f (at double speed), then walking back along the same path in reverse (also in double speed). Here, as t increases, the homotopy represents waiting at the start point for longer and longer, before starting to walk along the path (at the same speed), so we travel along less and less of the path before returning (so no rescaling is needed this time):

$$F(s,t) = \begin{cases} x_0 & s \in [0,\frac{t}{2}] \\ f(2s-t) & s \in [\frac{t}{2},\frac{1}{2}] \\ \overline{f}(2s-1+t) & s \in [\frac{1}{2},1-\frac{t}{2}] \\ x_0 & s \in [1-\frac{t}{2},1] \end{cases}$$

Since reversal is involutive, replacing f by $\overline{f}$ in the previous argument yields a homotopy $\overline{f} * f \overset{\partial}{\simeq} e$.

For associativity, a homotopy $(f * g) * h \overset{\partial}{\simeq} f * (g * h)$ is given by



For f, the argument is $\left(\frac{1+t}{4} - 0\right)^{-1}(s-0) = \frac{4s}{1+t}$; for g, $\left(\frac{2+t}{4} - \frac{1+t}{4}\right)^{-1}\left(s - \frac{1+t}{4}\right) = 4s - 1 - t$; and for h, $\left(1 - \frac{2+t}{4}\right)^{-1}\left(s - \frac{2+t}{4}\right) = \frac{4s-2-t}{2-t}$:

$$F(s,t) = \begin{cases} f(\frac{4s}{1+t}) & s \in [0,\frac{1+t}{4}] \\ g(4s-1-t) & s \in [\frac{1+t}{4},\frac{2+t}{4}] \\ h(\frac{4s-2-t}{2-t}) & s \in [\frac{2+t}{4},1] \end{cases}$$

■

38.5.5 The Fundamental Group

38.5.5.1 Path-Connected Spaces

A space X is *path-connected* if for every pair of points $x,y \in X$, there exists a path from x to y.

Theorem 38.5.10. *If X is path-connected, then for any two points $x_0,x_1 \in X$, we have an isomorphism of fundamental groups, $\pi_1(X,x_0) \cong \pi_1(X,x_1)$.*

Theorem 38.5.11. *For each path $h : I \to X$ from x_0 to x_1, define the map $\beta_h : \pi_1(X,x_0) \to \pi(X,x_1)$ by $[f] \mapsto [\overline{h} * f * h]$.*

Then,

$$\begin{aligned}
\beta_h\big([f] \bullet [g]\big) &= \beta_h\big([f * g]\big) \\
&= [\overline{h} * f * g * h] \\
&= [\overline{h} * f * h * \overline{h} * g * h] \\
&= [\overline{h} * f * h] \bullet [\overline{h} * g * h] \\
&= \beta_h\big([f]\big) \bullet \beta_h\big([g]\big)
\end{aligned}$$

so β_h is a group homomorphism for any path h. In particular, the map $\beta_{\overline{h}}$ induced by the reverse path is also a group homomorphism.

Because $h * \overline{h} \overset{\partial}{\simeq} e_{x_0}$ and $\overline{h} * h \overset{\partial}{\simeq} e_{x_1}$, we also have

$$\begin{aligned}
\beta_{\overline{h}} \circ \beta_h\big([f]\big) &= [h * \overline{h} * f * h * \overline{h}] \\
&= [f] \\
&= \mathrm{id}_{\pi_1(X,x_0)}\big([f]\big)
\end{aligned}$$

and similarly, $\beta_h \circ \beta_{\overline{h}} = \mathrm{id}_{\pi_1(X,x_1)}$, so β_h and $\beta_{\overline{h}}$ are inverse maps and hence form an isomorphism $\pi_1(X,x_0) \cong \pi_1(X,x_1)$.

Due to these isomorphisms, for path-connected spaces X, we may omit the basepoint and write just $\pi_1(X)$ for the fundamental group.

38.6 Covering Spaces

Let X be a topological space. A *covering* of X is a map $p : \tilde{X} \to X$ such that for every point $x \in X$, there exists an open neighbourhood $U_x \subseteq X$ of x whose preimage

$$p^{-1}[U_x] = \bigsqcup_{i \in I_x} V_i$$

is a disjoint union of open sets $(V_i)_{i \in I_x}$, and the restriction $p\big|_{V_i} : V_i \to U_x$ is a homeomorphism for every $i \in I_x$. Such an open set U_x is said to be *evenly covered* by p, and the open sets V_i are called the *sheets* of the covering.

If $p : \tilde{X} \to X$ is a covering, then the pair $(\tilde{X},p)$ is called a *covering space* or *cover* of X, and X is said to be the *base* of the covering.

Intuitively, a covering is a surjective map that acts locally like a projection of multiple copies of a space onto itself.

The preimage $p^{-1}\big[\{x\}\big]$ of any point x is called the *fibre* of x. A covering $p : \tilde{X} \to X$ is called an *n-fold covering* if the fibre $p^{-1}\big[\{x\}\big]$ consists of n points for all $x \in X$.

Example. For any $k \in \mathbb{N}$, the map $p_k : S^1 \to S^1$ defined by $z \mapsto z^k$ is a covering map. Given a point $z = \exp(2\pi i t) \in S^1$, we take the open neighbourhood $U = \big\{\exp(2\pi i s) : |s - t| < \varepsilon\big\}$ for some $0 < \varepsilon < \frac{1}{2k}$, which has preimage

$$p^{-1}[U] = \left\{ \sqrt[k]{\exp(2\pi i s)}, |s - t| < \varepsilon \right\}$$

$$= \bigcup_{0 \leq j < k} \left\{ \exp\left(2\pi i \frac{s+j}{k} \right) : |s-t| < \varepsilon \right\} < \varepsilon$$

These sets are all homeomorphic to $V = \big\{\exp(2\pi is/k) : |s-t| < \varepsilon\big\}$, and because $\frac{t+\varepsilon+j}{k} < \frac{t-\varepsilon+(j+1)}{k}$ for each j, they are all disjoint, so,

$$= \bigsqcup_{i=1}^{n} V$$

Intuitively, the preimage of the arc of length $2\varepsilon = \frac{1}{k}$ centred on z is the collection of arcs that each cover $\frac{1}{k}$th of the circle, centred on each root of z, and these arcs are disjoint as there are exactly k such roots evenly spaced along the circle.

This covering is also an k-fold covering map, as the fibre of any point $z = \exp(2\pi it)$ consists of k many kth roots of z – namely $\exp\big(2\pi i(t+j)/k\big)$, for $0 \leq j < k$. △

Example. The map $p_\infty : \mathbb{R} \to S^1$ defined by $x \mapsto \exp(2\pi ix)$ is a covering map. Given a point $z = \exp(2\pi it) \in S^1$, we take the open neighbourhood $U = \big\{\exp(2\pi is) : |s-t| < \varepsilon\big\}$ for some $0 < \varepsilon < 1$, which has preimage

$$\begin{aligned} p^{-1}[U] &= \bigcup_{j\in\mathbb{Z}} \{s+i : |s-t| < \varepsilon\} \\ &= \bigsqcup_{j\in\mathbb{Z}} V \end{aligned}$$

△

Two coverings $p : Y \to X$ and $q : Z \to X$ are *isomorphic* if they factor through each other. That is, there exist maps f and g such that

$$p = q \circ f \qquad \text{and} \qquad q = p \circ g$$

This also implies that f and g are inverse, so equivalently, p and q are isomorphic if there exists a homeomorphism $h : Y \to Z$ such that

$$\begin{tikzcd} Y \arrow[rr, "h", "\cong"'] \arrow[dr, "p"'] & & Z \arrow[dl, "q"] \\ & X & \end{tikzcd}$$

commutes.

Example. p_2 is isomorphic to p_{-2} via the homeomorphism $h(z) = z^{-1}$. △

Example. p_2 and p_3 are not isomorphic, as one is a 2-fold covering, and the other is a 3-fold covering. △

Let $p : \tilde{X} \to X$ be a covering of X. A *deck transformation* is a homeomorphism $\tau : \tilde{X} \to \tilde{X}$ such that $p \circ \tau = p$. That is, τ witnesses an automorphism of p. The set of all deck transformations of a cover p is denoted $\mathrm{Deck}(p)$, and has group structure under composition.

Example. The map $z \mapsto -z$ is a deck transformation for p_2. △

38.6.1 Liftings

Given a covering $p : \tilde{X} \to X$ and a map $f : Y \to X$, a *lift* of f is a map $\tilde{f} : Y \to \tilde{X}$ such that

$$\begin{tikzcd} & \tilde{X} \arrow[d, "p"] \\ Y \arrow[ur, "\tilde{f}"] \arrow[r, "f"'] & X \end{tikzcd}$$

commutes. That is, f factors through $\tilde{f}$.

Lemma 38.6.1. *Let $p : \tilde{X} \to X$ be a cover, and let $\tilde{f},\tilde{g} : Y \to \tilde{X}$ be continuous maps. Then,*

(i) $\tilde{f}$ is a lift of $p \circ \tilde{f}$;

(ii) If $\tilde{f} \simeq \tilde{g}$, then $p \circ \tilde{f} \simeq p \circ \tilde{g}$ ("homotopies descend");

(iii) If $\alpha,\beta : I \to X$ are paths with $\alpha(1) = \beta(0)$, then $p \circ (\alpha * \beta) = (p \circ \alpha) * (p \circ \beta)$ ("paths descend").

Proof.

(i) The diagram

$$\begin{tikzcd}
 & \tilde{X} \arrow[d, "p"] \\
Y \arrow[ur, "\tilde{f}"] \arrow[r, "p\circ\tilde{f}"'] & X
\end{tikzcd}$$

trivially commutes.

(ii) Let $F : Y \times I \to \tilde{X}$ be a homotopy between $f_0 = \tilde{f}$ and $f_1 = \tilde{g}$. Then, $p \circ F : Y \times I \to X$ is a homotopy between $p \circ f_0 = p \circ \tilde{f}$ and $p \circ f_1 = p \circ \tilde{g}$.

(iii) Expanding the definition of concatenation, we have

$$\begin{aligned}
\big(p \circ (\alpha * \beta)\big)(s) &= p \circ \begin{cases} \alpha(2s) & s \in [0,\frac{1}{2}] \\ \beta(2s-1) & s \in [\frac{1}{2},1] \end{cases} \\
&= \begin{cases} p \circ \alpha(2s) & s \in [0,\frac{1}{2}] \\ p \circ \beta(2s-1) & s \in [\frac{1}{2},1] \end{cases} \\
&= \big((p \circ \alpha) * (p \circ \beta)\big)(s)
\end{aligned}$$

■

38.6.2 Homotopy Lifting Property

Let $p : Z \to X$ be a continuous map. Then, p has the *homotopy lifting property* (HLP) if for any homotopy $F : Y \times I \to X$ and lift $g : Y \times \{0\} \to Z$ of f_0 (i.e. $f_0 = p \circ g$), there exists a unique homotopy $\tilde{F} : Y \times I \to Z$ such that

(i) $\tilde{f}_0 = g$;

(ii) $p \circ \tilde{F} = F$.

That is,

$$\begin{tikzcd}
Y \times \{0\} \arrow[r, "g"] \arrow[d, hook, "\iota"] & Z \arrow[d, "p"] \\
Y \times I \arrow[ur, dashed, "\tilde{F}"] \arrow[r, "F"] & X
\end{tikzcd}$$

commutes.

If we take $Y = \{*\}$ to be a singleton set, we may interpret the homotopies above as paths, and a lift $g : \{*\} \times \{0\} \to Z$ is simply a choice of a point in $p^{-1}\big[\{x_0\}\big]$:

Let $p : Z \to X$ be a continuous map. Then, p has the *path lifting property* (PLP) if for any path $f : I \to X$ with $f(0) = x_0$ and point $\tilde{x}_0 \in p^{-1}\big[\{x_0\}\big]$, there exists a unique path $\tilde{f} : I \to Z$ with $\tilde{f}(0) = \tilde{x}_0$ and $p \circ \tilde{f} = f$.

$$\begin{tikzcd}
\{*\} \times \{0\} \cong \{*\} \arrow[r, "g"] \arrow[d, hook, "\iota"] & Z \arrow[d, "p"] \\
\{*\} \times I \cong I \arrow[r, "f"] \arrow[ur, dashed, "\tilde{f}"] & X
\end{tikzcd}$$

38.6.2.1 The Local Homotopy Lifting Property

Let $\mathcal{U}$ be an open cover of a metric space (X,d).

The *diameter* of a subset $S \subseteq X$ is the least upper bound of the distance between any pair of points in that subset:

$$\operatorname{diam}(S) = \sup_{x,y \in S} d(x,y)$$

Recall that a number $\delta > 0$ is called a *Lebesgue number* for $\mathcal{U}$ if for every $x \in X$, there exists an open neighbourhood $U \in \mathcal{U}$ of x such that $\mathbb{B}(x,\delta) \subseteq U$.

Equivalently, $\delta > 0$ is a Lebesgue number for $\mathcal{U}$ if every subset $S \subseteq X$ with diameter at most $\operatorname{diam}(S) \leq \delta$ is contained within some member of the cover.

Lemma. *Let $\{I_\alpha\}_\alpha$ be an open cover of the unit interval I. Then, there exists a Lebesgue number for this cover. That is, there exists some $\delta > 0$ such that for every $S \subseteq I$ with diameter $\operatorname{diam}(S) \leq \delta$, we have $S \subseteq I_\alpha$ for some α.*

This is a special case of Lebesgue's number lemma (§37.5.3) applied to the unit interval.

Recall that, given a covering space $p : \tilde{X} \to X$, we can find a covering of X by evenly covered sets $\{U_\alpha\}_\alpha$ such that the preimage of each set U_α is a disjoint union of open sets $\{V_\alpha^\beta\}_\beta$

$$p^{-1}[U_\alpha] = \bigsqcup_\beta V_\alpha^\beta$$

and furthermore, the restrictions of the covering to each of these sets is a homeomorphism

$$p\big|_{V_\alpha^\beta} : V_\alpha^\beta \overset{\cong}{\to} U_\alpha$$

wth inverses denoted by $q_\alpha^\beta : U_\alpha \to V_\alpha^\beta$.

Let $F : Y \times I \to X$ be a homotopy and $g : Y \times \{0\} \to \tilde{X}$ be a lift of f_0, and suppose that the image of F is contained within an evenly covered subset $U_\alpha \subseteq X$. If the lift carries the domain of f_0 to one of the sheets V^β – that is, if $g(Y \times \{0\}) \subseteq V^\beta$ – then we can lift the whole homotopy F to a homotopy $\tilde{F} := q_\alpha^\beta \circ F$ that extends g.

Lemma 38.6.2. *Let $p : \tilde{X} \to X$ be a covering, and let $F : Y \times I \to X$ be a homotopy. Let $g : Y \times \{0\} \to \tilde{X}$ satisfy $p \circ g = f_0$. Then, for every $y_0 \in Y$, there exists an open neighbourhood $N \subseteq Y$ and a unique homotopy $\tilde{F}_N : N \times I \to \tilde{X}$ such that*

- *$p \circ \tilde{F}_N = F\big|_{N \times I}$;*
- *$\tilde{F}_N(-,0) = g\big|_{N \times \{0\}}$*

Moreover, if $M \subseteq Y$ is another such neighbourhood of y_0, then

$$\tilde{F}_M\big|_{(M\cap N)\times I} = \tilde{F}_N\big|_{(M\cap N)\times I} = \tilde{F}_{M\cap N}$$

Theorem 38.6.3. *Covering maps satisfy the homotopy lifting property*

Proof. Let $p : \tilde{X} \to X$ be a covering of X and $f : Y \to X$ be continuous. Cover $Y \times I$ with open sets $N_\alpha \times I$ as in the previous lemma.

This yields a family of lifts $\tilde{F}_{N_\alpha} : N_\alpha \times I \to \tilde{X}$ that coincide on the intersection of any two sets $N_i \times I$ and $N_j \times I$ in the cover, and hence we have a well-defined function $\tilde{F} : Y \times I \to \tilde{X}$ defined by piecing these lifts together. Since each local lift is continuous, $\tilde{F}$ is continuous by the pasting lemma, and is therefore itself a lift.

Uniqueness follows from the uniqueness of the homotopy given by the previous lemma. ■

38.6.3 The Fundamental Group of the Circle

Lemma 38.6.4. *The map $\Phi : \mathbb{Z} \to \pi_1(S^1,1)$ defined by $n \mapsto [\omega_n]$ is a group homomorphism.*

Proof. The map $\tilde{\omega}_n : I \to \mathbb{R}$ defined by $t \mapsto nt$ satisfies

$$\omega_n = p_\infty \circ \tilde{\omega}_n$$

so it is a lift of ω_n. Define the deck transformation $\tau : \mathbb{R} \to \mathbb{R}$ by $t \mapsto t+n$ and consider the composition $\tilde{\omega}_m \cdot (\tau_m \circ \tilde{\omega}_n)$. This composition is a path in $\mathbb{R}$ from 0 to $m+n$, and is therefore homotopic to $\tilde{\omega}_{m+n}$ (e.g. via the straight-line homotopy).

Then,

$$\begin{aligned}
\Phi(m+n) &= [\omega_{m+n}] \\
&= [p_\infty \circ \tilde{\omega}_{m+n}] \\
&= \big[p_\infty \circ \big(\tilde{\omega} \cdot (\tau \circ \tilde{\omega}_n)\big)\big] \\
&= [p_\infty \circ \tilde{\omega}_m \cdot p_\infty \circ \tau_m \circ \tilde{\omega}_n] \\
&= [p_\infty \circ \tilde{\omega}_m] \bullet [p_\infty \circ \tau_m \circ \tilde{\omega}_n] \\
&= [p_\infty \circ \tilde{\omega}_m] \bullet [p_\infty \circ \tilde{\omega}_n] \\
&= [\omega_m] \bullet [\omega_n] \\
&= \Phi(m) \bullet \Phi(n)
\end{aligned}$$

■

Theorem 38.6.5. *The map $\Phi : \mathbb{Z} \to \pi_1(S^1,1)$ defined by $n \mapsto [\omega_n]$ is a group isomorphism.*

Proof. By the path lifting property of covers, given a loop $\alpha \in S^1$, there exists a unique lift $\tilde{\alpha} : I \to \mathbb{R}$ such that

(*i*) $p \circ \tilde{\alpha} = \alpha$;

(*ii*) $\tilde{\alpha}(0) = 0$.

Since $\alpha(1) = 1$ and $p \circ \tilde{\alpha} = \alpha$, we have $\tilde{\alpha}(1) \in p^{-1}\big[\{1\}\big] = \mathbb{Z}$. Denote this value by $n := \tilde{\alpha}(1)$. We then have $\tilde{\alpha} \overset{\partial}{\simeq} \tilde{\omega}_n$ since both are paths from 0 to n in $\mathbb{R}$, with a homotopy given by the straight-line homotopy. Since homotopies descend,

$$\alpha = p_\infty \circ \tilde{\alpha} \overset{\partial}{\simeq} p_\infty \circ \omega_n = \omega$$

so $[\alpha] = [\omega_n]$ and Φ is surjective.

Now, suppose that $\Phi(n) = [\omega_n] = [e]$. That is, $\omega_n \overset{\partial}{\simeq} e$, given by a homotopy $F : I \times I \to S^1$ from $f_0 = \omega_n$ to $f_1 = e$.

Define a map $g : I \times \{0\} \to \mathbb{R}$ by $g(s,0) = \tilde{\omega}_n$. The cover p then lifts F to a homotopy $\tilde{F} : I \times I \to R$ from $\tilde{f}_0 = g$ to $p \circ \tilde{F} = F$. The other end of the homotopy $\tilde{f}_1$ then satisfies $p \circ \tilde{f}_1 = e$, the constant loop. Thus,

- $\tilde{f}_0(0) = 0$ since $\tilde{f}_0 = \tilde{\omega}_n$;
- $\tilde{f}_0(1) = n$ since $\tilde{f}_0 = \tilde{\omega}_n$;
- $\tilde{f}_t(0) \in \mathbb{Z}$ since $p_\infty \circ \tilde{f}_t(0) = f_t(0) = 1$;
- $\tilde{f}_t(1) \in \mathbb{Z}$ since $p_\infty \circ \tilde{f}_t(1) = f_t(1) = 1$;
- $\tilde{f}_1(s) \in \mathbb{Z}$ since $p_\infty \circ \tilde{f}_1(s) = e(s) = 1$;

As the continuous image of a connected space is connected, any continuous map that takes values in $\mathbb{Z} \subseteq \mathbb{R}$ must be constant. Thus,

$$\begin{aligned} 0 &= \tilde{f}_0(0) \\ &= \tilde{f}_t(0) \\ &= \tilde{f}_1(s) \\ &= \tilde{f}_t(1) \\ &= \tilde{f}_0(1) \\ &= n \end{aligned}$$

so Φ has trivial kernel. ■

38.7 Induced Homomorphisms

Recall that a pair of spaces (X,A) consists of two topological spaces satisfying $A \subseteq X$, and that a map of pairs $f : (X,A) \to (Y,B)$ is a continuous function $f : X \to Y$ such that $f(A) \subseteq B$, also called a pointed or based map when A and B are singletons.

The *induced homomorphism* of a pointed map $f : (X,x_0) \to (Y,y_0)$ is the map

$$\begin{aligned} f_* : \pi_1(X,x_0) &\to \pi_1(Y,y_0) \\ [\alpha] &\mapsto [f \circ \alpha] \end{aligned}$$

Lemma 38.7.1. *The induced homomorphism f_* is a group homomorphism.*

Proof. Let $\alpha \overset{\partial}{\simeq} \beta$, witnessed by $F : I \times I \to X$. Then, $G = f \circ F$ is a relative homotopy from $f \circ \alpha$ to $f \circ \beta$, so we have $f \circ \alpha \overset{\partial}{\simeq} f \circ \beta$, and the map is well-defined.

Now, let α,β be loops in $\pi(X,x_0)$. Then,

$$\begin{aligned} f_*\big([\alpha] \bullet [\beta]\big) &= f_*\big([\alpha * \beta]\big) \\ &= \big[f \circ (\alpha * \beta)\big] \\ &= \big[(f \circ \alpha) * (f \circ \beta)\big] \\ &= [f \circ \alpha] * [f \circ \beta] \\ &= f_*\big([\alpha]\big) \bullet f_*\big([\beta]\big) \end{aligned}$$

■

Example. Consider the covering map $p_2 : (S^1,1) \to (S^1,1)$ defined by $z \mapsto z^2$, and let ω_n be the loop defined by $\omega_n(s) = \exp(2\pi i n s)$. Then, $p_2 \circ \omega_n = \omega_{2n}$, so the induced homomorphism $(p_2)_*$ is defined by

$$(p_2)_*\big([\omega_n]\big) = [\omega_{2n}]$$

△

Theorem 38.7.2. *Induced homomorphisms satisfy the following properties:*

(*i*) $\big(\mathrm{id}_{(X,x_0)}\big)_* = \mathrm{id}_{\pi_1(X,x_0)}$*;*

(*ii*) *Given two pointed maps $f : (X,x_0) \to (Y,y_0)$ and $g : (X,x_0) \to (Y,y_0)$, we have,*

$$(g \circ f)_* = g_* \circ f_*$$

That is, the fundamental group is a functor $\pi_1 : \mathbf{Top}_* \to \mathbf{Grp}$, acting on objects by $(X,x_0) \mapsto \pi_1(X,x_0)$ and on morphisms by $f \mapsto f_*$.

Proof.

(*i*) Precomposing by the identity leaves the loop unchanged, and thus the fundamental group is unchanged.

(*ii*) Given a loop γ, we have

$$\begin{aligned}(g \circ f)_*\big([\gamma]\big) &= \big[(g \circ f) \circ \gamma\big] \\ &= \big[g \circ (f \circ \gamma)\big] \\ &= g_*\big([f \circ \gamma]\big) \\ &= (g_* \circ f_*)\big([\gamma]\big)\end{aligned}$$

■

Theorem 38.7.3. *If $f : (X,x_0) \to (Y,y_0)$ is an isomorphism, then $f_* : \pi_1(X,x_0) \to \pi_1(Y,y_0)$ is also an isomorphism.*

Proof. Follows from functoriality. That is,

$$\begin{aligned}\mathrm{id}_{\pi_1(X,x_0)} &= \big(\mathrm{id}_{(X,x_0)}\big)_* \\ &= \big(f \circ f^{-1}\big)_* \\ &= f_* \circ f_*^{-1}\end{aligned}$$

(and similarly with f and f^{-1} reversed). ■

It follows that the fundamental group of a path-connected space is a topological invariant: if $X \cong Y$, then $\pi_1(X) \cong \pi_1(Y)$.

38.8 Homotopy Invariance

Recall that, given a pair (X,A), a *retraction* is a map $r : X \to A$ such that $r\big|_A = \mathrm{id}_A$. Retractions and inclusions naturally fit together in a square,

$$\begin{tikzcd}
A \arrow[r, "\mathrm{id}_A"] \arrow[d, hook, "\iota"'] & A \arrow[d, hook, "\iota"] \\
X \arrow[ur, two heads, "r"] \arrow[r, "\iota \circ r"'] & X
\end{tikzcd}$$

noting that the upper triangle gives $r \circ \iota = \mathrm{id}_A$.

Theorem 38.8.1. *Let $r : X \to A$ be a retraction, $\iota : A \hookrightarrow X$ be the inclusion. Then, for any point $x_0 \in A$, the induced homomorphisms $r_* : \pi_1(X,x_0) \to \pi_1(A,x_0)$ and $\iota_* : \pi_1(A,x_0) \to \pi_1(X,x_0)$ have the following properties:*

(i) r_* is surjective and ι_* is injective;

(ii) If r is a deformation retract, then r_* and ι_* constitute an isomorphism.

Proof.

(i) Because $r \circ \iota = \mathrm{id}_{(A,x_0)}$, we have from functoriality of π_1 that $\mathrm{id}_{\pi_1(A,x_0)} = (r \circ \iota)_* = r_* \circ \iota_*$, so ι_* and r_* must be injective and surjective, respectively.

(ii) We have that r_* is surjective, so to establish an isomorphism, it suffices to show that r_* is also injective if it is a deformation retract.

Denote by $e_A : I \to A$ and $e_X : I \to X$ the constant loops at x_0 in A and X, respectively. Let $[\gamma] \in \pi_1(X,x_0)$, and suppose that $[\gamma] \in \ker(r_*)$, so $r_*\big([\gamma]\big) = [r \circ \gamma] = [e_A]$, or equivalently, $r \circ \gamma \overset{\partial}{\simeq} e_A$.

As $r \circ \gamma$ is a loop in $A \subseteq X$, postcomposing by the inclusion gives the loop $\iota \circ r \circ \gamma$ in X that is homotopic to e_X by the same homotopy that takes $r \circ \gamma$ to e_A.

Because r is a deformation retract, we have $\iota \circ r \overset{\partial}{\simeq} \mathrm{id}_X$ witnessed by a homotopy $F : X \times I \to X$ relative to A. Construct a new homotopy $G : I \times I \to X$ by $G(s,t) = F\big(\gamma(s),t\big)$ between $g_0 = \iota \circ r \circ \gamma$ and $g_1 = \gamma_1$. Note that G is a based homotopy (at the subspace $\{0 \sim 1\} \subset I$) since F is a homotopy relative to A, and $x_0 \in A$, so $g_t(0) = f_t(x_0) = x_0$ for all $t \in I$.

Then, we have $e_X \overset{\partial}{\simeq} \iota \circ r \circ \gamma \overset{\partial}{\simeq} \gamma$ so $[\gamma] = [e_X]$. It follows that r_* has trivial kernel and is thus injective.

Now, let $[\eta] \in \pi_1(X,x_0)$, and define a new based homotopy from F in the same way as before; $G(s,t) = F\big(\eta(s),t\big)$. Because $f_1(X) = r(X) = A$, the loop $g_1 = f_1 \circ \eta$ is contained within A, so g_1 is a loop in A and hence $[g_1] \in \iota_*\big(\pi_1(A,x_0)\big)$. Since G is a homotopy, $g_1 \overset{\partial}{\simeq} \eta$, so $[g_1] = [\eta]$, and ι_* is surjective.

■

In particular, this also implies that $(\iota \circ r)_* : \pi_1(X,x_0) \to \pi_1(X,x_0)$ is also an isomorphism for any deformation retract r.

We now extend this result to more general homotopy equivalences.

Theorem 38.8.2. *If $f : X \to Y$ is a homotopy equivalence, then for any $x_0 \in X$, the induced homomorphism $f_* : \pi_1(X,x_0) \to \pi\big(Y,f(x_0)\big)$ is an isomorphism.*

Proof. Let ■

This shows that not only is the fundamental group a *topological* invariant, but more generally a *homotopy* invariant: if $X \simeq Y$, then $\pi_1(X) \cong \pi_1(Y)$.

38.9 The Brouwer Fixed Point Theorem

In the previous section, we showed that retractions induce surjective homomorphisms. One simple application is as follows:

Theorem 38.9.1. *There is no retract from the the unit disk D^2 to the circle S^1.*

Proof. Such a retraction would imply a surjection $\pi_1(D^2,1) \twoheadrightarrow \pi_1(S^1,1)$, but $\pi_1(D^2,1) = 0$, while $\pi_1(S^1,1) \cong \mathbb{Z}$. ■

A more important consequence of this "no retract" theorem generalises the fact that a continuous function $f : I \to I$ has a fixed point. This is a straightforward consequence of the intermediate value theorem (in fact, the statement is true if f is only increasing, and not continuous, though this proof is more involved), but the generalisation to maps f from $I \times I \cong D^2$ to itself is surprisingly non-trivial.

Theorem (Brouwer Fixed Point Theorem)**.** *Every map $f : D^2 \to D^2$ has a fixed point.*

38.9.1 Applications

One application of the Brouwer fixed point theorem is to eigenvectors. The following result is a special case of the *Perron-Frobenius theorem*:

Theorem 38.9.2. *Let $\mathbf{A} \in \mathbb{R}^{3\times 3}$ be a matrix with only positive entries. Then, $\mathbf{A}$ has an eigenvector $\mathbf{v}$ with only positive entries.*

Another application of the Brouwer fixed point theorem is the famous *Borsuk-Ulam theorem*, but first, some additional theory is required for its proof.

38.9.1.1 Odd and Even Maps

Recall that an involution is an endofunction $f : X \to X$ that is its own inverse; $f\big(f(x)\big) = x$, or $f \circ f = \mathrm{id}_X$. One important example is the negation function, $f(x) = -x$ – but note that this only makes sense in spaces that are symmetric about the origin of their coordinate systems. For instance, $f(x) = -x$ does not make sense as function $I \to I$.

Let X and Y be spaces with negation. A map $f : X \to Y$ is *odd* if $f(-x) = -f(x)$, and *even* if $f(-x) = f(x)$ for all $x \in X$. Note that a map may be neither odd nor even.

Example.

- The zero map is the unique map that is simultaneously odd and even.
- The identity map is odd, as $\mathrm{id}(-x) = -x = -\,\mathrm{id}(x)$;
- The map $p_2 : S^1 \to S^1$ defined by $z \mapsto z^2$ is even, as $p_2(z) = z^2 = (-z)^2 = p_2(-z)$.
- The map $p_3 : S^1 \to S^1$ defined by $z \mapsto z^3$ is odd, as $p_3(-z) = -z^3 = -p_3(z)$.
- The (circular, hyperbolic) sine function is odd, while the (circular, hyperbolic) cosine function is even.
- The exponential function $\exp : \mathbb{R} \to \mathbb{R}$ is neither odd nor even.

△

Lemma 38.9.3. *The composition of,*

(*i*) *Two even functions is even;*

(*ii*) *Two odd functions is odd;*

(*iii*) *An even and odd function (in either order) is even;*

(*iv*) *Any function with an even function is even (but not the reverse).*

Proof. Let f be any function and suppose g is even. Then,

$$\begin{aligned}(f \circ g)(-x) &= f\big(g(-x)\big)\\ &= f\big(g(x)\big)\\ &= (f \circ g)(x)\end{aligned}$$

so $f \circ g$ is even. This covers (i), the reverse of (iii), and (iv).

For (ii), suppose f,g are odd. Then,

$$\begin{aligned}(f \circ g)(-x) &= f\big(g(-x)\big)\\ &= f\big(-g(x)\big)\\ &= -f\big(g(x)\big)\\ &= -(f \circ g)(x)\end{aligned}$$

and $f \circ g$ is odd.

For the other direction of (iii), suppose f is even and g is odd. Then,

$$\begin{aligned}(f \circ g)(-x) &= f\big(g(-x)\big)\\ &= f\big(-g(x)\big)\\ &= f\big(g(x)\big)\\ &= (f \circ g)(x)\end{aligned}$$

so $f \circ g$ is even. ■

38.9.2 Null-Homotopic Maps

A map $f : X \to Y$ is *null-homotopic* if it is free homotopic to a constant map. That is, if there exists a constant map e and a free homotopy $F : X \times I \to Y$ with $f_0 = f$ and $f_1 = e$.

A pointed map $f : (X,x_0) \to (Y,y_0)$ is *null-homotopic relative to the basepoint* if it is relatively homotopic to the constant map e_{y_0}. That is, there exists a homotopy $F : X \times I \to Y$ such that

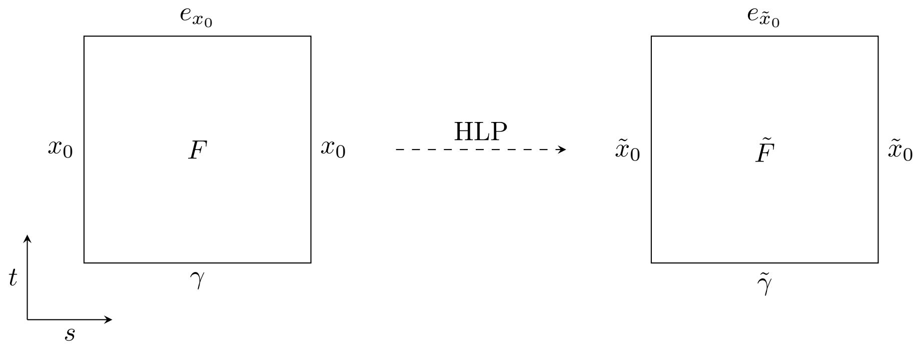
- $f_0 = f$;
- $f_1 = e_{y_0}$;
- $f_t(x_0) = y_0$ for all $t \in I$.

Conistent with the earlier notation for general relative homotopies, if f is null-homotopic relative to the basepoint x_0, we write $f \overset{x_0}{\simeq} e$.

Note that, even if the map $f : X \to Y$ is homotopic to e_{y_0}, and $x_0 \in X$ is such that $f(x_0) = y_0$, we do not necessarily have that the pointed map $f : (X,x_0) \to (Y,y_0)$ is null-homotopic, because a homotopy for the former need not preserve the pointedness of the intermediary maps f_t, while a null-homotopy of a pointed map further requires that $f_t(x_0) = y_0$ for all t.

Lemma 38.9.4. *Let $p : \tilde{X} \to X$ be a covering, and let $\gamma : I \to X$ be a loop such that $\gamma \overset{\partial}{\simeq} e_{x_0}$. Let $\tilde{x}_0 \in p^{-1}\big[\{x_0\}\big]$, and let $\tilde{\gamma}$ be the lift of γ with $\tilde{\gamma}(0) = \tilde{x}_0$. Then, $\tilde{\gamma} \overset{\partial}{\simeq} e_{\tilde{x}_0}$.*

Proof. Let $F : I \times I \to X$ be a homotopy between γ amd e_{x_0}. By the homotopy lifting property of coverings, there is a unique homotopy $\tilde{F} : I \times I \to \tilde{X}$ between $\tilde{\gamma}$ and $e_{\tilde{x}_0}$:



The left, right, and upper boundaries of the square on the right are all constant paths at x_0, so $\tilde{\gamma}$ is a loop at $\tilde{x}_0$, which is homotopic to $e_{\tilde{x}_0}$ via $\tilde{F}$. ■

Theorem 38.9.5. *If $f : S^1 \to S^1$ is odd, then f is not null-homotopic.*

Proof. WIP ■

Corollary 38.9.5.1. *If $f : S^2 \to \mathbb{R}^2$ is odd, then f has a root.*

Proof. WIP ■

38.9.2.1 The Borsuk-Ulam Theorem

Theorem (Borsuk-Ulam)**.** *For any continuous map $f : S^2 \to \mathbb{R}^2$, there exists a point $x \in S^2$ with $f(x) = f(-x)$.*

That is, for any continuous mapping of the sphere to $\mathbb{R}^2$, there exists two antipodal points for which the mapping has the same value.

One famous example of this theorem notes that mapping points on the Earth's surface to their temperature and atmospheric pressure can reasonably be assumed to be a continuous mapping, so the Borsuk-Ulam theorem states that at any time, there exist two antipodal points on the Earth's surface with equal temperature and atmospheric pressure.

Proof. Define $g : S^2 \to \mathbb{R}^2$ by

$$g(x) \coloneqq f(x) - f(-x)$$

We have $g(-x) = f(-x) - f(x) = -g(x)$, so g is odd, so by the previous corollary, g has a zero. That is, some $x \in S^2$ such that

$$\begin{aligned} g(x) &= 0 \\ f(x) - f(-x) &= 0 \\ f(x) &= f(-x) \end{aligned}$$ ■

38.9.3 Fundamental Groups of Product Spaces

Theorem 38.9.6. *Let (X,x_0) and (Y,y_0) be pointed spaces. Then,*

$$\pi_1(X \times Y, x_0 \times y_0) \cong \pi_1(X,x_0) \times \pi_1(Y,y_0)$$

That is, π_1 preserves binary products.

Proof. By the definition of the product topology, a map $Z \to X \times Y$ is continuous if and only if the components $p_1 \circ f$ and $p_2 \circ f$ are continuous, so a loop $\gamma : I \to X \times Y$ is equivalent to a pair of loops $\gamma_1 : I \to X$ and $\gamma_2 : I \to Y$.

Similarly, a homotopy F between loops in $X \times Y$ is equivalent to a pair of homotopies F_1 and F_2 between the equivalent loops in X and Y. That is, if $\alpha \overset{\partial}{\simeq} \beta$, then $p_1 \circ \alpha \overset{\partial}{\simeq} p_1 \circ \beta$, and $p_2 \circ \alpha \overset{\partial}{\simeq} p_2 \circ \beta$.

This induces a bijection $[\gamma] \mapsto \big([p_1 \circ \gamma],[p_2 \circ \gamma]\big)$, which gives the required isomorphism. ■

Example. The torus $T^2 = S^1 \times S^1$ with basepoint (1,1) has fundamental group

$$\pi_1\big(T^2,(1,1)\big) \cong \pi_1(S^1,1) \times \pi_1(S^1,1) \cong \mathbb{Z} \times \mathbb{Z}$$

△

Corollary 38.9.6.1. *By induction,*

$$\pi_1\left(\prod_{i=1}^{n}(X_i x_i)\right) = \prod_{i=1}^{n}\pi_1(X_i,x_i)$$

Example. The torus $T^n = \prod_{i=1}^{n} S^1$ has fundamental group

$$\pi_1\big(T^n\big) \cong \prod_{i=1}^{n}\pi_1(S^1) \cong \mathbb{Z}^n$$

△

Theorem 38.9.7. *For all $n \geq 2$, we have $\pi_1(S^n) \cong 0$.*

Proof. WIP ■

38.10 Galois Correspondence

Lemma 38.10.1. *Let $p : \tilde{X} \to X$ be a covering, and let $x_0 \in X$ and $\tilde{x}_0 \in p^{-1}\big[\{x_0\}\big]$. Then,*

(i) *The induced homomorphism $p_* : \pi_1(\tilde{X},\tilde{x}_0) \to \pi_1(X,x_0)$ is injective;*

(ii) *If $[\alpha] \in \pi_1(X,x_0)$, and $\tilde{\alpha}$ is the lift of α with $\tilde{\alpha}(0) = \tilde{x}_0$, then $\tilde{\alpha}$ is a loop if and only if $[\alpha] \in p_*\big(\pi_1(\tilde{X},\tilde{x}_0)\big)$.*

Proof.

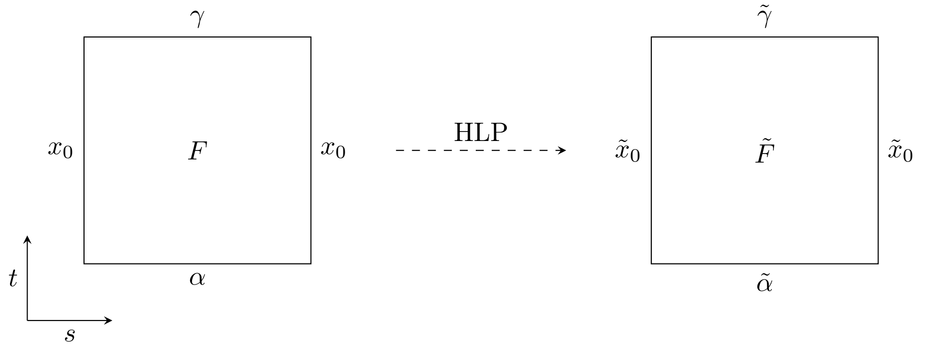
(i) Suppose p_* sends $[\tilde{\alpha}] \in \pi_1(\tilde{X},\tilde{x}_0)$ to the constant loop $[e_{x_0}]$. That is, $p \circ \tilde{\alpha} \overset{\partial}{\simeq} e_{x_0}$. Then by Lemma 38.9.4, $\tilde{\alpha} \overset{\partial}{\simeq} e_{\tilde{x}_0}$, so $[\tilde{\alpha}] = [e_{\tilde{x}_0}]$ and p_* has trivial kernel.

(ii) If $\tilde{\alpha}$ is a loop, then $p_*\big([\tilde{\alpha}]\big) = [p \circ \tilde{\alpha}] = [\alpha] \in p_*\big(\pi_1(\tilde{X},\tilde{x}_0)\big)$.

Conversely, suppose that $[\alpha] = p_*\big([\tilde{\gamma}]\big)$ for some $[\tilde{\gamma}] \in \pi_1(\tilde{X},\tilde{x}_0)$, so

$$\begin{aligned} \alpha &= p \circ \tilde{\alpha} \\ &\overset{\partial}{\simeq} p \circ \tilde{\gamma} \\ &= \gamma \end{aligned}$$

so there is some relative homotopy F from α to γ that lifts to a homotopy from $\tilde{\alpha}$ to $\tilde{\gamma}$:



The left and right boundaries of F are constant, so they lift to constant paths at $\tilde{x}_0$, so $\tilde{\alpha}$ is a loop (based at $\tilde{x}_0$) as required.

■

This shows that for any covering p, the image $p_*\big(\pi_1(\tilde{X},\tilde{x}_0)\big)$ is a subgroup of $\pi_1(X,x_0)$ that is isomorphic to $\pi_1(\tilde{X},\tilde{x}_0)$.

Example. The covering $p_2 : S^1 \to S^1$ induces the doubling map $n \mapsto 2n$, so

$$(p_2)_*\big(\pi_1(S^1,1)\big) \cong 2\mathbb{Z} \leq \mathbb{Z} \cong \pi_1(S^1,1)$$

△

Let $p : \tilde{X} \to X$ be a covering, and suppose X is connected. Then, the cardinality of the preimage of any point in X is called the *degree* of the covering:

$$\deg(p) \coloneqq \big|p^{-1}\big[\{x\}\big]\big|$$

Recall that, given a group G and a subgroup $H \leq G$, the *index* $[G : H]$ of H in G is the number of right (or left) cosets $G/H = \{Hg : g \in G\}$.

Lemma 38.10.2. *Let $p : \tilde{X} \to X$ be a covering and suppose that $\tilde{X}$ and X are path-connected. Let $x_0 \in X$ and $\tilde{x}_0 \in p^{-1}\big[\{x_0\}\big]$. Then,*

$$\deg(p) = \Big[\pi_1(X,x_0) : p_*\big(\pi_1(\tilde{X},\tilde{x}_0)\big)\Big]$$

Proof. WIP ■

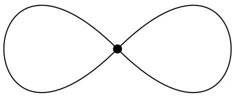
38.11 Wedge Sums

Let $\big((X_\alpha,x_\alpha)\big)_{\alpha\in\Lambda}$ be a collection of pointed spaces. The *wedge sum* of this collection is the "one-point union" of the spaces, defined as:

$$\bigvee_{\alpha\in\Lambda}(X_\alpha,x_\alpha) \coloneqq \bigsqcup_{\alpha\in\Lambda} X_\alpha\big/x_\alpha \sim x_\beta$$

That is, the disjoint union of each space with all the basepoints identified.

Example. The wedge sum of two pointed circles is the figure-eight graph:



$S^1 \vee S^1$

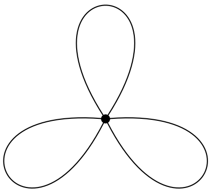
△

The identified point is a natural choice of basepoint for the wedge sum, and selecting this point makes the wedge sum associative and commutative (up to homeomorphism) over pointed spaces, as every basepoint is always identified to the same point, so associativity and commutativity follow from disjoint unions being associative and commutative.

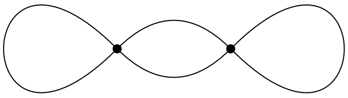
However, we may also treat the output as an ordinary topological space without any distinguished basepoint, in which case, the wedge sum is then *not* associative, as a new basepoint may be selected between applications.

If an expression involving wedge sums is not bracketed, we will assume that the natural basepoint is selected, so the resulting space is unambiguous and unique.

Example.



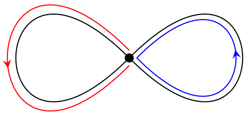
$S^1 \vee S^1 \vee S^1$



$(S^1 \vee S^1) \vee S^1$

In the first case, all three basepoints coincide, resulting in the bouquet of three circles. In the second, we were free to pick a basepoint distinct from the centre of the figure-eight, where the third circle was adjoined. △

Let us mark two loops a and b on $S^1 \vee S^1$:



$S^1 \vee S^1$

and denote their reverse paths by a^{-1} and b^{-1}.

Note that any loop in $S^1 \vee S^1$ can be decomposed into a string consisting of the symbols a, b, a^{-1}, and b^{-1}. For example,

$$aaba^{-1}b^{-1}a$$

corresponds to the loop that travels along a twice, b once, a backwards, b backwards, then a.

Some strings of this form may be *reduced* up to homotopy, as any substring consisting of a loop adjacent to its inverse is homotopic to the constant loop, which may then be removed from the string.

This structure is well-suited to be described by *free products*.

38.11.1 The Free Product of Groups

Let $\{G_\alpha\}_\alpha$ be a collection of groups. A *word* on these groups is a finite sequence $g_1 \cdots g_m$ of elements $g_i \in G_{\alpha_i}$, and m is the *length* of the word. The empty word of length 0 is denoted by ε. The *product* of two words is their concatenation

$$(g_1 \cdots g_m) * (h_1 \cdots h_n) = g_1 \cdots g_m h_1 \cdots h_n$$

A word is *reduced* if it does not contain the identity of any group, and if every pair of consecutive letters is not from the same group.

Given any word g on the groups $\{G_\alpha\}_\alpha$, we can *reduce* it to a reduced word g' by recursively removing all identity elements and replacing any consecutive elements g_i, g_{i+1} from the same group with their group product $g_i \cdot g_{i+1}$.

Let $*_\alpha G_\alpha$ be the set of reduced words on $\{G_\alpha\}_\alpha$. We can define an operation on this set as follows: given reduced words $g = g_1 \cdots g_m$ and $h = h_1 \dots h_n$, construct a new reduced word $g \bullet h$ by taking the concatenation $g * h$, then reduce the word recursively

$$g \bullet h = \begin{cases} gh & g_m \in G_\alpha, h_1 \in G_\beta, G_\alpha \neq G_\beta \\ g_1 \cdots g_{m-1}(g_m \cdot h_1)h_2 \cdots h_n & g_m, h_1 \in G_\alpha, g_m \cdot h_1 \neq \mathrm{id}_{G_\alpha} \\ g_1 \cdots g_{m-1} \bullet h_1 \cdots h_n & g_m, h_1 \in G_\alpha, g_m \cdot h_1 = \mathrm{id}_{G_\alpha} \end{cases}$$

Then, $(*_\alpha G_\alpha, \bullet)$ is a group called the *free product* of $\{G_\alpha\}_\alpha$, with identity ε, and the inverse of an element $g_1 \cdots g_m$ is given by $g_m^{-1} \cdots g_1^{-1}$.

Example. The free product of $\mathbb{Z}_2$ with itself is given by the semidirect product $\mathbb{Z}_2 * \mathbb{Z}_2 \cong \mathbb{Z} \rtimes \mathbb{Z}_2$. △

Example. If $G = \langle a \mid a^4 \rangle$ and $H = \langle b \mid b^5 \rangle$, then $G * H = \langle a, b \mid a^4 = b^5 \rangle$. △

Every group G_α is a subgroup of the free product $*_\alpha G_\alpha$ via the inclusion $\iota_\alpha : G_\alpha \hookrightarrow *_\alpha G_\alpha$ that maps each non-identity $g \in G_\alpha$ to the string g, and the identity to the empty string. The free product satisfies the following universal property:

Lemma 38.11.1. *Any pair of homomorphisms from groups G and H into K factor uniquely through the free product.*

That is, for any group homomorphisms $\varphi : G \to K$ and $\psi : H \to K$, there exists a unique homomorphism $\varphi * \psi : G * H \to K$ such that

$$\begin{tikzcd} G \arrow[r, hook, "\iota_1"] \arrow[dr, "\varphi"'] & G * H \arrow[d, dashed, "\varphi * \psi"] & H \arrow[l, hook', "\iota_2"'] \arrow[dl, "\psi"] \\ & K & \end{tikzcd}$$

commutes.

This holds more generally, with a collection of homomorphisms $\varphi_\alpha : G_\alpha \to K$ factoring uniquely through a map $*_\alpha \phi_\alpha : *_\alpha G_\alpha \to K$:

$$\begin{tikzcd} G_\alpha \arrow[r, hook, "\iota_\alpha"] \arrow[dr, "\varphi_\alpha"'] & *_\alpha G_\alpha \arrow[d, dashed, "*_\alpha \varphi_\alpha"] \\ & K \end{tikzcd}$$

38.12 The Seifert-van Kampen Theorem

Let X be a topological space and $\{U_\alpha\}_\alpha$ be an open cover with inclusion maps $\iota_\alpha : U_\alpha \hookrightarrow X$, and further suppose that the intersection $\bigcap_\alpha U_\alpha$ is non-empty.

Let $x_0 \in \bigcap_\alpha U_\alpha$, and consider the pointed spaces (U_α,x_0). The inclusion maps $\iota_\alpha : U_\alpha \to X$ induce homomorphisms between the fundamental groups based at x_0:

$$(\iota_\alpha)_* : \pi_1(U_\alpha,x_0) \to \pi_1(X,x_0)$$

which factor through the free product map

$$\Phi = *_\alpha(\iota_\alpha)_* : *_\alpha\pi_1(U_\alpha,x_0) \to \pi_1(X,x_0)$$

If the pairwise intersections $U_a \cap U_b$ are path-connected, then Φ is surjective; but in general, it is not injective, as loops in the intersections are counted twice in the free product.

The inclusions $\iota_{ab} : U_a \cap U_b \to U_a$ of the intersections then also induce maps between fundamental groups, completing the commutative diagram:

$$\begin{tikzcd}
& \pi_1(U_a,x_0) \arrow[dr, hook, "i_a"] \arrow[drr, bend left, "(\iota_a)_*"] & & \\
\pi_1(U_a \cap U_b,x_0) \arrow[ur, hook, "(\iota_{ba})_*"] \arrow[dr, hook, "(\iota_{ba})_*"'] & & *_\alpha\pi_1(U_\alpha,x_0) \arrow[r, dashed, "\Phi"] & \pi_1(X,x_0) \\
& \pi_1(U_b,x_0) \arrow[ur, hook, "i_b"'] \arrow[urr, bend right, "(\iota_b)_*"'] & &
\end{tikzcd}$$

In categorical language, $(\iota_a)_*$ and $(\iota_b)_*$ form a pushout for all a,b.

Now, note that every class $\omega \in \pi_1(U_a \cap U_b,x_0)$ is represented twice in $*_\alpha\pi_1(U_\alpha,x_0)$ as $(\iota_{ab})_*(\omega)$ and as $(\iota_{ba})_*(\omega)$. Define the set

$$V_{ab} = \left\{(i_{ab})_*(\omega)(i_{ba})_*(\omega)^{-1} : \omega \in \pi_1(U_a \cap U_b,x_0)\right\}$$

and define $V = \bigcup_{a,b} V_{ab}$. We then define N to be the normal closure of V. That is, the minimal normal subgroup N of $*_\alpha\pi_1(U_\alpha,x_0)$.

Theorem (Seifert-van Kampen)**.** *Let X be a topological space, $\{U_\alpha\}_\alpha$ be an open cover with non-empty intersection, and x_0 some point in $\bigcap_\alpha U_\alpha$. Then,*

(i) If the intersection $U_a \cap U_b$ is path-connected for all a,b, then the free product map

$$\Phi = *_\alpha(\iota_\alpha)_* : *_\alpha\pi_1(U_\alpha,x_0) \to \pi_1(X,x_0)$$

is surjective.

(ii) If in addition the intersection $U_a \cap U_b \cap U_c$ is path-connected for all a,b,c, then $\ker(\Phi) = N$ *and hence*

$$\pi_1(X,x_0) \cong *_\alpha\pi_1(U_\alpha,x_0)\big/N$$

Example. Consider the sphere S^n for $n \geq 2$, with the cover $\{U_1,U_2\}$ given by the sets obtained by deleting two distinct points from the sphere. The intersection is path-connected, so

$$\Phi : \pi_1(U_1,x_0) * \pi_1(U_2,x_0) \to \pi_1(S^n,x_0)$$

is surjective. The open sets U_1 and U_2 are also both homeomorphic to $\mathbb{R}^n$, which is contractible, so their fundamental groups, and hence the free product, are trivial, so $\pi_1(S^n,x_0)$ must also be trivial.

This argument fails for S^1 as the intersection $U_1 \cap U_2$ is disconnected. △

Example. Let $(X,x) = \bigvee_\alpha (X_\alpha,x_\alpha)$ be the wedge product with natural basepoint $x = [x_\alpha]$, and suppose that for every α, there exists a contractible open neighbourhood $N_\alpha \subseteq X_\alpha$ of x_α. Then, for each α, define $U_\alpha = X_\alpha \vee \bigvee_{\beta \neq \alpha} N_\beta$.

Each U_α is open in X, and the basepoint is contained in their intersection $\bigcup_\alpha U_\alpha$ by construction, and each pairwise intersection is $\bigvee_\alpha U_\alpha$, which deformation retracts to x, i.e. is contractible. It follows that the pairwise intersections have trivial fundamental groups, so by Seifert-van Kampen, we have

$$\pi_1\left(\bigvee_\alpha X_\alpha,x\right) \cong *_\alpha \pi_1(X_\alpha,x_\alpha)$$

△

38.13 CW Complexes

Given spaces X and Y, a subspace $A \subseteq X$, and a map $f : X \to Y$, we can form the space

$$X \cup_f Y := X \sqcup Y / \sim$$

where $\sim$ is the equivalence relation defined by $x \sim f(x)$ for all $x \in A$. This space is equipped with the quotient topology via the surjective map $X \sqcup Y \to X \cup_f Y$, where $X \sqcup Y$ has the disjoint union topology.

We will mostly be studying a important class of spaces built from this *gluing* process called *CW complexes* (where C stands for *closure-finite* and W for *weak topology*). Informally, these are spaces constructed by recursively gluing together discs of various dimensions.

Formally, we begin with a 0-*skeleton* consisting of a disjoint union $X^0 = \bigsqcup_i D_i^0$ of 0-discs, or 0-*cells.* Then, given an $(n-1)$-skeleton X^{n-1}, we construct the X^n by gluing a collection of n-*cells* (i.e. n-discs) D_α^n via *attaching maps* $\varphi_\alpha : \partial D_\alpha^n = S_\alpha^{n-1} \to X^{n-1}$.

That is, given the maps φ_j, we define the n-skeleton X^n to be the space

$$X^n = X^{n-1} \cup_{\bigsqcup_j \varphi_j} \bigsqcup_j D_j^n$$

The attaching maps of each D_j^n canonically extend to maps $\Phi_j : D_j^n \to X^n$ over the entire disk. This extension is called the *characteristic map* of the jth n-cell D_j^n.

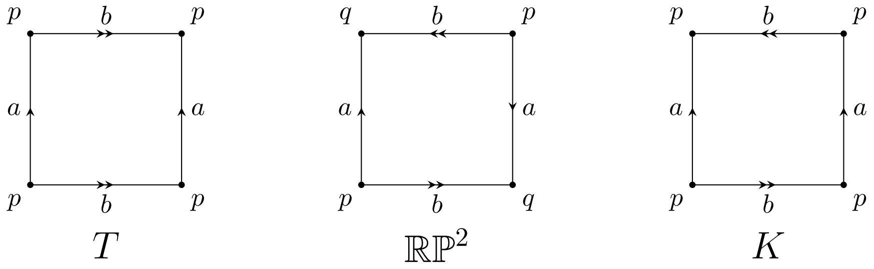
This recursion then either stops at some finite level n, yielding a CW complex $X := X^n$, or continuing infinitely with arbitrarily high dimensional discs, in which case we define $X := \bigcup_n X^n$, with a subset $U \subseteq X$ being open if and only if $U \cap X_n$ is open for all n (the "weak topology").

The *closure-finiteness* of a CW complex refers to the property that the closure of any open cell intersects with finitely many other cells.

Example. A one-dimensional CW complex is called a (topological) *graph*, consisting of a set X^0 of *vertices* and a set X^1 of *edges.* △

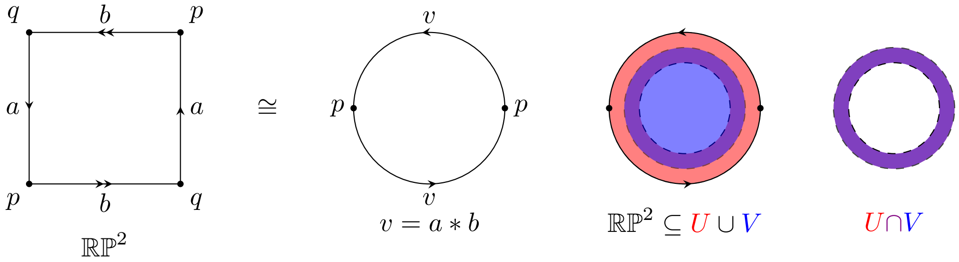
Example. The torus is a two-dimensional CW complex, given by a point $X^0 = \{p\}$, two 1-cells $X^1 = \{a,b\}$, and one 2-cell. The 1-cells are attached to the point p in the only way possible, yielding two circles attached to a point, then gluing the square to the two circles by mapping the upper and lower boundaries to one circle, and the left and right boundaries to the other.

This is often visualised by drawing the square and annotating the edges with arrows to indicate which edges are identified and in which orientation. The torus, real projective plane, and the Klein bottle can all be constructed as quotients of a square in this way:



△

These representations allow us to compute the fundamental groups of these objects. For instance, we can cover $\mathbb{RP}^2$ as follows:



Then, fix a basepoint $x_0 \in U \cap V$. The fundamental group $\pi_1(V,x_0)$ is trivial, as V is just a disk, while the intersection $U \cap V$ is homotopic to a circle, represented by some loop ω, so $\pi_1(U \cap V,x_0) = \langle [\omega] \rangle \cong \mathbb{Z}$. The set U is homeomorphic to a Möbius band, which deformation retracts, and is therefore homotopy equivalent to, its boundary γ homeomorphic to S^1, so $\pi_1(U,x_0) = \langle [\gamma] \rangle \cong \mathbb{Z}$.

Since $\pi_1(V,x_0)$ is trivial, $\pi_1(V,x_0) * \pi_1(U,x_0) \cong \pi_1(U,x_0) \cong \mathbb{Z}$. Because U encloses the boundary of a Möbius band, the inclusion $U \cap V \hookrightarrow U$ wraps twice around the circle the band retracts to, so $[\omega] \mapsto [\gamma]^2$. By Seifert-van Kampen, we then have

$$\pi_1(\mathbb{RP}^2,x_0) \cong \langle [\gamma] \rangle / \langle [\gamma]^2 \rangle \cong \mathbb{Z}/2\mathbb{Z}$$

38.13.1 Properties of CW Complexes

A topological space is *normal* if any two disjoint closed subsets have disjoint open neighbourhoods. This is related to the Hausdorff condition, which requires that every two distinct points have disjoint neighbourhoods. Note, however, that neither condition implies the other.

Lemma 38.13.1. *CW complexes are normal and Hausdorff.*

A topological space is *locally contractible* if for every $x \in X$ and every open neighbourhood $U \subseteq X$ of x, there exists an open neighbourhood $V \subseteq U$ of x that is contractible.

Lemma 38.13.2. *CW complexes are locally contractible.*

A *subcomplex* of a CW complex X is a space A that is a union of cells e_n^α in X such that the closure of each cell is also in A.

Lemma 38.13.3. *A compact topological subspace of a CW complex X is contained within a finite subcomplex.*

Lemma 38.13.4. *If $A \subseteq X$ is a subcomplex of X, then there exists an open set $U \subseteq X$ with $A \subseteq U$ such that U deformation retracts to A.*

Theorem 38.13.5. *For a path-connected CW complex X with $x_0 \in X^2$, the inclusion $X^2 \hookrightarrow X$ induces an isomorphism of fundamental groups $\pi_1(X^2,x_0) \cong x_1(X,x_0)$.*

Intuitively, this means that the fundamental group of a CW complex depends only on its 2-skeleton: loops cannot distinguish higher-dimensional topological properties.

Firstly, note that a map $\varphi : S^1 \to X$ induces a loop $f : I \to X$ based at $\varphi(1)$ via any parametrisation of the circle with I, i.e. $f(t) = \varphi\big(\exp(2\pi it)\big)$ (and intuitively, the image of a circle is precisely a loop).

In particular, the attaching map $\varphi_\alpha : S^1_\alpha \to X$ of any 2-cell D^2_α into a space X induces a loop in this way. While this loop may not be null-homotopic in X, it is certainly null-homotopic in

$$Y \coloneqq X \sqcup D^2_\alpha \big/ \big(x \sim \varphi_\alpha(x)\big)$$

after attaching the cell. If X is path-connected, then for any basepoint x_0, there is a path $h_\alpha : I \to X$ with from x_0 to $\varphi_\alpha(1)$, which induces a loop $\gamma_\alpha = h_\alpha * f_\alpha * \overline{h}_\alpha$ (here, $*$ is path concatenation, not free product!) in Y. The inclusion $X \hookrightarrow Y$ then induces a map of fundamental groups $\pi_1(X,x_0) \to \pi_1(Y,y_0)$, and the class $[\gamma_\alpha]$of every such loop γ_α is contained with the kernel of this map.

Theorem 38.13.6. *Let X be path-connected, and for a fixed n, let $\varphi^n_\alpha : S^{n-1}_\alpha \to X$ be a collection of attaching maps, and define*

$$Y \coloneqq X \sqcup \bigsqcup_\alpha D_{\alpha^n} \big/ x \sim \varphi^n_\alpha(x)$$

Let $x_0 \in X$ be a point. Then,

- *If $n = 2$, then*

$$\pi_1(Y,x_0) \cong \pi_1(X,x_0)/N$$

  where N is the normal subgroup generated by $[\gamma_\alpha]$ as defined above;

- *If $n > 2$, then*

$$\pi_1(Y,x_0) \cong \pi_1(X,x_0)$$

38.14 Generators and Relations

A *presentation* is a method of specifying a group G via a set S of *generators* – so that every element of the group may be expressed as a product of generators – and a set R of *relations* between those generators, and we write that G has presentation

$$\langle S \mid R \rangle$$

Informally, G is the "most general" or "freest" group generated by S constrained only by relations in R. Formally, G has presentation $\langle S \mid R \rangle$ if it is isomorphic to

$$G \cong \langle S \rangle \big/ \langle\langle R \rangle\rangle$$

where $\langle\langle R \rangle\rangle$ is the normal subgroup generated by R.

Example. The cyclic subgroup $\mathbb{Z}_n$ has presentation

$$\langle a \mid a^n = 1 \rangle$$

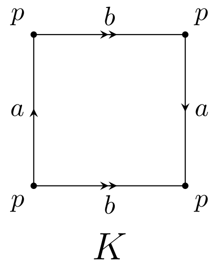
This may also be written as

$$\langle a \mid a^n \rangle$$

where the convention is that any terms without an equality symbol are taken to be equal to the group identity. △

A group is *finitely generated* if its set of generators S is finite; *finitely related* if its set of relators R is finite; and *finitely presented* if both S amd R are finite.

Example. Consider the Klein bottle K,

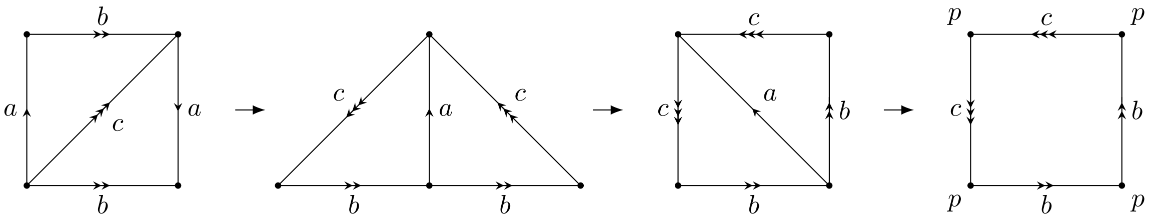


K

The 1-skeleton X^1 consists of the loops a and b, and the 0-skeleton is the single point p, so the generators are the classes corresponding to the cycles a and b, and the single relation is the loop that forms the boundary, given by $baba^{-1}$, so we have the presentation

$$\langle a,b \mid baba^{-1} \rangle$$

However, there are several cell structures on K. One rearrangement is as follows:



The resulting presentation is then

$$\langle b,c \mid b^2c^2 \rangle$$

This is of course obtainable purely group-theoretically by defining new elements in terms of old ones, but we also see that each presentation corresponds to a different way of describing a topological space. △

38.14.1 CW Complexes and Fundamental Groups

Recall that a topological graph X consists of a set of vertices X^0 and a set of edges $X^1 = \{S^0_\alpha\}_\alpha$ with attaching maps $\varphi_\alpha : S^0_\alpha \to X^0$ that assigns each interval $D^1_\alpha = I_\alpha$ to its endpoints. We also have the characteristic map $\Phi_\alpha : D^1_\alpha \to X^1$ that maps each interval to its image in the graph.

We use the term "edge" to refer both to the pair $(D^1_\alpha,\varphi_\alpha)$, which records combinatorial data, and the image $\Phi_\alpha(D^1_\alpha)$, as a topological subspace of the graph.

Given a topological graph X, an *edge-path* is a graph-theoretic path, i.e. a sequence or concatenation of connected edges in the graph

$$\gamma = e_1 * \cdots * e_n$$

where n is the *length* of the path; and an *edge-cycle* or *edge-loop* is a graph-theoretic cycle, i.e. a path that begins and ends at the same vertex.

Given a CW complex X and a point $x_0 \in X$, we can compute a presentation of the fundmaental group $\pi_1(X,x_0)$. As the path-components not containing x_0 are irrelevant to the fundamental group, we may replace X by the path-component that does contain x_0. We may also move the basepoint to lie in X^1 (or even X^0), as this component is path-connected. Finally, we may restrict to the 2-skeleton, and assume without loss of generality that $X = X^2$ is a path-connected two-dimensional CW complex.

Then,

- Given a maximal spanning tree of $T \subseteq X^1$ (found via, for example, Kruskal's or Prim's algorithm), let $\mathcal{A}$ be the set of edges not in the tree. By definition of a maximal spanning tree, pasting any edge $e \in \mathcal{A}$ into T yields a cycle, so $T \cup \{e\}$ contains a subgraph homotopic to a circle. The fundamental group of X^1 is then generated by these edge cycles:

$$\pi_1(X^1, x_0) \cong \mathop{\ast}_{e \in \mathcal{A}} \mathbb{Z}$$

  as every edge not in T yields a loop when adding it in T, and conversely, every loop in X^1 based at x_0 is homotopic to a combination of such edge-cycles.

- Let $e^2_\alpha \subseteq X^2$ be a 2-cell and $\varphi_\alpha : S^1_\alpha \to X^1$ be its attaching map. As usual, a map from the circle induces a loop $\gamma_\alpha(t)$ via parametrisation, and furthermore, this loop is homotopic to an edge-cycle. Let x_1 lie along this loop, and let g_α be a path from x_0 to x_1. Then,

$$\omega_\alpha = [g_\alpha * \gamma_\alpha * \overline{\gamma}] \in \pi_1(X^1, x_0)$$

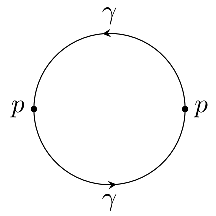
  corresponds to a reduced word u_α in $\mathcal{A}$. Define $U = \{u_\alpha\}_\alpha$ to be the set of these words. Then,

$$\pi_1(X, x_0) \cong \pi_1(X^1, x_0) / \langle\langle U \rangle\rangle$$

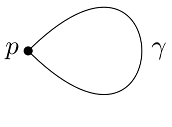
In more detail,

- Every cycle in the graph X^1 corresponds to a loop based at x_0 as we may travel from x_0 to the base of the loop (as X is path-connected), along the loop, then along the reverse path back to x_0 to close the loop. Each such cycle then corresponds to a generator of the fundamental group $\pi_1(X, x_0)$.
- Each loop in X may be represented as a combination of such cycles, corresponding to a reduced word in the generators of $\pi_1(X, x_0)$.
- A loop is null-homotopic if it is homotopic to a boundary of a 2-cell in X, and each such loop corresponds to a relation on the set of words in $\pi_1(X, x_0)$.

Example. Recall that one CW complex structure on $\mathbb{RP}^2$ is given by



The points and lines are identified together, so the 1-skeleton X^1 is just a loop,



and the spanning tree is consists of the unique edge in the graph. The fundamental group is then generated by this one cycle, whose homotopy class we denote by say, a. The relations are then all the cycles that are the boundary of a 2-cell. As can be seen in the original CW complex structure, one such loop is given by $\gamma * \gamma$, so it is represented by a^2. Thus, the fundamental group has presentation $\langle a \mid a^2 \rangle \cong \mathbb{Z}/2\mathbb{Z}$. △

Theorem 38.14.1. *For every group G, there exists a path-connected two-dimensional CW complex X_G such that*

$$\pi_1(X_G) \cong G$$

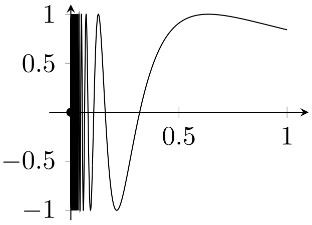
Example. Consider the cyclic group $G = \langle a \mid a^n \rangle \cong \mathbb{Z}/n\mathbb{Z}$. To construct a corresponding topological space, consider $X^1 = S^1$, and define $a = [\omega]$, where $\omega(t) = \exp(2\pi i t)$.

Then, we want an attaching map $\varphi : S^1 \to X^1$ for the 2-cell D^2 such that the induced loop γ (i.e. the boundary of the disc) satisfies $[\gamma] = a^n$. This is given via the n-fold covering $\varphi(z) = p_n(z) = z^n$. △

38.15 List of Useful (Counter)examples

- The *topologist's sine curve* is the subspace $T \subseteq \mathbb{R}^2$ defined by

$$T = \left\{ \left(x, \sin \frac{1}{x} \right) : x \in (0,1] \right\} \cup \{(0,0)\}$$

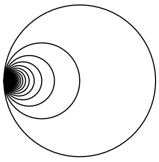


The topologist's sine curve is connected but not path-connected because the origin cannot be separated from the rest of the curve, but also cannot be connected to the rest of the curve via a path.

- Define $C_n \subseteq \mathbb{R}^2$ as the circle of radius $1/n$ centred at $(0,1/n)$. The *Hawaiian earring* is the union

$$H = \bigcup_{n \in \mathbb{Z}^+} C_n$$

equipped with the subspace topology.



The Hawaiian earring looks similar to the infinite wedge sum

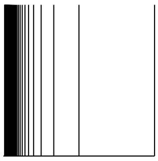
$$X = \bigvee_{n \in \mathbb{N}} S^1$$

but they are not homeomorphic:

  - The fundamental group $\pi_1(X)$ is countable, while $\pi_1(H)$ is uncountable.
  - The Hawaiian earring is compact, while the wedge sum is not.
  - In the Hawaiian earring, every open neighbourhood of the intersection point completely contains all but finitely many of the circles (i.e. an ε-ball around $(0,0)$ contains every circle whose radius is less than $\varepsilon/2$), while in the wedge sum, such a neighbourhood may contain no circles at all.

- The *topologist's comb* is the subset $C \subseteq \mathbb{R}^2$ defined by

$$C = \big(\{0\} \times [0,1]\big) \cup \left(\left\{\frac{1}{n} : n \in \mathbb{N}^+\right\} \times [0,1]\right) \cup \big([0,1] \times \{0\}\big)$$



  The comb space is contractible but does not deformation reformation to any point on the line segment $\{0\} \times [0,1]$.

- The *closed long ray* is the product of the first uncountable ordinal ω_1 with the half-open interval $[0,1)$:

$$L = \prod_{i \in \omega_1} [0,1)$$

  equipped with the lexicographical order topology. (Compare with the real number line, which can be constructed as the product of $\mathbb{N}$ copies of $[0,1)$.)

  The *open long ray* is obtained by removing $(0,0)$, and the *long line* is then obtained by gluing together two copies of the closed long ray at the origin.

  - The long rays and line are path-connected but not contractible.
  - The long rays and line are normal, Hausdorff, and sequentially compact, but not compact nor metrisable.

Chapter 39

Homology

> *"Of course, topologists don't care about any of this 'applied math' nonsense. They're just trying to find all the shapes."*
>
> — Milo Beckman, *Math Without Numbers*

The (fundamental) homotopy groups discussed in the previous chapter are a powerful invariant for some topological spaces, but they are unable to distinguish topological spaces in higher dimensions, and the higher dimensional analogues of nth homotopy groups become incredibly difficult to calculate – even for simple spaces like n-spheres, the homotopy groups are generally unknown. Instead, we study *homology* groups, which are slightly less powerful, but generally much easier to compute. But in exchange, we require significantly more preamble before we may develop much theory.

39.1 Preliminary Concepts

39.1.1 Note on Notation

In general, we will take the word "map" to mean a *continuous* function, and "space" to mean a *topological* space.

We will write $X \rightarrowtail Y$ to denote an injective map (more generally, a monomorphism) and $X \twoheadrightarrow Y$ for a surjective map (more generally, an epimorphism). In some texts, $X \hookrightarrow Y$ is used for monomorphisms, but here we reserve this symbol exclusively for inclusion maps. We use no special notation for projection maps.

We write $f^{-1}[X]$ for the preimage of a *set* X under a function f.

39.1.2 Common topological spaces

We list some standard topological spaces:

- The unit interval I is the subspace $I \coloneqq [0,1] \subset \mathbb{R}$.
- The (closed) n-disk D^n is the subspace of $\mathbb{R}^n$ defined by

$$D^n \coloneqq \left\{ \mathbf{x} \in \mathbb{R}^n : \sum_{i=1}^{n} x_i^2 \leq 1 \right\}$$

- The n-sphere S^n is the boundary of the $(n+1)$-disc

$$S^n \coloneqq \partial D^n = \left\{ \mathbf{x} \in \mathbb{R}^n : \sum_{i=1}^{n} x_i^2 = 1 \right\}$$

- The n-cube I^n is the n-fold product of the unit interval I:

$$I^n \coloneqq \prod_{i=1}^{n} I \cong \left\{ \mathbf{x} \in \mathbb{R}^n : 0 \leq x_i \leq 1 \right\}$$

- The n-torus T^n is the n-fold product of the 1-sphere:

$$T^n \coloneqq \prod_{i=1}^{n} S^1 \cong \mathbb{R}^n/\mathbb{Z}^n$$

39.1.3 Homotopies

If $f{,}g : X \to Y$ are maps between topological spaces X and Y, then a *homotopy* from f to g is a map $H : X \times I \to Y$ such that $H(-{,}0) = f$ and $H(-{,}1) = g$. This second variable is commonly denoted by t and called the *time* parameter. Intuitively, a homotopy parametrises (in the second variable) a continuous deformation of f to g.

If there exists a homotopy between f and g, we say they are *homotopic* and write $f \simeq g$ (or occasionally $f \simeq_H g$, if the particular homotopy is relevant).

Two topological spaces X and Y are *homotopy equivalent* if there exist maps $f : X \to Y$ and $g : Y \to X$ such that $g \circ f \cong \mathrm{id}_X$ and $f \circ g \cong \mathrm{id}_Y$.

A space is *contractible* if it is homotopy equivalent to a point.

39.1.4 Pairs

A *pair* $(X{,}A)$ consists of a topological space X and a subspace $A \subseteq X$. We denote the interior of A by A°, the closure of A by $\overline{A}$, and the boundary of A by $\partial A = \overline{A} \setminus A^\circ$. When $A = \{x\}$ is a single point, we instead write $(X{,}x)$, and call the pair a *pointed space* (we sometimes call X alone a pointed space with *basepoint* x).

A *map of pairs* $f : (X{,}A) \to (Y{,}B)$ is a continuous function $f : X \to Y$ such that $f(A) \subseteq B$. If A and B are points, then f is a *pointed* or *based* map. If $f{,}g : (X{,}A) \to (Y{,}B)$ are maps such that $f\big|_A = g\big|_A$, then a *homotopy relative to* A from f to g is a homotopy $H : X \times I \to Y$ such that $H(x{,}t) = f(x) = g(x)$ for all $x \in A$ and $t \in I$. That is, a homotopy relative to A is a homotopy that is constant over A. Again, if A and B are points, then H is a *pointed homotopy.*

Given a pair $(X{,}A)$, a *retraction* is a map $r : X \to X$ such that $r(X) = A$ and $r\big|_A = \mathrm{id}_A$. That is, a retraction is a (necessarily surjective) mapping from a space X onto a subspace A that preserves all points within that subspace. For instance, any non-empty space retracts to a point in the obvious way (just take the constant map). If a retraction exists, then A is a *retract* of X.

A *deformation retract* is a homotopy H relative to A between the identity id_X, and a retraction r. That is, $H(x{,}0) = x$, $H(x{,}1) \in A$, and $H(a{,}1) = a$ for every $x \in X$ and $a \in A$. A deformation retract captures the idea of continuously compressing a space onto a subspace: $H(-{,}0)$ is the identity on X and as the time parameter increases to 1, this mapping continuously shrinks down to the identity on A. Note that every deformation retract induces a retract $H(-{,}1) : X \to X$, but in general, retracts need not be deformation retracts – for instance, the constant map in any non-empty space is a retract, but is a deformation retract only if the space is contractible. Note also that a deformation retract induces a homotopy equivalence $A \simeq X$.

39.1.5 Quotient Spaces

Given a topological space X, we can endow a topology on any subset $A \subseteq X$ called the *subspace topology* by taking the open sets of A to be the open sets of X intersected with U.

This construction has the following universal property: if Y is any topological space and $f : Y \to A$ is a function, then f is continuous if and only if the composite

$$\begin{tikzcd} Y \arrow[r, "f"] & A \arrow[r, hook, "\iota"] & X \end{tikzcd}$$

is continuous as a function $Y \to X$, where ι is the canonical inclusion of A into X.

A similar construction can be performed for quotients; that is, a surjective set map $\pi : X \to B$. A topology can be placed on B by declaring that a set $U \subseteq B$ is open if and only if $\pi^{-1}[U]$ is open in X. This construction is called the *quotient topology* on B, and has the following universal property dual to that of subspaces: if Y is any topological spaec and $g : B \to Y$ is a function, then g is continuous if and only if the composite

$$\begin{tikzcd} X \arrow[r, "\pi"] & B \arrow[r, hook, "g"] & Y \end{tikzcd}$$

is continuous as a function $X \to Y$.

We sometimes prefer to describe a quotient as an equivalence class on X. This characterisation is equivalent to a surjective map $\pi : X \to B$, as given such a π, we may define an equivalence relation $\sim$ by $x \sim y$ if and only if $\pi(x) = \pi(y)$, so elements with the same image are identified under this relation. Conversely, given $\sim$ on X, we define B to be the set of equivalence classes and π to be the map defined by $x \mapsto [x]$.

If we have $A \subseteq X$, then we can define an equivalence relation $\sim$ such that $x \sim y$ if and only if both x,$y \in A$ or $x = y$. That is, every point in A is identified under $\sim$, and every point in $X \setminus A$ lies within its own singleton equivalence class. By a small abuse of notation, the resulting quotient space is denoted by X/A. Intuitively, this is the quotient space obtained by contracting all of A into a single point.

39.1.6 Gluing and CW Complexes

Given spaces X and Y, a subspace $A \subseteq X$, and a map $f : X \to Y$, we can form the space

$$X \cup_f Y := X \sqcup Y \big/ \sim$$

where $\sim$ is the equivalence relation defined by $x \sim f(x)$ for all $x \in A$. This space is equipped with the quotient topology via the surjective map $X \sqcup Y \to X \cup_f Y$, where $X \sqcup Y$ has the disjoint union topology.

We will mostly be studying a important class of spaces built from this *gluing* process called *CW complexes* (where C stands for closure-finite and W for weak topology). Informally, these are spaces constructed by recursively gluing together discs of various dimensions.

Formally, we begin with a 0-*skeleton* consisting of a disjoint union $X^0 = \bigsqcup_i D_i^0$ of 0-discs, or points. Then given an $(n-1)$-skeleton X^{n-1}, we glue a collection of n-discs $\{D_j^n\}$ via *attaching maps* $\varphi_j : \partial D_j^n = S_j^{n-1} \to X^{n-1}$.

That is, given the maps φ_j, we define the n-skeleton X^n to be the space

$$X^n = X^{n-1} \bigcup_{\bigsqcup_j \varphi_j} \bigsqcup_j D_j^n$$

The attaching maps of each D_j^n canonically extend to maps $\varphi : D_j^n \to X^n$, and the images of these maps are called the *n-cells* of X, and the extension of φ_j is called the *characteristic map* of this n-cell.

This recursion then either stops at some finite level n, yielding a CW complex $X := X^n$, or continuing infinitely with arbitrarily high dimensional discs, in which case we define $X := \bigcup_n X^n$, with a subset $U \subseteq X$ being open if and only if $U \cap X_n$ is open for all n.

39.1.7 Group Theory

39.1.7.1 Free Products

Let $\{G_\alpha\}_\alpha$ be a collection of groups. A *word* on these groups is a finite sequence $g_1 \cdots g_m$ of elements $g_i \in G_{\alpha_i}$, and m is the *length* of the word. The empty word of length 0 is denoted by ε. The *product* of two words is their concatenation

$$(g_1 \cdots g_m) * (h_1 \cdots h_n) = g_1 \cdots g_m h_1 \cdots h_n$$

A word is *reduced* if it does not contain the identity of any group, and if every pair of consecutive letters are not from the same group.

Given any word g on the groups $\{G_\alpha\}_\alpha$, we can *reduce* it to a reduced word g' by removing all identity elements and replacing any consecutive elements g_i,g_{i+1} from the same group with their group product $g_i \cdot g_{i+1}$.

Let $*_\alpha G_\alpha$ be the set of reduced words on $\{G_\alpha\}_\alpha$. We can define an operation on this set as follows: given reduced words $g = g_1 \cdots g_m$ and $h = h_1 \ldots h_n$, construct a new reduced word $g \bullet h$ by taking the concatenation $g * h$, then reduce the word recursively

$$g \bullet h = \begin{cases} gh & g_m \in G_\alpha, h_1 \in G_\beta,\, G_\alpha \neq G_\beta \\ g_1 \cdots g_{m-1}(g_m \cdot h_1)h_2 \cdots h_n & g_m, h_1 \in G_\alpha,\, g_m \cdot h_1 \neq \mathrm{id}_{G_\alpha} \\ g_1 \cdots g_{m-1} \bullet h_1 \cdots h_n & g_m, h_1 \in G_\alpha,\, g_m \cdot h_1 = \mathrm{id}_{G_\alpha} \end{cases}$$

Then, $(*_\alpha G_\alpha, \bullet)$ is a group called the *free product* of $\{G_\alpha\}_\alpha$, with identity ε, and the inverse of an element $g_1 \cdots g_m$ is given by $g_m^{-1} \cdots g_1^{-1}$.

39.1.7.2 Cokernels

Given a group homomorphism $\phi : A \to B$ of abelian groups, we have two fundamental subgroups, given by the image $\mathrm{im}(\phi) \leq B$, and the kernel $\ker(T) \trianglelefteq A$. A third fundamental subspace is given by the *cokernel*, defined as the quotient

$$\mathrm{coker}(\phi) \coloneqq B / \mathrm{im}(\phi)$$

For intuition on this definition, note that this definition makes sense for linear maps and vector spaces. A linear map $T : A \to B$ is a way to transform A into B. The kernel can be viewed as the space of elements in A that are "destroyed" by T. Then, the cokernel can be viewed as the space of elements in B that are "created" by T, in the sense that A is mapped to $\mathrm{im}(A) \subseteq B$, so any other element in B is new.

39.1.7.3 Smith Normal Form and the Structure Theorem for Finitely Generated Abelian Groups

Given two finite-dimensional vector spaces V and W over a field K and any linear map $T : V \to W$, there exist bases of V and W with respect to which the matrix of T is a block matrix of the form

$$\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}$$

where r is the rank of T. A slightly weaker result holds if instead of vector spaces, we work with finitely generated free modules over $\mathbb{Z}$. Note that these modules are exactly the finitely generated abelian groups, so we phrase this theorem in terms of groups.

Given any group homomorphism $\phi : \mathbb{Z}^n \to \mathbb{Z}^m$, there exist bases of $\mathbb{Z}^n$ and $\mathbb{Z}^m$ and a diagonal matrix of the form

$$D = \begin{bmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_r \end{bmatrix}$$

with *invariants* $d_i \in \mathbb{Z}^+$ and $d_1 \mid d_2 \mid d_3 \mid \ldots \mid d_r$ (where $\mid$ is the divides relation), such that the matrix of ϕ with respect to these bases is a block matrix of the form

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}$$

That is, for any $m \times n$ matrix M with entries in $\mathbb{Z}$, there exist change of basis matrices $P \in \mathbb{Z}^{m \times m}$ and $Q \in \mathbb{Z}^{n \times n}$ such that $PMQ = \Sigma$ is of the above form, called the Smith normal form of M.

This form is convenient for calculating the kernel and cokernel of φ. First note that $r \leq \min(n{,}m)$. Then,

$$\ker \varphi \cong \mathbb{Z}^{n-r}$$

(note, $n - r$ is the number of zero columns) and

$$\operatorname{coker} \varphi \cong \left(\bigoplus_{i=1}^{r} \mathbb{Z}/d_r \right) \oplus \mathbb{Z}^{m-r}$$

(note, $m - r$ is the number of zero rows).

Since every finitely generated abelian group is the cokernel of some map, we see that every finitely generated abelian group must be of this form:

$$A \cong \left(\bigoplus_{i=1}^{r} \mathbb{Z}/d_r \right) \oplus \mathbb{Z}^{k}$$

and we call k the *rank* of A, written as $k = \operatorname{rk}_{\mathbb{Z}}(A)$.

39.2 Introduction

The goal of algebraic topology is to translate questions in topology into questions in algebra, most commonly by constructing algebraic invariants of topological spaces. That is, given a space X, we wish to construct an algebraic structure $A(X)$ such that if spaces X and Y are homeomorphic (or just homotopy equivalent), then the associated algebraic objects are isomorphic:

$$X \simeq Y \quad \longrightarrow \quad A(X) \cong A(Y)$$

For an algebraic invariant $A(-)$ to be useful, we require that:

1. it is “easy” to compute $A(-)$ and to tell when the algebraic objects are not isomorphic;
2. the algebraic invariant is “fine” enough in that $A(X) \ncong A(Y)$ often, for non-homeomorphic X and Y.

We have previously constructed the fundamental group (or first homotopy group) $\pi_1(-)$. This invariant takes a pointed space (X,x) as input and returns the group of homotopy classes of pointed maps $(S^1,*) \to (X,x)$, or loops in X based at x, under the operation of path concatenation.

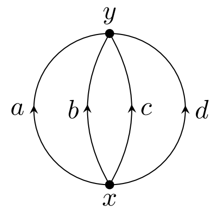
The fundamental group is a complete invariant for compact surfaces, but fails to capture the topology of higher-dimensions, only detecting the 1-dimensional hole structure of a space. For instance, the fundamental groups of $\mathbb{R}^3$ and $\mathbb{R}^4$ are both trivial and cannot be used to distinguish them. Fundamental groups are also in general non-abelian, so it is difficult to determine whether two such groups are non-isomorphic.

One natural generalisation of this is to consider pointed maps not just from the circle S^1, but from n-spheres into a pointed map: given a pointed space (X,x), the *nth homotopy group* $\pi_n(X,x)$ is the group of homotopy classes of pointed maps $(S^n,*) \to (X,x)$. The nth homotopy groups are a much more fine invariant, and can tell apart many topological spaces, but they are extremely difficult to compute. Even for simple spaces like spheres, the homotopy groups are generally unknown.

Here, we study a more computable alternative: *homology groups*. These invariants are abelian groups and are easier to distinguish and compute than homotopy groups – for instance, the homology groups of spheres are all known – but conversely, they contain less information than homotopy groups.

39.2.1 Homology

Consider the following graph, X_1, consisting of two 0-cells connected with four oriented 1-cells:



The fundamental group of X_1 consists of loops formed by sequences of edges, starting and ending at some fixed basepoint. For instance, at the basepoint x, one possible loop is given by ab^{-1}, travelling along a, then along b in reverse direction. Another loop is given by $ad^{-1}bc^{-1}ac^{-1}$. Because these loops must be continuous paths, the fundamental group is generally non-abelian.

To simplify, let us consider what happens if we abelianise this group. For example, the loops ab^{-1} and $b^{-1}a$ are equal if we allow a to commute with b^{-1}. Note that these loops are really the same circle, just with a different basepoint – x for ab^{-1} and y for $b^{-1}a$. Choosing a new basepoint in a loop just cyclically permutes its edges, so we no longer have to consider pointed spaces: instead of loops, we have *cycles*, independent of a choice of basepoint.

Now working with an abelian group, we swap to additive notation, so cycles are $\mathbb{Z}$-linear combinations of edges. More generally, a (1-)*chain* is any such linear combination of edges. Even more generally, for any cell complex X, an n*-chain* in X is a linear combination of n-cells – that is, the group of n-chains in a space X with k n-cells $\{c_i\}_{i=1}^k$ is the free abelian group $C_n(X) = \bigoplus_{i=1}^k \mathbb{Z} \cdot c_i \cong \mathbb{Z}^k$ on the basis $\{c_i\}$.

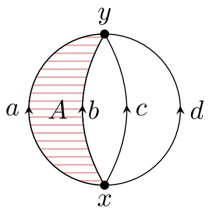
Note that not all 1-chains may be interpreted as paths, as endpoints do not have to match up. For instance $a + b$ is a chain, but not a meaningful path. In any case, the order of concatenation of edges is immaterial in an 1-chain. Note also that chains may have multiple decompositions into cycles: for instance, $(a - b) + (c - d) = (a - d) + (c - b)$, so more generally, we define a *cycle* to be any chain that has at least one decomposition into a cycle of the previous geometric sense. How can we determine when a chain admits such a decomposition?

In a geometric cycle, interpreted as a path, every vertex is entered and exited the same number of times. In the above graph, given a chain $\alpha a + \beta b + \gamma c + \delta d$, the net number of times y is entered is $\alpha + \beta + \gamma + \delta$, and similarly, the net number of times x is entered is $-\alpha - \beta - \gamma - \delta$. For a chain to be a cycle, we require that these quantities are simultaneously zero, so in the above graph, a chain is a cycle if and only if $\alpha + \beta + \gamma + \delta = 0$.

Let C_1 be the free abelian group with basis a,b,c,d, and C_0 the free abelian group with basis x,y. Elements of C_1 are then linear combinations of edges, which are exactly the 1-chains, and similarly, elements of C_0 are linear combinations of vertices, or 0*-chains*.

We define the *(1st) boundary* homomorphism $\partial_1 : C_1 \to C_0$ by sending each basis element (edge) to the vertex at its head, minus its vertex at the tail. For instance, for the graph above, every edge is sent to $y - x$, as every edge points from x to y. Then, the action of this homomorphism on a chain $\alpha a + \beta b + \gamma c + \delta d$ is given by $(\alpha + \beta + \gamma + \delta)y - (\alpha + \beta + \gamma + \delta)x$. Thus, the cycles are precisely the kernel of ∂_1. It is easy to verify that $a - b$, $b - c$, and $c - d$ form a basis for this kernel – so every cycle in X_1 is a linear combination of these three cycles. In this way, this kernel captures the geometric information that the graph X_1 has three "(1-dimensional) holes".

Let us expand the graph by attaching a 2-cell, A, along the cycle $a - b$ to produce a 2-dimensional cell complex X_2.



We similarly define the group C_2 to be the free abelian group with basis A. We can also define another boundary operator $\partial_2 : C_2 \to C_1$, but this requires a choice of orientation for A.

If we regard A as being oriented clockwise, its boundary is then the cycle $a - b$. This cycle now no longer encloses a hole as it did in X_1, as it can be linearly contracted to a point over A. This suggests that we form a quotient of the group of cycles in the previous example by factoring out the subgroup generated by $a - b$. For instance, the cycles $a - c$ and $b - c$ would now be equivalent in this quotient, consistent with them being homotopic in X_2. This quotient group is exactly $\ker \partial_1 / \operatorname{im} \partial_2$ – the 1-cycles modulo those that are boundaries of 2-cells. This quotient group is the *homology group* $H_1(X_2)$. In this case, $H_1(X_2)$ is free abelian on 2 generators, corresponding to filling A having removed one of the three holes.

We can also compute $H_1(X_1)$ by taking $C_2 = 0$ to be the trivial group, as there are no 2-cells in X_1, and ∂_2 to be the trivial homomorphism; so we have $H_1(X_1) = \ker \partial_1 / \operatorname{im} \partial_2 = \ker \partial_1$ is free abelian on 3 generators, corresponding to our three 1-dimensional holes.

We could attach another 2-cell along the same cycle $a - b$, forming a kind of hollow banana shape in X_3. $H_1(X_3)$ is unchanged, but now ∂_2 has a non-trivial kernel – the group generated by the spherical 2-cycle $A - B$. Just as the three cycles in X_1 detected 1-dimensional holes, the presence of the 2-cycle $A - B$ indicates the existence of a 2-dimensional hole – the missing interior of this sphere. We could expand this cell complex again, attaching a 3-cell C along $A - B$. This creates a new chain group C_3, and we can define a boundary homomorphism $\partial_3 : C_3 \to C_2$ by sending C to $A - B$ (note that this again depends on a choice of orientation for C). Now, $H_2(X_4) = \ker \partial_2 / \operatorname{im} \partial_3$ is trivial, as this 2-hole has now been filled in.

39.3 Simplicial Homology

The general pattern is now clear: for any cell complex, we have chain groups $C_n(X)$ of n-chains in X, and boundary operators $\partial_n : C_n(X) \to C_{n-1}(X)$, from which we define the nth homology group $H_n(X) = \ker \partial_n / \operatorname{im} \partial_{n+1}$.

The difficulty is in how we define ∂_n in general. For $n = 1$, this is not hard: the boundary of an edge is the vertex at its head minus the vertex at its tail. However, for arbitrary n, this becomes rather complicated.

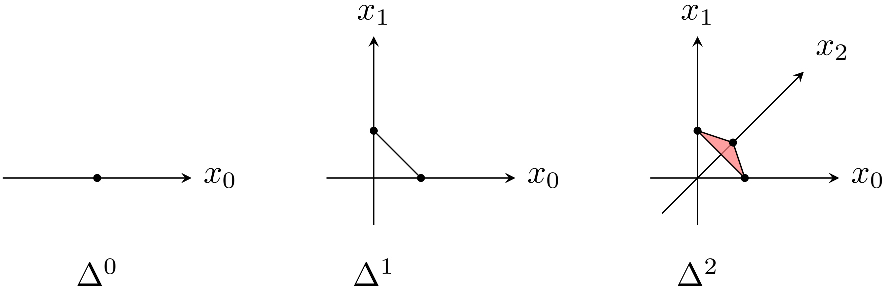
For now, we start with a simplified version of homology called *simplicial homology*, where we deal only with topological spaces that can be expressed in terms of *simplices* which admit an easier notion of boundary mapping.

39.3.1 Δ-Complexes

The *standard n-simplex* $\Delta^n \subseteq \mathbb{R}^{n+1}$ is the subspace

$$\Delta^n := \left\{ \mathbf{x} \in \mathbb{R}^{n+1} : x_i \geq 0, \sum_{i=0}^{n} x_i = 1 \right\}$$

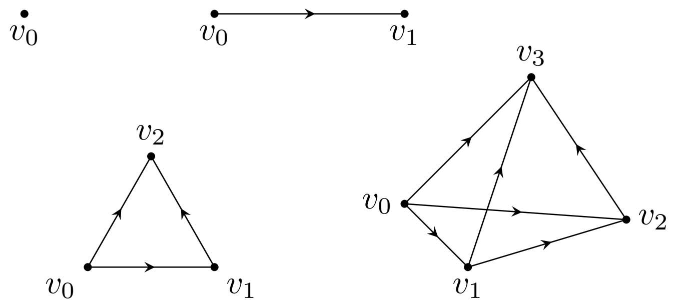
whose vertices $v_0, v_1, \ldots, v_n$ are the unit vectors along the coordinate axes. More generally, any homeomorphic space will also be called an n-simplex and labelled Δ^n.



The standard n-simplex for $n = 0,1,2$

From now, on we suppress the coordinate axes and draw general (non-standard) simplices (particularly as drawing axes in > 3 dimensions is rather difficult!).

For the purposes of simplicial homology, it is important to keep track of the ordering of these vertices, so we also refer to a simplex by an ordered list of its vertices: $[v_0, \ldots, v_n]$. This representation also has a side effect of determining the orientation of its edges $[v_i, v_j]$ according to increasing subscripts. When drawing simplices, the convention is to annotate the edges with an arrow pointing in ascending order of vertices:



Deleting one of the $n+1$ vertices of an n-simplex $\Delta^n = [v_0, \ldots, v_n]$, say v_j, the remaining n vertices span an $(n-1)$-simplex $[v_0, \ldots, \widehat{v_j}, \ldots, v_n]$ (where $\widehat{\ }$ indicates omission) called the jth *face* of $[v_0, \ldots, v_n]$, denoted by $\partial_j \Delta^n$.

More concretely, in terms of the standard n-simplex, the jth face of Δ^n is the subspace $\partial_j \Delta^n \subseteq \Delta^n$ of points whose jth coordinate is zero. That is,

$$\begin{aligned} \partial_j \Delta^n &= \{\mathbf{x} \in \Delta^n : x_j = 0\} \\ &= \left\{ \mathbf{x} \in \mathbb{R}^{n+1} : x_j = 0, x_i \geq 0 \sum_{i=0}^{n} x_i = 1, \right\} \end{aligned}$$

Note that, geometrically, the jth face is the one *opposite* the deleted jth vertex.

For instance, in the above simplices: the 0th face of Δ^1 is the vertex v_1, and the 1st face is v_0; the 0th face of Δ^2 is the edge $[v_1, v_2]$, the 1st face the edge $[v_0, v_2]$, and the 2nd face the edge $[v_0, v_1]$; and the 0th face of Δ^3 is the triangle $[v_1, v_2, v_3]$, etc.

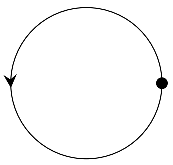
The *boundary* $\partial \Delta^n$ of Δ^n is then the union of its $n+1$ faces. Note that the unique face $\partial_0 \Delta^0$ of Δ^0 is the empty set, so $\partial \Delta^0 = \varnothing$.

A Δ-*complex* X is a topological space[*] defined inductively as follows:

1. Start with a collection of 0-simplices, or points. This is the 0-skeleton X^0.
2. Inductively, the n-skeleton X^n is obtained from X^{n-1} by attaching n-simplices Δ^n_α where each face $\partial_i \Delta^n_\alpha$ is identified with an $(n-1)$-simplex Δ^{n-1}_β in X^{n-1}.
3. If k is the minimal k such that $X^k = X^{k+1}$, i.e., there are no m-cells added for any $m > k$, then $X = X^k$ has dimension k.

   More generally, $X = \bigcup_{n \in \mathbb{N}} X^n$, in which case, a subspace $U \subseteq X$ is open if and only if $U \cap X^n \subseteq X^n$ is open for all n.

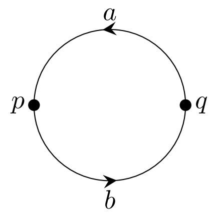
Example. Start with a single point $X^0 = \Delta^0$, and attach a single 1-simplex Δ^1 in the only possible way. That is, both boundary points are identified with the 0-skeleton:



△

Example. Now start with two points $X^0 = \{p, q\}$, and attach two 1-simplices a, b as follows:

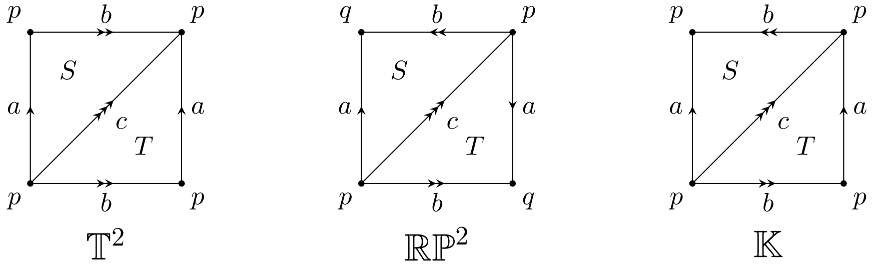
[*] More properly, a Δ-complex *structure on* a space X.



with the arrows indicating that the 0th face of a is identified with p, the 1st face with q; and the 0th face of b is identified with q, and the 1st face with p.

Note that both of these spaces are homeomorphic to a circle, so a topological space can have multiple Δ-complex structures. △

Example. The torus $\mathbb{T}^2$, real projective plane $\mathbb{RP}^2$, and the Klein bottle $\mathbb{K}$ can all be constructed as quotients of a square by identifying opposite edges. They can all be constructed as Δ-complexes as follows:



△

A Δ-complex is essentially just combinatorial data. That is, it is determined up to homeomorphism by the sets of n-simplices, S_n, $n \geq 0$, together with the attaching rules – namely, the *face maps* $d_i^n : S_n \to S_{n-1}$, $0 \leq i \leq n$, specifying that $\partial_i \Delta^n_\alpha$ is identified with $\Delta^{n-1}_{d_i^n(\alpha)}$. These maps are not arbitrary, but satisfy the relation

$$d_i^{n-1} \circ d_j^n = d_{j-1}^{n-1} \circ d_i^n$$

whenever $i < j$.

Writing the simplex as $[v_0,\ldots,v_i,\ldots,v_j,\ldots,v_n]$, this relation is just saying that removing v_j, then v_i, should be the same as removing v_i, then v_j; removing v_i first means that v_j is the $(j-1)$th vertex in the intermediary simplex.

Such a collection of combinatorial data $S = (S_\bullet, d_\bullet^\bullet)$ is called a Δ-*set* or *semi-simplicial set.*

Given a Δ-set S, we denote the associated Δ-complex, called its *geometric realisation*, by $|S|$. More precisely, a Δ-complex is really a topological space X equipped with a homeomorphism $X \cong |S|$ for some Δ-set S, the latter of which is then called a Δ-complex structure on X. As in the earlier examples of circles, a topological space can admit distinct Δ-complex structures.

Example. In the torus $\mathbb{T}^2$ above, we have

$$S_0 = \{p\}, \qquad S_1 = \{a,b,c\}, \qquad S_2\{S,T\}$$

with face maps

$$\begin{aligned} d_0^2(S) &= b, & d_0^2(T) &= a, \\ d_1^2(S) &= c, & d_1^2(T) &= c, \\ d_2^2(S) &= a, & d_2^2(T) &= b; \end{aligned} \qquad \begin{aligned} d_0^1(a) = d_0^1(b) = d_0^1(c) = p, \\ d_1^1(a) = d_1^1(b) = d_1^1(c) = p; \end{aligned} \qquad d_0^0(p) = \varnothing$$

△

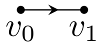
39.3.2 Simplicial Homology

Previously, we defined some homology groups on simple CW complexes in terms of certain *boundary operators*. For arbitrary CW complexes, defining these operators for higher dimensions is tricky, but for Δ-complexes, the situation is a little easier.

- For a 0-simplex, i.e. a point,

  the boundary is empty.

- For a 1-simplex,

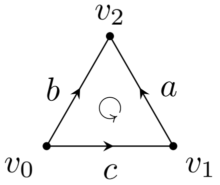
  

  we define its oriented boundary to be the formal difference between the vertices at its head and its tail, just like we did for CW complexes.

  $$v_1 - v_0$$

  Note that the choice of orientation is arbitrary, and $v_0 - v_1$ would work just as well.

- For a 2-simplex,

  we define its oriented boundary as

  $$a - b + c$$

  Again, this choice is arbitrary, and orienting clockwise works equally well.

- Similarly, for a general n-simplex s, the boundary is then the alternating sum of its faces:

  $$\partial_n(s) = \sum_{i=0}^{n}(-1)^i d_i^n(s)$$
  $$\partial_n\big([v_0, \dots, v_n]\big) = \sum_{i=0}^{n}(-1)^i [v_0, \dots, \widehat{v_i}, \dots, v_n]$$

  where $\widehat{\ }$ indicates omission.

From this point, the theory is entirely the same as in the introduction:

Let S be a Δ-set.

- The group of *n-chains* in S is the free abelian group on S_n, denoted by $\Delta_n(S)$.
- The *boundary operator* $\partial : \Delta_n(S) \to \Delta_{n-1}(S)$ is the homomorphism given on the generators $s \in S_n$ by the formula above, noting that $\Delta_{-1}(S)$ is the trivial group 0, and that ∂_0 is the zero map.
- The group of *n-cycles* $Z_n(S)$ is the kernel of the nth boundary operator,

  $$Z_n(S) \coloneqq \ker(\partial_n)$$

- The group of *n-boundaries* $B_n(S)$ is the image of the $(n+1)$th boundary operator,

$$B_n(S) \coloneqq \operatorname{im}(\partial_{n+1})$$

- The *nth simplicial homology group* $H_n^{\Delta}(S)$ is the group of n-cycles modulo those that are boundaries,

$$H_n^{\Delta}(S) \coloneqq \frac{Z_n(S)}{B_n(S)} = \frac{\ker(\partial_n)}{\operatorname{im}(\partial_{n+1})}$$

  and elements of this group are called *homology classes*.

Of course, for this last expression to hold, we require that $B_n(S) \subseteq Z_n(S)$ for all n and all Δ-sets S.

Lemma 39.3.1. *Let S be a Δ-set. Then, $\partial_n \circ \partial_{n+1} = 0$. Equivalently, $B_n(S) \subseteq Z_n(S)$.*

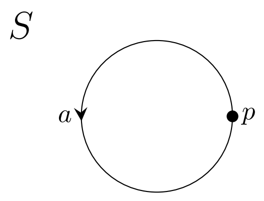
Proof. The equivalence of the two statements is clear from the definition of cycles and boundaries.

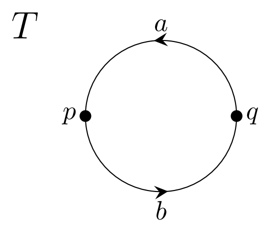
Let $s \in S_{n+1}$. Then,

$$\begin{aligned}
\partial_n \partial_{n+1}(s) &= \partial_n \left(\sum_{j=0}^{n+1} (-1)^j d_j^{n+1}(s) \right) \\
&= \sum_{i=0}^{n} \sum_{j=0}^{n+1} (-1)^{i+j} d_i^n d_j^{n+1}(s) \\
&= \sum_{0 \le i < j \le n+1} (-1)^{i+j} d_i^n d_j^{n+1}(s) \quad + \quad \sum_{0 \le j \le i \le n} (-1)^{i+j} d_i^n d_j^{n+1}(s) \\
&= \sum_{0 \le i < j \le n+1} (-1)^{i+j} d_{j-1}^n d_i^{n+1}(s) \quad + \quad \sum_{0 \le j \le i \le n} (-1)^{i+j} d_i^n d_j^{n+1}(s) \\
&= \sum_{0 \le i < j \le n+1} (-1)^{i+j-1} d_j^n d_i^{n+1}(s) \quad + \quad \sum_{0 \le j \le i \le n} (-1)^{i+j} d_i^n d_j^{n+1}(s) \\
&= 0
\end{aligned}$$

■

Example. We calculate the homology groups of the following Δ-sets.





For S, we have boundary operators

$$\begin{aligned} \partial_1 &: \mathbb{Z}a = \Delta_1(S) \to \Delta_0(S) = \mathbb{Z}p \\ \partial_0 &: \mathbb{Z}p = \Delta_0(S) \to \Delta_{-1}(S) = 0 \end{aligned}$$

respectively defined on the generators a and p by

$$\begin{aligned} \partial_1(a) &= p - p = 0 \\ \partial_0(p) &= 0 \end{aligned}$$

with all other boundary operators trivially the zero map, as there are no simplices of any other dimension. These boundary operators both vanish, so $Z_1(S) = Z_0(S) = \mathbb{Z}$, and $B_n(S) = 0$ for all n. So, $H_0(S) = H_1(S) = \mathbb{Z}/0 \cong \mathbb{Z}$.
For T, we have boundary operators

$$\begin{aligned} \partial_1 &: \mathbb{Z}a \oplus \mathbb{Z}b = \Delta_1(S) \to \Delta_0(S) = \mathbb{Z}p \oplus \mathbb{Z}p \\ \partial_0 &: \mathbb{Z}p \oplus \mathbb{Z}q = \Delta_0(S) \to \Delta_{-1}(S) = 0 \end{aligned}$$

defined by

$$\begin{aligned} \partial_1(a) &= p - q, & \partial_0(p) &= 0, \\ \partial_1(b) &= q - p; & \partial_0(q) &= 0 \end{aligned}$$

which we can represent more compactly by

$$\begin{tikzcd} \cdots \arrow[r] & 0 \arrow[r] & \mathbb{Z}a \oplus \mathbb{Z}b \arrow[r, "\partial_1"] & \mathbb{Z}p \oplus \mathbb{Z}q \arrow[r, "\partial_0"] & 0 \end{tikzcd}$$

$$\begin{array}{c} \\ p \\ q \end{array} \begin{array}{c} \begin{array}{cc} a & b \end{array} \\ \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \end{array} \qquad \begin{array}{c} \\ * \end{array} \begin{array}{c} \begin{array}{cc} p & q \end{array} \\ \begin{bmatrix} 0 & 0 \end{bmatrix} \end{array}$$

with all other n-chains trivial. By inspection, we find that $\ker(\partial_1) = \mathbb{Z}(a+b)$ and $\operatorname{im}(\partial_1) = \mathbb{Z}(p-q)$, but we can do this more generally by examining the Smith normal form of the matrix associated with ∂_1:

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

Along the diagonal, we have a single 1, so $\operatorname{im}(\partial_1) \cong \mathbb{Z}$, and the remaining zero row gives $\ker(\partial_1) \cong \mathbb{Z}^{2-1} = \mathbb{Z}$. The zero matrix for ∂_0 also gives $\ker(\partial_0) = \mathbb{Z}p \oplus \mathbb{Z}q \cong \mathbb{Z}^2$.

$$\begin{aligned} H_0(T) &= \frac{\ker(\partial_0)}{\operatorname{im}(\partial_1)} & H_1(T) &= \frac{\ker(\partial_1)}{\operatorname{im}(\partial_2)} \\ &= \frac{\mathbb{Z}^2}{\mathbb{Z}} & &= \frac{\mathbb{Z}}{0} \\ &= \mathbb{Z} & &= \mathbb{Z} \end{aligned}$$

and all other homology groups 0. △

We just saw that these two Δ-sets S and T, which have the same geometric realisation, have the same homology groups. This is not a coincidence. Though it is not clear at all at this point, it turns out that homology is an invariant of the geometric realisation.

If X is a topological space with a Δ-complex structure $|S| \cong X$, we define its *nth simplicial homology group* to be

$$H_n^{\Delta}(X) \coloneqq H_n(S)$$

Note that, once we have the chain groups and boundary operators, computing the simplicial homology is an entirely mechanical process:

Algorithm 12 Simplicial Homology

1. Determine the matrix of each boundary operator

$$\partial_n\big([v_0,\dots,v_n]\big) = \sum_{i=0}^{n}(-1)^i[v_0,\dots,\widehat{v_i},\dots,v_n]$$

2. Determine the Smith normal form for each boundary operator.
3. For each pair of matrices

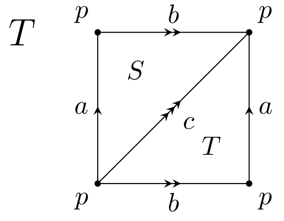
$$\mathbb{Z}^{\ell} \xrightarrow{\mathbf{A}} \mathbb{Z}^m \xrightarrow{\mathbf{B}} \mathbb{Z}^n$$

with $\mathbf{BA} = 0$, we have

$$\frac{\ker \mathbf{B}}{\operatorname{im} \mathbf{A}} \cong \left(\bigoplus_{i=1}^{r} \mathbb{Z}/d_i\right) \oplus \mathbb{Z}^{m-a-b}$$

where $\{d_i\}_{i=1}^{r}$ are the invariants of $\mathbf{A}$ (i.e. the diagonal elements), and $a = \operatorname{rank}(\mathbf{A})$ and $b = \operatorname{rank}(\mathbf{B})$ (i.e. the number of invariants of $\mathbf{A}$ and $\mathbf{B}$, respectively).

Example. We compute the simplicial homology of the torus



We have the chain of boundary operators

$$\begin{tikzcd}
\cdots \arrow[r] & 0 \arrow[r] & \mathbb{Z}S \oplus \mathbb{Z}T \arrow[r, "\partial_2"] & \mathbb{Z}a \oplus \mathbb{Z}b \oplus \mathbb{Z}c \arrow[r, "\partial_1"] & \mathbb{Z}p \arrow[r, "\partial_0"] & 0
\end{tikzcd}$$

$$\begin{array}{c|cc} & S & T \\ \hline a & -1 & 1 \\ b & -1 & 1 \\ c & 1 & -1 \end{array} \qquad\qquad \begin{array}{c|ccc} & a & b & c \\ \hline p & 0 & 0 & 0 \end{array}$$

where ∂_2 has Smith normal form

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

and ∂_1 already in Smith nornal form, so we have $\text{rank}(A) = 1$ and $\text{rank}(B) = 0$, giving,

$$\begin{aligned} H_1(T) &= \left(\bigoplus_{i=1}^{1} \mathbb{Z}/1\right) \oplus \mathbb{Z}^{3-1-0} \\ &= \mathbb{Z}/1 \oplus \mathbb{Z}^2 \\ &= \mathbb{Z}^2 \end{aligned}$$

For $H_0(T)$, ∂_1 has no invariants; ∂_1 has rank 0; and ∂_0 is the zero map which also has rank 0, so

$$\begin{aligned} H_0(T) &= \mathbb{Z}^{1-0-0} \\ &= \mathbb{Z} \end{aligned}$$

and for $H_2(T)$, ∂_3 is the zero map with no invariants, so

$$\begin{aligned} H_2(T) &= \mathbb{Z}^{1-0-0} \\ &= \mathbb{Z} \end{aligned}$$

All other boundary maps are trivial, so all other homology groups are 0, giving

$$H_n^{\Delta}(T) \cong \begin{cases} \mathbb{Z} \oplus \mathbb{Z} & n = 1 \\ \mathbb{Z} & n = 0,2 \\ 0 & n \geq 3 \end{cases}$$

△

39.3.3 Chain Complexes

So far, we have been computing homology groups of Δ-sets, first by mapping them to free abelian groups of n-chains with boundary operators between them, before finally computing the homology groups in terms of these operators.

$$S \quad \mapsto \quad \big(\Delta_\bullet(S), \partial_\bullet\big) \quad \mapsto \quad H_\bullet(S)$$

Algebraically, this middle step takes the form of sequences of abelian groups with homomorphisms between them,

$$\cdots \to \Delta_{n+1}(S) \xrightarrow{\partial_{n+1}} \Delta_n(S) \xrightarrow{\partial_n} \Delta_{n-1}(S) \to \cdots \to \Delta_1(S) \xrightarrow{\partial_1} \Delta_0(S) \xrightarrow{\partial_0} 0$$

It will be useful to discuss this structure independently from the specific situation here, as this process will recur repeatedly in the future in multiple different contexts.

A *chain complex* $C_\bullet = (C_\bullet, \partial_\bullet)$ is a family of abelian groups $(C_n)_{n\in\mathbb{Z}}$ equipped with maps called *differentials* $\partial_n : C_n \to C_{n-1}$,

$$\cdots \to C_{n+1} \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \to \cdots$$

such that $\partial_n \circ \partial_{n+1} = 0$ for each n.

In general, if we only define C_n for $n \in [a,b]$, then it is understood that $C_n = 0$ for all $n \notin [a,b]$, and we call such a chain complex *bounded* (or *bounded above/below*, if instead defined on a half-infinite interval), or say that the chain complex is *concentrated in degrees* $[a,b]$.

Example. Let S be a Δ-set. Then, the collection of chain groups in S and the boundary operators between them form a (bounded below) chain complex $\Delta_\bullet(S)$ called the *simplicial chain complex* associated with S. △

Let $C_\bullet$ be a chain complex.

- The *n-cycles* are
$$Z_n(C_\bullet) = \ker(\partial_n)$$
- The *n-boundaries* are
$$B_n(C_\bullet) = \operatorname{im}(\partial_{n+1})$$

and since $\partial_n \circ \partial_{n+1} = 0$, we have $B_n \subseteq Z_n$, so

- The *nth homology group* is
$$H_n(C_\bullet) \coloneqq \frac{Z_n}{B_n} = \frac{\ker(\partial_n)}{\operatorname{im}(\partial_{n+1})}$$
- If $H_n(C_\bullet) = 0$, or equivalently, if $\operatorname{im}(\partial_{n+1}) = \ker(\partial_n)$, then we say that $C_\bullet$ is *exact in degree n*. If $C_\bullet$ is exact in all degrees, then we say that $C_\bullet$ is *exact.*

Let $(C_\bullet,\partial_\bullet)$ and $(D_\bullet,\partial'_\bullet)$ be chain complexes. A *chain map* $f_\bullet : C_\bullet \to D_\bullet$ is a family of maps $f_n : C_n \to D_n$, such that

$$\begin{tikzcd}
\cdots \arrow[r] & C_{n+1} \arrow[r, "\partial_{n+1}"] \arrow[d, "f_{n+1}"] & C_n \arrow[r, "\partial_n"] \arrow[d, "f_n"] & C_{n-1} \arrow[r] \arrow[d, "f_{n-1}"] & \cdots \\
\cdots \arrow[r] & D_{n+1} \arrow[r, "\partial'_{n+1}"] & D_n \arrow[r, "\partial'_n"] & D_{n-1} \arrow[r] & \cdots
\end{tikzcd}$$

commutes for all n. That is,
$$\partial'_n \circ f_n = f_{n-1} \circ \partial_n$$

Lemma 39.3.2. *A chain map $f_\bullet : C_\bullet \to C'_\bullet$ restricts to maps*

- *$f_n : Z_n(C_\bullet) \to Z_n(C'_\bullet)$;*
- *$f_n : B_n(C_\bullet) \to B_n(C'_\bullet)$,*

and hence induces maps $f_n : H_n(C_\bullet) \to H_n(C'_\bullet)$.

Proof. If $\partial_n(\alpha) = 0$, then $\partial'_n\big(f_n(\alpha)\big) = f_{n-1}\big(\partial_n(\alpha)\big) = 0$, so f_n sends cycles to cycles. Also, f_n sends boundaries to boundaries as $f_{n-1}\big(\partial_n(\beta)\big) = \partial'_n\big(f_n(\beta)\big)$. Hence, f_n induces homomorphisms in homology. ■

Let $S = (S_\bullet,d_\bullet)$ and $T = (T_\bullet,d'_\bullet)$ be two Δ-sets. A *map of Δ-sets* $f_\bullet : S \to T$ is a family of maps $f_n : S_n \to T_n$ such that every square in

$$\begin{tikzcd}
\cdots \arrow[r, shift left=2] \arrow[r, shift left=1] \arrow[r] \arrow[r, shift right=1] \arrow[r, shift right=2] & C_{n+1} \arrow[r, shift left=1.5, "d_\bullet^{n+1}"] \arrow[r, shift left=0.5] \arrow[r, shift right=0.5] \arrow[r, shift right=1.5] \arrow[d, "f_{n+1}"] & C_n \arrow[r, shift left=1, "d_\bullet^{n}"] \arrow[r] \arrow[r, shift right=1] \arrow[d, "f_n"] & C_{n-1} \arrow[r, shift left=0.5] \arrow[r, shift right=0.5] \arrow[d, "f_{n-1}"] & \cdots \\
\cdots \arrow[r, shift left=2] \arrow[r, shift left=1] \arrow[r] \arrow[r, shift right=1] \arrow[r, shift right=2] & T_{n+1} \arrow[r, shift left=1.5, "d'^{n+1}_\bullet"] \arrow[r, shift left=0.5] \arrow[r, shift right=0.5] \arrow[r, shift right=1.5] & T_n \arrow[r, shift left=1, "d'^{n}_\bullet"] \arrow[r] \arrow[r, shift right=1] & T_{n-1} \arrow[r, shift left=0.5] \arrow[r, shift right=0.5] & \cdots
\end{tikzcd}$$

commutes. That is, for all $0 \leq i \leq n$,

$$f_n \circ d_i^n = d_i^{n+1} \circ f_{n+1}$$

whenever both sides of the equation are defined.

Lemma 39.3.3. *A map of Δ-sets $f_\bullet : S \to T$ induces a chain map $f_\bullet : \Delta_\bullet(S) \to \Delta_\bullet(T)$.*

Proof. Define $f_n : \Delta_n(S) \to \Delta_n(T)$ on generators $s \in S_n$ by

$$\begin{aligned} \mathbb{Z}S_n &\longrightarrow \mathbb{Z}T_n \\ s &\longmapsto f_n(s) \end{aligned}$$

Then,

$$\begin{aligned} \partial'_n \circ f_n(s) &= \partial'_n\big(f_n(s)\big) \\ &= \sum_{i=0}^{n}(-1)^i d_i'^n\big(f_n(s)\big) \\ &= \sum_{i=0}^{n}(-1)^i f_{n-1}\big(d_i(s)\big) \\ &= f_{n-1}\left(\sum_{i=0}^{n}(-1)^i d_i(s)\right) \\ &= f_{n-1} \circ \partial_n(s) \end{aligned}$$

■

Combining the previous two results, we have,

Corollary 39.3.3.1. *Every map of Δ-sets induces a map in simplicial homology.*

39.4 Singular Homology

In the previous section, we defined the simplicial homology for Δ-complexes. That is, spaces equipped with homeomorphisms to a Δ-set. There are two main problems with this result. Firstly, topological spaces often do not have an obvious Δ-complex structure – and some topological spaces admit no such structure at all. Secondly, even if a given space admits a Δ-complex structure, it may not be unique, and we haven't yet proven that simplicial homology is independent of choice of Δ-complex structure.

We now present an alternative theory of homology that avoids these difficulties, and will eventually allow us to prove the independence mentioned above.

Let X be a topological space and let $n \geq 0$. A *singular n-simplex* in X is a continuous map $\sigma : \Delta^n \to X$.

Example.

1. A singular 0-simplex is a function $\sigma : \Delta^0 \cong \{*\} \to X$. Such a function just picks out a point $x \in X$ and we sometimes identify the two.
2. A singular 1-simplex is a function $\sigma : \Delta^1 \cong [0,1] \to X$, which is just a path in X from $\sigma(0)$ to $\sigma(1)$.
3. If X is a Δ-complex, then any n-simplex in X can be viewed as the image of a simplex Δ^n under a function into X, i.e. a singular n-simplex in X.

△

Now, recall the definition of the oriented boundary of a simplex in a Δ-complex:

$$\partial_n\big([v_0,\ldots,v_n]\big) = \sum_{i=0}^{n}(-1)^i[v_0,\ldots,\widehat{v_i},\ldots,v_n]$$

Because a singular simplex $\sigma : \Delta^n = [v_0, \ldots, v_n] \to X$ is a map, the ith "face" of this simplex is just the restriction of σ to the ith face of the standard simplex. So, the oriented boundary of $\sigma : \Delta^n = [v_0, \ldots, v_n] \to X$ is

$$\partial_n(\sigma) = \sum_{i=0}^{n} (-1)^i \sigma\big|_{\partial_i \Delta^n}$$

where ∂_i is the boundary operator for standard simplices found earlier. That is,

$$= \sum_{i=0}^{n} (-1)^i \sigma\big|_{[v_0, \ldots, \widehat{v_i}, \ldots, v_n]}$$

Note that these faces that σ is restricted to are themselves homeomorphic to the standard $(n-1)$-simplex, so we can view each restriction as a singular $(n-1)$-simplex themselves, and thus this expression becomes a formal linear combination of singular $(n-1)$-simplices in X, i.e., an element of C_{n-1}.

Example. Let $\sigma : \Delta^2 = [v_0, v_1, v_2] \to X$ be a singular simplex. Then, the boundary of σ is

$$\partial(\sigma) = \sigma\big|_{[v_1, v_2]} - \sigma\big|_{[v_0, v_2]} + \sigma\big|_{[v_0, v_1]}$$

△

Let X be a topological space and $n \geq 0$.

- The group of *singular n-chains* in X is the free abelian group on the singular n-simplices, denoted by $C_n(X) \coloneqq \mathbb{Z} \cdot \{\sigma : \Delta^n \to X\}$.
- The *boundary operator* $\partial : C_n(X) \to C_{n-1}(X)$ is the homomorphism given on the generators $\sigma \in C_n(X)$ by the alternating sum of faces as above.

Note that the groups $C_n(X)$ are usually infinite, and frequently uncountable, as there are many ways to map a standard simplex into a space.

The same proof as for Δ-sets then translates across to singular simplices:

Lemma 39.4.1. *Let X be a topological space. Then, $\partial_n \circ \partial_{n+1} : C_{n+1}(X) \to C_{n-1}(X)$ is the zero map.*

That is, $(C_\bullet(X), \partial_\bullet)$ forms a chain complex called the *singular chain complex* associated with X. The *singular homology groups* of X are then the homology groups of this chain complex, i.e.,

$$H_n(X) \coloneqq H_n\big(C_\bullet(X)\big)$$

Example. Let $X = \{*\}$ be a point. For each $n \geq 0$, there is a unique singular n-simplex given by the constant map $c_n : \Delta^n \to X$ at the unique point of X, so the singular chain complex is

$$\cdots \to \mathbb{Z} \xrightarrow{\partial_n} \mathbb{Z} \to \cdots \to \mathbb{Z} \xrightarrow{\partial_2} \mathbb{Z} \xrightarrow{\partial_1} \mathbb{Z} \xrightarrow{\partial_0} 0$$

Now,

$$\begin{aligned} \partial_n(c_n) &= \sum_{i=0}^{n} (-1)^i c_n|_{\partial_i \Delta^n} \\ &= \sum_{i=0}^{n} (-1)^n c_{n-1} \\ &= \begin{cases} c_{n-1} & n \geq 0 \text{ even} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

and hence the differentials are:

$$\cdots \to \mathbb{Z} \xrightarrow{\cong} \mathbb{Z} \xrightarrow{0} \mathbb{Z} \xrightarrow{\cong} \mathbb{Z} \xrightarrow{0} \mathbb{Z} \xrightarrow{0} 0$$

At degree zero, we have

$$\mathbb{Z} \xrightarrow{0} \mathbb{Z} \xrightarrow{0} 0$$

so $H_0(X) = \frac{\ker(0)}{\operatorname{im}(0)} = \frac{\mathbb{Z}}{0} = \mathbb{Z}$. At all other even degrees, we have

$$\mathbb{Z} \xrightarrow{\cong} \mathbb{Z} \xrightarrow{0} \mathbb{Z}$$

so $H_n(X) = \frac{\ker(0)}{\operatorname{im}(\cong)} = \frac{\mathbb{Z}}{\mathbb{Z}} = 0$, and at odd degrees, we have

$$\mathbb{Z} \xrightarrow{0} \mathbb{Z} \xrightarrow{\cong} \mathbb{Z}$$

so $H_n(X) = \frac{\ker(\cong)}{\operatorname{im}(0)} = \frac{0}{0} = 0$. So,

$$H_n\big(\{*\}\big) = \begin{cases} \mathbb{Z} & n = 0 \\ 0 & n \neq 0 \end{cases}$$

△

We see that, even for the most trivial of topological spaces, the singular chain complex is decidedly non-trivial. For larger spaces, the number of singular n-simplices quickly becomes unmanageable, and direct computation is rarely a feasible strategy.

Eventually, we will develop some theory that will enable the computation of singular homology groups without having to work with the singular chain complex directly, but for now, we give some interpretations of H_0 and H_1 in topological spaces.

39.4.1 Reduced Homology

It is often convenient for us to have a version of homology for which the one-point space has trivial homology groups in every dimension.

Let $\pi : X \to *$ be the unique morphism to the point. The *reduced homology* of X is defined as

$$\tilde{H}_n(X) := \ker\big(H_n(X) \xrightarrow{\pi_*} H_n(*)\big)$$

Equivalently, reduced homology can also be characterised as the homology of the augmented chain complex

$$\cdots \to C_2(X) \xrightarrow{\partial_2} C_1(X) \xrightarrow{\partial_1} C_0(X) \xrightarrow{\varepsilon} \mathbb{Z} \to 0$$

where $\varepsilon\left(\sum_i n_i\sigma_i\right) = \sum_i n_i$. Note that by convention, we only consider reduced homology of non-empty spaces X, or else various pathologies can arise.

Since $\varepsilon \circ \partial_1 = 0$, ε vanishes on $\operatorname{im}(\partial_1)$ and hence induces a map $H_0(X) \to \mathbb{Z}$ with kernel $\tilde{H}_0(X)$, so $H_0(X) \cong \tilde{H}_0(X) \oplus \mathbb{Z}$. Also note that, by construction, $H_n(X) \cong \tilde{H}_n(X)$ for $n > 0$.

39.4.2 Low-Degree Interpretation

Two n-chains x and y are *homologous* if they are in the same equivalence class or *homology class* in $H_n(X) = \frac{Z_n(X)}{B_n(X)}$, that is, if they differ by a boundary (i.e. an element of $B_n(X) = \operatorname{im}(\partial_{n+1})$) – or equivalently, if their formal difference $x - y$ or $y - x$ is itself a boundary – and we write $x \sim y$ to denote this relation.

Example. Let x and y be singular 0-simplices (points) in a space X, and suppose they lie in the same path-connected component. Let $\gamma : \Delta^1 \cong [0{,}1] \to X$ be a singular 1-simplex with $\gamma(0) = x$ and $\gamma(1) = y$, i.e., a path from x to y. Then,

$$\partial_1(\gamma) = y - x$$

and hence x and y are homologous, as they differ by the boundary of γ. △

Lemma 39.4.2. *Let X be a non-empty and path-connected* topological space. Then, $H_0(X) = \mathbb{Z}$.*

Proof. By definition,

$$H_0(X) := \frac{Z_0\big(C_\bullet(X)\big)}{B_0\big(C_\bullet(X)\big)} = \frac{\ker(\partial_0)}{\operatorname{im}(\partial_1)}$$

The differential $\partial_0 : C_1(X) \to 0$ maps into the trivial group, so the kernel $\ker(\partial_0) = C_0(X)$ is the entire group. The idea is now to define a homomorphism from $C_0(X)$ to $\mathbb{Z}$ with kernel $B_0\big(C_\bullet(X)\big) = \operatorname{im}(\partial_1)$, which, combined with the first isomorphism theorem, the result will follow.

Define the *degree homomorphism* $\deg : C_0(X) \to \mathbb{Z}$ by sending every basis element $x \in X$ (i.e. singular 0-simplex) to $1 \in \mathbb{Z}$.

Because X is non-empty, there exists at least one basis element $x \in X$ which maps to the generator $1 \in \mathbb{Z}$, so deg is surjective.

We also have $B_0(X) \subseteq \ker(\deg)$, since the boundary $\partial_1(\gamma) \in B_0(X)$ of any singular 1-simplex has degree

$$\deg\big(\partial_1(\gamma)\big) = \deg\big(\gamma(1) - \gamma(2)\big) = 1 - 1 = 0$$

so $\partial_1(\gamma) \in \ker(\deg)$. The reverse containment $\ker(\deg) \subseteq B_0(X)$ also holds:

Let $L = \sum_{x\in X} \lambda_x \cdot x \in \ker(\deg)$ be a 0-chain whose degree vanishes. Then,

$$\begin{aligned} L &= \sum_{x\in X} \lambda_x \cdot x \\ &= \sum_{\substack{y\in X\\ \lambda_y>0}} \lambda_y \cdot y + \sum_{\substack{z\in Z\\ \lambda_z<0}} \lambda_z \cdot z \\ &= \sum_{\substack{y\in Y\\ \lambda_y>0}} \lambda_y \cdot y - \sum_{\substack{z\in Z\\ \lambda_z<0}} (-\lambda_z) \cdot z \end{aligned}$$

Since $\deg(L) = 0$, these two sums are equal, so we can pair up terms from each sum and write

$$L = \sum_i (y_i - z_i)$$

for some (possibly repeated) points $y_i{,}z_i \in X$. Since X is path-connected, there is a path γ_i from y_i to z_i for all i, so $y_i - z_i = \partial_1(\gamma_i) \in B_0(X)$, and hence $L \in B_0(X)$.

So, $\ker(\deg) = B_0(X)$. The first isomorphism theorem then gives

$$H_0(X) := \frac{\ker(\partial_0)}{\operatorname{im}(\partial_1)} = \frac{C_0(X)}{B_0(X)} = \frac{C_0(X)}{\ker(\deg)} \cong \operatorname{im}(\deg) = \mathbb{Z}$$

as required. ■

To interpret the 0th singular homology group for general spaces, we need the following intuitive fact:

* We will assume that path-connected spaces are non-empty, and will not mention it from this point onwards.

Theorem 39.4.3. *Let X be a topological space and $(X_\alpha)_{\alpha\in\Lambda}$ its path-connected components. Then,*

$$H_n(X) = \bigoplus_{\alpha\in\Lambda} H_n(X_\alpha)$$

Proof. If $\sigma : \Delta^n \to X$ is a singular n-simplex, then its image is path-connected and thus lies entirely in one of the X_α. That is, we have $C_n(X) \cong \bigoplus_{\alpha\in\Lambda} C_n(X_\alpha)$. Moreover, the oriented boundary of σ is a linear combination of $(n-1)$-simplices, all of which also lie in X_α, so ∂_n is the sum of the boundary operators for each X_α. That is, $C_\bullet(X) = \bigoplus_{\alpha\in\Lambda} C_\bullet(X_\alpha)$ as chain complexes. This decomposition therefore passes to cycles and boundaries, and eventually to homology. ■

Corollary 39.4.3.1. *Let X be a topological space. Then, $H_0(X) = \mathbb{Z}\pi_0(X)$. That is, $H_0(X)$ is the free abelian group with generators the path-connected components of X.*

Proof. By definition, $\pi_0(X)$ is the collection of path-connected components of X. Applying the previous two results then yields the desired result. ■

There is a similarly close relation between the first homology group $H_1(X)$ and the first homotopy group $\pi_1(X,x)$, as any path $f : [0,1] \to X$ can also be interpreted as a singular 1-simplex. In particular, if f is a loop a a path, then it is also a cycle as a singular 1-simplex, since $\partial_1(f) = f(1) - f(0) = 0$.

Lemma 39.4.4. *For any topological space X and all loops $f,g \in \pi_1(X,x)$,*

(i) If f is a constant path, then $f \sim 0$. That is, f is a boundary.

(ii) If $f \simeq g$, then $f \sim g$.

(iii) $f \cdot g \sim f + g$ where $\cdot$ on the left is path concatenation.

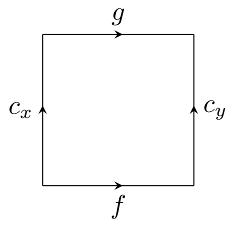
(iv) $f^{-1} \sim -f$ where f^{-1} is the reverse path of f.

Proof.

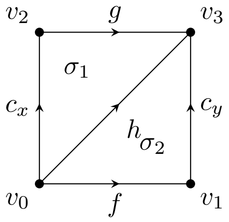
(*i*) If f has constant value x, then the singular 2-simplex $\sigma : \Delta^2 = [v_0,v_1,v_2] \to X$ with constant value x has boundary

$$\begin{aligned}\partial_2(\sigma) &= \sigma\big|_{[v_1,v_2]} - \sigma\big|_{[v_0,v_2]} + \sigma\big|_{[v_0,v_1]} \\ &= f - f + f \\ &= f\end{aligned}$$

(*ii*) Let $H : [0,1]^2 \to X$ be the homotopy between f and g relative to their endpoints:



where c_x and c_y are the constant maps at x and y, respectively. By subdividing the square into two triangles $[v_0,v_1,v_3]$ and $[v_0,v_2,v_3]$, we obtain a pair of singular 2-simplices in X:

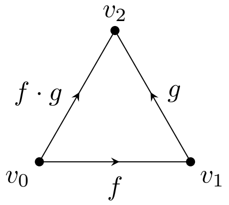


with $h(t) = H(t,t)$ the diagonal. Then,

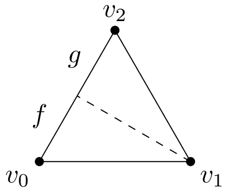
$$\begin{aligned}\partial_2(\sigma_2 - \sigma_1) &= \partial_2(\sigma_2) - \partial_2(\sigma_1)\\ &= (\gamma_2 - \gamma + c_x) - (c_y - \gamma + \gamma_1)\\ &= (\gamma_2 - \gamma_1) + (c_x - c_y)\end{aligned}$$

By property (i), the constant maps c_x and c_y are boundaries, so $f \sim g$.

(iii) Let f and g be two loops based at x, and consider the following 2-simplex:



Now, define $\sigma : \Delta^2 \to X$ to be the orthogonal projection of $\Delta^2 = [v_0,v_1,v_2]$ onto the edge $[v_0,v_2]$ composed with by $f \cdot g : [v_0,v_2] \to X$:



Then, the three faces of σ are $\sigma\big|_{[v_0,v_1]} = f$, $\sigma\big|_{[v_1,v_2]} = g$, and $\sigma\big|_{[v_0,v_2]} = f \cdot g$, so σ has boundary

$$\partial(\sigma) = g - f \cdot g + f$$

and we have $f \cdot g \sim f + g$.

(iv) Using the previous properties, $f + \overline{f} \sim f \cdot \overline{f} \simeq c_x \sim 0$.

■

Together, these properties imply that the map

$$h_1 : \pi_1(X,x) \to H_1(X)$$

defined by $h_1\big([\gamma]\big) = [\gamma]$ – where the brackets on the left mean homotopy class, and those on the right mean homology class – is a group homomorpism.

Property (i) shows that identities are mapped to identities, (ii) shows that this map is well defined, and (iii) shows that $h_1\big([f] \cdot [g]\big) = [f] + [g]$. Note, however, that this homomorphism is generally not an isomorphism, as $H_1(X)$ is abelian, while $\pi_1(X,x)$ is generally not.

For a group G, the *commutator subgroup* $[G,G]$ is the normal subgroup generated by the elements $ghg^{-1}h^{-1}$ for $g,h \in G$. The *abelianisation* of G, denoted G^{ab} is then the quotient $G/[G,G]$.

Example. If G is abelian, then $ghg^{-1}h^{-1} = \mathrm{id}_G$ for all $g,h \in G$, so the commutator subgroup is trivial and hence $G^{\mathrm{ab}} = G$. △

Lemma 39.4.5. *The abelianisation of a free product is the direct sum of the abelianisations. That is,*

$$(G * H)^{\mathrm{ab}} \cong G^{\mathrm{ab}} \oplus H^{\mathrm{ab}}$$

Example. If $G = \mathbb{Z} * \mathbb{Z}$, then $G^{\mathrm{ab}} = \mathbb{Z} \oplus \mathbb{Z}$. △

If $G = \langle S \mid R \rangle$ is a presentation, then the abelianisation is given by adjoining the commutator $[x,y]$ to the relations, for all generators $x,y \in S$.

Example. If G is given by

$$G = \langle x,y \mid x^3 = y^5 \rangle$$

then the abelianisation has presentation:

$$G^{\mathrm{ab}} = \langle x,y \mid 3x = 5y, x + y = y + x \rangle$$

Then, because x and y now commute, every element of G may be expressed in the form $ax + by$, and is equal to the identity precisely when $a = 3k$ and $b = -5k$ for some $k \in \mathbb{Z}$, so $G^{\mathrm{ab}} \cong \mathbb{Z}^2/\mathbb{Z}(3,-5) \cong \mathbb{Z}$. △

By construction, G^{ab} is abelian, and is in fact universal with respect to this property. That is, if $\phi : G \to A$ is a morphism to an abelian group A, then there exists a unique morphism $\bar{\phi} : G^{\mathrm{ab}} \to A$ such that the following diagram commutes:

$$\begin{tikzcd}
G \arrow[r, "\phi"] \arrow[d, two heads, "\iota"'] & A \\
G^{\mathrm{ab}} \arrow[ur, "\bar{\phi}"'] &
\end{tikzcd}$$

where $\iota : G \to G^{\mathrm{ab}}$ is the quotient map.

Lemma 39.4.6. *For every group homomorphism $\phi : G \to A$, $[G,G] \subseteq \ker(\phi)$.*

This gives another strategy for finding abelianisations of groups:

Example. Consider the symmetric group, S_n. The sign function $\mathrm{sgn} : S_n \to \{-1,1\} \cong \mathbb{Z}/2$ defined by

$$\sigma \mapsto \begin{cases} +1 & \sigma \text{ is even} \\ -1 & \sigma \text{ is odd} \end{cases}$$

Because $\mathbb{Z}/2$ is abelian, the commutator subgroup $[S_n,S_n]$ is contained in the kernel $\ker(\mathrm{sgn}) = A_n$. We also have $A_n \subseteq [S_n,S_n]$, since any two transpositions are conjugate in S_n, since $\sigma(i,j)\sigma^{-1} = \big(\sigma(i),\sigma(j)\big)$.

So, all transpositions are sent to the same element in $(S_n)^{\mathrm{ab}}$. Because S_n is generated by transpositions, all non-identity elements are identified in the abelianisation, so $(S_n)^{\mathrm{ab}} \cong \mathbb{Z}/2$. △

This universal property also implies that the map $h_1 : \pi_1(X,x) \to H_1(X)$ sending homotopy classes to homology classes factors uniquely through a morphism

$$\bar{h}_1 : \pi_1(X,x)^{\mathrm{ab}} \to H_1(X)$$

Theorem 39.4.7. *For any path-connected space X, $H_1(X) \cong \pi_1(X,x)^{\mathrm{ab}}$. More specifically, the isomorphism is given by the induced map $\bar{h}_1$.*

Proof. Since X is path-connected, we may choose for every point $y \in X$ some fixed path η_y from x to y. To any path $\gamma : \Delta^1 \to X$, we associate the following loop based at x:

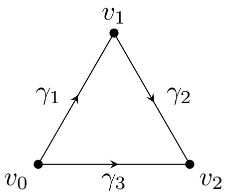
$$\gamma \mapsto \eta_{y(0)} \cdot \gamma \cdot \eta_{\gamma(1)}^{-1}$$

i.e. the loop that travels along the previously fixed path η from x to the starting point of γ, along the path γ, then along the fixed η path in reverse direction back to x.

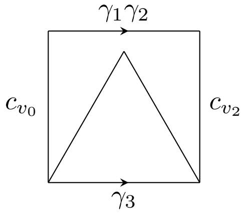
This association linearly extends from these generators to a group homomorphism $C_1(X) \to \pi_1(X,x)^{\mathrm{ab}}$ which we can restrict to $Z_1(X)$,

$$g : Z_1(X) \to \pi_1(X,x)^{\mathrm{ab}}$$

Now, consider a singular 2-simplex $\sigma : \Delta^2 \to X$:



Then, $\gamma_1 \cdot \gamma_2$ is homotopic to γ_3 relative to their endpoints. To see this, we can embed this 2-simplex into the square and project vertically onto the 2-simplex $[v_0,v_2]$:



It follows that in $\pi_1(X,x)^{\mathrm{ab}}$ we have

$$\begin{aligned} g\big(\partial_2(\sigma)\big) &= g(\gamma_1 + \gamma_2 - \gamma_3) \\ &= g(\gamma_1) + g(\gamma_2) - g(\gamma_3) \\ &= [\eta_{v_0} \cdot \gamma_1 \cdot \eta_{v_1}^{-1}] + [\eta_{v_1} \cdot \gamma_2 \cdot \eta_{v_2}^{-1}] - [\eta_{v_0} \cdot \gamma_3 \cdot \eta_{v_2}^{-1}] \\ &= [\eta_{v_0} \cdot \gamma_1 \cdot \eta_{v_1}^{-1}] + [\eta_{v_1} \cdot \gamma_2 \cdot \eta_{v_2}^{-1}] + [\eta_{v_2} \cdot \gamma_3^{-1} \cdot \eta_{v_0}^{-1}] \\ &= [\eta_{v_0} \cdot \gamma_1 \cdot \eta_{v_1}^{-1} \cdot \eta_{v_1} \cdot \gamma_2 \cdot \eta_{v_2}^{-1} \cdot \eta_{v_2} \cdot \gamma_3^{-1} \cdot \eta_{v_0}^{-1}] \\ &= [\eta_{v_0} \cdot \gamma_1 \cdot \gamma_2 \cdot \gamma_3^{-1} \cdot \eta_{v_0}^{-1}] \\ &= [\eta_{v_0}\eta_{v_0}^{-1}] \\ &= 0 \end{aligned}$$

so g passes to $\bar{g} : H_1(X) \to \pi_1(X,x)^{\mathrm{ab}}$.

Let $\gamma \in \pi_1(X,x)^{\mathrm{ab}}$ be a loop based at x. Then,

$$\begin{aligned} \bar{g} \circ \bar{h}\big([\gamma]\big) &= \bar{g}\big([\gamma]\big) \\ &= [\eta_x \cdot \gamma \cdot \eta_x^{-1}] \end{aligned}$$

$$\begin{aligned} &= [\eta_x] + [\gamma] - [\eta_x] \\ &= [\gamma] \end{aligned}$$

so $\bar{g} \circ \bar{h}_1 = \text{id}$.

Now, let $L = \sum_i \lambda_i \gamma_i$, $\lambda_i \in \mathbb{Z}$ be a 1-cycle in $Z_1(X)$. With relabelling and allowing repetition of paths in the sum, we may assume each λ_i is ± 1 (e.g. $2\gamma_1 - 3\gamma_2 = \gamma_1 + \gamma_1 - \gamma_2 - \gamma_2 - \gamma_2$).

Even further, by property (iv), we can eliminate all the negative coefficients by replacing any γ_i with γ_i^{-1} if necessary, and thus take $\lambda_i = 1$ for all i, so $L = \sum_i \gamma_i$.

Because $\partial(L) = 0$, if some γ_i is not a loop, then there must be some γ_j such that the chain $\gamma_i \cdot \gamma_j$ is valid as a path, i.e., $\gamma_i(1) = \gamma_j(0)$. By (iii), we may replace $\gamma_i + \gamma_j$ in the sum by $\gamma_i \cdot \gamma_j$. Repeating this relabelling, we can reduce L to a single loop γ in X, based at say, y. Then,

$$\begin{aligned} \bar{h}_1 \circ \bar{g}\big([\gamma]\big) &= [\eta_y \cdot \gamma \cdot \eta_y^{-1}] \\ &= [\eta_y] + [\gamma] - [\eta_y] \\ &= [\gamma] \end{aligned}$$

so $\bar{h} \circ \bar{g} = \text{id}$, and hence the maps constitute an isomorphism $\pi_1(X,x)^{\text{ab}} \cong H_1(X)$. ■

Corollary 39.4.7.1. *If X is simply connected (and hence path-connected and non-empty), then $H_1(X) = 0$.*

Intuitively, all loops in a simply connected space can contract to a point, so the space has no one-dimensional holes, and hence the first homology vanishes.

Corollary 39.4.7.2. *$H_1(S^1) = \mathbb{Z}$, with a generator given by the homology class of the obvious surjective map $\gamma_1 : \Delta^1 \to S^1$ identifying the end points.*

39.5 Fundamental Theorems

So far, we have only examined singular homology in degrees 0 and 1. For instance, we still haven't computed the higher homology groups of even basic spaces, such as the circle or higher n-spheres. As noted earlier, the singular chain complex is much too large to admit any manual computation, so here, we prove two fundamental theorems that allow us to compute the singular homology of topological spaces without directly using the singular chain complex.

39.5.1 Homotopy Invariance

Given a continuous map $f : X \to Y$, we can transform a singular n-simplex in X into a singular n-simplex in Y by postcomposing the singular n-simplex $\sigma : \Delta^n \to X$ by f to obtain the composition $f \circ \sigma : \Delta^n \to Y$. We can extend this to a group homomorphism $f_\sharp : C_n(X) \to C_n(Y)$ by linearly extending

$$\begin{aligned} f_\sharp\left(\sum_i n_i \sigma_i\right) &= \sum_i n_i f_\sharp(\sigma_i) \\ &= \sum_i n_i (f \circ \sigma_i) \end{aligned}$$

How do these maps act on boundaries? Expanding the definition, we have,

$$f_\sharp\big(\partial(\sigma)\big) = f_\sharp\left(\sum_i (-1)^i \sigma\big|_{[v_0,\ldots,\widehat{v_i},\ldots v_n]}\right)$$

$$\begin{aligned}
&= \sum_i (-1)^i f_\sharp \left(\sigma\big|_{[v_0,\ldots,\widehat{v_i},\ldots v_n]}\right) \\
&= \sum_i (-1)^i f \circ \sigma\big|_{[v_0,\ldots,\widehat{v_i},\ldots v_n]} \\
&= \partial(f \circ \sigma) \\
&= \partial\big(f_\sharp(\sigma)\big)
\end{aligned}$$

so, the following diagram commutes:

$$\begin{tikzcd}
\cdots \arrow[r] & C_{n+1}(X) \arrow[r, "\partial"] \arrow[d, "f_\sharp"] & C_n(X) \arrow[r, "\partial"] \arrow[d, "f_\sharp"] & C_{n-1}(X) \arrow[r] \arrow[d, "f_\sharp"] & \cdots \\
\cdots \arrow[r] & D_{n+1}(X) \arrow[r, "\partial"] & D_n(X) \arrow[r, "\partial"] & D_{n-1}(X) \arrow[r] & \cdots
\end{tikzcd}$$

That is, the $f_\sharp$ assemble into a chain map $f_\bullet : C_\bullet(X) \to C_\bullet(Y)$, which then induces a map in homology (Lemma 39.3.2).

Lemma 39.5.1. *Let $f : X \to Y$ be a continuous map. Then, there are induced maps in homology*

$$f_* : H_n(X) \to H_n(Y)$$

satisfying:

(i) $(f \circ g)_* = f_* \circ g_*$*;*

(ii) $(\mathrm{id}_X)_* = \mathrm{id}_{H_n(X)}$.

*That is, $H_n(-) :$ **Top** $\to$ **Ab** is a functor.*

Proof. The construction is as above. Functoriality follows from the associativity of the composition $\Delta^n \xrightarrow{\sigma} X \xrightarrow{g} Y \xrightarrow{f} Z$ and the definition of an identity map. ■

Theorem 39.5.2 (Homotopy Invariance)**.** *Suppose $f,g : X \to Y$ are homotopic. Then,*

$$f_* = g_* : H_n(X) \to H_n(Y)$$

Corollary 39.5.2.1. *If $X \simeq Y$ are homotopy equivalent, then $H_n(X) \cong H_n(Y)$ are isomorphic.*

Proof. Let $f : X \to Y$ have homotopy inverse $g : Y \to X$. Then,

$$\begin{aligned}
f_* \circ g_* &= (f \circ g)_* & \qquad g_* \circ f_* &= (g \circ f)_* \\
&= (\mathrm{id}_Y)_* & &= (\mathrm{id}_X)_* \\
&= \mathrm{id}_{(H_n(X))} & &= \mathrm{id}_{H_n(X)}
\end{aligned}$$

so $f_* : H_n(X) \to H_n(Y)$ and $g_* : H_n(Y) \to H_n(X)$ are inverse. ■

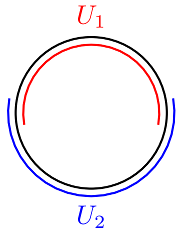
Corollary 39.5.2.2. *Let X be a contractible space. Then,*

$$H_n(X) = \begin{cases} \mathbb{Z} & n = 0 \\ 0 & n \neq 0 \end{cases}$$

Proof. We have previously computed the homology of a point, in this example. If X is contractible, then it is homotopy equivalent to the point, and hence has the same homology. ■

Example. We therefore know the homology of real Euclidean space $\mathbb{R}^n$, complex n-space $\mathbb{C}^n$, the unit ball D^n, the unit cube I^n, n-simplices Δ^n,etc. and every other contractible space. △

By itself, homotopy invariance is not very powerful. For instance, the sphere S^n is not contractible, nor is it homotopy equivalent to some other space whose homology we can compute (at least for $n \geq 1$). However, we can cover the sphere with (contractible) k-balls:



More generally, for any manifold X, we can find an open covering $\{U_i\}_i$ such that each U_i is homeomorphic to $\mathbb{R}^n$ for some n, and therefore contractible. If we can express the homology of X in terms of the homology of the U_i, then we could apply Homotopy Invariance. Our second fundamental theorem allows just that.

39.5.2 The Mayer–Vietoris Long Exact Sequence

We recall some material from homotopy theory:

Theorem (Seifert–van Kampen)**.** *Let $X = U_1 \cup U_2$ be the union of two path-connected open subspaces $j_1 : U \hookrightarrow X$, $j_2 : U_2 \hookrightarrow X$ such that $U_1 \cap U_2$ is path-connected. Then, we have a map*

$$\pi_1(U_1) * \pi_1(U_2) \xrightarrow{(j_1)_* * (j_2)_*} \pi_1(X)$$

where

1. *$(j_1)_* * (j_2)_*$ is surjective;*
2. *its kernel is the normal subgroup generated by elements of the form $i(\gamma) = (i_1)_*(\gamma)(i_2)_*(\gamma)^{-1}$, where $i_j : U_1 \cap U_2 \hookrightarrow U_j$ are the canonical inclusion maps.*

We have this setup:

$$\begin{tikzcd}
U_1 \cap U_2 \arrow[r, hook, "j_2"] \arrow[d, hook, "i_1"] & U_2 \arrow[d, hook, "i_2"] \\
U_1 \arrow[r, hook, "j_1"] & X
\end{tikzcd}$$

What do the induced maps in homology look like? Along the upper path, we have

$$(j_2)_* \circ (i_2)_* = (j_2 \circ i_2)_*$$

and along the lower path, we have

$$(j_1)_* \circ (i_1)_* = (j_1 \circ i_1)_*$$

by functoriality of homology. However, both maps are just the canonical inclusion of $U_1 \cap U_2$ into X, so these must be equal.

Now, passing to homology, we abelianise the groups above (Lemma 39.4.5 is useful here) to obtain:

$$H_1(U_1 \cap U_2) \xrightarrow{\left((i_1)_*, -(i_2)_*\right)} H_1(U_1) \oplus H_1(U_2) \xrightarrow{(j_1)_* + (j_2)_*} H_1(X)$$

where

1. $j \coloneqq (j_1)_* + (j_2)_*$ is surjective;
2. its kernel is precisely the image of $i \coloneqq \big((i_1)_*, -(i_2)_*\big)$.

To clarify how these maps act, i is a map into a product, so $i(\sigma) = \big((i_1)_*(\sigma), -(i_2)_*(\sigma)\big)$ is a pair, while $(j_1)_* + (j_2)_*$ is a map out of a product, and it acts on each component as $j(\sigma,\tau) = (j_1)_*(\sigma) + (j_2)_*(\tau)$.

The second point above is equivalent to saying that this chain complex is exact in the middle. Note that to facilitate this, we had to add a negative sign in the map i; without it, the composition would be twice the map induced by including $U_1 \cap U_2$ into X. Of course, this negative sign could be attached to any of the four maps above; it is just convention that we put it in the second component of the first map here.

If we extend the chain complex by a 0 to the right, then the first point says precisely that the chain complex is also exact at $H_1(X)$. However, as we will see, this does not typically hold when $U_1 \cap U_2$ is not path-connected.

In this, we require that U_1, U_2, and $U_1 \cap U_2$ are all path-connected, but this is rather restrictive: we still cannot apply this to the circle. However, it turns out that we can drop these assumptions in homology:

Theorem (Mayer–Vietoris Long Exact Sequence)**.** *Let $X = U_1 \cup U_2$ be the union of two open subspaces $j_1 : U \hookrightarrow X$, $j_2 : U_2 \hookrightarrow X$. Then, there are connecting homomorphisms $\partial : H_n(X) \to H_{n-1}(U_1 \cap U_n)$ such that*

$$\cdots \to H_{n+1}(X) \xrightarrow{\partial} H_n(U_1 \cap U_2) \xrightarrow{((i_1)_*, -(i_2)_*)} H_n(U_1) \oplus H_n(U_2) \xrightarrow{(j_1)_* + (j_2)_*} H_n(X) \xrightarrow{\partial} H_{n-1}(U_1 \cap U_2) \to \cdots$$

is an exact chain complex.

Recall that in a chain complex, the composition of any two maps in the sequence is the zero map. Equivalently, the image of each morphism is contained in the kernel of the next. Exactness means that the converse also holds; the image of each morphism is precisely the kernel of the next. Equivalently, its homology vanishes in each degree.

Note that, unlike for Seifert–van Kampen, the subspaces need not be open; the only requirement is that their interiors jointly cover X.

Example. Consider the circle S^1 with the same covering as previously.

The Mayer–Vietoris long exact sequence on either side of $H_n(X)$ is then

$$\cdots \to H_n(U_1) \oplus H_n(U_2) \to H_n(S^1) \xrightarrow{\partial} H_{n-1}(U_1 \cap U_2) \to \cdots$$

We will consider this situation in degrees $n \geq 2$ (so the lowest degree homology group possibly involved is $H_1(U_1 \cap U_2)$). U_1 and U_2 are contractible, so their homology is trivial, and similarly, their intersection consists of two contractible path-connected components, and homology splits across path-connected components, so the last term also vanishes, leaving:

$$\begin{tikzcd}\cdots \arrow[r] & H_1(X) \arrow[r, "\partial"] & H_0(U_1 \cap U_2) \arrow[r, "\big((i_1)_*{,} -(i_2)_*\big)"] & H_0(U_1) \oplus H_0(U_2) \arrow[r, "(j_1)_* + (j_2)_*"] & H_0(S^1) \arrow[r] & 0\end{tikzcd}$$

By exactness, $H_n(S^1) = \ker(\partial) = \operatorname{im}(f) = 0$, so the homology of S^1 vanishes in degrees $n \geq 2$. For degree $n = 0$, we note that S^1 is path-connected, so $H_0(S^1) = \mathbb{Z}$; and for degree $n = 1$, we have $H_1(S^1) = \pi_1(S^1)^{\text{ab}} = \mathbb{Z}$.

Overall, we have.

$$H_n(S^1) = \begin{cases} \mathbb{Z} & n = 0,1 \\ 0 & n \neq 0,1 \end{cases}$$

△

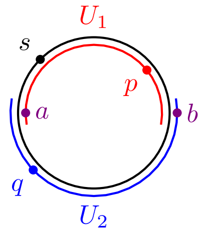
Example. We compute some of the maps at the end of the Mayer–Vietoris long exact sequence for S^1:

$$\begin{tikzcd}
\cdots \arrow[r] & H_1(S^1) \arrow[r, "\partial"] & H_0(U_1 \cap U_2) \arrow[r, "\big((i_1)_*, -(i_2)_*\big)"] \arrow[d, phantom, "\parallel"] & H_0(U_1) \oplus H_0(U_2) \arrow[r, "(j_1)_* + (j_2)_*"] \arrow[d, phantom, "\parallel"] & H_0(S^1) \arrow[r] \arrow[d, phantom, "\parallel"] & 0 \\
 & & \mathbb{Z} \oplus \mathbb{Z} & \mathbb{Z} \oplus \mathbb{Z} & \mathbb{Z} &
\end{tikzcd}$$

Let us label the generators more explicitly in the long exact sequence:

$$\cdots \to H_1(S^1) \xrightarrow{\partial} \mathbb{Z}a \oplus \mathbb{Z}b \xrightarrow{\big((i_1)_*, -(i_2)_*\big)} \mathbb{Z}p \oplus \mathbb{Z}q \xrightarrow{(j_1)_* + (j_2)_*} \mathbb{Z}s \xrightarrow{0} 0$$

Because the zeroth homology groups are free abelian on the set of path-connected components, a generator is just a choice of point in each component. So, a and b are points in the intersection $U_1 \cap U_2$, with one in each component; p and q are points in U_1 and U_2, respectively; and s is some point in S^1. For instance,



The induced map $(i_1)_*$ sends the generators a and b to the generator p of $\mathbb{Z}p$, and similarly, $(i_2)_*$ sends the generators a and b to the generator q of $\mathbb{Z}q$. However, the map given by Mayer–Vietoris is given by $\big((i_1)_*, -(i_2)_*\big)$, so whenever we map into the second component, $\mathbb{Z}q$, we have a negative in the matrix:

$$\begin{array}{c c} & \begin{array}{cc} a & b \end{array} \\ \begin{array}{c} p \\ q \end{array} & \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} \end{array}$$

The next map then sends both p and q to s, so the matrix is given by

$$\begin{array}{c c} & \begin{array}{cc} p & q \end{array} \\ s & \begin{bmatrix} 1 & 1 \end{bmatrix} \end{array}$$

(Again, the placement of the negative sign is arbitrary; we could have equally negated the first row of the first matrix, or either column of the second matrix, and the sequence would still be exact.) △

Corollary 39.5.2.3. *Let $k \geq 1$. Then,*

$$H_n(S^k) = \begin{cases} \mathbb{Z} & n = 0,k \\ 0 & n \neq 0,k \end{cases}$$

Proof. The sphere S^k can be written as the union of the upper and lower hemispheres (plus some extra space to overlap) U_1 and U_2 respectively. The Mayer–Vietoris long exact sequence is then

$$\cdots \to H_n(U_1) \oplus H_n(U_2) \to H_n(S^k) \xrightarrow{\partial} H_{n-1}(U_1 \cap U_2) \to H_{n-1}(U_1) \oplus H_{n-1}(U_2) \to \cdots$$

Each hemisphere is contractible, and the intersection $U_1 \cap U_2$ is homotopy equivalent to S^{k-1}, so in degrees $n \geq 2$, this reduces to

$$\cdots \to 0 \to H_n(S^k) \xrightarrow{\partial} H_{n-1}(S^k) \to 0 \to \cdots$$

so $H_n(S^k) \cong H_{n-1}(S^{k-1})$. We also have $H_0(S^k) \cong \mathbb{Z}$, since S^k is path-connected, and $H_1(S^k) = 0$ for $k \geq 2$, since S^k is simply connected. We have also already computed the homology for $k = 1$ in a previous example.

So, by induction on $k \geq 2$, we have

$$H_k(S^k) \underset{\cong}{\xrightarrow{\partial}} H_{k-1}(S^{k-1}) = \mathbb{Z}$$

and zero elsewhere. Along with the base cases above, this completes the proof.

■

Using reduced homology, the previous corollary becomes:

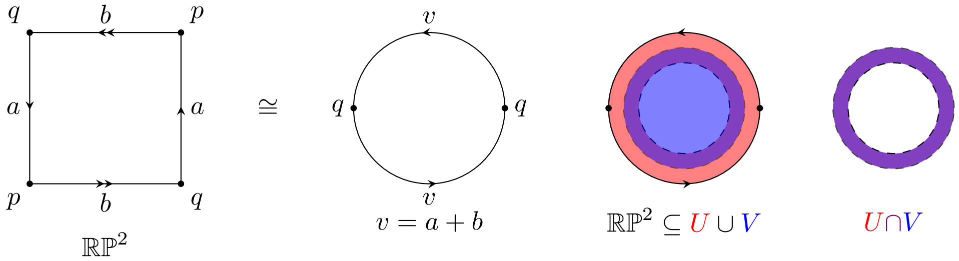
Corollary 39.5.2.4. *Let $k \geq 0$. Then,*

$$\tilde{H}_n(S^k) = \begin{cases} \mathbb{Z} & n = k \\ 0 & n \neq k \end{cases}$$

Theorem 39.5.3. *If U_1 and U_2 have non-empty intersection, then there is a sequence in reduced homology that agrees with the ordinary Mayer–Vietoris sequence in positive degrees, and ends as:*

$$\cdots \xrightarrow{\partial} \tilde{H}_0(U_1 \cap U_2) \xrightarrow{\left((i_1)_*, -(i_2)_*\right)} \tilde{H}_0(U_1) \oplus \tilde{H}_0(U_2) \xrightarrow{(j_1)_* + (j_2)_*} \tilde{H}_0(X) \xrightarrow{0} 0$$

Example. We compute the simplicial homology of $\mathbb{RP}^2$:



V is contractible, so,

$$H_n(V) = \begin{cases} \mathbb{Z} & n = 0 \\ 0 & n \geq 1 \end{cases}$$

U is homeomorphic to a Möbius band, which deformation retracts to S^1, and is thus homotopy equivalent to S^1. $U \cap V$ deformation retracts to, and is therefore homotopy equivalent to, S^1 also. So,

$$H_n(U) = H_n(U \cap V) = H_n(S^1) = \begin{cases} \mathbb{Z} & n = 0,1 \\ 0 & n \geq 2 \end{cases}$$

Because $\mathbb{RP}^2$ is non-empty and path-connected, we have $H_0(\mathbb{RP}^2) = \mathbb{Z}$, and because $\mathbb{RP}^2$ is 2-dimensional, $H_n(\mathbb{RP}^2) = 0$ for $n \geq 3$.

Thus, the Mayer-Vietoris long exact sequence is

$$\begin{tikzcd}
0 \arrow[r, "\partial"] & H_2(U\cap V) \arrow[r] & H_2(U)\oplus H_2(V) \arrow[r] & H_2(\mathbb{RP}^2) \arrow[dll, "\partial"] & \\
& H_1(U\cap V) \arrow[r] & H_1(U)\oplus H_1(V) \arrow[r] & H_1(\mathbb{RP}^2) \arrow[dll, "\partial"] & \\
& H_0(U\cap V) \arrow[r] & H_0(U)\oplus H_0(V) \arrow[r] & H_0(\mathbb{RP}^2) \arrow[r, "\partial"] & 0
\end{tikzcd}$$

which reduces to

$$\begin{tikzcd}
0 \arrow[r, "\partial"] & 0 \arrow[r] & 0\oplus 0\cong 0 \arrow[r] & H_2(\mathbb{RP}^2) \arrow[dll, "\partial"] & \\
& \mathbb{Z} \arrow[r] & \mathbb{Z}\oplus 0\cong \mathbb{Z} \arrow[r] & H_1(\mathbb{RP}^2) \arrow[dll, "\partial"] & \\
& \mathbb{Z} \arrow[r] & \mathbb{Z}\oplus\mathbb{Z} \arrow[r] & H_0(\mathbb{RP}^2) \arrow[r, "\partial"] & 0
\end{tikzcd}$$

or shorter still, to

$$0\to H_2(\mathbb{RP}^2)\xrightarrow{g}\mathbb{Z}\xrightarrow{f}\mathbb{Z}\xrightarrow{h}H_1(\mathbb{RP}^2)\xrightarrow{i}\mathbb{Z}\xrightarrow{j}\mathbb{Z}\oplus\mathbb{Z}\to\mathbb{Z}\to 0$$

Because U encloses the boundary of a Möbius band, the inclusion $U\cap V\to U$ wraps twice around the circle that the Möbius band retracts to, so $f:\mathbb{Z}\to\mathbb{Z}$ is given by $z\mapsto 2z$, which is injective. By exactness at the first $\mathbb{Z}$, $\operatorname{im} g=\ker f=0$, but also by exactness at $H_2(\mathbb{RP}^2)$, g is injective, so $H_2(\mathbb{RP}^2)\cong\operatorname{im} g=0$.

j is defined by $z\mapsto(z,-z)$, which is injective, so $\ker j=0$, and by exactness, $\operatorname{im} i=\ker j=0$. Then, by exactness at $H_1(\mathbb{RP}^2)$, h is surjective, so $H_1(\mathbb{RP}^2)\cong\mathbb{Z}/\ker h=\mathbb{Z}/\operatorname{im} f=\operatorname{coker} h=\mathbb{Z}/2$ by the first isomorphism theorem.

$$H_n(\mathbb{RP}^2)=\begin{cases}\mathbb{Z} & n=0\\ \mathbb{Z}/2 & n=1\\ 0 & n\geq 2\end{cases}$$

△

39.5.3 Applications

Corollary 39.5.3.1. *$S^{k-1}=\partial_k D^k$ is not a retract of D^k.*

Proof. Suppose there is a retraction $r:D^k\to S^{k-1}$, so $r\big|_{S^k}=r\circ\iota=\mathrm{id}_{S^k}$. We then have the induced maps in homology $(r\circ\iota)_*=\mathrm{id}_*$:

$$\mathbb{Z}=\tilde{H}_{k-1}(S^{k-1})\xrightarrow{\iota_*}\tilde{H}_{k-1}(D^k)\xrightarrow{r_*}\tilde{H}_{k-1}(S^{k-1})=\mathbb{Z}$$

but D^k is contractible, so $\tilde{H}_{k-1}(D^k)=0$, which r_* cannot surject onto $\mathbb{Z}$ from. ■

Corollary 39.5.3.2 (Brouwer's Fixed-Point Theorem)**.** *Every continuous map $f : D^k \to D^k$ has a fixed point. That is, a point $x \in D^k$ such that $f(x) = x$.*

In dimension $k = 1$, this is saying that a continuous map $f : [0,1] \to [0,1]$ necessarily has a fixed point. In dimension $k = 2$, this implies that if you have a map of an area within the bounds of that area, then there is a point on that map directly above the point it represents on the Earth; this holds even if the map is folded up or crumpled into a ball, as these transformations are continuous. In dimension $k = 3$, this implies that if you stir a cup of coffee continuously, then there is a molecule whose position is unchanged after the stirring.

Proof. Suppose otherwise, that $f(x) \neq x$ for all x. For each x, consider the line connecting the distinct points $f(x)$ and x. Starting at $f(x)$ and traveling towards x, this ray intersects S^{k-1} at exactly one point x' as D^k is convex. Define a map $g : D^k \to S^{k-1}$ by $x \mapsto x'$. If x is already on the boundary, then $g(x) = x$, so g is a retraction, contradicting the previous corollary. ■

We have another result from Brouwer:

Corollary (Invariance of Domain)**.** *If $k \neq \ell$, then $\mathbb{R}^k \ncong \mathbb{R}^\ell$.*

Proof. Let $k \neq \ell$, and suppose $f : \mathbb{R}^k \to \mathbb{R}^\ell$ is a homeomorphism. Then, removing a point from $\mathbb{R}^k$ yields

$$S^{k-1} \simeq \mathbb{R}^k \setminus \{0\} \cong \mathbb{R}^\ell \setminus \{f(0)\} \simeq S^{\ell-1}$$

But then,

$$\mathbb{Z} = \tilde{H}_{k-1}(S^{k-1}) = \tilde{H}_{k-1}\big(\mathbb{R}^k \setminus \{0\}\big) \overset{f_*}{\cong} \tilde{H}_{k-1}\big(\mathbb{R}^\ell \setminus \{f(0)\}\big) = \tilde{H}_{k-1}(S^{\ell-1}) = 0$$

■

39.6 Proof of Fundamental Theorems

39.6.1 Homotopy Invariance

Theorem (Homotopy Invariance)**.** *Suppose $f,g : X \to Y$ are homotopic. Then,*

$$f_* = g_* : H_n(X) \to H_n(Y)$$

The strategy for the proof is as follows:

1. Let $H : X \times [0,1] \to Y$ be the homotopy between f and g. Using the *prism operator* we produce a *chain homotopy* between the chain maps f_* and g_*.
2. We show that chain homotopic chain maps induce the same maps in homology.

39.6.1.1 Chain Homotopy

Let $a_\bullet,b_\bullet : (C_\bullet,\partial) \to (C'_\bullet,\partial')$ be two chain maps. A *chain homotopy* from $a_\bullet$ to $b_\bullet$ is a collection of morphisms

$$\eta_n : C_n \to C'_{n+1}$$

such that, in this (non-commutative!) diagram,

$$\begin{tikzcd}
\cdots \arrow[r, "\partial_{n+2}"] & C_{n+1} \arrow[r, "\partial_{n+1}"] \arrow[d, "b_{n+1}-a_{n+1}"'] & C_n \arrow[r, "\partial_n"] \arrow[d, "b_n-a_n"] \arrow[dl, "\eta_n"'] & C_{n-1} \arrow[r, "\partial_{n-1}"] \arrow[d, "b_{n-1}-a_{n-1}"] \arrow[dl, "\eta_{n-1}"'] & \cdots \\
\cdots \arrow[r, "\partial'_{n+2}"] & C'_{n+1} \arrow[r, "\partial'_{n+1}"] & C'_n \arrow[r, "\partial'_n"] & C'_{n-1} \arrow[r, "\partial'_{n-1}"] & \cdots
\end{tikzcd}$$

where the red path is equal to the sum of the blue and cyan paths. That is,

$$b_n - a_n = \partial'_{n+1} \circ \eta_n + \eta_{n-1} \circ \partial_n$$

for all $n \in \mathbb{Z}$. We say that $a_\bullet$ and $b_\bullet$ are *chain homotopic* if there exists a chain homotopy between them.

Lemma 39.6.1. *Let $a_\bullet$ and $b_\bullet$ be chain homotopic. Then their induced maps in homology are equal:*

$$a_n = b_n : H_n(C_\bullet) \to H_n(C'_\bullet)$$

Proof. Let $c \in Z_n(C_\bullet) = \ker(\partial_n)$ be an n-cycle. Then,

$$\begin{aligned} b_n(c) - a_n(c) &= \partial'_{n+1} \circ \eta_n(c) + \eta_{n-1} \circ \partial_n(c) \\ &= \partial'_{n+1} \circ \eta_n(c) \\ &= \partial'_{n+1}\big(\eta_n(c)\big) \end{aligned}$$

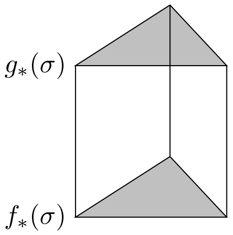
so $b_n(c) - a_n(c)$ is a boundary (of $\eta_n(c)$), so they are homologous. ■

39.6.1.2 Prism Operators

Given a singular n-simplex $\sigma : \Delta^n \to X$ and a homotopy $H : X \times I \to Y$ between f and g, we can compose them into a continuous map

$$H \circ (\sigma \times \mathrm{id}_I) : \Delta^n \times I \to Y$$

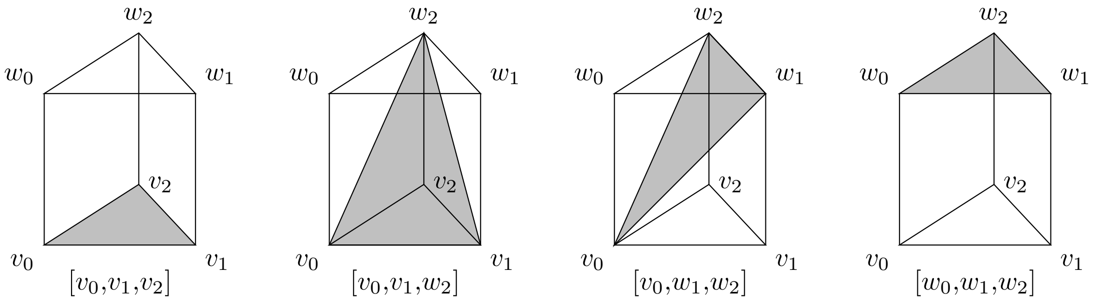
This is a homotopy from $H \circ \big(\sigma \times \{0\}\big) = f \circ \sigma = f_*(\sigma)$ to $H \circ \big(\sigma \times \{1\}\big) = g \circ \sigma = g_*(\sigma)$, which can be visualised as a prism (for $n = 2$):



The goal is now to produce an $(n+1)$-chain in Y from this data.

Denote the lower simplex by $[v_0, \ldots, v_n]$, and the upper simplex by $[w_0, \ldots, w_n]$. The idea is to generate a sequence of n-simplices that starts from the lower simplex, and ends at the upper simplex, by incrementally moving a vertex v_i up to w_i, starting with v_n, and working backwards to v_0.

So, the first step is to move $[v_0, \ldots, v_n]$ to $[v_0, \ldots, v_{n-1}, w_n]$; the second step is to move this simplex up to $[v_0, \ldots, v_{n-2}, w_{n-1}, w_n]$; etc.



$[v_0, v_1, v_2]$ $[v_0, v_1, w_2]$ $[v_0, w_1, w_2]$ $[w_0, w_1, w_2]$

More generally, we move $[v_0,\ldots,v_i,w_{i+1},\ldots,w_n]$ to $[v_0,\ldots,v_{i-1},w_i,\ldots,w_n]$. The region between two successive n-simplices is precisely the $(n+1)$-simplex $[v_0,\ldots,v_i,w_i,\ldots,w_n]$, which has $[v_0,\ldots,v_i,w_{i+1},\ldots,w_n]$ as its lower face, and $[v_0,\ldots,v_{i-1},w_i,\ldots,w_n]$ as its upper face.

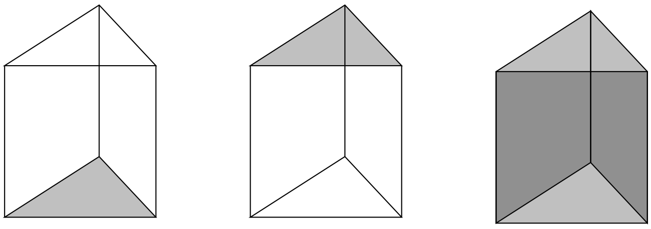
The *prism operator* $P : C_n(\Delta \times I) \to C_{n+1}(\Delta^n \times I)$ returns the alternating sum of these $(n+1)$-simplices:

$$P(\Delta^n) = \sum_{i=0}^{n} (-1)^i [v_0,\ldots,v_i,w_i,\ldots,w_n]$$

Lemma 39.6.2. *For all $n \geq 0$,*

$$\partial P(\Delta^n) = [w_0,\ldots,w_n] - [v_0,\ldots,v_n] - P(\partial \Delta^n)$$

Geometrically, the left side of this equation represents the boundary of the prism, while the three terms on the right represent these three pieces of the boundary:



$\Delta^n \times \{0\}$ $\Delta^n \times \{1\}$ $\partial\Delta^n \times I$

Proof. First, we have

$$\begin{aligned}\partial_{n+1} P(\Delta^n) &= \sum_{j \leq i} (-1)^{i+j} [v_0,\ldots,\widehat{v_j},\ldots,v_i w_i,\ldots,w_n] \\ &+ \sum_{j \geq i} (-1)^{i+j+1} [v_0,\ldots,v_i,w_i,\ldots,\widehat{w_j}\ldots,w_n]\end{aligned}$$

Consider the terms with $i = j$. Here, we have

$$\sum_{i=0}^{n} [v_0,\ldots,v_{i-1} w_i,\ldots,w_n] - \sum_{i=0}^{n} [v_0,\ldots,v_i,w_{i+1},\ldots,w_n]$$

of which, all but two terms cancel out, leaving

$$[w_0,\ldots,w_n] - [v_0,\ldots,v_n]$$

For the remaining terms with $i \neq j$, we apply the prism operator to each face $[v_0,\ldots,\widehat{v_j},\ldots,v_n]$ of Δ^n:

$$\begin{aligned}P\big([v_0,\ldots,\widehat{v_j},\ldots,v_n]\big) &= \sum_{i<j} (-1)^i [v_0,\ldots,v_i,w_i,\ldots,\widehat{w_j},\ldots,w_n] \\ &+ \sum_{j<i} (-1)^{i+1} [v_0,\ldots,\widehat{v_j},\ldots,v_i,w_i,\ldots,w_n]\end{aligned}$$

hence, taking the alternating sum over all j, we have,

$$\begin{aligned}P(\partial_n \Delta^n) &= \sum_{i<j} (-1)^{i+j} [v_0,\ldots,v_i,w_i,\ldots,\widehat{w_j},\ldots,w_n] \\ &+ \sum_{j<i} (-1)^{i+j+1} [v_0,\ldots,\widehat{v_j},\ldots,v_i,w_i,\ldots,w_n]\end{aligned}$$

which is precisely the negative of the terms in the first equation yet to be accounted for (that is, those with $i \neq j$). ■

Proof of Homotopy Invariance. Let $H : X \times I \to Y$ be a homotopy from f to g, and let $\sigma : \Delta^n \to X$ be a singular n-simplex in X. This induces a map on $(n+1)$-chains

$$\big(H \circ (\sigma \times \mathrm{id}_I)\big)_* : C_{n+1}(\Delta^n \times I) \to C_{n+1}(Y)$$

Define the chain map $\eta_n : C_n(X) \to C_{n+1}(Y)$ by

$$c \mapsto \big(H \circ (c \times \mathrm{id}_I)\big)_*\big(P(\Delta^n)\big)$$

Then,

$$\begin{aligned}
\partial\eta_n(\sigma) &= \partial\big(H \circ (\sigma \times \mathrm{id}_I)\big)_*\big(P(\Delta^n)\big) \\
&= \big(H \circ (\sigma \times \mathrm{id}_I)\big)_*\big(\partial P(\Delta^n)\big) \\
&= \big(H \circ (\sigma \times \mathrm{id}_I)\big)_*\big([w_0,\dots,w_n] - [v_0,\dots,v_n] - P(\partial\Delta^n)\big) \\
&= g_*(\sigma) - f_*(\sigma) - \eta_{n-1}\partial(\sigma)
\end{aligned}$$

so η is a chain homotopy from f_* to g_*, so they induce equal maps in homology. ■

39.6.2 Mayer-Vietoris

Theorem (Mayer–Vietoris Long Exact Sequence)**.** *Let $X = U_1 \cup U_2$ be the union of two open subspaces $j_1 : U \hookrightarrow X$, $j_2 : U_2 \hookrightarrow X$. Then, there are connecting homomorphisms $\partial : H_n(X) \to H_{n-1}(U_1 \cap U_n)$ such that*

$$\cdots \to H_{n+1}(X) \xrightarrow{\partial} H_n(U_1 \cap U_2) \xrightarrow{((i_1)_*, -(i_2)_*)} H_n(U_1) \oplus H_n(U_2) \xrightarrow{(j_1)_* + (j_2)_*} H_n(X) \xrightarrow{\partial} H_{n-1}(U_1 \cap U_2) \to \cdots$$

is an exact chain complex.

As with the proof of Homotopy Invariance, the proof of Mayer–Vietoris involves a topological step followed by a purely algebraic step:

1. A *short exact sequence of chain complexes* induces a long exact sequence in homology.
2. The relevant short exact sequence of chain complexes associated with the cover $X = U_1 \cup U_2$ is:

$$0 \to C_\bullet(U_1 \cap U_2) \to C_\bullet(U_1) \oplus C_\bullet(U_2) \to C_\bullet(U_1 + U_2) \to 0$$

   where $C_\bullet(U_1 + U_2) \subseteq C_\bullet(X)$ is a subcomplex with the same homology groups. Showing that these chain complexes have the same homology is the most involved part of the proof, requiring a topological process called *barycentric subdivision.*

39.6.2.1 Short Exact Sequences of Chain Complexes

A *short exact sequence of chain complexes*

$$0 \to A_\bullet \xrightarrow{f_\bullet} B_\bullet \xrightarrow{g_\bullet} C_\bullet \to 0$$

is a pair of chain maps $f_\bullet$ and $g_\bullet$ such that

$$0 \to A_n \xrightarrow{f_n} B_n \xrightarrow{g_n} C_n \to 0$$

is a short exact sequence of abelian groups for all $n \in \mathbb{Z}$.

Example. Let $(B_\bullet, \partial_\bullet)$ be any chain complex. Define a new chain complex $A_\bullet$ by

$$A_n \coloneqq \begin{cases} B_n & n > 0 \\ \ker(\partial_0) & n = 0 \\ 0 & n = 0 \end{cases}$$

and where the differentials are the restrictions of $\partial_\bullet$ to $A_\bullet$. We may then define $C_\bullet$ levelwise as $C_n = B_n/A_n$ with the induced differentials. The resulting short exact sequence $0 \to A_\bullet \to B_\bullet \to C_\bullet 0$ of chain complex looks as follows:

$$\begin{tikzcd}
 & \vdots \arrow[d] & \vdots \arrow[d] & \vdots \arrow[d] & \\
0 \arrow[r] & B_2 \arrow[r, equal] \arrow[d, "\partial_2"] & B_2 \arrow[r] \arrow[d, "\partial_2"] & 0 \arrow[r] \arrow[d] & 0 \\
0 \arrow[r] & B_1 \arrow[r, equal] \arrow[d, "\partial_1"] & B_1 \arrow[r] \arrow[d, "\partial_1"] & 0 \arrow[r] \arrow[d] & 0 \\
0 \arrow[r] & \ker(\partial_0) \arrow[r, hook] \arrow[d] & B_0 \arrow[r, two heads] \arrow[d, "\partial_0"] & B_0/\ker(\partial_0) \arrow[r] \arrow[d, "\partial_0"] & 0 \\
0 \arrow[r] & 0 \arrow[r, hook] \arrow[d] & B_{-1} \arrow[r, equal] \arrow[d, "\partial_{-1}"] & B_{-1} \arrow[r] \arrow[d, "\partial_{-1}"] & 0 \\
0 \arrow[r] & 0 \arrow[r, hook] \arrow[d] & B_{-2} \arrow[r, equal] \arrow[d] & B_{-2} \arrow[r] \arrow[d] & 0 \\
 & \vdots & \vdots & \vdots &
\end{tikzcd}$$

△

In the following proof, it will be helpful to have the following commutative diagram of a general short exact sequence of chain complexes

$$\begin{tikzcd}
 & \vdots \arrow[d] & \vdots \arrow[d] & \vdots \arrow[d] & \\
0 \arrow[r] & A_{n+1} \arrow[r, "f_{n+1}"] \arrow[d, "\partial"] & B_{n+1} \arrow[r, "g_{n+1}"] \arrow[d, "\partial"] & C_{n+1} \arrow[r] \arrow[d, "\partial"] & 0 \\
0 \arrow[r] & A_n \arrow[r, "f_n"] \arrow[d, "\partial"] & B_n \arrow[r, "g_n"] \arrow[d, "\partial"] & C_n \arrow[r] \arrow[d, "\partial"] & 0 \\
0 \arrow[r] & A_{n-1} \arrow[r, "f_{n-1}"] \arrow[d] & B_{n-1} \arrow[r, "g_{n-1}"] \arrow[d] & C_{n-1} \arrow[r] \arrow[d] & 0 \\
 & \vdots & \vdots & \vdots &
\end{tikzcd}$$

where all the rows are exact and the columns are chain complexes.

Theorem 39.6.3. *Let*

$$0 \to A_\bullet \xrightarrow{f_\bullet} B_\bullet \xrightarrow{g_\bullet} C_\bullet \to 0$$

be a short exact sequence of chain complexes. Then, there are connecting homomorphisms $\partial : H_n(C_\bullet) \to H_{n-1}(A_\bullet)$ such that

$$\begin{tikzcd}
\cdots \arrow[r, "\partial"] & H_n(A_\bullet) \arrow[r, "f_*"] & H_n(B_\bullet) \arrow[r, "g_*"] & H_n(C_\bullet) \arrow[dll, rounded corners, "\partial"] & \\
 & H_{n-1}(A_\bullet) \arrow[r, "f_*"] & H_{n-1}(B_\bullet) \arrow[r, "g_*"] & H_{n-1}(C_\bullet) \arrow[r, "\partial"] & \cdots
\end{tikzcd}$$

is a long exact sequence in homology.

Proof. Let $c \in C_n$ be a cycle, i.e. $c \in \ker(\partial)$. By exactness at C_n, g_n is surjective, so $c = g_n(b)$ for some $b \in B_n$.

The element $\partial(b) \in B_{n-1}$ is in $\ker(g_{n-1})$ since, by commutativity of the lower right square, $g_{n-1}\big(\partial(b)\big) = \partial\big(g_n(b)\big) = \partial(c) = 0$. Then, by exactness at B_{n-1}, $\ker(g_{n-1}) = \operatorname{im}(f_{n-1})$, so $\partial(b) = f_n(a)$ for some $a \in A_{n-1}$.

$$\begin{tikzcd}
 & b \arrow[r, mapsto, "g_n"] \arrow[d, mapsto, "\partial"] & c \arrow[d, mapsto, "\partial"] \\
a \arrow[r, mapsto, "f_{n-1}"] & \partial(b) \arrow[r, mapsto, "g_{n-1}"] & \partial(c) = 0
\end{tikzcd}$$

We claim that a is a cycle. First, apply the differential to a, then travel along f_{n-2}:

$$\begin{tikzcd}
 & b \arrow[r, mapsto, "g_n"] \arrow[d, mapsto, "\partial"] & c \arrow[d, mapsto, "\partial"] \\
a \arrow[r, mapsto, "f_{n-1}"] \arrow[d, mapsto, "\partial"] & \partial(b) \arrow[r, mapsto, "g_{n-1}"] \arrow[d, mapsto, "\partial"] & \partial(c) = 0 \\
\partial(a) \arrow[r, mapsto, "f_{n-2}"] & f_{n-2}\big(\partial(a)\big) = \partial\big(\partial(b)\big) = 0 &
\end{tikzcd}$$

By commutativity of this lower square, $f_{n-2}\big(\partial(a)\big) = \partial\big(f_{n-1}(a)\big) = \partial\big(\partial(b)\big) = 0$, and by exactness at A_{n-2}, f_{n-2} is injective, so $\partial(a) = 0$, and a is a cycle, thus defining a homology class $[a]$.

We define the connecting homomorphism, $\partial : H_n(C) \to H_{n-1}(A)$ by $[c] \mapsto [a]$. This is well-defined since:

- The element a is uniquely determined by ∂b since f_{n-1} is injective.
- Suppose we chose a different preimage b' of c in the first step. Repeating the previous construction, we have

$$\begin{tikzcd}
 & b' \arrow[r, mapsto, "g_n"] \arrow[d, mapsto, "\partial"] & c \\
a' \arrow[r, mapsto, "f_{n-1}"] & \partial(b') &
\end{tikzcd}$$

  Then, we have $g_n(b') = c = g_n(b)$, so $g_n(b' - b) = g_n(b') - g_n(b) = 0$, giving $b' - b \in \ker(g_n)$. By exactness at B_n, $\ker(g_n) = \operatorname{im}(f_n)$, so $b' - b = f_n(\sigma)$ for some $\sigma \in A_n$. Applying the differential to σ, we obtain the square:

$$\begin{tikzcd}
\sigma \arrow[r, mapsto, "f_n"] \arrow[d, mapsto, "\partial"] & b - b' \arrow[r, mapsto, "g_n"] \arrow[d, mapsto, "\partial"] & 0 \\
\partial(\sigma) \arrow[r, mapsto, "f_{n-1}"] & \partial(b) - \partial(b') &
\end{tikzcd}$$

  Now, consider $a - a'$. By construction, $f_{n-1}(a) = \partial(b)$ and $f_{n-1}(a') = \partial(b')$, so $f_{n-1}(a - a') = \partial(b) - \partial(b')$. By injectivity of f_{n-1}, $\partial(\sigma) = a - a'$. Thus, a and a' are homologous, so $[a] = [a']$.
- A different choice of c within its homology class would have the form $c + \partial(c')$ for some $c \in C_{n+1}$. By exactness at C_{n+1}, g_{n+1} is surjective, so $c' = g_n(b')$ for some $b' \in B_{n+1}$.

$$\begin{tikzcd}
b' \arrow[r, mapsto, "g_{n+1}"] \arrow[d, mapsto, "\partial"] & c' \arrow[d, mapsto, "\partial"] \\
\partial(b') \arrow[r, mapsto, "g_n"] & \partial(c')
\end{tikzcd}$$

  Then, $\partial(c') = \partial\big(g_{n+1}(b')\big) = g_n\big(\partial(b')\big)$, so $c + \partial(c') = g_n(b) + g_n\big(\partial(b')\big) = g_n\big(b + \partial(b')\big)$. So, b is replaced by $b + \partial(b')$, which leaves $\partial(b)$ and therefore also a unchanged.

This map is a group homomorphism since if $[c_1] \mapsto [a_1]$ and $[c_2] \mapsto [a_2]$ via b_1 and b_2 as above, then $g_n(b_1 + b_2) = g_n(b_1) + g_n(b_2) = c_1 + c_2$, and $f_{n-1}(a_1 + a_2) = f_{n-1}(a_1) + f_{n-1}(a_2) = \partial(b_1) + \partial(b_2) = \partial(b_1 + b_2)$, so $[c_1] + [c_2] \mapsto [a_1] + [a_2]$.

It remains to verify that the induced sequence in homology

$$\cdots \to H_{n+1}(C_\bullet) \xrightarrow{\partial} H_n(A_\bullet) \xrightarrow{f_*} H_n(B_\bullet) \xrightarrow{g_*} H_n(C_\bullet) \xrightarrow{\partial} H_{n-1}(A_\bullet) \to \cdots$$

is exact.

- Exactness at $H_n(B_\bullet)$:
  - $\operatorname{im}(f_*) \subseteq \ker(g_*)$. This is immediate, since $g_\bullet \circ f_\bullet = 0$ as chain maps, so $g_* \circ f_* = 0$ in homology.
  - $\ker(g_*) \subseteq \operatorname{im}(f_*)$. Let $[b] \in \ker(g_*)$ so $g(b) = \partial(c')$ for some $c' \in C_{n+1}$. Since g is surjective, $c' = g(b')$ for some $b' \in B_{n+1}$. Since $\partial\big(g(b')\big) = \partial(c') = g(b)$, we have $g\big(b - \partial(b')\big) = g(b) - g\big(\partial(b')\big) = g(b) - \partial\big(g(b')\big) = 0$.

    So, $b - \partial(b') = f(a)$ for some $a \in A_n$. This a is a cycle since $f\big(\partial(a)\big) = \partial\big(f(a)\big) = \partial\big(b - \partial(b')\big) = \partial b = 0$, and f is injective. Thus, $f_*\big([a]\big) = \big[b - \partial(b')\big] = [b]$, so f_* surjects onto $\ker(g_*)$.
- Exactness at $H_n(A_\bullet)$:
  - $\operatorname{im}(g_*) \subseteq \ker(\partial)$. $\partial \circ j_* = 0$ since $\partial(b) = 0$ by the definition of ∂.
  - $\ker(\partial) \subseteq \operatorname{im}(g_*)$. If c represents a homology class in $\ker(\partial)$, then $a = \partial(a')$ for some $a' \in A_n$. Then, $b - f(a')$ is a cycle since $\partial\big(b - f(a')\big) = \partial(b) - \partial\big(f(a')\big) = \partial(b) - f(a) = 0$, and $g\big(b - f(a')\big) = g(b) - g\big(f(a')\big) = g(b) = c$, so $g_*\big(\big[b - f(a')\big]\big) = [c]$, so g_* surjects onto $\ker(\partial)$.
- Exactness at $H_n(C_\bullet)$:
  - $\operatorname{im}(\partial) \subseteq \ker(f_*)$. $f_* \circ \partial = 0$ since $f_* \circ \partial$ maps $[c]$ to $\big[\partial(b)\big] = 0$.
  - $\ker(f_*) \subseteq \operatorname{im}(\partial)$. If $a \in A_{n-1}$ is a cycle such that $f(a) = \partial(b)$ for some $b \in B_n$, then $g(b)$ must be a cycle, since $\partial\big(g(b)\big) = g\big(\partial(b)\big) = g\big(f(a)\big) = 0$, so $\partial\big(\big[g(b)\big]\big) = [a]$.

■

Let $U_1,U_2 \subseteq X$ be two subspaces, not necessarily open. We write $C_n(U_1 + U_2)$ for the subgroup of $C_n(X)$ consisting of n-chains that can be written as the sum of n-chains in U_1 and n-chains in U_2:

$$C_n(U_1 + U_2) := \left\{ \sum_{i=0}^{m} \lambda_i \sigma_i : \lambda_i \in \mathbb{Z}, \sigma_i \in C_n(U_1) \cup C_n(U_2) \right\}$$

The boundary of an n-chain in U_ℓ is an $(n-1)$-chain in U_ℓ, so the differentials in $C_\bullet(X)$ restrict to $C_\bullet(U_1 + U_2)$, so $C_\bullet(U_1 + U_2)$ is a sub-chain complex. That is, the inclusion maps $C_n(U_1 + U_2) \hookrightarrow C_n(X)$ define a chain map.

Theorem 39.6.4. *Let $j_\ell : U_\ell \hookrightarrow X$ and $i_\ell : U_1 \cap U_2 \hookrightarrow U_\ell$ be the canonical inclusion maps for $\ell = 1,2$. Then, there is a short exact sequence of chain complexes*

$$0 \to C_\bullet(U_1 \cap U_2) \xrightarrow{\big((i_1)_*, -(i_2)_*\big)} C_\bullet(U_1) \oplus C_\bullet(U_2) \xrightarrow{(j_1)_* + (j_2)_*} C_\bullet(U_1 + U_2) \to 0$$

Proof.

- The subgroup $C_n(U_1 + U_2)$ is precisely the image of $(j_1)_* + (j_2)_*$, so this map is surjective.
- It suffices to check that one of the components of $\big((i_1)_*, -(i_2)_*\big)$ is injective, and indeed, $(i_1)_*$ is an inclusion and is hence injective.

- The composition is given by

$$\begin{aligned}\big((j_1)_* + (j_2)_*\big) \circ \big((i_1)_*, -(i_2)_*\big) &= (j_1)_*(i_1)_* - (j_2)_*(i_2)_* \\ &= (j_1 \circ i_1)_* - (j_2 \circ i_2)_* \\ &= 0\end{aligned}$$

since $j_1 \circ i_1 = j_2 \circ i_2$ are both the inclusion $k : U_1 \cap U_2 \hookrightarrow X$. So, $\operatorname{im}(i) \subseteq \ker(j)$.

Conversely, let $(c_1,c_2) \in \ker(j)$. That is, $j(c_1,c_2) = (j_1)_*(c_1) + (j_2)_*(c_2) = 0$, then $(j_1)_*(c_1) = -(j_2)_*(c_2)$. The left side of this equation is a chain whose simplices are contained in U_1, while the right side is a chain whose simplices are contained in U_2. So, this equality implies that all of these simplices are in the intersection $U_1 \cap U_2$, so there exists a chain $c \in C_n(U_1 \cap U_2)$ such that $k_*(c) = (j_1)_*(c_1) = -(j_2)_*(c_2)$.

Then, $(j_1)_*(c_1) = k_*(c) = (j_1)_*\big((i_1)_*(c)\big)$, so $c_1 = (i_1)_*(c)$ by injectivity of $(j_1)_*$, and similarly, $c_2 = -(i_2)_*(c)$. So, $i(c) = (c_1,c_2)$, so $\ker(j) \subseteq \operatorname{im}(i)$.

■

39.6.3 Barycentric Subdivision

This is the last theorem required to prove Mayer–Vietoris, but its proof requires us to first develop some more machinery.

Theorem 39.6.5. *Let $U_1,U_2 \subseteq X$ be subspaces, not necessarily open. If their interiors jointly cover X, then the inclusion $\iota : C_\bullet(U_1 + U_2) \hookrightarrow C_\bullet(X)$ induces an isomorphism in homology. That is,*

$$H_n\big(C_\bullet(U_1 + U_2)\big) \overset{\iota_*}{\cong} H_n\big(C_\bullet(X)\big)$$

Let $\sigma : \Delta^n \to X$ be a singular n-simplex. Since the interiors of U_1 and U_2 cover X, we have an open cover

$$\sigma^{-1}[U_1^\circ] \cup \sigma^{-1}[U_2^\circ]$$

of Δ^n. Let $A_i = \Delta^n \setminus \sigma[U_i^\circ]$, $i = 1,2$. be the closed complements, and define a function $f : \Delta^n \to \mathbb{R}$,

$$f(x) = \frac{d(x,A_1) + d(x,A_2)}{2}$$

as the average distance of x to each of the two closed subsets. (If one of the A_i is empty, then $\sigma(\Delta^n) \subseteq U_i$, and we are done.) By compactness of Δ^n, this function attains a minimum δ, which is necessarily positive, or otherwise we wouldn't have a cover.

One can then verify that every simplex $[w_0, \dots ,w_n] \subseteq \Delta^n$ of diameter less than δ is entirely contained in one of the $\sigma^{-1}[U_i^\circ]$. That is, that δ is a Lebesgue number (§37.5.3) for the given open cover of Δ^n.

The point is that, if we can divide Δ^n into simplices of diameter less than δ, then the restriction of σ to each of these lies in $C_n(U_1 + U_2)$. *Barycentric subdivision* is a systematic method of dividing simplices into smaller ones such that the diameter tends to 0 as the process is iterated.

Given a Euclidean space V and some elements $v_0 \dots ,v_n \in V$, the *linear simplex* $[v_0, \dots ,v_n]$ is the subspace

$$\left\{\sum_{i=0}^{n} x_i v_i : x_i \geq 0, \sum_{i=0}^{n} x_i = 1\right\} \subseteq V$$

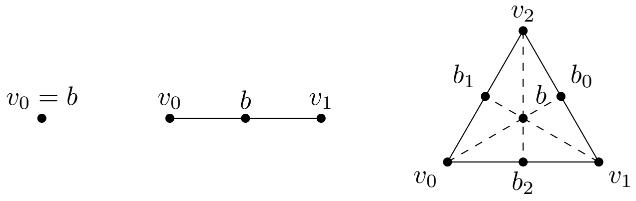
Example. The standard n-simplex arises from taking the standard basis vectors in $V = \mathbb{R}^{n+1}$. △

If the difference vectors $[v_i,v_j]$ are linearly independent, then a linear simplex on n vertices will be homeomorphic to Δ^n (though generally not isometric). For this reason, we will also require linear simplices to satisfy this condition.

The *barycentre* of a simplex $[v_0,\dots,v_n]$ is the point

$$b := \frac{1}{n+1}\sum_{i=0}^{n} v_i$$

Example.



△

As suggested by the above, can find the barycentre recursively. Given the barycentre b_i on the ith face $[v_0,\dots,\widehat{v_i},\dots,v_n]$, let ℓ_i be the line connecting b_i and v_i. Then, the barycentre of $[v_0,\dots,v_n]$ is the intersection of all these lines ℓ_i.

Given a linear $(n-1)$-simplex $[w_1,\dots,w_n] \subseteq \Delta^n$, we define its *barycentric cone* as:

$$b[w_1,\dots,w_n] := [b,w_1,\dots,w_n] \subseteq \Delta^n$$

We extend this to linear combinations of linear $(n-1)$ simplices.

The *barycentric subdivision* $S(\Delta^n) \in C_n(\Delta^n)$ of Δ^n is defined by induction on n:

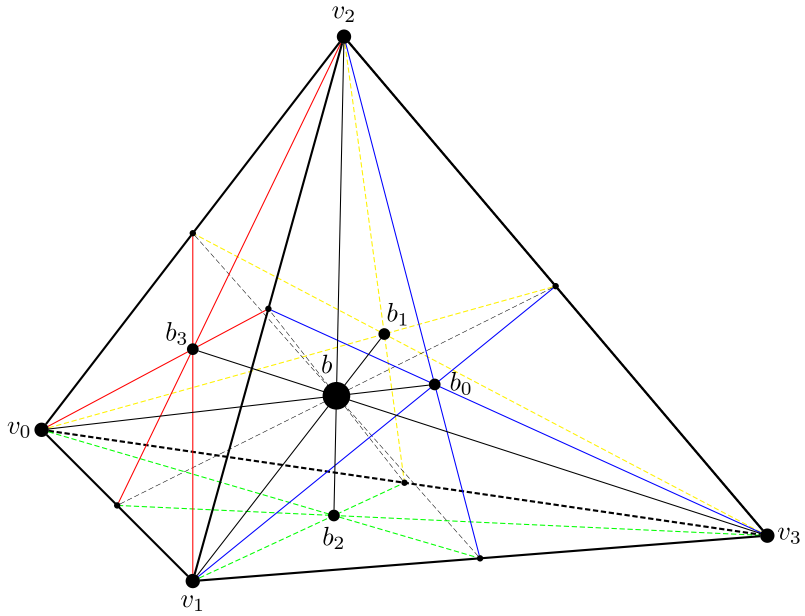
- We define $S(\Delta^0) = \Delta^0$.
- For $n > 0$,

$$\begin{aligned} S(\Delta^n) &:= bS\partial\Delta^n \\ &= \sum_{i=0}^{n}(-1)^i bS(\partial_i\Delta^n) \end{aligned}$$

Note that by induction, $S(\partial_i\Delta^n)$ is itself a linear combination of linear simplices, so this construction is well defined.

Example. Compare with the examples above:

- $S(\Delta^0) = [v_0]$;
- $S(\Delta^1) = bS[v_1] - bS[v_0] = [b,v_1] - [b,v_0]$;
- $S(\Delta^2) = [b,b_0,v_2] - [b,b_0,v_1] - [b,b_1,v_2] + [b,b_1,v_0] + [b,b_2,v_1] - [b,b_2,v_0]$;
- In dimension 3, we give a picture instead of the lengthy formula (and ignoring the signs):



△

The signs are chosen such that the boundary of $S(\Delta^n)$ is (the subdivision of) the boundary of Δ^n. That is, all the internal boundaries cancel out. For instance,

$$\begin{aligned}\partial S(\Delta^1) &= \partial\big([b,v_1] - [b,v_0]\big)\\ &= v_1 - b - v_0 + b\\ &= v_1 - v_0\\ &= \partial\Delta^1\end{aligned}$$

We now prove this in general:

Lemma 39.6.6.

(i) $\partial b(\sigma) + b\partial(\sigma) = \sigma$ *for every linear simplex* σ*;*

(ii) $\partial S(\Delta^n) = S(\partial\Delta^n)$.

If we ignore the signs, the first identity says something intuitively clear: the boundary of the cone consists of the base and the cones on its faces.

Proof.

(i) Let $\sigma = [w_1,\dots,w_n]$. Then,

$$\begin{aligned}\partial b[w_1,\dots,w_n] &= \partial[b,w_1,\dots,w_n]\\ &= \sum_{i=0}^{n}(-1)^i\partial_i[b,w_1,\dots,w_n]\\ &= [w_1,\dots,w_n] - b\partial[w_1,\dots,w_n]\end{aligned}$$

(ii) We induct using *(i)*. For $n = 0$, we have zero on both sides. For $n > 0$, we have

$$\begin{aligned}\partial S(\Delta^n) &= \partial bS(\partial\Delta^n)\\ &= (\mathrm{id} - b\partial)S(\partial\Delta^n)\end{aligned}$$

$$\begin{aligned} &= S\partial\Delta^n - bS(\partial^2\Delta^n) \\ &= S(\partial\Delta^n) \end{aligned}$$

where we used (i) in the second equality, and induction in the third.

■

Let X be a topological space and $\sigma : \Delta^n \to X$ be a singular n-simplex. The *barycentric subdivision* of σ is then the n-chain

$$S(\sigma) = \sigma_* S(\Delta^n) \in C_n(X)$$

Linearly extending, we obtain a homomorphism

$$S : C_n(X) \to C_n(X)$$

Corollary 39.6.6.1. *$S : C_\bullet(X) \to C_\bullet(X)$ is a chain map.*

Proof. We have

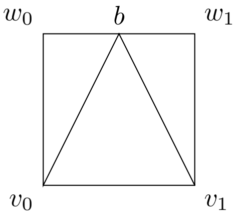
$$\begin{aligned} \partial S(\sigma) &= \partial\sigma_* S(\Delta^n) \\ &= \sigma_*\partial S(\Delta^n) \\ &= \sigma_* S(\partial\Delta^n) \\ &= \sum_{i=0}^{n}(-1)^i\sigma_* S(\partial_i\Delta^n) \\ &= \sum_{i=0}^{n}(-1)^i S(\partial_i\sigma) \\ &= S(\partial\sigma) \end{aligned}$$

so $\partial S = S\partial$, as required. ■

Theorem 39.6.7. *The barycentric subdivision is chain homotopic to the identity:*

$$S \simeq \mathrm{id} : C_\bullet(X) \to C_\bullet(X)$$

Proof. To construct the chain homotopy $T : C_n(X) \to C_{n+1}(X)$ from S to id, the idea is to split the prism $\Delta^n \times [0,1]$ into $(n+1)$-simplices such that the lower face remains intact (as the identity), while the upper face is subdivided:



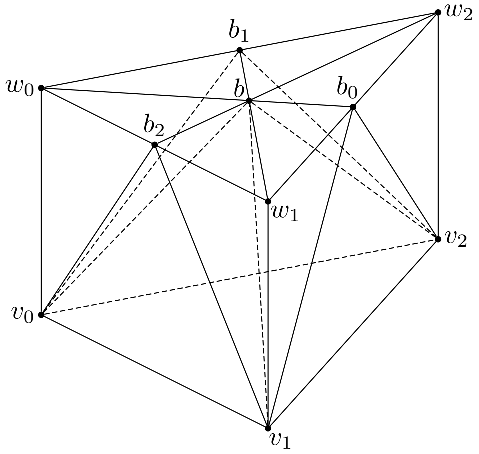
$\Delta^1 \times [0,1]$

Let us write Δ^n_0 for the lower face $\Delta^n \times \{0\}$, Δ^n_1 for the upper face $\Delta^n \times \{1\}$, and b for the barycentric cone of Δ^n_1. Then, we set, recursively,

$$T(\Delta^n) := b\Delta^n_0 - bT\partial\Delta^n_0 \in C_{n+1}(\Delta^n \times [0,1])$$

So, for example,

- For $n = 0$, we have $T(\Delta^0) = [b,v_0] = [w_0,v_0]$;
- For $n = 1$, we have $T(\Delta^1) = [b,v_0,v_1] - [b,w_1,v_1] + [b,w_0,v_0]$;
- For $n = 2$, we have (omitting signs):



We verify the following identity reminiscent of the chain homotopy requirement:

$$\partial T(\Delta^n) + T(\partial\Delta_0^n) = \Delta_0^n - S(\Delta_1^n)$$

For $n = 0$, both sides are $[v_0] - [w_0]$, and for $n > 0$, we have

$$\begin{aligned}
\partial T(\Delta^n) &= \partial b\Delta_0^n - \partial bT(\partial\Delta_0^n) \\
&= \partial_0^n - b\partial(\Delta_0^n) - T\partial(\Delta_0^n) + b\partial T(\partial\Delta_0^n) \\
&= \partial_0^n - T(\partial\Delta_0^n) + b\big(\partial T(\Delta_0^n) - \partial\Delta_0^n\big) \\
&= \partial_0^n - T(\partial\Delta_0^n) - b\big(S(\partial\Delta_1^n) + T(\partial^2\Delta_0^n)\big) \\
&= \partial_0^n - T(\partial\Delta_0^n) - S\Delta_1^n
\end{aligned}$$

Given a singular n-simplex $\sigma : \Delta^n \to X$, let $\sigma' : \Delta^n \times [0,1] \to \Delta^n \to X$ be composition of σ with the projection onto the first component. If we define $T : C_n(X) \to C_{n+1}(X)$ by $\sigma \mapsto \sigma'_* T(\Delta^n)$, then this formula implies that T defines a chain homotopy $S \simeq \mathrm{id}$, as required. ■

The last thing to prove is that the diameters of the simplices in $S^k\Delta^n$ tend to *zero* as $k \to \infty$.

Lemma 39.6.8. *Let* $[w_0, \ldots ,w_n]$ *be a simplex in the barycentric subdivision of* $[v_0, \ldots ,v_n]$*. Then,*

$$\operatorname{diam}\big([w_0, \ldots ,w_n]\big) \leq \frac{n}{n+1} \operatorname{diam}\big([v_0, \ldots ,v_n]\big)$$

Proof. If $n = 0$, then $[w_0] = [v_0]$ both have diameter 0, so suppose $n > 0$. We start with the following observation about any linear simplex $[x_0, \ldots ,x_n]$:

For every point x, its maximum distance to points in the simplex is attained at a vertex x_i.

To see this, let $y \in [x_0, \ldots ,x_n]$ be such that $\|x - y\|$ is maximal so $y = \sum_i t_i x_i$ with $\sum_i t_i = 1$ and $t_i \geq 0$. Then,

$$\|x - y\| = \left\| x - \sum_i t_i x_i \right\|$$

$$\begin{aligned}
&= \left\| \sum_i t_i(x - x_i) \right\| \\
&\leq \sum_i t_i \|x - x_i\| \\
&\leq \max \|x - x_i\|
\end{aligned}$$

with equality if and only if y is one of the vertices x_i with $\|x - x_i\|$ maximal.

In particular, applying this observation twice, we see that the diameter of $[w_0, \ldots, w_n]$ is the length of the longest edge $[x,y]$. We have two cases:

- If neither of x and y are the barycenter b, then they must be vertices of a simplex in the barycentric subdivision of one of the faces $[v_0, \ldots, \widehat{v_i}, \ldots, v_n]$. By induction, we have

$$\begin{aligned}
\operatorname{diam}\bigl([w_0, \ldots, w_n]\bigr) &= \|x - y\| \\
&\leq \frac{n-1}{n} \operatorname{diam}\bigl([v_0, \ldots, \widehat{v_i}, \ldots, v_n]\bigr) \\
&\leq \frac{n}{n+1} \operatorname{diam}\bigl([v_0, \ldots, v_n]\bigr)
\end{aligned}$$

- If, say, $x = b$, then y lies on some face of $[v_0, \ldots, v_n]$, and furthermore, the observation above gives that y is one of the vertices v_i of that face. Let b_i be the barycentre of $[v_0, \ldots, \widehat{v_i}, \ldots, v_n]$. That is,

$$b_i = \frac{1}{n} \sum_{i \neq j} v_j$$

Then,

$$\begin{aligned}
b &= \frac{1}{n+1} \sum_j v_j \\
&= \frac{1}{n+1} v_i + \frac{n}{n+1} b_i
\end{aligned}$$

It follows that

$$\begin{aligned}
\operatorname{diam}\bigl([w_0, \ldots, w_n]\bigr) &= \|v_i - b\| \\
&\leq \frac{n}{n+1} \|v_i - b_i\| \\
&\leq \frac{n}{n+1} \operatorname{diam}\bigl([v_0, \ldots, v_n]\bigr)
\end{aligned}$$

■

Exercise. Verify this in low dimensions.

Proof of Mayer-Vietoris. By Theorem 39.6.3, the short exact sequence in Theorem 39.6.4 induces a long exact sequence in homology:

$$\cdots \to H_{n+1}\bigl(C_\bullet(U_1 + U_2)\bigr) \xrightarrow{\partial} H_n(U_1 \cap U_2) \xrightarrow{i} H_n(U_1) \oplus H_n(U_2) \xrightarrow{j} H_n\bigl(C_\bullet(U_1 + U_2)\bigr) \xrightarrow{\partial} H_{n-1}(U_1 \cap U_2) \to \cdots$$

Then, using Theorem 39.6.5, we may replace $H_n\bigl(C_\bullet(U_1 + U_2)\bigr)$ by $H_n\bigl(C_\bullet(X)\bigr) =: H_n(X)$. ■

39.7 Applications

39.7.1 Fundamental Classes for Spheres

We have seen that the homology of the k-sphere S^k, $k \geq 1$, is given by

$$\tilde{H}_n(S^k) = \begin{cases} \mathbb{Z} & n = k \\ 0 & \text{otherwise} \end{cases}$$

A generator of $\tilde{H}_k(S^k)$ is called a *fundamental class.* Our goal is to explicitly describe these fundamental classes for all spheres by giving a cycle whose homology class is such a generator.

Example. We can view the circle S^1 as the quotient of the interval or 1-simplex. The singular 1-simplex given by the quotient map

$$\sigma : \Delta^1 \cong [0,1] \to [0,1]/0 \sim 1 \cong S^1$$

has boundary

$$\partial\sigma = * - * = 0$$

where $* : \Delta^1 \to S^1$ is the constant path at the identified point. Thus, σ is a 1-cycle which generates $H_1(S^1)$, i.e., the fundamental class of S^1. △

However, if we try to extend this construction to S^2, we run into a problem. The 2-sphere S^2 can be obtained by quotienting the 2-simplex by its boundary. The singular 2-simplex given by the quotient map

$$\sigma : \Delta^2 \to \Delta^2/\partial\Delta^2 \cong S^2$$

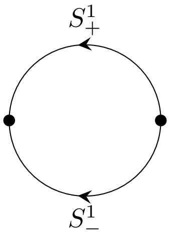
then has boundary

$$\partial\sigma = * - * + * = * \neq 0$$

so σ is not a cycle.

This construction gives a fundamental class for S^k if and only if k is odd. We describe below a construction that works in all dimensions.

Instead of taking S^1 to be the quotient of a single 1-simplex, we instead split S^1 into upper and lower hemispheres:



or,

$$S^1 \cong \frac{\Delta^1 \sqcup \Delta^1}{\sim}$$

where $\sim$ identifies the two 0th faces and two 1st faces of the two Δ^1.

Let $\sigma_+,\sigma_- : \Delta^1 \to S^1$ be the singular simplices that pick out the upper and lower hemispheres, respectively. Then, $\sigma_+ - \sigma_-$ is a cycle that represents a fundamental class.

Consider the identity map $\mathrm{id}_{\Delta^{k+1}}$ on Δ^{k+1}. This is a singular $(k+1)$-simplex in Δ^{k+1}, so we can apply the boundary operator to it:

$$\partial(\mathrm{id}_{\Delta^{k+1}}) = \sum_{i=0}^{k+1}(-1)^i[v_0, \dots, \widehat{v_i}, \dots, v_n]$$

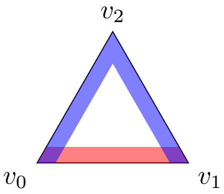
This is a linear combination of k-simplices, each in the geometric boundary $\partial\Delta^{k+1}$ of Δ^{k+1}, so it is an element of the chain group $C_{k+1}(\partial\Delta^{k+1}) \leq C_{k+1}(\Delta^{k+1})$.

Theorem 39.7.1. *The chain $\partial(\mathrm{id}_{\Delta^{k+1}})$ is a cycle in $C_k(\partial\Delta^{k+1})$. Moreover, it is a generator in homology.*

Proof. $\partial(\mathrm{id}_{\Delta^{k+1}})$ is already a boundary, so $\partial\big(\partial(\mathrm{id}_{\Delta^{k+1}})\big) = 0$. So, $\partial(\mathrm{id}_{\Delta^{k+1}})$ is a cycle in $C_k(\partial\Delta^{k+1})$.

Now, we induct on k. If $k = 0$, the statement is clear, so assume $k > 1$.

Let U_1 be an open neighbourhood of the last face $d_{k+1}\Delta^{k+1}$ which deformation retracts onto this face, and let U_2 be an open neighbourhood of the remaining faces $\bigcup_{0\leq i\leq k} d_i\Delta^{k+1}$ which deformation retracts onto this union. Also choose these subspaces such that their intersection $U_1 \cap U_2$ deformation retracts to the boundary of the final face $\partial d_{k+1}\Delta^{k+1} = \partial[v_0, \dots, v_k]$, and such that their union $U_1 \cup U_2$ deformation retracts to the entire boundary $\partial\Delta^{k+1} = \delta[v_0, \dots, v_{k+1}]$. One such selection is illustrated below for $k = 1$.



The Mayer–Vietoris long exact sequence for U_1 and U_2 covering $X \coloneqq U_1 \cup U - 2$ is:

$$\cdots \to \tilde{H}_k(U_1) \oplus \tilde{H}_k(U_2) \to \tilde{H}_k(U_1 \cup U_2) \xrightarrow{\partial} \tilde{H}_{k-1}(U_1 \cap U_2) \to \tilde{H}_{k-1}(U_1) \oplus \tilde{H}_{k-1}(U_2) \to \cdots$$

Because U_1 and U_2 deformation retract to a face, which is a simplex, which is contractible, the outer terms vanish, so we have an isomorphism

$$0 \to \tilde{H}_k(U_1 \cup U_2) \overset{\partial}{\cong} \tilde{H}_{k-1}(U_1 \cap U_2) \to 0$$

By the induction hypothesis, $\tilde{H}_{k-1}(U_1 \cap U_2) \cong \mathbb{Z}$, so it suffices to show that $\partial(\mathrm{id}_{\Delta^{k+1}})$ maps to a generator in $\tilde{H}_{k-1}(U_1 \cap U_2)$ under the connecting homomorphism ∂.

First, take the relevant segment of the short exact sequence of chain complexes:

$$\begin{tikzcd}
 & C_k(U_1) \oplus C_k(U_2) \arrow[r, "(j_1)_* + (j_2)_*"] \arrow[d, "\partial"] & C_k(U_1 + U_2) \\
C_{k-1}(U_1 \cap U_2) \arrow[r, "\big((i_1)_*, -(i_2)_*\big)"] & C_{k-1}(U_1) \oplus C_{k-1}(U_2) &
\end{tikzcd}$$

starting with $\partial(\mathrm{id}_{\Delta^{k+1}}) \in C_k(U_1 + U_2)$. This cycle has a lift in $C_k(U_1) \oplus C_k(U_2)$, given by

$$\left((-1)^{k+1}[v_0, \dots, v_k], \sum_{i=1}^{k}(-1)^i[v_0, \dots, \widehat{v_i}, \dots, v_{k+1}]\right)$$

which maps down to

$$\left((-1)^{k+1}\partial\big([v_0,\ldots,v_k]\big),\sum_{i=1}^{k}(-1)^i\partial\big([v_0,\ldots,\widehat{v_i},\ldots,v_{k+1}]\big)\right)$$

So, we are looking for the unique $(k-1)$-cycle σ in $U_1\cap U_2$ satisfying

$$\big((i_1)_*(\sigma),-(i_2)_*(\sigma)\big)=\left((-1)^{k+1}\partial\big([v_0,\ldots,v_k]\big),\sum_{i=1}^{k}(-1)^i\partial\big([v_0,\ldots,\widehat{v_i},\ldots,v_{k+1}]\big)\right)$$

By comparing the first components, it is clear that $\sigma=(-1)^{k-1}\partial\big([v_0,\ldots,v_k]\big)$ works. ■

Let S^k_+ and S^k_- be the upper and lower hemispheres of S^k. Choose homeomorphisms

$$\sigma_+:\Delta^k\xrightarrow{\cong}S^k_+\qquad\sigma_-:\Delta^k\xrightarrow{\cong}S^k_-$$

such that

- σ_+ and σ_- both map the boundary $\partial\Delta^k$ homeomorphically onto the equator $S^k_+\cap S^k_-$;
- the composition $\partial\Delta^k\xrightarrow{\sigma_+}S^k_+\cap S^k_-\xrightarrow{(\sigma_-)^{-1}}\partial\Delta^k$ is the identity.

Corollary 39.7.1.1. *The chain $\sigma_+-\sigma_-\in C_k(S^k)$ is a cycle, and represents a fundamental class for S^k.*

Proof. We have seen this for $k=0{,}1$ above, so assume $k\geq 2$.

The second requirement on σ_+ and σ_- say that their boundaries are the same, so $\partial(\sigma_+-\sigma_-)=\partial(\sigma_+)-\partial(\sigma_-)=0$, so $\sigma_+-\sigma_-$ is a cycle.

Now, choose open neighbourhoods U_+ and U_- of the two hemispheres which deformation retract to the hemispheres, and whose intersection $U_-\cap U_+$ deformation retracts onto the equator. Then, the Mayer–Vietoris long exact sequence for U_1 and U_2 covering $S^k=U_1\cup U_2$ is:

$$\cdots\to H_k(U_+)\oplus H_k(U_-)\to H_k(S^k)\to H_{k-1}(U_+\cap U_-)\to H_{k-1}(U_+)\oplus H_{k-1}(U_-)\to\cdots$$

The subspaces are contractible, so their homology vanishes, leaving an isomorphism

$$0\to H_k(S^k)\overset{\partial}{\cong}H_{k-1}(U_+\cap U_-)\to 0$$

So, it suffices to prove that $\sigma_+-\sigma_-\in H_k(S^k)$ maps to a generator under the connecting homomorphism. Again, we take the relevant segment of the short exact sequence of chain complexes:

$$\begin{tikzcd}
 & C_k(U_-)\oplus C_k(U_+) \arrow[r, "(j_1)_*+(j_2)_*"] \arrow[d, "\partial"] & C_k(U_++U_-) \\
C_{k-1}(U_+\cap U_-) \arrow[r, "\big((i_1)_*{,}-(i_2)_*\big)"] & C_{k-1}(U_+)\oplus C_{k-1}(U_-) &
\end{tikzcd}$$

and chase $\sigma_+-\sigma_-$

$$\begin{tikzcd}
 & (\sigma_+,-\sigma_-) \arrow[r, mapsto, "(j_1)_*+(j_2)_*"] \arrow[d, mapsto, "\partial"] & \sigma_+-\sigma_- \\
\partial(\sigma_+)=\partial(\sigma_-) \arrow[r, mapsto, "\big((i_1)_*{,}-(i_2)_*\big)"] & \big(\partial(\sigma_+),-\partial(\sigma_-)\big) &
\end{tikzcd}$$

So, the connecting homomorphism maps $\sigma_+ - \sigma_-$ to $\partial(\sigma_+)$. Then, by construction, $U_+ \cap U_- \simeq S^k_+ \cap S^k_- \overset{\sigma_+}{\cong} \partial\Delta^k$, and $\partial(\mathrm{id}_{\Delta^k})$ is a generator of $H_{k-1}(\partial\Delta^k)$. ■

39.7.2 Jordan Curve Theorem

Recall that a *Jordan curve* is a simple closed curve in $\mathbb{R}^2$. Informally, the *Jordan curve theorem* states that

> *Every jordan curve splits the plane into two regions.*

One of the two regions is necessarily bounded and is thus interpreted as the *interior*, while the other region is necessarily unbounded and is thus interpreted as the *exterior*. The Jordan curve is then the boundary of each of these regions.

This is intuitively clear for any reasonably nice curve, but is difficult to interpret for, say, fractal curves.

Theorem 39.7.2 (Jordan Curve Theorem). *Let $\gamma : S^1 \rightarrowtail \mathbb{R}^2$ be an injective continuous map with image $C \subseteq \mathbb{R}^2$. Then,*

$$H_n(\mathbb{R}^2 \setminus C) = \begin{cases} \mathbb{Z}^2 & n = 0 \\ \mathbb{Z} & n = 1 \\ 0 & n > 1 \end{cases}$$

Because $\mathbb{R}^2 \setminus C$ is locally path-connected, the part $H_0(\mathbb{R}^2 \setminus C) = \mathbb{Z}^2$ is saying precisely that the complement of C has two path-connected components.

We translate the problem as follows: let $\mathbb{R}^2 \hookrightarrow \mathbb{R}^2 \cup \{\infty\} \cong S^2$ be the one-point compactification of $\mathbb{R}^2$. Then,

$$H_n(S^2 \setminus C) = \begin{cases} \mathbb{Z}^2 & n = 0 \\ 0 & n > 0 \end{cases}$$

To prove this, we need a few more lemmata.

Lemma 39.7.3. *Let $\kappa : [0,1] \rightarrowtail S^2$ be an injective continuous map with image $D \subseteq \mathbb{R}^2$. Then,*

$$H_n(S^2 \setminus D) = \begin{cases} \mathbb{Z} & n = 0 \\ 0 & n > 0 \end{cases}$$

Proof of Jordan Curve Theorem. We compute the homology of $S^2 \setminus C$. Let S^1_+ and S^1_- be the upper and lower semicircles in S^1 such that $S^1_+ \cap S^1_- = S^0$. Now, apply Mayer–Vietoris in reduced homology with

- $U_+ \coloneqq S^2 \setminus \gamma(S^1_+)$;
- $U_- \coloneqq S^2 \setminus \gamma(S^1_-)$;
- $X \coloneqq U_+ \cup U_2 = S^2 \setminus \gamma(S^0)$;
- $U_+ \cap U_1 = S^2 \setminus C$.

As S^1_+ and S^1_- are homeomorphic to [0,1], we have the homology groups of U_+ and U_- from the previous lemma, and in reduced homology, these vanish in all degrees. Also, X is the twice-punctured 2-sphere – the punctured 2-sphere is homeomorphic to the plane $\mathbb{R}^2$, and the punctured plane is homotopy equivalent to the circle S^1 – so $X \simeq S^1$, and these homology groups also vanish in degrees $n > 1$.

So, the long exact sequence is zero everywhere past $H_1(X)$. The end of the sequence is then given by:

$$\cdots \to 0 \to \tilde{H}_1(X) \to \tilde{H}_0(U_+ \cap U_-) \to 0 \oplus 0 \to 0 \to 0$$

so $\tilde{H}_0(S^2 \setminus C) = \tilde{H}_0(U_+ \cap U_-) \cong \tilde{H}_1(X) = \mathbb{Z}$. So, $H_0(S^2 \setminus C) = \mathbb{Z}^2$. ■

39.7.3 Relative Homology

Let $A \subseteq X$ be a subspace, and $\iota : A \hookrightarrow X$ be the canonical inclusion map. Then, there is an induced inclusion between chain groups, $C_n(A) \hookrightarrow C_n(X)$, and these inclusions assemble into a chain map $C_\bullet(A) \to C_n(X)$.

However, the induced map in homology $\iota_* : H_n(A) \to H_n(X)$ is not injective in general (nor surjective). For instance, if $A = S^1$ and $X = D^2$, then $H_1(A) = \mathbb{Z}$ cannot inject into $H_1(X) = 0$.

We define the group of *relative singular n-chains* $C_n(X,A)$ as the quotient $C_n(X)/C_n(A)$.

$$C_n(A) \hookrightarrow C_n(X) \to \frac{C_n(X)}{C_n(A)} =: C_n(X,A)$$

Because $C_\bullet(A)$ is a sub-chain complex of $C_\bullet(X)$, $C_\bullet(X,A)$ also inherits the structure of a chain complex called the *relative singular chain complex*, with the differentials $\partial : C_n(X,A) \to C_{n-1}(X,A)$ induced from $\partial : C_n(X) \to C_{n-1}(X)$. The quotient maps $C_n(X) \to C_n(X,A)$ also assemble into a chain map $C_\bullet(X) \to C_\bullet(X,A)$.

The *relative homology* of the pair (X,A) is then given by the homology of the relative singular chain complex:

$$H_n(X,A) := H_n\big(C_\bullet(X,A)\big)$$

- Call an n-chain $c \in C_n(X)$ a *relative n-cycle* if $\partial(c) \in C_{n-1}(A)$. For example, a singular n-simplex $\sigma : \Delta^n \to X$ is a relative n-cycle if the image of the boundary $\partial\Delta^n$ is contained in A.
- Call an n-chain $c \in C_n$ a *relative n-boundary* if it is homologous to some n-chain in A. That is, if there exists $a \in C_n(A)$ such that the difference $c - a = \partial w$ is the boundary of some $(n+1)$-chain $w \in C_{n+1}(X)$. Note that every relative n-boundary is a relative n-cycle since $\partial c = \partial a$.

By construction,

$$H_n(X,A) \cong \frac{\text{relative } n\text{-cycles}}{\text{relative } n\text{-boundaries}}$$

Intuitively, the relative homology $H_n(X,A)$ measures the homology of X with A "discarded".

Corollary 39.7.3.1. *There is an exact sequence in relative homology:*

$$\begin{tikzcd}
\cdots \arrow[r, "\partial"] & H_n(A) \arrow[r] & H_n(X) \arrow[r] & H_n(X,A) \arrow[dll, rounded corners, "\partial"] & \\
& H_{n-1}(A) \arrow[r] & H_{n-1}(X) \arrow[r] & H_{n-1}(X,A) \arrow[r, "\partial"] & \cdots
\end{tikzcd}$$

Proof. By construction, there is a short exact sequence of chain complexes

$$0 \to C_\bullet(A) \to C_\bullet(X) \to C_\bullet(X,A) \to 0$$

so the claim follows from Theorem 39.6.3. ■

We can also describe the connecting homomorphism explicitly:

If $[z] \in H_n(X,A)$ is represented by a relative cycle $z \in C_n(X)$, then the connecting homomorphism is defined by

$$\partial[z] = [\partial z]$$

Because z is a relative cycle, its boundary ∂z is contained in A, so this class $[\partial z]$ is an element of $H_{n-1}(A)$.

Theorem 39.7.4 (Excision)**.** *Let $Z \subseteq A \subseteq X$, with $\bar{Z} \subseteq A^\circ$. Then,*

$$H_n(X,A) \cong H_n(X \setminus Z, A \setminus Z)$$

Intuitively, the relative homology ignores the interior of A, so we may excise a portion Z, with minor restrictions.

Recall that a *topological manifold of dimension k* is a Hausdorff space such that every point has an open neighrbourhood homeomorphic to $\mathbb{R}^k$. Every smooth manifold is a topological manifold.

Corollary 39.7.4.1. *Let M be a k-dimensional topological manifold and let $x \in M$ be a point. Then,*

$$H_n(M, M \setminus x) \cong H_n(\mathbb{R}^k, \mathbb{R}^k \setminus *) \cong \begin{cases} \mathbb{Z} & n = k \\ 0 & n \neq k \end{cases}$$

That is, relative singular homology is able to detect the dimension of a manifold.

Proof. Let $U \ni x$ be an open neighbourhood of x homeomorphic to $\mathbb{R}^k$. Then, Excision gives the first isomorphism with $X = M$, $A = M \setminus x$, and $Z = M \setminus U$. For the second isomorphism, consider the long exact sequence of the pair $(\mathbb{R}^k, \mathbb{R}^k \setminus *)$. ■

Corollary 39.7.4.2 (Invariance of Domain II)**.** *Let $U \subseteq \mathbb{R}^k$ and $V \subseteq \mathbb{R}^\ell$ be non-empty open subsets. If $U \cong V$, then $k = \ell$.*

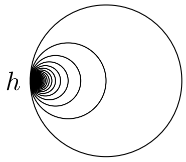
A pair (X,A) is *good* if:

(*i*) $A \subseteq X$ is closed;

(*ii*) there exists an open neighbourhood $V \supseteq A$ which deformation retracts onto A.

Example. If X is a CW-complex, then (X,A) is good for any subcomplex A. △

Example. Consider the Hawaiian earring H with $h \in H$ the distinguished point where all the circles meet.

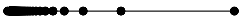


Then, (H,h) is not a good pair, since any open neighbourhood of h contains infinitely many circles and cannot be contractible; and in particular, cannot deformation retract to h. △

Theorem 39.7.5. *Let (X,A) be a good pair. Then, the quotient map $X \to X/A$ induces isomorphisms*

$$H_n(X,A) \cong H_n(X/A, A/A) \cong \tilde{H}_n(X/A)$$

Example. Let $X = [0,1]$ be the interval, and $A = \left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots, 0\right\} \subseteq X$.



(X,A) is not a good pair, because $(X/A, A/A)$ would then also be a good pair. But $(X/A, A/A) \cong (H,h)$. △

Example. $(\Delta^k, \partial\Delta^k)$ is a good pair for any k, so

$$H_n(\Delta^k, \partial\Delta^k) \cong \tilde{H}_n(S^k) \cong \begin{cases} \mathbb{Z} & n = k \\ 0 & n \neq k \end{cases}$$

△

39.8 Degrees

Let $f : S^k \to S^k$ be a continuous map. Then, the induced map in kth homology,

$$f_* : \tilde{H}_k(S^k) \to \tilde{H}_k(S^k)$$

is a group homomorphism $\mathbb{Z} \to \mathbb{Z}$. Such a homomorphism is determined entirely by the image of the generator $1 \mapsto d$, and thus acts by multiplication by d. This integer d is called the *degree* of f, written as $\deg(f)$.

Lemma 39.8.1. *Let $f,g : S^k \to S^k$. Then,*

(*i*) $\deg(\mathrm{id}_{S^k}) = 1$*;*

(*ii*) $\deg(g \circ f) = \deg(g) \cdot \deg(f)$*;*

(*iii*) *if $f \simeq g$, then* $\deg(f) = \deg(g)$*;*

(*iv*) *if f is a homotopy equivalence, then* $\deg(f) = \pm 1$*;*

(*v*) *if f is not surjective, then* $\deg(f) = 0$*.*

Proof.

(*i*) The identity induces the identity in homology.

(*ii*) $(g \circ f)_* = g_* \circ f_*$.

(*iii*) By homotopy invariance, f and g induce the same maps in homology, so they have the same degree.

(*iv*) If f is a homotopy equivalence, then it induces an isomorphism in homology. The only possible images for the generator 1 are then the generators 1 and -1.

(*v*) Let $x \in S^k$ be outside the image of f. Then, f factors as

$$\begin{tikzcd}
S^k \arrow[rr, "f"] \arrow[dr, "f"'] & & S^k \\
& S^k \setminus x \arrow[ur, hook, "\iota"'] &
\end{tikzcd}$$

Then in reduced homology, f_* factors through $\tilde{H}_k(S^k \setminus x) = 0$ since $S^k \setminus x$ is contractible.

$$\begin{tikzcd}
\mathbb{Z} \arrow[rr, "f_*"] \arrow[dr] & & \mathbb{Z} \\
& 0 \arrow[ur] &
\end{tikzcd}$$

so $\deg(f) = 0$.

■

Example. Consider an endomorphism on the 0-sphere:

$$\begin{tikzcd}
S^0 \arrow[r, "f"] \arrow[d, phantom, "\cong" rotate=90] & S^0 \arrow[d, phantom, "\cong" rotate=90] \\
\{a,b\} & \{a,b\}
\end{tikzcd}$$

There are only four possible maps:

$$(a,b) \mapsto \begin{cases} (a,a) \\ (b,b) \\ (a,b) \\ (b,a) \end{cases}$$

The first two maps are not surjective, so they have degree $\deg(f) = 0$. The third map is the identity, so in this case, $\deg(f) = 1$. For the final map, consider the reduced homology:

$$\begin{aligned} \tilde{H}_n(S^0) &= \ker\Big(H_0(S^0) \xrightarrow{\pi} H_0(*)\Big) \\ &= \ker\Big(\mathbb{Z}a \oplus \mathbb{Z}b \xrightarrow{\pi} \mathbb{Z}*\Big) \end{aligned}$$

$a - b$ is a generator of $\tilde{H}_0(S^0)$ since $\pi(a-b) = * - * = 0$, so $\tilde{H}_0(S^0) = \mathbb{Z}(a-b)$. Then,

$$f_*(a-b) = b - a = -(a-b)$$

so $\deg(f) = -1$. △

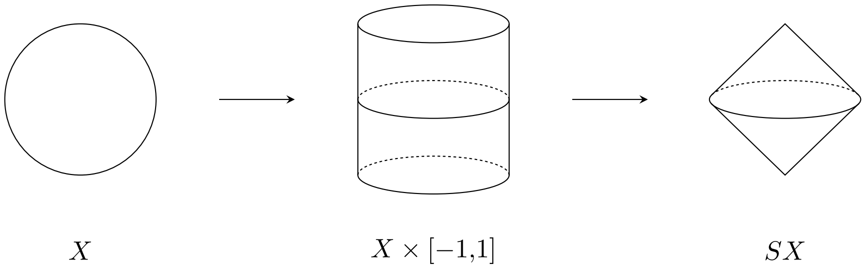
Example. Consider S^1 as a subset of the complex numbers, and let $f : S^1 \to S^1$ be defined by $z \mapsto z^n$ for some integer n. The loop $\sigma : [0,1] \to S^1$ defined by $t \mapsto e^{2\pi i t}$ represents a generator of $H_1(S^1) \cong \tilde{H}_1(S^1)$. Then, $f_*\big([\sigma]\big) = [f \circ \sigma]$ is represented by the loop $t \mapsto e^{2\pi i n t}$ which is homologous to $n\sigma$, so $\deg(f) = n$. △

Theorem 39.8.2. *Let $k \geq 1$. For every integer $n \in \mathbb{Z}$, there exists a map $f : S^k \to S^k$ of degree n.*

Proof. The *suspension* SX of a space X is the space

$$X \times [-1,1]/\sim$$

where $(x,1) \sim (y,1)$ and $(x,-1) \sim (y,-1)$ for all $x,y \in X$.



Taking the upper and lower cones, plus some extra space for overlap:

$$C_+X := X \times (-\varepsilon,1]/X \times 1 \qquad C_-X := X \times [-1,\varepsilon)/X \times -1$$

we have two open subspaces that jointly cover X, and their intersection deformation retracts to X. These subspace are also contractible, so their homology vanishes, and Mayer–Vietoris gives an isomorphism

$$H_{k+1}(SX) \overset{\partial}{\cong} H_k(X)$$

Then, any map $f : X \to Y$ induces a map $Sf : SX \to SY$, and we have a commutative square

$$\begin{tikzcd} H_{k+1}(SX) \arrow[r, "\partial"] \arrow[d, "Sf_*"] & H_k(X) \arrow[d, "f_*"] \\ H_{k+1}(SY) \arrow[r, "\partial"] & H_k(Y) \end{tikzcd}$$

Applying this with $X = Y = S^{k-1}$, and noticing the suspension of S^{k-1} is homeomorphic to S^k, we have $\deg(Sf) = \deg(f)$. We can then reduce inductively to $k = 1$. ■

39.8.1 Antipodes

Lemma 39.8.3. *Let $S^k \subseteq \mathbb{R}^{k-1}$ be the unit circle. Let $f : S^k \to S^k$ be the reflection in a hyperplane through the origin. Then $\deg(f) = -1$.*

Proof. Let $H \subseteq \mathbb{R}^{k-1}$ be the fixed hyperplane. It splits the sphere S^k into two hemispheres S^k_+ and S^k_-. Fix some homeomorphism $\sigma_+ : \Delta^k \to S^k_+$, and set $\sigma_- = f \circ \sigma_+$.

Because f is the identity on H, the composition

$$\partial\Delta^k \xrightarrow{\sigma_+} S^k_+ \cap S^k_- \xrightarrow{(\sigma_-)^{-1}} \partial\Delta^k$$

is the identity:

$$\begin{aligned}(\sigma_-)^{-1} \circ \sigma_+ &= (f \circ \sigma_+)^{-1} \circ \sigma_+ \\ &= (\sigma_+)^{-1} \circ \sigma_+ \\ &= \mathrm{id}_{\partial\Delta^k}\end{aligned}$$

so Corollary 39.7.1.1 applies. So, $[\sigma_+ - \sigma_-]$ generates $\tilde{H}_k(S^k)$, and

$$f_*\big([\sigma_+ - \sigma_-]\big) = [f \circ \sigma_+] - [f \circ \sigma_-] = [\sigma_-] - [\sigma_+] = -[\sigma_+ - \sigma_-]$$

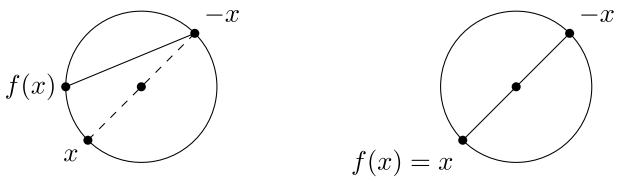
so $\deg(f) = -1$. ■

Theorem 39.8.4. *Let $T : \mathbb{R}^{k+1} \to \mathbb{R}^{k+1}$ be a orthogonal linear transformation. It restricts to a homeomorphism $f : S^k \to S^k$. Then $\deg(f) = \det(T)$.*

Corollary 39.8.4.1. *Let $f : S^k \to S^k$ be the antipodal map $x \mapsto -x$. Then, $\deg(f) = (-1)^{k+1}$.*

Corollary 39.8.4.2. *If $f : S^k \to S^k$ has no fixed points, then $\deg(f) = (-1)^{k+1}$.*

Proof. We show that f is homotopic to the antipodal map. The line through $f(x)$ and $-x$ passes through the origin if and only if $f(x)$ and $-x$ are antipodal. That is, if $f(x) = x$.



Since f has no fixed points, this cannot be the case, so the line $tf(x) + (1-t)(-x)$ connecting the two points, parametrised by t, is never zero. So, dividing by its norm yields an element of S^1, so

$$(t,x) \mapsto \frac{tf(x) + (1-t)(-x)}{\|tf(x) + (1-t)(-x)\|}$$

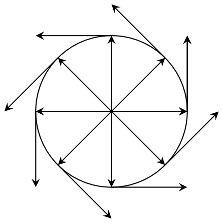
is a homotopy from the antipodal map to f. ■

Recall that a vector field on S^k is a continuous map $v : S^k \to \mathbb{R}^{k+1}$. A vector field is a *tangent vector field* if $v(x)$ is orthogonal to x for all $x \in S^k$.

Example. The constant map $v(x) = 0$ is a tangent vector field on every sphere since the zero vector is orthogonal to every vector. △

We are interested in tangent vector fields that vanish nowhere.

Example. On the 1-sphere,



is a non-vanishing tangent vector field. This construction generalises to other odd-dimensional spheres as:

$$x = (x_1, \ldots ,x_{2m}) \mapsto v(x) - (-x_2,x_1, - x_4,x_3, \ldots , - x_{2m},x_{2m-1})$$

△

Do there exist non-vanishing tangent vector fields on *even*-dimensional spheres?

Corollary 39.8.4.3 (Hairy Ball Theorem)**.** *Every tangent vector field on an even-dimensional sphere vanishes at some point.*

Proof. Suppose for a contradiction that $v : S^k \to \mathbb{R}^{k+1}$ is a non-vanishing tangent vector field with k even. Because $v(x)$ and x are non-zero and orthogonal, they are linearly independent, so the map

$$S^k \times [0,1] \ni (x,t) \mapsto \cos(\pi t)x + \sin(\pi t)v(x) \in \mathbb{R}^{k+1}$$

cannot vanish. So, we can divide by its norm to obtain a homotopy $S^k \times [0,1] \to S^k$ from the identity, at $t = 0$, to the antipodal map, at $t = 1$.

But, this is impossible, as identity has degree 1, while the antipodal map has degree $(-1)^{k+1} = -1$. ■

39.8.2 Local Degrees

Let $k \geq 1$, and let $f : S^k \to S^k$ be a continuous map, and let $y \in S^k$ such that its preimage $f^{-1}(y) = \{x_1, \ldots ,x_n\}$ is finite. We may choose disjoint open balls $U_i \subseteq S^k$ around the x_i.

So, f induces a map of pairs $(U_1,U_1 \setminus x_i) \hookrightarrow (S^k,S^k \setminus y)$, and hence by excision, a map in homology

$$H_k(S^k,S^k \setminus x_i) \cong H_k(U_i,U_i \setminus x_i) \xrightarrow{f_*} H_k(S^k,S^k \setminus y)$$

Now, recall the long exact sequence in relative homology:

$$\cdots \to H_n(A) \to H_n(X) \to H_n(X,A) \to H_{n-1}(A) \to \cdots$$

We have $A = S^k \setminus x_i$, which is contractible, so the outer terms vanish and we have an isomorphism

$$H_k(S^k) \cong H_k(S^k,S^k \setminus x_i)$$

Similarly, on the right we have $A = S^k \setminus y$, which is also contractible, so we again have an isomorphism

$$H_k(S^k) \cong H_k(S^k,S^k \setminus y)$$

So, the whole composition

$$\begin{tikzcd}
H_k(S^k,S^k \setminus x_i) \arrow[r, "\cong"] & H_k(U_i,U_i \setminus x_i) \arrow[r, "f_*"] & H_k(S^k,S^k \setminus y) \\
H_k(S^k) \arrow[u, "\cong"] \arrow[rr, "f\big|_{x_i}"] & & H_k(S^k) \arrow[u, "\cong"]
\end{tikzcd}$$

can be viewed as an endomorphism

$$f\big|_{x_i} : H_k(S^k) \to H_k(S^k)$$

for each x_i. $H_k(S^k) \cong \mathbb{Z}$, so these maps are given by multiplication by some integer, called the *local degree* of f at x_i, denoted by $\deg(f|_{x_i})$.

Theorem 39.8.5. *In the situation above,*

$$\deg(f) = \sum_{i=1}^{n} \deg(f|_{x_i})$$

Example. Let $p(z) \in \mathbb{C}[z]$ be a non-constant polynomial interpreted as a map $\mathbb{C} \to \mathbb{C}$. It extends to a continuous map on the one-point compactification

$$f_p : S^2 \cong \mathbb{C} \cup \{\infty\} \to \mathbb{C} \cup \{\infty\} \cong S^2$$

Let $w \in \mathbb{C}$ be such that $p'(z_i) \neq 0$ for all $z_i \in f_p^{-1}\big[\{w\}\big]$. Such a w always exists as p' vanishes at finitely many points. So, f_p is invertible around each z_i, so $\deg(f_p|_{z_i}) = \pm 1$. In fact, since polynomials are orientation preserving, we must have $\deg(f_p|_{z_i}) = 1$. So,

$$\deg(f_p) = \sum_{i=1}^{n} \deg(f_p|_{z_i}) = n = \deg(p)$$

△

39.9 Manifolds

We have already recalled the notion of a (topological) manifold: a Hausdorff space such that every point has an open neighbourhood homeomorphic to $\mathbb{R}^k$, for some fixed k called the *dimension* of the manifold.

Example.

- Euclidean space $\mathbb{R}^k$ is itself a k-manifold.
- The sphere S^k is a k-manifold: for any point that isn't the north pole, take the open neighbourhood to be the entire sphere minus the north pole, which is homeomorphic to $\mathbb{R}^k$ via sterographic projection. For the north pole, take the open neighbourhodo to be the entire sphere minus the south pole, which is again homeomorphic to $\mathbb{R}^k$ via sterographic projection.
- Any open subspace of a k-manifold is itself a k-manifold.
- The torus $\mathbb{T}^2$, Klein bottle $\mathbb{K}$, and $\mathbb{RP}^2$ are all 2-manifolds.

△

Example. The interval [0,1] is not a manifold, since no open neighbourhood of 0 or 1 is homeomorphic to $\mathbb{R}^k$ for any k. (Instead, it is a *manifold with boundary*, which we will not discuss.) △

Example. A 0-manifold is any space with the discrete topology (i.e. every set is open): for an open neighourhood of a point x to be homeomorphic to $\mathbb{R}^0 \cong \{*\}$, it must be a singleton set, namely $\{x\}$, so the topology is discrete. △

Theorem 39.9.1. *Up to homeomorphism, the only connected compact* 1*-manifold is* S^1*.*

Proof sketch. Let M be a connected compact 1-manifold. By assumption, there are open subsets $U_1, \ldots, U_n \subseteq M$ all homeomorphic to $\mathbb{R}^1$. Choose n to be minimal, noting that $n > 1$, or else $M = U_1 \cong \mathbb{R}$ is not compact. Since M is connected, we may assume, after relabelling if necessary, that $U_1 \cap U_2 \neq \varnothing$. If $\pi_0(U_1 \cap U_2) = *$, then $U_1 \cup U_2 \cong \mathbb{R}^1$, contradicting minimality.

Next, one shows that the only other possiblity is $\pi_0(U_1 \cap U_2) = * \amalg *$, in which case $M = U_1 \cup U_2 \cong S^1$. (In particular, $n = 2$.) ■

Recall that a *covering* of a topological space X is a map $p : \tilde{X} \to X$ such that for every point $x \in X$, there exists an open neighbourhood $U_x \subseteq X$ of x whose preimage

$$p^{-1}[U_x] = \bigsqcup_{i \in I_x} V_i$$

is a disjoint union of open sets $(V_i)_{i \in I_x}$, and the restriction $p\big|_{V_i} : V_i \to U_x$ is a homeomorphism for every $i \in I_x$. Such an open set U_x is said to be *evenly covered* by p, and the open sets V_i are called the *sheets* of the covering. If $p : \tilde{X} \to X$ is a covering, then the pair $(\tilde{X}, p)$ is called a *covering space* or *cover* of X, and X is said to be the *base* of the covering.

If X is connected, then the indexing set I_x does not depend on X. If $|I| = n$, then we say that $p : \tilde{X} \to X$ is an *n-fold* or *n-sheeted* cover.

Intuitively, a covering is a surjective map that acts locally like a projection of multiple copies of a space onto itself.

Example. For any $k \in \mathbb{N}$, the map $p_k : S^1 \to S^1$ defined by $z \mapsto z^k$ is a covering map. The preimage of the arc of length $\frac{1}{k}$ centred on z is the collection of arcs that each cover $\frac{1}{k}$th of the circle, centred on each root of z, and these arcs are disjoint as there are exactly k such roots evenly spaced along the circle.

This covering is also an k-fold covering map, as the fibre of any point $z = \exp(2\pi i t)$ consists of k many kth roots of z – namely $\exp\bigl(2\pi i(t + j)/k\bigr)$, for $0 \leq j < k$. △

Example. The map $p_\infty : \mathbb{R} \to S^1$ defined by $x \mapsto \exp(2\pi i x)$ is a covering map. Given a point $z = \exp(2\pi i t) \in S^1$, we take the open neighbourhood $U = \bigl\{\exp(2\pi i s) : |s - t| < \varepsilon\bigr\}$ for some $0 < \varepsilon < 1$, which has preimage

$$\begin{aligned} p^{-1}[U] &= \bigcup_{j \in \mathbb{Z}} \{s + i : |s - t| < \varepsilon\} \\ &= \bigsqcup_{j \in \mathbb{Z}} V \end{aligned}$$

△

Lemma 39.9.2. *Let X be a k-manifold, and let $p : Y \to X$ be a covering space. Then, Y is also a k-manifold.*

Proof. Let $y_1, y_2 \in Y$ be distinct points, with images $x_1 = p(y_1) = x_1$ and $x_2 = p(y_2)$. If $x_1 \neq x_2$, then y_1 and y_2 are in different fibres: there exist disjoint open neighbourhoods $U_i \subseteq X$ of x_i since X is Hausdorff; then, $V_i := p^{-1}[U_i]$ are disjoint open neighbourhoods of the y_i.

If $x_1 = x_2$, then y_1 and y_2 are in the same fibre, but different sheets, since $y_1 \neq y_2$ and coverings are homeomorphic on sheets: choose an evenly covered neighbourhood $U \subseteq X$ of $x_1 = x_2$; then, y_1 and y_2 are in disjoint open sheets of U.

So, Y is Hausdorff.

Let $y \in Y$ have image $x = p(y)$. By assumption, there exists an evenly covered neighbourhood $U \subseteq X$ of x, so y is in a sheet V_y of U. Since U is open in X, it is also a manifold, so there is an open neighbourhood $V \subseteq U$ of x homeomorphic to $\mathbb{R}^k$. Then, $p^{-1}[V] \cap V_y \cong V \cong \mathbb{R}^k$ is an open neighbourhood of y. ■

Given a map $f : (Y{,}y) \to (X{,}x)$ between pointed spaces and a covering $p : (\tilde{X}{,}\tilde{x}) \to (X{,}x)$, when does a lift $\tilde{f} : (Y{,}y) \to (\tilde{X}{,}\tilde{x})$ exist?

$$\begin{tikzcd}
 & (\tilde{X},\tilde{x}) \arrow[d, "p"] \\
(Y,y) \arrow[ur, dashed, "\tilde{f}"] \arrow[r, "f"'] & (X,x)
\end{tikzcd}$$

Lemma 39.9.3. *If X, $\tilde{X}$, and Y are connected manifolds, then the lift $\tilde{f}$ exists if and only if $f_*\big(\pi_1(Y{,}y)\big) \subseteq p_*\big(\pi_1(\tilde{X}{,}\tilde{x})\big)$. Moreover, such a lift is unique.*

Two coverings $p : Y \to X$ and $q : Z \to X$ are *isomorphic* if they factor through each other. That is, there exist maps f and g such that

$$p = q \circ f \qquad \text{and} \qquad q = p \circ g$$

This also implies that f and g are inverse, so equivalently, p and q are isomorphic if there exists a homeomorphism $h : Y \to Z$ such that

$$\begin{tikzcd}
Y \arrow[rr, "h", "\cong"'] \arrow[dr, "p"'] & & Z \arrow[dl, "q"] \\
 & X &
\end{tikzcd}$$

commutes.

Example. p_2 is isomorphic to p_{-2} via the homeomorphism $h(z) = z^{-1}$. △

Example. p_2 and p_3 are not isomorphic, as one is a 2-fold covering, and the other is a 3-fold covering. △

Let $p : \tilde{X} \to X$ be a covering of X. A *deck transformation* is a homeomorphism $\tau : \tilde{X} \to \tilde{X}$ such that $p \circ \tau = p$. That is, τ witnesses an automorphism of p. The set of all deck transformations of a cover p is denoted $\operatorname{Deck}(p)$, and has group structure under composition.

Example. The map $z \mapsto -z$ is a deck transformation for p_2. △

Theorem 39.9.4 (Galois Theory for Covering Spaces)**.** *Let X be a connected manifold. Then, there is a bijection*

$$\{\textit{connected covering spaces of } X\}\big/{\cong} \qquad \leftrightarrow \qquad \{\textit{subgroups of } \pi_1(X)\}\big/\textit{conjugacy}$$

This bijection sends a covering space $p : \tilde{X} \to X$ to the conjugacy class of subgroups $p_*\big(\pi_1(Y)\big) \subseteq \pi_1(X)$. The trivial subgroup corresponds to the *universal cover* $\overline{X} \to X$ – the connected and simply connected covering space of X unique up to isomorphism. Moreover, $\pi_1(X)$ is the group $\operatorname{Aut}_X(\overline{X})$ of isomorphisms, or *deck transformations*, $\overline{X} \to \overline{X}$.

Example. Since p_∞ is a covering space, and $\mathbb{R}$ is connected and simply connected, it is the universal cover. The group of deck transformations is then an infinite cyclic group generated by the map $x \mapsto x+1$. Hence, (if we didn't already know) $\pi_1(S^1) = \mathbb{Z}$. △

The index of the subgroup in $\pi_1(X)$ corresponds to the number of sheets in the covering. So, for example, the n-sheeted covers $p_n : S^1 \to S^1$ correspond to the subgroups $n\mathbb{Z} \leq \mathbb{Z}$.

39.9.1 Orientations

Let V be a non-zero finite-dimensional real vector space. Recall that two bases define the same orientation if the change of basis transformation from one to the other has positive determinant. This defines an equivalence relation whose two equivalence classes are the two possible orientations of V.

Conversely, we can think of invertible transformations with positive determinant as *orientation-preserving*, and those with negative determinant as *orientation-reversing.*

Example. The determinant of any reflection is -1, so reflections reverse orientation. △

Since manifolds locally appear as finite-dimensional vector spaces, we should expect that orientations can be generalised, at least locally, to manifolds.

Recall that, given a k-manifold M and any point $x \in M$, then the relative homology group at x is given by Corollary 39.7.4.1 to be infinite cyclic, and can therefore be identified with:

$$H_k(M,M \setminus x) \cong \mathbb{Z} \cong H_{k-1}(S^{k-1})$$

The choice of generator of $\mathbb{Z}$ is exactly analogous to the choice of basis in a vector space.

Vector Spaces	k-Manifolds
V	$H_k(M,M \setminus x) \cong \mathbb{Z}$
basis	generator
linear transformation	endomorphism of $\mathbb{Z}$
orientation preserving (det > 0)	deg $= 1$
orientation reversing (det < 0)	deg $= -1$

A *local orientation* of M at x is a choice of one of the two generators of $H_k(M,M \setminus x) \cong \mathbb{Z}$.

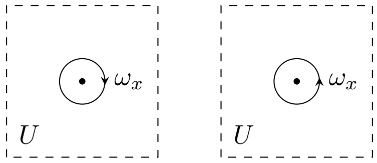
Example. Let M be a 2-manifold, and let $U \ni x$ be an open neighbourhood of x that is homeomorphic to $\mathbb{R}^2$. The long exact sequence for relative homology of the pair $(U, U \setminus x)$ is:

$$\cdots \to H_{n+1}(U) \to H_{n+1}(U,U \setminus x) \to H_n(U \setminus x) \to H_n(U) \to \cdots$$

The open neighbourhood $U \cong \mathbb{R}^2$ is contractible, so the outer terms vanish in degrees $n \geq 1$, and we have an isomorphism in the middle. Also, $U \setminus x$ deformation retracts onto a small circle around x, so,

$$H_2(U,U \setminus x) \cong H_1(U \setminus x) \cong H_1(S^1)$$

Choosing a local orientation ω_x at x therefore amounts to choosing in which direction to traverse this circle:



△

In the above, choosing a local orientation at x also determines the local orientation of every other point y contained in a small neighbourhood around x. In some manifolds, say $\mathbb{R}^2$, this extends to the entire space: we can pick the clockwise or counterclockwise orientation globally.

This is not true on the open Möbius band; if we choose a local orientation and try to "transport" it along a loop around the band, we end up with the opposite orientation after having traversed the band once.

Let $B \subseteq M$ be a subset of a k-manifold. We say that B is a *small open* (resp. *closed*) *ball* if it has an open neighbourhood $U \supseteq B$ homeomorphic to $\mathbb{R}^k$ via, say f, such that $f(B)$ is an open (resp. closed) ball of finite radius.

$$\begin{tikzcd}
U \arrow[r, "f", "\cong"'] \arrow[d, phantom, "\cup" description] & \mathbb{R}^k \arrow[d, phantom, "\cup" description] \\
B \arrow[r, "f\big|_B", "\cong"'] & B(x,r)
\end{tikzcd}$$

The point of this definition is that by excision, then applying the homeomorphisms above:

$$H_k(M,M \setminus B) \cong H_k(U,U \setminus B) \overset{f_*}{\cong} H_k\big(\mathbb{R}^k,\mathbb{R}^k \setminus B(x,r)\big)$$

Then, the long exact sequence in relative homology for the pair $\big(\mathbb{R}^k,\mathbb{R}^k \setminus B(x,r)\big)$ is:

$$\cdots \to H_k(\mathbb{R}^k) \to H_k\big(\mathbb{R}^k,\mathbb{R}^k \setminus B(x,r)\big) \to H_{k-1}\big(\mathbb{R}^k \setminus B(x,r)\big) \to H_{k-1}(\mathbb{R}^k) \to \cdots$$

but $\mathbb{R}^k$ is contractible, so the outer terms vanish, and we have the isomorphism

$$H_k\big(\mathbb{R}^k,\mathbb{R}^k \setminus B(x,r)\big) \cong H_{k-1}\big(\mathbb{R}^k \setminus B(x,r)\big)$$

and finally, $\mathbb{R}^k \setminus B(x,r)$ deformation retracts to the boundary $\partial B(x,r)$, giving

$$H_{k-1}\big(\mathbb{R}^k \setminus B(x,r)\big) \cong H_{k-1}\big(\partial B(x,r)\big) \cong H_{k-1}(S^{k-1}) \cong \mathbb{Z}$$

which is infinite cyclic. Chaining these all together, we have

$$H_k(M,M \setminus B) \cong H_{k-1}\big(\partial B(x,r)\big)$$

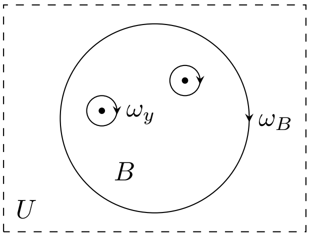
so we can think of a generator here as an orientation of the boundary of B.

Now, for every point $y \in B$, we then get an induced local orientation through the canonical inclusion map

$$H_k(M,M \setminus B) \xrightarrow[\cong]{\mathrm{sp}_y} H_k(M,M \setminus y)$$

because $M \setminus B \subseteq M \setminus y$.

A family of local orientations $(\omega_y)_{y \in B}$ is *consistent* if there is a generator $\omega_B \in H_k(M,M \setminus B)$ such that $\mathrm{sp}_y(\omega_B) = \omega_y$ for all $y \in B$.



An *orientation* of a k-manifold M is a family of local orientations $(\omega_x)_{x \in M}$ which are locally consistent. That is, for all $x \in M$, there exists a small open ball B such that the local orientations $(\omega_y)_{y \in B}$ are consistent.

M is *orientable* if it admits an orientation and is *non-orientable* otherwise.

Example. The k-sphere $M = S^k$ is orientable. For $k = 0$, S^0 is

Choose a generator $\omega \in H_k(S^k)$. Then, for each point $x \in S^k$, the long exact sequence in relative homology for the pair $(S^k, S^k \setminus x)$ is

$$\cdots \to H_k(S^k \setminus x) \to H_k(S^k) \to H_k(S^k, S^k \setminus x) \to H_{k-1}(S^k \setminus x) \to \cdots$$

$S^k \setminus x$ is contractible, so the outer terms vanish, so the map in the centre is an isomorphism:

$$H_k(S^k) \xrightarrow{f_x} H_k(S^k, S^k \setminus x)$$

and hence this map induces local orientations $\omega_x \coloneqq f_x(\omega)$ at each point $x \in S^k$. These local orientations are also locally consistent, since the map above factors through $H_k(S^k, S^k \setminus B)$ via inclusion for any small open ball B around x. △

We define the *orientation bundle* $\tilde{M}$ to be the set of pairs (x, ω_x), where $x \in M$ and ω_x is a local orientation at x. This set is equipped with the map $\pi : \tilde{M} \to M$ that projects to the first coordinate. We can put a topology on this set using this map.

If $B \subseteq M$ is a small open ball, then we have seen that there are precisely two collections of local orientations $(\omega_y)_{y \in B}$ that are locally consistent in B. In other words,

$$\begin{aligned} \pi^{-1}[B] &= \big(y, \mathrm{sp}_y(\omega_B)\big)_{y \in B} \sqcup \big(y, \mathrm{sp}_y(-\omega_B)\big)_{y \in B} \\ &\cong B_+ \sqcup B_- \end{aligned}$$

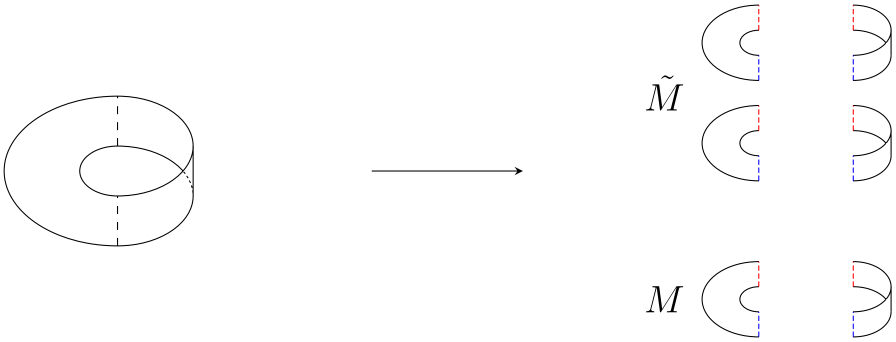
where $B_\pm \overset{\pi}{\cong} B$. We define the topology on $\tilde{M}$ to be generated by the sets $B+$ and B_- for all small open balls $B \subseteq M$.

From this, we have that $\pi : \tilde{M} \to M$ is a 2-fold covering, as, by construction, the preimage of any small open ball consists of two open sets homeomorphic to B under π.

Example. Let M be the open Möbius band

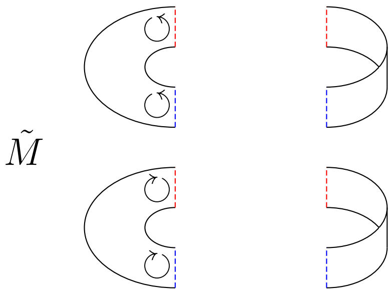
$$M \coloneqq [0,1] \times (0,1) / \sim$$

where $(0,y) \sim (1,1-y)$ for all $y \in (0,1)$.

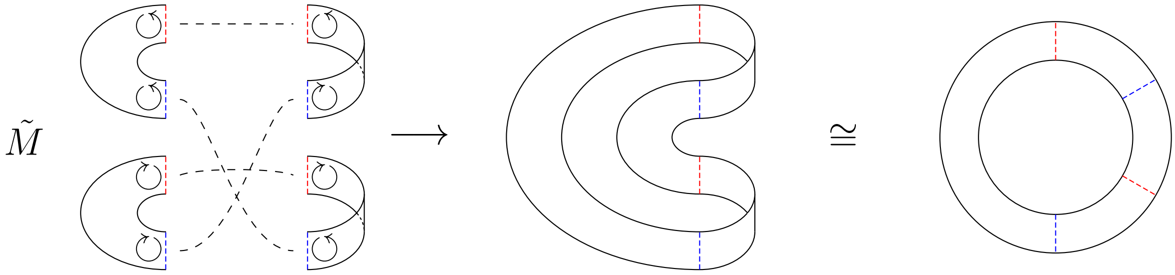


If we cut the Möbius band along these two dashed lines, then we can take the two halves, plus some extra space to overlap, to be two small open balls that jointly cover M. $\tilde{M}$ is a two-fold cover, so we have the setup on the right, where the two copies of M in $\tilde{M}$ have different local orientations.

Pick some orientation in the upper left piece, next to the red boundary. We can transport this orientation down to the blue boundary. Also, the lower left piece must have opposite orientation.



Now, the red boundary at the top must glue to one of the other red boundaries; suppose it glues to the upper right piece. So, this piece shares the same local orientation, but this orientation is reversed by the twist as we reach the blue boundary. Again, the piece below must have opposite local orientations, so altogether, we have:



So $\tilde{M} \cong S^1 \times (0,1)$. △

Lemma 39.9.5. *Giving an orientation of M is equivalent to giving a continuous section to π.*

Proof. Giving a section $\omega : M \to \tilde{M}$ (not necessarily continuous) amounts to choosing, for each $x \in M$, a local orientation ω_x at x, since the first component must be the identity.

The map ω is continuous if and only if for each small open ball $B \subseteq M$, $\omega^{-1}[B_+]$ and $\omega^{-1}[B_-]$ are open in M, where $B_+ \sqcup B_- := B \sqcup B = \pi^{-1}[B]$.

Since these preimages are disjoint and jointly cover B, this condition is equivalent to $\omega(B) = B_+$ or $\omega(B) = B_-$,which means precisely that the local orientations $(\omega_y)_{y \in B}$ are consistent. ■

Theorem 39.9.6. *Let M be a k-manifold. Then, its orientation bundle $\tilde{M}$ is an orientable k-manifold. Furthermore, this orientation is natural, and the deck transformation $(x,\omega_x) \mapsto (x, -\omega_x)$ reverses this orientation.*

Let $x \in M$ and choose a local orientation ω_x. A path in M from x to y has a unique lift to $\tilde{M}$ starting at (x,ω_x) and ending at (y,ω_y) for some ω_y. In other words, this path determins a unique local orientation at y.

Corollary 39.9.6.1. *If M is a connected manifold, then:*

- *either, $\tilde{M}$ is connected, and M is non-orientable;*
- *or, $\tilde{M} \cong M \sqcup M$, and M admits precisely two orientations.*

Example. We have seen that the orientation bundle of the Möbius band is homeomorphic to $S^1 \times (0,1)$, which is connected. Hence, the Möbius band is not orientable. △

Corollary 39.9.6.2. *Any simply connected manifold is orientable.*

Theorem 39.9.7. *Let $k \geq 1$. Then, $\mathbb{RP}^k$ is orientable if and only if k is odd.*

39.9.2 Surfaces

We define a *surface* here to mean a compact connected 2-manifold. (In particular, a surface is non-empty.)

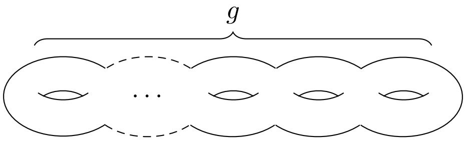
Example. S^2, $\mathbb{T}^2$, $\mathbb{K}$, and $\mathbb{RP}^2$ are surfaces. △

Let S_1 and S_2 be two surfaces, and let $D_i \subseteq S_i$ be two small closed disks. We can glue $S_1 \setminus D_1^\circ$ and $S_2 \setminus D_2^\circ$ along $\partial D_1 \cong \partial D_2$. The resulting space is called the *connected sum* $S_1 \# S_2$.

The connected sum operation is associative, commutative, and unital on the set of homeomorphism types of surfaces, with the unit being given by the 2-sphere S^2.

Example. The g-holed torus Σ_g can be obtained as the connected sum of g tori:

$$\Sigma_g = \underbrace{\mathbb{T}^2 \# \cdots \# \mathbb{T}^2}_{g}$$



△

Example. $\mathbb{K} \cong \mathbb{RP}^2 \# \mathbb{RP}^2$. △

Up to homeomorphism, every surface is one of:

(*i*) Σ_g, $g \geq 0$: the integer g is called the *genus* of the surface;

(*ii*) N_h, $h \geq 1$; the integer h is called the *non-orientable genius* of the surface.

Example. The torus $\mathbb{T}^2$ is of the first type, and has genus 1. The real projective plane $\mathbb{RP}^2$ is of the second type, and has non-orientable genus 1. △

Within each subtype, orientable and non-orientable genus behave well with respect to connected sums. That is,

$$\Sigma_a \# \Sigma_b \cong \Sigma_{a+b} \qquad N_a \# N_b \cong N_{a+b}$$

However, the connected sum of the torus $\mathbb{T}^2$ and the real projective plane $\mathbb{RP}^2$ is

$$\mathbb{T}^2 \# \mathbb{RP}^2 \cong N_3$$

Theorem 39.9.8. *The set of surfaces up to homeomorphism forms a commutative monoid with the connected sum, isomorphic to the monoid with presentation*

$$\langle t,r \mid t + r = 3r \rangle$$

where t represents $\mathbb{T}^2$, and r represents $\mathbb{RP}^2$.

39.9.3 Homology and Orientation of Surfaces

Recall that a compact 0-manifold is just a finite discrete set, so the 0th homology group classifies them completely (is a *complete invariant*). Compact 1-manifolds are just finite disjoint unions of circles, so H_0 is also classifies them. H_1 can also distinguish 1-manifolds from 0-manifolds, so (H_0,H_1) is a complete invariant for compact manifolds of dimension at most 1.

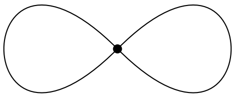
This pattern continues into dimension 2: we will show that (H_0,H_1,H_2) is a complete invariant for compact manifolds of dimension at most 2.

Let $\big((X_\alpha,x_\alpha)\big)_{\alpha\in\Lambda}$ be a collection of pointed spaces. Recall that the *wedge sum* of this collection is the "one-point union" of the spaces, defined as:

$$\bigvee_{\alpha\in\Lambda}(X_\alpha,x_\alpha) := \bigsqcup_{\alpha\in\Lambda} X_\alpha/x_\alpha \sim x_\beta$$

That is, the disjoint union of each space with all the basepoints identified.

Example. The wedge sum of two pointed circles is the figure-eight graph:



$S^1 \vee S^1$

△

Lemma 39.9.9. *If each pair in $\big((X_\alpha,x_\alpha)\big)_{\alpha\in\Lambda}$ is a good pair, then*

$$\tilde{H}_n\left(\bigvee_{\alpha\in\Lambda}(X_\alpha,x_\alpha)\right) = \bigoplus_{a\in\Lambda}\tilde{H}_n(X_\alpha)$$

Proof. We prove it for the binary case with good pairs (X,x) and (Y,y).

As (X,x) and (Y,y) are good pairs, there exist open and contractible $A \subseteq X$ and $B \subseteq Y$ such that $x \in A$ and $y \in B$.

Let $U = A \cup Y$ and $V = X \cup B$, noting that U,V are open in $X \vee Y$, being disjoint unions of sets open in each part. We also have that $U \cap V = (A \cup Y) \cap (X \cup B) = A \cup B$ is contractible, and $U \cup V = X \cup Y$. Then, the reduced Mayer-Vietoris long exact sequence is:

$$\cdots \to \tilde{H}_n(U \cap V) \to \tilde{H}_n(U) \oplus \tilde{H}_n(V) \to \tilde{H}_n(X \vee Y) \to \tilde{H}_{n-1}(U \cap V) \to \cdots$$

The reduced homology of contractible spaces is trivial, so we have

$$\cdots \to 0 \to \tilde{H}_n(U) \oplus \tilde{H}_n(V) \to \tilde{H}_n(X \vee Y) \to 0 \to \cdots$$

and hence

$$\tilde{H}_n(U) \oplus \tilde{H}_n(V) \cong \tilde{H}_n(X \vee Y)$$

for all n. Because A and B are contractible, U and V deformation retract to, and are homotopy equivalent to, Y and X, respectively, which yields

$$\tilde{H}_n(X) \oplus \tilde{H}_n(Y) \cong \tilde{H}_n(X \vee Y)$$

■

Theorem 39.9.10. *The homology of the g-holed torus is*

$$H_n(\Sigma_g) = \begin{cases} \mathbb{Z} & n = 0,2 \\ \mathbb{Z}^{2g} & n = 1 \\ 0 & n \geq 3 \end{cases}$$

Corollary 39.9.10.1. *The surfaces Σ_g, $g \geq 0$, are orientable.*

Theorem 39.9.11. *The homology of N_h is*

$$H_n(N_h) = \begin{cases} \mathbb{Z} & n = 0 \\ \mathbb{Z}^{h-1} \oplus \mathbb{Z}/2 & n = 1 \\ 0 & n \geq 2 \end{cases}$$

Corollary 39.9.11.1. *The surfaces N_h, $h > 0$, are non-orientable.*

39.10 Comparison

At this point, we still have not proved that simplicial homology is an invariant of geometric realisation. That is, that $H_n^{\Delta}(X)$ is independent from the choice of Δ-complex structure on X.

We will show that the simplicial homology of a space is isomorphic to the singular homology, regardless of the choice of Δ-complex structure.

39.10.1 Simplicial = Singular

Let X be a topological space with a Δ-complex structure $\big(T,f : |T| \cong X\big)$. Every n-simplex $s \in T$ induces a canonical continuous map $\Delta^n \to X$, which we will also denote by s.

This extends to a homomorphism $\Delta_n(T) \to C_n(X)$ from between simplicial and singular chain groups, and in fact to a chain map $\Delta_\bullet(T) \to C_\bullet(X)$, since the boundary operator is defined in the same way in both chains.

Recall that we defined the simplicial homology groups as:

$$H_n^{\Delta}(X) \coloneqq H_n(Y)\Big(\coloneqq H_n\big(\Delta_\bullet(T)\big)\Big)$$

Theorem 39.10.1. *The induced map $H_n^{\Delta}(X) \to H_n(X)$ is an isomorphism.*

Proof. Equivalently, the goal is to show that $H_n(T) \to H_n\big(|T|\big)$ is an isomorphism.

Given a Δ-set A and a sub-δ-sets $B \subseteq A$, we define the relative simplicial chain complex by $\Delta_\bullet(A,B) \coloneqq \Delta_\bullet(A)/\Delta_\bullet(B)$, and similarly, we define the relative homology as:

$$H_n(A,B) \coloneqq H_n\big(\Delta_\bullet(A,B)\big)$$

As usual, we have an induced long exact sequence in relative homology:

$$\cdots \to H_{n+1}(A,B) \xrightarrow{\partial} H_n(B) \to H_n(A) \to H_n(A,B) \xrightarrow{\partial} H_{n-1}(B) \to \cdots$$

Now, apply this with $A = T^k$ and $B = T^{k-1}$, the Δ-sets of simplices in T of dimension at most k and $(k-1)$, respectively. This gives a morphism of exact sequences

$$\begin{tikzcd}
\cdots \arrow[r] & H_{n+1}(T^k,T^{k-1}) \arrow[r] \arrow[d] & H_n(T^{k-1}) \arrow[r] \arrow[d] & H_n(T^k) \arrow[r] \arrow[d] & H_n(T^k,T^{k-1}) \arrow[r] \arrow[d] & H_{n-1}(T^{k-1}) \arrow[r] \arrow[d] & \cdots \\
\cdots \arrow[r] & H_{n+1}\big(|T^k|,|T^{k-1}|\big) \arrow[r] & H_n\big(|T^{k-1}|\big) \arrow[r] & H_n\big(|T^k|\big) \arrow[r] & H_n\big(|T^k|,|T^{k-1}|\big) \arrow[r] & H_{n-1}\big(|T^{k-1}|\big) \arrow[r] & \cdots
\end{tikzcd}$$

When $k = 0$, $|T^0|$ is a discrete topological space on the set T^0, and the map $H_n(T^0) \to H_n(|T^0|)$ is clearly an isomorphism. For $k > 0$, by induction and the five lemma, the middle vertical arrow in the diagram above is an isomorphism if the first and fourth arrows are isomorphisms.

We will show that $H_n(T^k,T^{k-1}) \to H_n\big(|T^k|,|T^{k-1}|\big)$ is indeed an isomorphism by identifying both sides independently with the free abelian group on the set T_k of k-simplices, and then observing that generators are mapped to generators in the obvious way.

Note that $\Delta_\bullet(T^{k-1})$ is a chain complex in degrees $k-1,k-2,\dots,0$, and that $\Delta_\bullet(T^k)$ is essentially the same chain complex, just with an additional term $\mathbb{Z}T_k$ in degree k. It follows that $\Delta_\bullet(T^k,T^{k-1})$ is the chain complex with $\mathbb{Z}T_k$ concentrated in degree k. In particular, this gives:

$$H_n(T^k,T^{k-1}) = \begin{cases} \mathbb{Z}T_k & n = k \\ 0 & n \neq k \end{cases}$$

On the other side, by construction of the geometric realisation we have homeomorphisms

$$\begin{aligned} \frac{|T^k|}{|T^{k-1}|} &\cong \frac{\Delta^k \times T_k}{\partial\Delta^k \times T^k} \\ &\cong \bigvee_{T_k} \frac{\Delta^k}{\partial\Delta^k} \\ &\cong \bigvee_{T_k} S^k \end{aligned}$$

So,

$$\begin{aligned} H_n\big(|T^k|,|T^{k-1}|\big) &\cong H_n\big(|T^k|/|T^{k-1}|\big) \\ &\cong H_n\left(\bigvee_{T_k} S^k\right) \\ &\cong \bigoplus_{T_k} H_n(S^k) \\ &= \begin{cases} \mathbb{Z}T_k & n = k \\ 0 & n \neq k \end{cases} \end{aligned}$$

with generators corresponding to elements $s \in T_k$ via the relative cycles $s : \Delta^K \to |T^k|$. It follows that the homomorphism identifies these two groups.

Now, if $T = T^k$ is finite dimensional, then we are already done. Otherwise, let $z \in Z_n(T)$ be an n-cycle whose image in $H_n\big(|T|\big)$ vanishes. That is, there exists $\tau \in C_{n+1}\big(|T|\big)$ with $\partial(\tau) = z$. Now, every compact subspace of $|T|$ is contained within some $|T^k|$, so $\tau \in C_{n+1}\big(|T^k|\big)$ for some $k > n$, so z maps to zero in $H_n\big(|T^k|\big)$. Then, $z = 0 \in H_n(T^k) = H_n(T)$, so the map $H_n(T) \to H_n\big(|T|\big)$ is injective.

Similarly, let $s \in Z_n\big(|T|\big)$ be an n-cycle. As in the previous case, $\sigma \in Z_n\big(|T^k|\big)$ for some $k > n$, and similarly, $[\sigma]$ has a preimage in $H_n(T^k) = H_n(T)$, so the map $H_n(T) \to H_n\big(|T|\big)$ is surjective. ■

Corollary 39.10.1.1. *The simplicial homology $H^\Delta_\bullet(X)$ depends only on X and not on any Δ-complex structure.*

Corollary 39.10.1.2. *If X has a Δ-complex structure with simplices in dimensions at most k, then $H_n(X) = 0$ for all $n > k$.*

39.10.2 CW Complexes

Recall that a *CW complex* ("*Closure finite*, *Weak topology*") is a topological space* X obtained as follows:

(*i*) Start with a 0-*skeleton* consisting of a disjoint union $X^0 = \bigsqcup_i D_i^0$ of 0-discs (i.e. points), or 0-*cells*.

(*ii*) Given an $(n-1)$-skeleton X^{n-1}, we construct the X^n by gluing a collection of n-*cells* (i.e. n-discs) D_α^n via *attaching maps* $\varphi_\alpha : \partial D_\alpha^n = S_\alpha^{n-1} \to X^{n-1}$:

$$X^n := X^{n-1} \sqcup \textstyle\bigsqcup_\alpha D_\alpha^n / {\sim}$$

where $x \sim \varphi_\alpha(x)$ for all $x \in S_\alpha^{n-1}$.

This recursion then either stops at some finite level n, yielding a CW complex $X := X^n$ of *dimension* n, or continues infinitely, in which case we define $X := \bigcup_{n\in\mathbb{N}} X^n$, with a subspace $U \subseteq X$ being open if and only if $U \cap X^n$ is open in X^n for all n.

For each n-cell D_α^n, we define the *characteristic map* $\Phi_\alpha : D_\alpha^n \to X$ to be the composition

$$D_\alpha^n \hookrightarrow X^{n-1} \sqcup \bigsqcup_\alpha D_\alpha^n \xrightarrow{q} X^n \hookrightarrow X$$

where q is the quotient map induced by φ_α, identifying $x \sim \varphi_\alpha(x)$ for all $x \in S_\alpha^{n-1}$, and the other two maps are the canonical inclusions.

Every Δ-complex is a CW complex. The main difference between the two is that the attaching maps for an n-cell (D^n) in a CW complexes may be *any* continuous map into anywhere in the $(n-1)$-skeleton, while in a Δ-complex, the attaching maps must glue each face of the n-cell (Δ^n) to an $(n-1)$-simplex already in the complex.

In particular, this means that CW complexes may "skip" dimensions, and add no cells in a particular step, but add more cells after. This cannot be the case for a Δ-complex, because the face maps after a skipped step would have no simplices to attach to.

Example. The sphere S^k for $k > 0$ admits a CW complex structure with a single 0-cell, and a single k-cell, where the attaching map is the unique map sending $\partial D^k = S^{k-1}$ to the unique point of the 0-cell. △

Example. A 1-dimensional CW complex is the same thing as a 1-dimensional Δ-complex, and both can be identified with a topological graph. △

Example. Real projective k-space can be constructed as the quotient $\mathbb{RP}^k \cong S^k/(x \sim -x)$ of the k-sphere under the antipodal map. Alternatively, it is the quotient of one of the two hemispheres D^k with antipodal points on the boundary $\partial D^k = S^{k-1}$ identified. But, this boundary is precisely $\mathbb{RP}^{k-1}$, so $\mathbb{RP}^k$ can be obtained by attaching a k-cell to $\mathbb{RP}^k$ along the quotient map $S^{k-1} \to \mathbb{RP}^{k-1}$.

Inductively, it follows that $\mathbb{RP}^k$ has a CW structure with exactly one cell in each dimension $0,1,\dots,k$. △

Example. If we continue this process, we can construct the infinite real projective space $\mathbb{RP}^\infty := \bigcup_{k\in\mathbb{N}} \mathbb{RP}^k$ as a CW complex with a single cell in each dimension. △

39.10.3 Cellular Homology

We would like to define a homology theory for CW complexes, similar to simplicial homology for Δ-complexes. As before, we can take the group of n-chains $C_n^{\mathrm{CW}}(X)$ to be the free abelian group on the n-cells, but cells in a CW complex may be attached in much more complicated ways than in Δ-complexes, so there isn't an obvious way to define the oriented boundary of an n-cell.

* As with Δ-complexes, this only describes a CW complex *structure* on a space X, of which there can be many.

Lemma 39.10.2. *Let X be a CW complex with n-cells $(D^n_\alpha)_{\alpha\in\Lambda}$, $n > 0$. Then,*

$$H_k(X^n,x^{n-1}) \cong \begin{cases} \bigoplus_{\alpha\in\Lambda}\mathbb{Z} & k = n \\ 0 & k \neq n \end{cases}$$

That is, the nth relative homology of (X^n,X^{n-1}) is the free abelian group on the set of n-cells.

Proof. (X^n,X^{n-1}) is a good pair, so by Theorem 39.7.5,

$$H_k(X^n,X^{n-1}) \cong \tilde{H}(X^n/X^{n-1})$$

The boundaries of the n-cells in X^n are glued into the X^{n-1} skeleton, so in this quotient, their boundaries are all identified together, so

$$X^n/X^{n-1} \cong \bigvee_{\alpha\in\Lambda} D^n_\alpha/\partial D^n_\alpha \cong \bigvee_{\alpha\in\Lambda} S^n$$

Relative homology splits across wedge sums, so

$$\tilde{H}\left(\bigvee_{\alpha\in\Lambda} S^n\right) \cong \bigoplus_{\alpha\in\Lambda} \tilde{H}_k(S^n) \cong \begin{cases} \bigoplus_{\alpha\in\Lambda}\mathbb{Z} & k = n \\ 0 & k \neq n \end{cases}$$

■

We can describe this isomorphism explicitly as follows. Choose a homeomorphism $f : \Delta^n \xrightarrow{\cong} D^n$. Then, we have a continuous map given by the composition

$$\Delta \xrightarrow{f} D^n \xrightarrow{\Phi_\alpha} X^n$$

which is in fact a relative cycle for the pair (X^n,X^{n-1}), and its relative homology class generates the copy of $\mathbb{Z}$ corresponding to D^n_α.

Lemma 39.10.3. *Let X be a CW complex. Then,*

(i) $H_n(X^k) = 0$ *for all* $n > k$.

In particular, the homology of X vanishes in all degrees $n > \dim(X)$;

(ii) $H_n(X^k) \cong H_n(X)$ *for all* $n < k$.

More specifically, the map $\iota_* : H_n(X^k) \to H_n(X)$ induced by the inclusion $\iota : X^k \hookrightarrow X$ is an isomorphism if $n < k$, and is surjective if $n = k$.

Proof.

(*i*) Consider the long exact sequence for relative homology of the pair (X^k,X^{k-1}):

$$\cdots \to H_{n+1}(X^k,X^{k-1}) \to H_n(X^{k-1}) \to H_n(X^k) \to H_n(X^k,X^{k-1}) \to \cdots$$

From the previous lemma, the left relative homology group vanishes for $n + 1 \neq k$, and the right group vanishes for $n \neq k$. So, both outer terms vanish in degrees $n \neq k,k-1$, and we have an isomorphism in the middle. So, if $n > k$,

$$H_n(X^k) \cong H_n(X^{k-1}) \cong \cdots \cong H_n(X^0) \cong 0$$

(ii) Suppose X has finite dimension d. If $n < k$, then we have

$$H_n(X^k) \cong H_n(X^{k+1}) \cong H_n(X^{k+2}) \cong \cdots \cong H_n(X^d) = H_n(X)$$

so $H_n(X^k) \cong H_n(X)$.

Otherwise, if $n = k$, then only the right group vanishes, and we only have a surjection in this degree, so

$$H_n(X^k) \twoheadrightarrow H_n(X^{k+1}) \cong H_n(X^{k+2}) \cong \cdots \cong H_n(X^d) = H_n(X)$$

so $H_n(X^k) \twoheadrightarrow H_n(X)$.

If X is infinite-dimensional, the proof is complicated and is omitted.

■

We can now describe cellular homology. We will take the relative homology groups

$$H_{n+1}(X^{n+1},X^n) \longrightarrow H_n(X^n,X^{n-1}) \longrightarrow H_{n-1}(X^{n-1},X^{n-2})$$

to be the chain groups. By Lemma 39.10.2, these are all free abelian on the set of cells of the matching dimension. The task is now to construct these differentials.

We construct the long exact sequences for the three chain groups, and they fit together into a diagram:

$$\begin{tikzcd}
 & & & & 0 & \\
 & 0 \arrow[dr] & & H_n(X^{n+1}) \arrow[ur] & & \\
 & & H_n(X^n) \arrow[ur] \arrow[dr, "\beta"] & & & \\
 & H_{n+1}(X^{n+1},X^n) \arrow[ur, "\alpha"] \arrow[rr, "d_{n+1}"] & & H_n(X^n,X^{n-1}) \arrow[rr, "d_n"] \arrow[dr, "\gamma"'] & & H_{n-1}(X^{n-1},X^{n-2}) \\
 & & & & H_{n-1}(X^{n-1}) \arrow[ur, "\delta"'] & \\
 & & & 0 \arrow[ur] & &
\end{tikzcd}$$

The leftmost zero is $H_n(X^{n-1})$, which vanishes by Lemma 39.10.3; the upper zero is $H_n(X^{n+1},X^n)$, which vanishes by Lemma 39.10.2; and the lower zero is $H_{n-1}(X^{n-2})$, which vanishes by Lemma 39.10.3.

Also, note that the group at the top $H_n(X^{n+1})$ is isomorphic to $H_n(X)$ via Lemma 39.10.3.

We define the differentials to be the compositions

$$d_{n+1} \coloneqq \beta \circ \alpha \qquad\qquad d_n \coloneqq \delta \circ \gamma$$

This defines a chain complex, since the composition $d_n \circ d_{n+1}$ factors through β and γ, and these are consecutive maps in a long exact sequence, so their composition is zero, and hence $d_n \circ d_{n+1} = 0$, as required.

Let X be a CW complex. We define the *cellular chain complex* $C_\bullet^{\mathrm{CW}}(X)$ by

$$\cdots \to C_{n+1}^{\mathrm{CW}} = H_{n+1}(X^{n+1},X^n) \xrightarrow{d_{n+1}} C_n^{\mathrm{CW}}(X) = H_n(X^n,X^{n-1}) \xrightarrow{d_n} \cdots$$

The *cellular homology groups* are the homology of this chain complex:

$$H_n^{\mathrm{CW}}(X) \coloneqq H_n\big(C_\bullet^{\mathrm{CW}}(X)\big)$$

Lemma 39.10.4. *The cellular homology group $H_n^{CW}(X)$ is canonically isomorphic to the singular homology group $H_n(X)$.*

Proof.

$$\begin{aligned}
H_n^{\mathrm{CW}}(X) &\coloneqq \frac{\ker(d_n)}{\mathrm{im}(d_{n+1})} \\
&\cong \frac{\ker(\gamma)}{\mathrm{im}(d_{n+1})} && \text{[by injectivity of } \delta\text{]} \\
&\cong \frac{\mathrm{im}(\beta)}{\mathrm{im}(d_{n+1})} && \text{[exactness at } C_n^{\mathrm{CW}}(X)\text{]} \\
&\cong \frac{\mathrm{im}(\beta)}{\mathrm{im}(\beta \circ \alpha)} \\
&\cong \frac{H_n(X^n)}{\mathrm{im}(\beta \circ \alpha)} && \text{[by injectivity of } \beta\text{]} \\
&\cong \frac{H_n(X^n)}{\mathrm{im}(\alpha)} \\
&\cong \mathrm{coker}(\alpha) \\
&\cong H_n(X^{n+1}) \\
&\cong H_n(X)
\end{aligned}$$

■

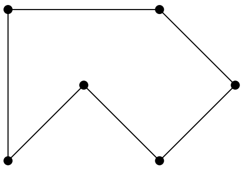
Corollary 39.10.4.1. *For any CW complex X, there are canonical isomorphisms $H_n^{CW}(X) \cong H_n(X)$.*

Theorem 39.10.5. *The boundary operator for cellular homology is given by*

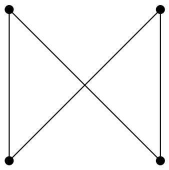
- *in degree $n = 1$:*
$$d_1(D_\alpha^1) = \varphi$$
- *in degrees $n > 1$:*
$$d_n(D_\alpha^n) = \sum_\beta d_{\alpha\beta} D_\beta^{n-1}$$
where
$$d_{\alpha\beta} = \deg\left(\Delta_{\alpha\beta} : S_\alpha^{n-1} \xrightarrow{\varphi_\alpha} X^{n-1} \xrightarrow{\pi_\beta} S_\beta^{n-1}\right)$$

39.11 The Euler Characteristic

A *plane graph* is a finite 1-dimensional CW complex embedded in the real plane $\mathbb{R}^2$. Equivalently, it is a finite graph in the plane in which the edges do not cross.



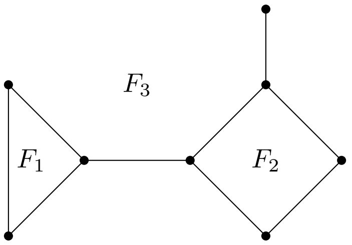
a plane graph



not a plane graph

Note that this is distinct from the notion of a *planar graph* in graph theory. Some finite graphs, for instance, the complete graph K_5 or the complete bipartite graph $K_{3,3}$, do not admit an embedding into $\mathbb{R}^2$, so they are *nonplanar*. The example on the right above *does* admit an embedding into $\mathbb{R}^2$ as a square, so it is *planar*, but the particular embedding shown is not a *plane* graph.

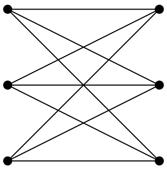
A *face* of a plane graph $\mathscr{G}$ is a connected component of $\mathbb{R}^2 \setminus \mathscr{G}$:



Theorem 39.11.1 (Euler)**.** *For a plane graph with v vertices, e edges, and f faces,*

$$v - e + f = 2$$

Example. The complete bipartite graph



has $v = 6$ vertices and $e = 9$ edges. Suppose that $K_{3,3}$ admits an embedding into $\mathbb{R}^2$. Because it is bipartite, all cycles have even length, so every face has at least 4 edges. Clearly, every edge meets at most two faces, so, $4f \leq 2e = 18$, which simplifies to $f \leq 4$. Hence,

$$v - e + f \leq -3 + 4 = 1$$

contradicting Euler's formula, so $K_{3,3}$ is nonplanar. △

By taking the one-point compactification of $\mathbb{R}^2$, a planar graph yields a CW complex structure on the $\mathbb{R}^2 \cup \{\infty\} \cong S^2$ with v many 0-cells, e 1-cells, and f 2-cells.

This result is non-trivial to prove (compare with the Jordan curve theorem), and is a special case of the *Schoenflies' theorem*, which we will assume without proof.

Given a topological space X that admits a finite CW complex structure, we define the *Euler characteristic* to be the alternating sum

$$\chi(X) := \sum_{n\in\mathbb{N}} (-1)^n \big|\{n\text{-cells in } X\}\big|$$

of the number of n-cells in X.

Theorem 39.11.2. *The Euler characteristic is independent of choice of CW complex structure.*

Let A be a finitely generated abelian group. It decomposes as $A = F \oplus T$, the direct sum of the torsion-free part $T \cong \mathbb{Z}^r$, which is necessarily free abelian, and a finite abelian group F. The integer r is well-defined and is called the *rank* of A, denoted $\mathrm{rk}_{\mathbb{Z}}(A)$. (See §39.1.7.3.)

Alternatively, this rank is given by the dimension of the $\mathbb{Q}$-vector space $A \otimes_{\mathbb{Z}} \mathbb{Q}$ since $\mathbb{Z}^r \otimes_{\mathbb{Z}} \mathbb{Q} = \mathbb{Q}^r$, and $F \otimes_{\mathbb{Z}} \mathbb{Q} = 0$.

Theorem 39.11.3. *For a short exact sequence $0 \to A_1 \to A_2 \to A_3 \to 0$ of finitely generated abelian groups,*

$$\mathrm{rk}_{\mathbb{Z}}(A_1) - \mathrm{rk}_{\mathbb{Z}}(A_2) + \mathrm{rk}_{\mathbb{Z}}(A_3) = 0$$

Corollary 39.11.3.1. *Let $C_\bullet$ be a chain complex with finitely many non-zero terms, all of which are finitely generated abelian groups. Then,*

$$\sum_n (-1)^n \operatorname{rk}_{\mathbb{Z}}(C_n) = \sum_n (-1)^n \operatorname{rk}_{\mathbb{Z}}\big(H_n(C_\bullet)\big)$$

Let X be a space with only finitely many non-zero homology groups, all of which are finitely generated abelian groups. Then, its Euler characteristic is defined as

$$\chi(X) \coloneqq \sum_{n\in\mathbb{N}} (-1)^n \operatorname{rk}_{\mathbb{Z}}\big(H_n(X)\big)$$

Example.

$$\chi(S^k) = \begin{cases} 2 & k \text{ even} \\ 0 & k \text{ odd} \end{cases}$$

△

Theorem 39.11.4. *Let $X = U \cup V$ and either*

- *X is a CW complex and U,V are subcomplexes;*
- *or X is any topological space and U,$V \subseteq X$ are open subsets.*

Then, if $\chi(U)$, $\chi(V)$, and $\chi(U \cap V)$ are all defined, then so is $\chi(X)$, and it is given by

$$\chi(X) = \chi(U) + \chi(V) - \chi(U \cap V)$$

Corollary 39.11.4.1. *We have*

$$\chi(\Sigma_g) = 2 - 2g$$
$$\chi(N_h) = 2 - h$$

The Euler characteristic of a surface can therefore attain any integer less than or equal to 2, with equality only in the case of the 2-sphere.

Every even non-positive integer is the Euler characteristic of precisely two surfaces, one orientable and one non-orientable. In particular, surfaces are completely classified by

- whether they are orientable or not, and
- their Euler characteristic.

Theorem 39.11.5. *Let X and Y be finite CW complexes. Then so is $X \times Y$,*

$$\chi(X \times Y) = \chi(X) \cdot \chi(Y)$$

Theorem 39.11.6. *Let $p : Y \to X$ be an n-fold covering, and suppose that X is a finite CW complex. Then, Y is also a finite CW complex, and we have*

$$\chi(Y) = n \cdot \chi(X)$$

39.12 Homology Theories

So far, we have defined three homology theories:

- Simplicial;
- Singular;
- Cellular.

Importantly, we have seen that they are all the “same”. This is not a coincidence: everything that behaves “like a homology theory” is in fact the same.

It turns out that very few axioms are required to characterise homology theories. It won’t be surprising that many of the important theorems we have seen will appear as axioms. However, these axioms are stated in the language of *category theory.*

39.12.1 Categories

For a more thorough overview, see §51.

A *category* $\mathscr{C}$ consists of:

- A class $\mathrm{ob}(\mathscr{C})$ of *objects* in $\mathscr{C}$.
- For all (ordered) pairs of objects $A,B \in \mathrm{ob}(\mathscr{C})$, a class $\hom(A,B)$ of *morphisms* from A to B, called the *hom-set* of morphisms from A to B, also sometimes written $\mathscr{C}(A,B)$ or $\hom_{\mathscr{C}}(A,B)$ (particularly useful if multiple categories are in use). If $f \in \hom(A,B)$, we write $f : A \to B$ or $A \xrightarrow{f} B$.

  The collection of all of these classes is the hom-set of $\mathscr{C}$, and is written $\hom(\mathscr{C})$.
- For any three objects $A,B,C \in \mathrm{ob}(\mathscr{C})$, a binary operation, $\circ : \hom(A,B) \times \hom(B,C) \to \hom(A,C)$, $(g,f) \mapsto g \circ f$, called *composition*, such that,
  - (*associativity*) if $f : A \to B$, $g : B \to C$, and $h : C \to D$, then $h \circ (g \circ f) = (h \circ g) \circ f$;
  - (*identity*) for every object $X \in \mathrm{ob}(\mathscr{C})$, there exists a morphism $\mathrm{id}_X : X \to X$ called the *identity morphism* on X, such that every morphism $f : A \to X$ satisfies $\mathrm{id}_X \circ f = f$, and every morphism $g : X \to B$ satisfies $g \circ \mathrm{id}_X = g$.

Example. We list a few simple examples of categories:

(i) The prototypical example of a category is the category of sets and set functions, **Set**. Identity morphisms are identity functions, and associativity follows from basic properties of set functions.

(ii) The category of groups and group homomorphisms, **Grp**. Every group is a set with extra structure, and every group homomorphism is a set function that happens to preserve this structure, so associativity and identity are inherited from **Set**.

(iii) Similarly, collections of sets with extra structure and maps that preserve that structure generally form categories; for instance, the categories of:

- Monoids and monoid homomorphisms, **Mon**;
- Rings and ring homomorphisms, **Ring**;
- Metric spaces and non-expansive maps, **Met**;
- C^p-manifolds and p-times differentiable maps $\mathbf{Man}^p$;
- Pointed sets and based maps, $\mathbf{Set}_\bullet$;

- Measurable spaces and measurable functions, **Mea**;
- Vector spaces and linear maps over a fixed field K, $\mathbf{Vect}_K$; etc.

More relevant to this chapter are perhaps the categories of:

- Topological spaces and continuous maps, **Top**;
- Topological spaces and homotopy classes of continuous maps, **hTop**;
- Pointed topological spaces and continuous based maps, $\mathbf{Top}_\bullet$;
- Δ-sets and Δ-set maps, Δ**Set**;
- Chain complexes and chain maps, **Cpx**.

(iv) For any topological space X, its fundamental groupoid $\Pi_1(X)$ is a category. Its objects are points in X, and morphisms are homotopy classes of paths, with composition given by path concatenation.

△

Suppose we have objects A and B in a category, and morphisms f from A to B and g from B to A such that the following diagram is commutative.

$$\begin{tikzcd} A \arrow[loop left, "\mathrm{id}_A"] \arrow[r, shift left, "f"] & B \arrow[l, shift left, "g"] \arrow[loop right, "\mathrm{id}_B"] \end{tikzcd}$$

That is, $f \circ g = \mathrm{id}_B$ and $g \circ f = \mathrm{id}_A$. f and g are then *isomorphisms* – morphisms with inverses – and we alternatively label g as f^{-1}. If an isomorphism between A and B exists, we say that A and B are *isomorphic*, and we denote this relation as $A \cong B$.

Example. This recovers the following notions in some familiar categories:

- In **Grp**: group isomorphisms;
- In **Top**: homeomorphisms;
- In **Man**: diffeomorphisms;
- In $\mathbf{Vect}_K$: linear isomorphisms.

△

Let $\mathscr{C}$ and $\mathscr{D}$ be categories. A *functor*, $F : \mathscr{C} \to \mathscr{D}$, consists of two parts: a mapping on objects, and a mapping on morphisms, that follow two constraints. $F : \mathrm{ob}(\mathscr{C}) \to \mathrm{ob}(\mathscr{D})$ associates each object X in $\mathscr{C}$ to an object, $F(X)$, in $\mathscr{D}$.

$$X \mapsto F(X)$$

Similarly, the map $F : \hom(\mathscr{C}) \to \hom(\mathscr{D})$ associates each morphism $f : X \to Y$ in $\mathscr{C}$ to a morphism $F(f) : F(X) \to F(Y)$ in $\mathscr{D}$ such that:

- $F(\mathrm{id}_X) = \mathrm{id}_{F(X)}$ for every object X in $\mathscr{C}$;
- $F(g \circ f) = F(g) \circ F(f)$ for all morphisms $f : X \to Y$ and $g : Y \to Z$ in $\hom(\mathscr{C})$.

That is, the functor preserves identity morphisms and composition of morphisms.

A more concise way to phrase this is, for every pair of objects $A,B \in \mathrm{ob}(\mathscr{C})$, the functor F induces a mapping $F_{A,B} : \hom_{\mathscr{C}}(A,B) \to \hom_{\mathscr{D}}(F(A),F(B))$ that respects the structure of the categories.

$$\begin{tikzcd} A \arrow[rr, "f"] & {} \arrow[d, mapsto] & B \\ F(A) \arrow[rr, "F(f)"] & {} & F(B) \end{tikzcd}$$

Example. We have already seen many functors:

- $H_n : \mathbf{Top} \to \mathbf{Ab}$;
- $H_n : \mathbf{Cpx} \to \mathbf{Ab}$;
- $C_\bullet : \mathbf{Top} \to \mathbf{Cpx}$;
- $(-)^{\mathrm{ab}} : \mathbf{Grp} \to \mathbf{Ab}$;
- $\pi_1 : \mathbf{Top}_\bullet \to \mathbf{Ab}$.

In fact, the first three of these compose into an commutative triangle:

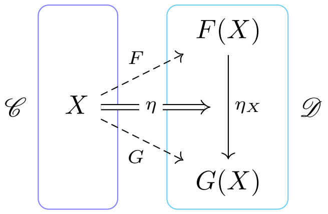
$$\begin{tikzcd}
\mathbf{Top} \arrow[r, "C_\bullet"] \arrow[dr, "H_n"'] & \mathbf{Cpx} \arrow[d, "H_n"] \\
 & \mathbf{Ab}
\end{tikzcd}$$

△

Example. Let Δ denote the *simplex category*, which has linearly ordered sets as objects and order homomorphisms as morphisms. Then, a simplicial set is a functor $\Delta^{\mathrm{op}} \to \mathbf{Set}$, where Δ^{op} is the *opposite category* to Δ, which has all morphisms reversed. △

Given categories and functors $\mathscr{C} \overset{F}{\underset{G}{\rightrightarrows}} \mathscr{D}$, a *natural transformation* is a mapping $\begin{tikzcd} \mathscr{C} \arrow[r, bend left, "F"] \arrow[r, bend right, "G"'] \arrow[r, phantom, "\Downarrow\eta"] & \mathscr{D} \end{tikzcd}$ or $\eta : F \Rightarrow G$ between functors.

The functors F and G map objects and morphisms in $\mathscr{C}$ to objects and morphisms in $\mathscr{D}$, so to define a mapping $F \Rightarrow G$, we want to associate the images of objects and morphisms under F to their images under G. For objects, this just means that if X is in $\mathscr{C}$, then $F(X)$ should be associated with $G(X)$ – this is just a morphism in $\hom_{\mathscr{D}}(F(X), G(X))$. So, the natural transformation η associates each object $X \in \mathrm{ob}(\mathscr{C})$ to a morphism $h_X : F(X) \to G(X)$ called the *component* of η at X.



However, there could be many morphisms $F(X) \to G(X)$ we could choose. We need a way of selecting these components that is consistent throughout the whole category.

Consider a morphism $f : A \to B$ in $\mathscr{C}$. Under F and G, we have the images $F(f) : F(A) \to F(B)$ and $G(f) : G(A) \to G(B)$. Along with the components $\eta_A : F(A) \to G(A)$ and $\eta_B : F(B) \to G(B)$, this completes the square

$$\begin{tikzcd}
 & A \arrow[r, dashed, "F"] \arrow[dd, "f"] \arrow[rr, dashed, bend left=50, "G"] \arrow[r, Rightarrow, bend left, "\eta"] & F(A) \arrow[r, "\eta_A"] \arrow[dd, "F(f)"] \arrow[ddr, dashed, "(\exists!)?"] & G(A) \arrow[dd, "G(f)"] & \\
\mathscr{C} & & & & \mathscr{D} \\
 & B \arrow[r, dashed, "F"] \arrow[rr, dashed, bend right=50, "G"'] \arrow[r, Rightarrow, bend right, "\eta"'] & F(B) \arrow[r, "\eta_B"] & G(B) &
\end{tikzcd}$$

In this diagram, there are two paths from $F(A)$ to $G(B)$, namely, $\eta_B \circ F(f)$, and $G(f) \circ \eta_A$, and without any further conditions on the components of η, these paths may be distinct. However, if we require that these paths are equal – that the diagram commutes – then this forces our selection of components to be consistent throughout the whole category. This coherency condition is called the *naturality* requirement.

So, overall, a natural transformation a natural transformation $\eta : F \Rightarrow G$ between functors $F,G : \mathscr{C} \to \mathscr{D}$ is a collection of morphisms $\big(F(X) \xrightarrow{\eta_X} G(X)\big)_{X \in \mathrm{ob}(\mathscr{C})}$ indexed by the objects of $\mathscr{C}$ such that the following diagram commutes:

$$\begin{tikzcd}
A \arrow[d, "f"] & F(A) \arrow[r, "\eta_A"] \arrow[d, "F(f)"] & G(A) \arrow[d, "G(f)"] \\
B & F(B) \arrow[r, "\eta_B"] & G(B)
\end{tikzcd}$$

That is, $\eta_B \circ F(f) = G(f) \circ \eta_A$ for all $f : A \to B$ in $\hom(\mathscr{C})$.

39.12.2 Axioms

We denote by $\mathbf{CW}_2$ the category of *CW pairs*. Its objects are pairs (X,Y) of CW complexes X and Y where $Y \subseteq X$ is a subcomplex, and its morphisms $(X,Y) \to (X',Y')$ are continuous maps $f : X \to X'$ such that $f(Y) \subseteq Y'$.

A *homology theory* is a collection of functors $(n \geq 0)$,

$$h_n : \mathbf{CW}_2 \to \mathbf{Ab}$$

and natural transformations with components

$$\partial_n : h_n(X,Y) \to h_{n-1}(Y) \coloneqq h_{n-1}(Y,\varnothing)$$

satisfying the following axioms:

- (i) **Homotopy Invariance**: If $f \simeq G : X \to Y$, then $h_n(f) = h_n(g) : h_n(X) \to h_n(Y)$.
- (ii) **Excision**: If $X = U_1 \cup U_2$, then the inclusion $\iota : (U_1,U_1 \cap U_2) \to (X,U_2)$ induces an isomorphism $h_n(U_1,U_1 \cap Y_2) \xrightarrow{\iota_*} (X,U_2)$.
- (iii) **Exactness**: The sequence

$$\cdots \to h_{n+1}(X,Y) \xrightarrow{\partial_{n+1}} h_n(Y) \to h_n(X) \to h_n(X,Y) \xrightarrow{\partial_n} h_{n-1}(Y) \to \cdots$$

is exact for every pair (X,Y).

(*iv*) **Additivity**: For any collection (X_α,Y_α), the map $\oplus_\alpha h_n(X_\alpha,Y_\alpha) \to h_n\big(\bigsqcup_\alpha (X_\alpha,Y_\alpha)\big)$ is an isomorphism.

(*v*) **Dimension**: $h_n(*) = \delta_{n0}\mathbb{Z}$.

The naturality square for ∂_n looks like:

$$\begin{tikzcd}
(X,Y) \arrow[d, "f"] & h_n(X,Y) \arrow[r, "\partial_n"] \arrow[d, "h_n(f)"] & h_{n-1}(Y) \arrow[d, "h_{n-1}(f)"] \\
(X',Y') & h_n(X',Y') \arrow[r, "\partial_n"] & h_{n-1}(Y')
\end{tikzcd}$$

Theorem 39.12.1. *Singular homology defines a homology theory.*

Proof. Homotopy invariance is §39.6.1, Excision is Theorem 39.7.5, Additivity is Theorem 39.4.3, Exactness is Corollary 39.7.3.1, and Dimension was computed in an example in §39.4. ■

Fix a homology theory h_n. Let us see what we can compute just from the axioms. As before, we define the reduced homology group as $\tilde{h}_n(X) := \ker\big(h_n(X) \to h_n(*)\big)$.

Example.

- For each triple $X \supseteq Y \supseteq Z$, the sequence

$$\cdots \to h_{n+1}(X,Y) \xrightarrow{\partial_{n+1}} h_n(Y,Z) \to h_n(X,Z) \to h_n(X,Y) \xrightarrow{\partial_n} h_{n-1}(Y,Z) \to \cdots$$

  is exact. This is a diagram chase, starting with the Exactness axiom.

- If X is contractible, then $\tilde{h}_n(X) = 0$ for all n: the map $h_n(X) \to h_n(*)$ is an isomorphism by the Homotopy Invariance axiom.

- Consider the exact sequence above for the triple $(D^k,\partial D^k,*)$:

$$\cdots \to \tilde{h}_n(D^k) \to h_n(D^k,S^{k-1}) \to \tilde{h}_{n-1}(S^{k-1}) \to \tilde{h}_{n-1}(D^k)$$

  As the outer terms vanish, we have an isomorphism in the middle, and by Excision, Additivity, and Dimension, we have:

$$\tilde{h}_n(S^k) \xleftarrow{\sim} h_n(D^k,S^{k-1}) \xrightarrow{\sim} \tilde{h}_{n-1}(S^{k-1}) \xleftarrow{\sim} \cdots \xrightarrow{\sim} \tilde{h}_{n-k}(S^0) = \delta_{nk}\mathbb{Z}$$

- From Excision, we have:

$$h_n(X^k,X^{k-1}) \xrightarrow{\sim} \tilde{h}_n(X^k/X^{k-1}) \xrightarrow{\sim} \tilde{h}_n\left(\bigvee_\alpha S^k_\alpha\right) \xleftarrow{\sim} \bigoplus_\alpha \tilde{h}_n(S^k_\alpha) = \bigoplus_\alpha \delta_{nk}\mathbb{Z}$$

  where the leftward isomorphisms are deduced from Additivity and Excision, and where α ranges over the k-cells in X.

△

In fact, *everything* we have computed (in terms of homology) in this chapter can be derived from just the axioms:

Theorem 39.12.2. *If (h_n,∂_n) is a homology theory, then there are natural isomorphisms*

$$h_n(X,Y) \cong H_n(x,y)$$

These natural isomorphisms commute with the connecting homomorphisms ∂_n, so in a suitable category of homology theories, all objects are isomorphic.

This theorem implies that we can compute the homology of a space using whatever homology theory we like, so we can use whichever theory is simplest for any given space. However, it also implies that if two spaces cannot be distinguished by singular homology, then they cannot be distinguished by *any* homology theory. Given that the goal of algebraic topology is to distinguish topological spaces, this may be somewhat disappointing.

However, it is possible to weaken or modify the homology axioms to generate other theories.

39.12.3 Coefficients

Let A be an abelian group. A *homology theory with coefficients in* A is a collection of functors $h_n : \mathbf{CW}_2 \to \mathbf{Ab}$ together with natural transformations ∂_n satisfying the same axioms as for an ordinary homology theory, but with the dimension axiom replaced by $h_n(*) = \delta_{n0}A$.

For example, we can replace the singular chain complex $C_\bullet(X)$ by the tensor product $C_\bullet(X) \otimes_{\mathbb{Z}} A$. That is, in degree n, we have the group

$$C_n(X; A) \coloneqq \bigoplus_{\sigma:\Delta^n \to X} A[\sigma]$$

with the "same" differential as before. Then, we set $H_n(X; A) \coloneqq H_n\big(C_\bullet(X; A)\big)$, and similarly for relative homology. This *singular homology with coefficients in* A satisfies these modified axioms.

One particularly useful group to consider is $A = \mathbb{Z}/2$.

Example. The cellular chain complex for $\mathbb{RP}^k$ with $\mathbb{Z}/2$-coefficients is:

$$0 \to \mathbb{Z}/2 \xrightarrow{0} \mathbb{Z}/2 \to \cdots \to \mathbb{Z}/2 \xrightarrow{0} \mathbb{Z}/2 \xrightarrow{0} \mathbb{Z}/2 \to 0$$

so that

$$H_n(\mathbb{RP}^k; \mathbb{Z}/2) = \begin{cases} \mathbb{Z}/2 & 0 \leq n \leq k \\ 0 & n > k \end{cases}$$

and

$$H_n(\mathbb{RP}^\infty; \mathbb{Z}/2) = \mathbb{Z}/2$$

for all $n \geq 0$. △

However, for distinguishing more spaces, this modification of homology theory is not of much help:

Theorem 39.12.3. *If $H_n(X) \cong H_n(Y)$ for all n, then $H_n(X; A) \cong H_n(Y : A)$ for all homology theories with coefficients in A. More specifically, there is an isomorphism*

$$H_n(X; A) \cong \big(H_n(X) \otimes_{\mathbb{Z}} A\big) \oplus \operatorname{Tor}\big(H_{n-1}(X), A\big)$$

39.12.4 Generalised Homology Theories

A *generalised* or *extraordinary homology theory* is a collection of functors $h_n : \mathbf{CW}_2 \to \mathbf{Ab}$ together with natural transformations ∂_n satisfying the same axioms as for an ordinary homology theory except for (possibly) the dimension axiom.

Some examples of generalised homology theories include *(co)bordism* and *stable homotopy*. These theories have been studied extensively, but many questions remain open.

However, one thing that is clear is that no uniqueness theorem holds for generalised homology theories: these two examples (and many others) are genuinely distinct from ordinary homology.

39.12.4.1 Stable Homotopy

The homotopy groups for $n > 0$ of a pointed space are given by homotopy classes of pointed maps $S^n \to X$:

$$\pi_n(X,x) = [S^n,X]$$

For $n \geq 2$, it turns out that these groups are abelian.

So far, we have seen that homotopy equivalences and homeomorphisms induce isomorphisms in homology, and very few instances (mainly surfaces) where the converse holds. The following theorem illustrates how fine of an invariant homotopy groups are:

Theorem 39.12.4 (Whitehead's Theorem)**.** *Let $f : X \to Y$ be a continuous map between CW complexes. If it induces a bijection on connected components and*

$$f_* : \pi_n(X,x) \to \pi_n\big(Y,f(x)\big)$$

is an isomorphism for all $x \in X$ and all n, then f is a homotopy equivalence.

Given this, it is perhaps not surprising that computing these homotopy groups is very difficult in general. As previously mentioned, the higher homotopy groups even for spheres are generally unknown.

Example (Hopf Fibration)*.* Consider the map $S^3 \to \mathbb{C} \cup \{\infty\}$ defined by

$$(x_1,x_2,x_3,x_4) \mapsto \frac{x_1 + ix_2}{x_3 + ix_4}$$

which maps to ∞ when the denominator vanishes. Identifying the one-point compactification $\mathbb{C} \cup \{\infty\}$ with the 2-sphere, this defines a map $\eta : S^3 \to S^2$, called the *Hopf fibration.* Over each point in S^2, the fibre of η consists of a circle S^1, and for any pair of distinct points in S^2, these circles pass through each other precisely once.

What is the relation with homotopy groups? One might think that just as $\pi_n(S^k) = 0$ for $n < k$, we would also have $\pi_n(S^k) = 0$ for $n > k$, with only $\pi_k(S^k) = \mathbb{Z}$ being non-zero.

However, it turns out that the Hopf fibration is not null-homotopic, and in fact,

$$\pi_3(S^2) \cong \mathbb{Z}[\eta]$$

is generated by the class of the Hopf fibration. △

However, there is a kind of stability that arises in these higher homotopy groups at a certain point.

Given two pointed spaces (X,x_0) and (Y,y_0), we define the *smash product* $X \wedge Y$ as the quotient

$$X \wedge Y := X \times Y \big/ (x,y_0) \sim (x_0,y)$$

where $x \in X$ and $y \in Y$.

In the product $X \times Y$, we can identify X and Y with the subspaces $X \times \{y_0\}$ and $\{x_0\} \times Y$, respectively. These subspaces intersect only at the point (x_0,y_0), so the union of these subspaces can be identified with the wedge sum $X \vee Y$. In particular, $\{x_0\} \times Y$ in $X \times Y$ is identified with Y in $X \vee Y$, and similarly for $X \times \{y_0\}$ and X; and in $X \vee Y$, the subspaces X and Y intersect at the single point $x_0 \sim y_0$. So, the smash product is also given by the quotient

$$X \wedge Y := X \times Y \big/ X \vee Y$$

The *reduced suspension* ΣX of a pointed space (X,x_0) is the quotient

$$\Sigma X := X \times I \big/ \big(X \times \{0\}\big) \cup \big(X \times \{1\}\big) \cup \big(\{x_0\} \times I\big)$$

This is equivalent to taking the ordinary suspension SX, then collapsing the line $\{x\} \times I$ connecting the two suspension points to a single point. The reduced suspension is also naturally a pointed space, with the basepoint given by the equivalence class of $(x_0,0)$.

It can be shown that the reduced suspension of X is homeomorphic to the smash product $X \wedge S^1$. More generally, the k-fold iterated reduced suspension is homeomorphic to the smash product

$$\Sigma^k \cong X \wedge S^k$$

Theorem 39.12.5 (Freudenthal Suspension Theorem)**.** *Let X be a pointed CW complex. Then, the suspension map*

$$\pi_{n+k}(\Sigma^k X) = [S^{n+k},\Sigma^k X] \to [\Sigma S^{n+k},\Sigma^{k+1} X] = [S^{n+k+1},\Sigma^{k+1} X] = \pi_{n+k+1}(\Sigma^{k+1},X)$$

is an isomorphism for $k \gg n$.

In fact, there are precise bounds on k,n (depending on X) for this map to become an isomorphism.

The theorem says that the group $\pi_{n+k}(\Sigma^k X)$ is independent of k for large k. So, we can define a set of groups based on this limiting behaviour as k becomes large:

Let X be a pointed CW complex. The *stable homotopy groups* are

$$\pi_n^s(X) := \varinjlim_{k \to \infty} \pi_{n+k}(\Sigma^k X)$$

Note that these make sense even for $n < 0$, and that these are all abelian.

These groups record only *stable* phenomena at the level of homotopy groups. Because of this, they are somewhat easier to compute than the original "unstable" homotopy groups, although still (too) difficult. For instance, the stable homotopy groups $\pi_n^s(*)$ are still largely unknown outside of small values of n, though it is known that infinitely many of them are non-zero.

Theorem 39.12.6. *Stable homotopy groups π_n^s define a generalised homology theory.*

Note that this theory violates the dimension axiom in an extreme way: $\pi_n^s(*)$ is non-zero for infinitely many n.

Exercise. The groups $\pi_n^s(*)$ are difficult to compute, but assuming these, would it then be possible to compute $\pi_n^s(X)$ for any CW complex X? An obvious idea would be to try reconstruct the uniqueness theorem for ordinary homology theories, and to construct a cellular chain complex with terms given by stable homotopy groups by spheres. What goes wrong?

Hint: note that $\pi_\bullet^s(S^n) = \pi_{\bullet-n}^s(*)$.

39.13 Exercises

1. Let A be abelian and $B_1,B_2 \leq A$ be subgroups. Prove that $B_1 \cong B_2$ does not imply $A/B_1 \cong A/B_2$.

   For instance, $2\mathbb{Z},3\mathbb{Z} \leq \mathbb{Z}$, and $2\mathbb{Z} \cong 3\mathbb{Z}$, but $\mathbb{Z}/2\mathbb{Z} \ncong \mathbb{Z}/3\mathbb{Z}$.

   Remark. The point is that this naïve notion of isomorphism for subgroups is too weak! They are isomorphic *as groups*, but not *as subgroups of A*: the isomorphism should also respect how these subgroups embed into A. In this case, the isomorphism between the two subgroups does not factor through the inclusions maps in both directions, so they are not isomorphic as subgroups. (See also *slice category* and §52.6.)

2. Prove that every free abelian group is isomorphic to a direct sum of copies of $\mathbb{Z}$.
3. Prove that the abelian group $\mathbb{Q}$ is not free abelian.
4. Compute the simplicial homology groups of the standard n-simplex Δ^n.
5. Let G,H be groups. Prove that $(G * H)^{\mathrm{ab}} \cong G^{\mathrm{ab}} \oplus H^{\mathrm{ab}}$.
6. Construct a Δ-complex structure on S^n and hence compute the simplicial homology groups.
7. Recall that the suspension SX of a space X is the quotient

   $$X \times [-1,1]/\sim$$

   where $(x,1) \sim (y,1)$ and $(x,-1) \sim (y,-1)$ for all $x,y \in X$. Prove that

   $$H_n(X) \cong \begin{cases} H_{n+1}(SX) & n \geq 1 \\ H_1(SX) \oplus \mathbb{Z} & n = 0 \end{cases}$$

   and $H_0(SX) \cong \mathbb{Z}$. (Or in reduced homology, $\tilde{H}_n(X) \cong \tilde{H}_{n+1}(SX)$ for all $n \geq -1$.)

8. Let $C_\bullet$ and $D_\bullet$ be chain complexes and $f_\bullet : C_\bullet \to D_\bullet$ be a chain map. Prove that the component map $f_n : C_n \to D_n$ induces a homomorphism

   $$f_n : H_n(C_\bullet) \to H_n(D_\bullet)$$

   for all $n \in \mathbb{Z}$.

9. Let X be a Δ-complex. Show that the natural map

   $$\Delta_\bullet(X) \to C_\bullet(X)$$

   from the simplicial to singular chain groups, given by regarding each n-simplex in the Δ-complex of X as a singular n-simplex, is a chain map.

10. Prove that chain homotopy is an equivalence relation on the set of chain maps between two fixed chain complexes.
11. Prove that chain homotopy respects composition. That is, if $f,f' : C_\bullet \to D_\bullet$ and $g,g' : D_\bullet \to E_\bullet$ are chain maps with $f \sim f'$ and $g \sim g'$, then $f \circ g \sim f' \circ g'$.
12. Let $(C_\bullet,d_\bullet)$ be the chain complex concentrated in degrees $[0,1]$ with $\mathbb{Z}$ and $d_1(n) = 2n$; and let $(D_\bullet,d'_\bullet)$ be the chain complex concentrated in degrees $[0,1]$ with $\mathbb{Z}^2$ and $d'_1(n,m) = (n-m,n+m)$. Prove that $C_\bullet$ and $D_\bullet$ are chain homotopy equivalent.
13. Show that Homotopy Invariance and the Mayer–Vietoris long exact sequence holds in reduced homology.
14. Compute the singular homology groups of $\mathbb{RP}^2$ using Homotopy Invariance and Mayer–Vietoris.

15. Compute the singular homology groups of m-punctured Euclidean n-space, $n \geq 2$, $m \geq 1$.

16. Let $X \supseteq A \supseteq B$ be topological (sub)spaces. Prove that there exists a long exact sequence

$$\begin{tikzcd}
\cdots \arrow[r] & H_n(A,B) \arrow[r] & H_n(X,B) \arrow[r] & H_n(X,A) \arrow[dll, rounded corners] & \\
& H_{n-1}(A,B) \arrow[r] & H_{n-1}(X,B) \arrow[r] & H_{n-1}(X,A) \arrow[r] & \cdots
\end{tikzcd}$$

in relative homology.

17. In this exercise, we generalise the Jordan Curve theorem to higher dimensions.

(a) Let $\gamma : I^k \to S^n$ be injective and continuous, with $k < n$. Show that

$$H_i\big(S^n \setminus \gamma(I^k)\big) \cong \begin{cases} \mathbb{Z} & i = 0 \\ 0 & i \neq 0 \end{cases}$$

Hint: Write $I^k = [0,1] \times I^{k-1} = [0,\frac{1}{2}] \times I^{k-1} \cup [\frac{1}{2},1] \times I^{k-1}$ and show that any non-trivial homology glass on $S^n \setminus \gamma(I^k)$ is also non-trivial on either $S^n \setminus \gamma\big([0,\frac{1}{2}] \times I^{k-1}\big)$ or on $S^n \setminus \gamma\big([\frac{1}{2},1] \times I^{k-1}\big)$. Repeat to contain a nested sequence of intervals $I_0 = [0,1] \supset I_1 \supset \cdots$ and continue by induction on k.

(b) Let $\gamma : S^k \to S^n$ be injective and continuous, with $k < n$. Show that

$$H_i\big(S^n \setminus \gamma(S^k)\big) \cong \begin{cases} \mathbb{Z}^2 & i = 0, \quad k = n-1 \\ \mathbb{Z} & i = 0, n-k-1, \quad k < n-1 \\ 0 & \text{else} \end{cases}$$

(c) Let $\gamma : S^{n-1} \to \mathbb{R}^n$ be injective and continuous, with $k < n$. Show that

$$H_i\big(\mathbb{R}^n \setminus \gamma(S^{n-1})\big) \cong \begin{cases} \mathbb{Z}^2 & i = 0 \\ \mathbb{Z} & i = n-1 \\ 0 & \text{else} \end{cases}$$

18. Let $p \in \mathbb{C}[z]$ be a non-constant complex polynomial. This defines a map $\mathbb{C} \to \mathbb{C}$, which we may extend to a map to the one-point compactification

$$p : \mathbb{C} \cup \{\infty\} \to \mathbb{C} \cup \{\infty\}$$

by setting $p(\infty) = \infty$.

(a) Verify that p is a continuous map.

(b) Noting that $\mathbb{C} \cup \{\infty\} \cong S^2$, show that the topological degree of p is equal to its polynomial degree.

Now, let $p = \sum_{i=0}^{n} a_i x^i \in \mathbb{R}[x]$ be a non-constant real polynomial, and similarly extend this to the one-point compactification $\mathbb{R} \cup \{\infty\} \cong S^1$.

(c) Show that

$$\deg(p) = \begin{cases} 0 & n \text{ even} \\ \operatorname{sgn} a_n & n \text{ odd} \end{cases}$$

19. In this exercise, we use degree theory to prove the fundamental theorem of algebra. That is, that every non-constant polynomial in $\mathbb{C}[z]$ has a root.

    (a) Suppose $n > 0$, and let
    $$p(z) = z^n + a_{n-1}z^{n-1} + \cdots a_1 z + a_0$$
    and define
    $$p_t(z) = z^n + t(a_{n-1}z^{n-1} + \cdots a_1 z + a_0)$$
    for $t \in [0,1]$. Let B_R be the open ball of radius R centred on the origin, and let $S^1_R := \partial B_R$ be its boundary. Show that if R is sufficiently large, then p_t has no root on S^1_R for any $t \in [0,1]$, so that
    $$F : S^1 \times [0,1] \to S^1 : (z,t) \mapsto \frac{p_t(Rz)}{p_t Rz}$$
    is well-defined and continuous.

    (b) Denote by F_0 and F_1 the maps $S^1 \to S^1$ obtained by restricting F to $S^1 \times \{0\}$ and $S^1 \times \{1\}$, respectively. Compute the degree of F_0.

    (c) Show that if p has no root in B_r, then F_1 has degree 0.

    (d) Conclude that p must have a root in B_R, and thus has a root in $\mathbb{C}$.

20. Show that the Klein bottle is the connected sum
$$K \cong \mathbb{RP}^2 \# \mathbb{RP}^2$$

21. Let $X = S^1 \vee S^1 \vee S^2$.

    (a) Show that X has the same homology groups as the torus $\mathbb{T}^2$ for all $n \in \mathbb{Z}$.

    (b) Show that there is no map $f : X \to \mathbb{T}^2$ that induces isomorphisms in homology in all degrees.

22. Compute the homology groups of $\mathbb{T}^n = \prod_n S^1$ and $S^n \times S^m$ for $n,m \geq 1$.

23. Let S_1 and S_2 be surfaces. Show that $S_1 \# S_2$ is orientable if and only if S_1 and S_2 are both orientable.

24. Let X be a CW complex or a Δ-complex. Show that the n and $(n-1)$-skeletons (X^n,X^{n-1}) form a good pair.

25. Let Σ_g be the g-holed torus, and set $\Sigma_0 = S^2$. Show that there exists an n-fold covering
$$\pi : \Sigma_h \to \Sigma_g$$
if and only if $h = n(g-1)+1$.

26. Define $\mathbb{R}^\infty$ to be the union
$$\mathbb{R}^\infty := \bigcup_{n\in\mathbb{N}} \mathbb{R}^n$$
where $\mathbb{R}^n \subset \mathbb{R}^{n+1}$ as the subspace with $(n+1)$th coordinate being zero. Similarly, define
$$D^\infty := \bigcup_{n\in\mathbb{N}} D^n \qquad S^\infty := \bigcup_{n\in\mathbb{N}} S^n$$

    (a) Show that S^∞ is contractible.

    (b) The proof of the Brouwer fixed point theorem does not apply to D^∞. Find a map $D^\infty \to D^\infty$ that does not have a fixed point.

27. Prove any/all of the results in the next section.

39.14 Results from Homological Algebra

39.14.1 Common Exact Sequences

If:

- $$0 \to A \xrightarrow{f} B \to 0$$

  is exact, then $A \cong B$; exactness at A gives $\ker(f) = \operatorname{im}(0) = 0$, so f is injective; and exactness at B gives $\operatorname{im}(f) = \ker(0) = B$, so f is surjective.

- $$0 \xrightarrow{f} A \xrightarrow{g} 0$$

  is exact, then $A \cong 0$, as $0 = \operatorname{im}(f) = \ker(g) = A$.

- $$0 \to A \to B \to C \to 0$$

  is exact and C is free abelian (or projective), then $B \cong A \oplus C$.

- $$0 \to A \xrightarrow{a} B \xrightarrow{f} C \xrightarrow{b} D \to 0$$

  is exact, then $A \cong \ker(f)$ and $D \cong \operatorname{coker}(f)$: exactness at A gives $\ker(a) = \operatorname{im}(0) = 0$, so a is injective and hence $A \cong \operatorname{im}(a)$; by exactness at B, $A \cong \operatorname{im}(a) = \ker(f)$.

39.14.2 Splitting Lemma

Lemma 39.14.1. *Given a short exact sequence*

$$0 \to A \xrightarrow{q} B \xrightarrow{r} C \to 0$$

the following statements are all equivalent:

(*i*) ***Left split**: There exists a morphism $t : B \to A$ such that $t \circ q = \mathrm{id}_A$;*

(*ii*) ***Right split**: There exists a morphism $u : C \to B$ such that $r \circ u = \mathrm{id}_C$;*

(*ii*) *There is an isomorphism $h : B \to A \oplus C$, such that $h \circ q = \iota_1 : A \hookrightarrow A \oplus C$ is the canonical inclusion mapping, and $r \circ h^{-1} = \pi_2 : A \oplus C \to C$ is the canonical projection mapping.*

If any of these statements hold, then the sequence is called a *split exact sequence*, or the sequence is said to *split*.

39.14.3 Five Lemma

Lemma 39.14.2 (Five Lemma)**.** *For the following commutative diagram,*

$$\begin{tikzcd}
A \arrow[r, "a"] \arrow[d, "\alpha"] & B \arrow[r, "b"] \arrow[d, "\beta"] & C \arrow[r, "c"] \arrow[d, "\gamma"] & D \arrow[r, "d"] \arrow[d, "\delta"] & E \arrow[d, "\varepsilon"] \\
A' \arrow[r, "a'"] & B' \arrow[r, "b'"] & C' \arrow[r, "c'"] & D' \arrow[r, "d'"] & E'
\end{tikzcd}$$

if the two rows are exact, β and δ are isomorphisms; α is an epimorphism; and ε is a monomorphism, then γ is an isomorphism.

39.14.4 Nine Lemma

Lemma 39.14.3 (Nine Lemma). *In the following commutative diagram,*

$$\begin{tikzcd}
& 0 \arrow[d] & 0 \arrow[d] & 0 \arrow[d] & \\
0 \arrow[r] & A_1 \arrow[r] \arrow[d] & B_1 \arrow[r] \arrow[d] & C_1 \arrow[r] \arrow[d] & 0 \\
0 \arrow[r] & A_2 \arrow[r] \arrow[d] & B_2 \arrow[r] \arrow[d] & C_2 \arrow[r] \arrow[d] & 0 \\
0 \arrow[r] & A_3 \arrow[r] \arrow[d] & B_3 \arrow[r] \arrow[d] & C_3 \arrow[r] \arrow[d] & 0 \\
& 0 & 0 & 0 &
\end{tikzcd}$$

if all columns and the lower two rows are exact, then the top row is exact as well. Similarly, if all columns and the upper two rows are exact, then the bottom row is exact as well.

The diagram is symmetric about the diagonal, so rows and columns may be interchanged in the above as well.

39.14.5 Snake Lemma

Lemma 39.14.4 (Snake Lemma). *In the following commutative diagram,*

$$\begin{tikzcd}
& A \arrow[r, "f"] \arrow[d, "a"] & B \arrow[r, "g"] \arrow[d, "b"] & C \arrow[r] \arrow[d, "c"] & 0 \\
0 \arrow[r] & A' \arrow[r, "f'"] & B' \arrow[r, "g'"] & C' &
\end{tikzcd}$$

if the rows are exact, then there is a connecting homomorphism $\partial : \ker(c) \to \operatorname{coker}(a)$ *and an exact sequence*

$$\ker(a) \to \ker(b) \to \ker(c) \xrightarrow{\partial} \operatorname{coker}(a) \to \operatorname{coker}(b) \to \operatorname{coker}(c)$$

Moreover, if f *is a monomorphism, then so is* $\ker(a) \to \ker(b)$*; and if* g' *is an epimorphism, then so is* $\operatorname{coker}(b) \to \operatorname{coker}(c)$.

Chapter 40

Cohomology

"It has been said that Poincaré did not invent topology, but that he gave it wings. This is surely true, and verges on understatement. His six great topological papers created, almost out of nothing, the field of algebraic topology."
— Donal O'Shea, *The Poincaré Conjecture: In Search of the Shape of the Universe*

Chapter 41

Manifolds

Chapter 42

Differential Geometry

Chapter 43

Hyperbolic Geometry

Chapter 44

Introduction to Vector Calculus

> *"Using the chain rule is like peeling an onion: you have to deal with each layer at a time, and if it is too big, you will start crying."*
>
> — Unattributed

44.1 Curves & Parametrisation

A function $\mathbf{r}$ is *vector-valued* if it maps a number $t \in \mathbb{R}$ to a vector $\mathbf{r}(t) \in \mathbb{R}^n$, where $n \geq 2$.

If $n = 1$, then the function is instead *scalar-valued.*

One way to represent a vector-valued function is to write it in terms of its scalar-valued components, such as $\mathbf{r}(t) = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k}$, where x, y and z are scalar-valued and $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ are some basis vectors of $\mathbb{R}^3$.

Let $\mathbf{r} : (I \subseteq \mathbb{R}) \to \mathbb{R}^n$ be a function. Then, the set $C = \{\mathbf{r}(t) : t \in I\}$ is said to be a *curve* that is *parametrised* by $\mathbf{r}$. The expression $\mathbf{r}(t)$ can be viewed as the position of a particle at time t, tracing out the curve as t varies; then, $\mathbf{r}'(t)$ represents the particle's velocity, $\mathbf{r}''(t)$ represents acceleration, and so on.

Note that the parametrisation of a curve is not unique.

Example. The two functions $\mathbf{r}$,$\mathbf{s} : \mathbb{R} \to \mathbb{R}^2$ defined by

$$\mathbf{r}(t) = (t,t)$$
$$\mathbf{s}(t) = (t^3,t^3)$$

parametrise the same curve

$$\Delta = \{(t,t) : t \in \mathbb{R}\}$$

△

Additionally, curves may also have a specified *direction* or *orientation* – a specified direction in which a parametrisation traces out the curve. If a direction is specified, the curve is *oriented*, and *non-oriented* otherwise.

Example. The two functions $\mathbf{r}$,$\mathbf{s} : [0,1] \to \mathbb{R}^2$ defined by

$$\mathbf{r}(t) = (t,t)$$
$$\mathbf{s}(t) = (1 - t,1 - t)$$

both parametrise the curve

$$\{(t,t) : t \in [0,1]\}$$

but $\mathbf{r}$ starts from (0,0) and ends at (1,1), while $\mathbf{s}$ starts from (1,1) and ends at (0,0) – they parametrise the same curve, but with different orientations. △

Example. Parametrise the semi-circle $x^2 + y^2 = 4$, $y \geq 0$, in the anticlockwise direction.

First, we rearrange the equation, solving for one of the variables:

$$\begin{aligned} x^2 + y^2 &= 4 \\ y^2 &= 4 - x^2 \\ y &= \sqrt{4 - x^2} \end{aligned}$$

Now, let $x = t$, so we have

$$\mathbf{r}(t) = \left(t, \sqrt{4 - t^2}\right)$$

We also need bounds for t; $x^2 + y^2 = 4$, so x should range from -2 to 2, giving $t \in [-2,2]$, but now we need to verify that the curve is being traced out in the right direction. We have $\mathbf{r}(-2) = (-2,0)$, so we appear to be starting at the wrong side. We can amend this with the following equation:

$$\mathbf{r}(t) = \left(-t, \sqrt{4 - t^2}\right), \qquad t \in [-2,2]$$

△

More generally, to reverse the direction of a parametrisation, $\mathbf{r}(t) = (x(t), y(t)), t \in [a, b]$, let $\mathbf{s}(t) = (x(a + b - t), y(a + b - t)), t \in [a, b]$. In this case, $a = -b$, so we just replace t with $-t$.

Another valid parametrisation for this example is given by:

$$\mathbf{r}(t) = \big(2\cos(t), 2\sin(t)\big), \qquad t \in [0,\pi]$$

If $\mathbf{r}(t)$ is infinitely differentiable, $\mathbf{r}(t)$ parametrises a *smooth* curve.

A curve parametrised by $\mathbf{r}(t), t \in [a,b]$ is *closed* or is a *loop* if $\mathbf{r}(a) = \mathbf{r}(b)$.

A curve that does not intersect itself is *embedded* or *simple*. A curve being embedded is equivalent to its parametrisation being injective (except possibly at the endpoints, if the curve is closed).

44.2 Vector Calculus

To differentiate a vector-valued function, we simply differentiate each scalar-valued component.

If $\mathbf{r}(t)$ is a parametrisation of a curve, then $\mathbf{r}'(c)$ is a tangent vector to the curve at the point $t = c$.

If $\mathbf{r}'(t) \neq 0$ for all t, then $\mathbf{r}$ is a *regular* parametrisation. Conversely, a curve is said to be regular if there exists a regular parametrisation of it.

Vector differentiation is *linear* (see §33.2.1 for a general overview of linearity and §33.6 for the linearity of differentiation).

If $f(t)$ is a scalar-valued function and $\mathbf{u}(t)$, $\mathbf{v}(t)$ are vector-valued functions, then,

- $f(t)\mathbf{u}(t) = f(t)\mathbf{u}'(t) + f'(t)\mathbf{u}(t)$;
- $\mathbf{u}(t) \cdot \mathbf{v}(t) = \mathbf{u}(t) \cdot \mathbf{v}'(t) + \mathbf{u}'(t) \cdot \mathbf{v}(t)$;

- $\mathbf{u}(t) \times \mathbf{v}(t) = \mathbf{u}(t) \times \mathbf{v}'(t) + \mathbf{u}'(t) \times \mathbf{v}(t)$;
- $\mathbf{u}(f(t)) = f'(t)\mathbf{u}'(f(t))$.

The notation $\mathbf{u}(t)\mathbf{v}(t)$ should be avoided, as it is unclear as to which product (cross or dot) is being applied.

Let $\mathbf{r}(t)$ be a vector-valued function such that $\|\mathbf{r}(t)\| = C$, where C is a constant. Then, $\mathbf{r}(t)$ and $\mathbf{r}'(t)$ are orthogonal (have zero dot product).

Proof.

$$\begin{aligned} \mathbf{r} \cdot \mathbf{r} &= ||\mathbf{r}||^2 \\ \frac{d}{dt}(\mathbf{r} \cdot \mathbf{r}) &= \frac{d}{dt}\mathbf{r}^2 \\ \mathbf{r} \cdot \mathbf{r}' + \mathbf{r}' \cdot \mathbf{r} &= 0 \\ \mathbf{r}' \cdot \mathbf{r} &= 0 \end{aligned}$$

■

If a curve is parametrised by $\mathbf{r}(t)$, the length of the curve between two given values of t, a and b, is given by,

$$\int_a^b ||\mathbf{r}'||\, dt$$

with $b \geq a$. Note that this returns a scalar, but in contrast:

The *arc length function* is given by the same integral, but with a new variable in the top bounds:

$$\mathbf{s}(t) = \int_a^t ||\mathbf{r}'(u)||\, du$$

This is a function, dependent on t.

The *arc length parametrisation* of $\mathbf{r}(t)$ is denoted $\mathbf{r}(s)$, where s is the arc length function. If $\mathbf{r}(t)$ is regular, then $||\mathbf{r}'(s)|| = 1$. This function may also be called the *unit-speed parametrisation.*

44.2.1 Curvature & Torsion

To measure curvature, we define a quantity as follows:

$$\kappa(s) = ||\mathbf{r}''(s)||$$

where $\mathbf{r}(s)$ is a unit speed parametrisation.

The greater the value of $\kappa(s)$, the more curvature the curve has.

Example. Let $\mathbf{r}(t) = \mathbf{a}t + \mathbf{b}$, where $||\mathbf{a}|| = 1$. Find the curvature.

$\mathbf{r}'(t) = \mathbf{a}, ||\mathbf{r}'(t)|| = ||\mathbf{a}|| = 1$, so $\mathbf{r}(t)$ is a unit speed parametrisation, and we may use our curvature equation.

$\kappa(s) = ||\mathbf{r}''(t)|| = 0$, so a line has zero curvature, as we'd might expect. △

The curvature of a circle with radius $R > 0$ has constant curvature $\frac{1}{R}$ at all points. The quantity $\frac{1}{\kappa(a)}$ is called the *radius of curvature*, and represents the radius of the circle that best approximates $\mathbf{r}(s)$ at $s = a$. Such a circle is called the *osculating circle* of the curve at that point.

Curvature is independent of the choice of parametrisation: it is an intrinsic property of a curve.

For a non-unit-speed regular parametrisation, we alternatively have:

$$\begin{aligned}\kappa(t) &= \frac{||\mathbf{r}'(t) \times \mathbf{r}''(t)||}{||\mathbf{r}'(t)||^3} \\ &= \frac{\mathbf{T}'(t)}{\mathbf{r}'(t)}\end{aligned}$$

where $\mathbf{T}(t)$ is the unit tangent.

44.2.2 Principle Normal & Binormal Vectors

As discussed earlier, the tangent vector of a curve is given by the derivative of its parametrisation. The *unit tangent*, $\mathbf{T}(t)$, is the normalised tangent vector.

We define the principle normal vector, $\mathbf{N}(s)$ as the vector that satisfies,

$$\mathbf{r}''(s) = \kappa(s)\mathbf{N} \text{ or equivalently, } \mathbf{T}'(s) = \kappa(s)\mathbf{N}(s), \text{ or } \kappa(s) = \mathbf{T}'(s) \cdot \mathbf{N}(s)$$

If $\kappa(s) = 0$, then the normal is undefined. The normal is perpendicular to the tangent, and points towards the centre of the osculating circle of the curve.

The *binormal* is defined as:

$$\mathbf{B}(s) = \mathbf{T}(s) \times \mathbf{N}(s)$$

$\mathbf{T}$, $\mathbf{N}$, and $\mathbf{B}$ are always orthogonal and form an orthonormal basis of $\mathbb{R}^3$. The basis $\{\mathbf{T},\mathbf{N},\mathbf{B}\}$ is called the *Frenet-Serret frame.*

We define *torsion*, τ as,

$$\mathbf{B}'(s) = -\tau\mathbf{N}(s)$$

or,

$$\tau = -\mathbf{B}'(s) \cdot \mathbf{N}(s)$$

Torsion is also independent of the choice of parametrisation.

Most of the results from the previous two sections can be compactly summarised as a matrix equation as follows:

$$\begin{aligned}\begin{bmatrix}\mathbf{T}' \\ \mathbf{N}' \\ \mathbf{B}'\end{bmatrix} &= \begin{bmatrix}0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0\end{bmatrix}\begin{bmatrix}\mathbf{T} \\ \mathbf{N} \\ \mathbf{B}\end{bmatrix} \\ &= \begin{bmatrix}\kappa\mathbf{N} \\ -\kappa\mathbf{T} + \tau\mathbf{B} \\ -\tau\mathbf{N}\end{bmatrix}\end{aligned}$$

where $\mathbf{T}$, $\mathbf{N}$, and $\mathbf{B}$ are understood to be functions of the unit-speed function of t.

44.3 Multivariable Scalar-Valued Functions

A *multivariable scalar-valued function* maps vectors to scalars. For example, $f(x,y) = \sqrt{x^2 + y^2}$ is a multivariable scalar-valued function.

For a general function, $f : \mathbb{R}^3 \to \mathbb{R}$, a set of points $f(x,y,\cdots) = C$ where C is a constant is called a *level set.* When $n = 2$, level sets are also called *contour lines.* When $n = 3$, level sets are also called *isosurfaces.* For the function above, we may also represent f as a *surface* in $\mathbb{R}^3$ by setting $z = f(x,y)$.

Multivariable functions require *multiple* derivatives to fully describe rates of change in every direction – one for each dimension. For this, we use *partial derivatives.*

The partial derivative of a function $f(x,y,z)$ with respect to x, is variously written as,

$$\frac{\partial f}{\partial x}, f_x, \partial_x f$$

Other notations exist, but these are the main ones we will use.

The second-order partial derivative of f with respect to x is written as,

$$\frac{\partial^2 f}{\partial x^2}, f_{xx}, \partial_{xx} f, \partial_x^2 f$$

and the second-order mixed derivative of f with respect to x, then y is given by,

$$\frac{\partial^2 f}{\partial y \partial x}, f_{xy}, \partial_{yx} f, \partial_y \partial_x f$$

Let f, g be functions of $(x,y,\cdots)$. Then,

$$\frac{\partial}{\partial x}(fg) = f\frac{\partial g}{\partial x} + g\frac{\partial f}{\partial x}$$

Let f be a function of (x,y), and x, y be functions of t. Then,

$$\frac{df}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}$$

Let $f : \mathbb{R}^n \to \mathbb{R}$. We define the *gradient* of f, denoted $\text{grad}\, f$ or ∇f, as the vector $(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \frac{\partial f}{\partial x_3}, \ldots, \frac{\partial f}{\partial x_n})$.

∇ is the *grad operator*, and is effectively a vector full of differential operators. ∇ takes a scalar, $f(x,y,z)$, and returns a vector, ∇f.

The *directional derivative* of f in the direction of a unit vector, $\mathbf{v}$ is given by $D = \mathbf{v} \cdot \nabla f$. The directional derivative is the rate of change of f in the direction of $\mathbf{v}$.

44.3.1 Linear Approximations

Recall that the linear Taylor approximation of a function $f(x)$ about a point $x = a$ is given by $f(a) + f'(a)(x - a)$. We can similarly approximate a surface in $\mathbb{R}^3$ as a plane using partial derivatives.

A function $f(x,y)$ at the point (a,b) has a *linear approximation* given by

$$f(a,b) + \left[\frac{\partial f}{\partial x}(a,b)\right](x - a) + \left[\frac{\partial f}{\partial y}(a,b)\right](y - b)$$

This generalises to scalar valued functions of any number of variables. In general, the linear approximation of a function, $f : \mathbb{R}^n \to \mathbb{R}$ about a point $\mathbf{a}$, is given by,

$$f(\mathbf{x}) = f(\mathbf{a}) + [\nabla f(\mathbf{a})](\mathbf{x} - \mathbf{a})$$

where $\nabla f(\mathbf{a})$ is ∇f evaluated at $\mathbf{a}$.

Example. Find the linear approximation of $f(x,y) = \sqrt{1 - x^2 - y^2}$ about the point (0,0).

Setting $z = f(x,y)$, we have $z = \sqrt{1 - x^2 - y^2}$, so $x^2 + y^2 + z^2 = 1$, which is the upper half of the sphere of radius 1 centred on the origin, so we should expect our linear approximation to be a horizontal plane parallel to the $x - y$ plane.

$$\begin{aligned}\nabla f &= \left(-\frac{x}{\sqrt{1 - x^2 - y^2}}, -\frac{y}{\sqrt{1 - x^2 - y^2}}\right) \\ \nabla f(0,0) &= 0 \\ f(0,0) &= 1\end{aligned}$$

so $f(x,y)$ is approximately 1 near (0,0). This corresponds to the plane $z = 1$ approximating the hemisphere of radius 1. △

We can also find the normal vector to a curve at a given point. To do this, we can use the fact that $\nabla f(a,b)$ is normal to $f(a,b)$.

Example. Let $f(x,y) = x + 2\sin(x + y)$. Find the normal to the curve $f(x,y) = 0$ at the point (0,0).

$$\begin{aligned}\nabla f &= (1 + 2\cos(x + y), 2\cos(x + y)) \\ \nabla f(0,0) &= (3,2)\end{aligned}$$

so the normal is (3,2), or any vector along the line $y = \frac{2x}{3}$. △

44.3.2 Critical Points

A *critical point* of a function, $f(x)$ is a point where $\frac{df}{dx} = 0$. Critical points can be classified using the *second derivative test.*

Suppose $f'(x) = 0$ at $x = a$. If,

- $f''(a) > 0$, the critical point is a local minimum;
- $f''(a) < 0$, the critical point is a local maximum;
- $f''(a) = 0$, the test is inconclusive.

Note: Say $x = a$ is a local minimum/maximum point or that $f(a)$ is a local maximum.

Now, we can extend this definition to functions of two variables. A function, $f(x,y)$ has a critical point at (a,b) if $\nabla f(a,b) = 0$. That is, *every* partial derivative has to evaluate to 0 at the given point. We can similarly classify these critical points using the *Hessian matrix* and the *second partial derivative test.*

$$\mathbf{H} = \begin{bmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{bmatrix}$$

Let $D = \det(\mathbf{H}) = f_{xx}f_{yy} - f_{xy}^2$ ($f_{xy} = f_{yx}$ by Young's theorem), and suppose $\nabla f(x,y) = 0$ at (a,b). If,

- $D > 0$ and $f_{xx} > 0$ at (a,b), then (a,b) is a local minimum point;
- $D > 0$ and $f_{xx} < 0$ at (a,b), then (a,b) is a local maximum point;
- $D < 0$ at (a,b), then (a,b) is a saddle point;
- $D = 0$ at (a,b), then the test is inconclusive.

If it is easier to calculate, you may check f_{yy} instead of f_{xx} for the first two cases.

Performing this test on a single-variable function just gives the standard second-derivative test.

For functions or three or more variables, the determinant alone does not provide sufficient information to classify the critical point. Instead, we check the eigenvalues of the Hessian.

Suppose we have $f : \mathbb{R}^n \to \mathbb{R}$ with $n \geq 3$, and that $\nabla f(\mathbf{x}) = 0$ at $\mathbf{x} = \mathbf{a}$. If the Hessian has,

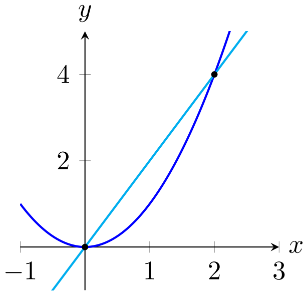
- all eigenvalues positive at $\mathbf{a}$, then $\mathbf{a}$ is a local minimum point;
- all eigenvalues negative at $\mathbf{a}$, then $\mathbf{a}$ is a local maximum point;
- both positive and negative eigenvalues at $\mathbf{a}$, then $\mathbf{a}$ is a saddle point;
- other; the test is inconclusive.

44.4 Integration

Most of these methods are best explained through example.

44.4.1 Double Integration

Example. Write down the area between $y = x^2$ and $y = 2x$ as a double integral in two different orders. △



We observe that the points of intersection are (0,0) and (2,4), which will be helpful for some of our integration bounds.

Let's do the order $dy\,dx$ first. Working from the outermost integral to the innermost integral: We're looking at dx first – the change in x. x varies between 0 and 2, so our outermost integral should go from 0, up to 2. Now, how does y vary? On the graph, x^2 is below $2x$, so we have x^2 as our lower bound and $2x$ as our upper. Overall, we have

$$\int_0^2 \int_{x^2}^{2x} 1\,dy\,dx$$

Doing the other order is very similar. y varies from 0 to 4, so our outermost integral goes from 0 to 4. Now, how does x vary? Well, the line $y = 2x$ is "below" the line $y = x^2$ (taking right to be the positive "upwards" direction), so we should go from $y = 2x$ to $y = x^2$. But we're integrating with respect to x, so we can't have x in our integration bounds, so rearrange for y to get $x = \frac{y}{2}$ and $x = \sqrt{y}$ as our bounds. Overall, we have

$$\int_0^4 \int_{\frac{y}{2}}^{\sqrt{y}} 1\,dx\,dy$$

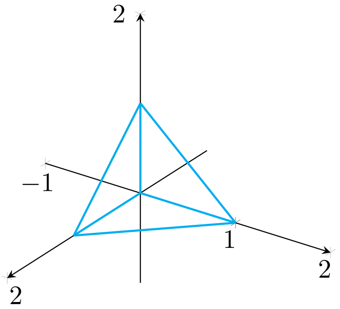
Note that, for a double integral for working out area, your outermost bounds should always be numbers and not functions of any of the variables being integrated with respect to. If the sole integrand is 1, it also may be tastefully omitted.

More generally, the double integral of some function $\iint\limits_R f(x,y)\,dA$ evaluated over some region, R, represents the volume between $f(x,y)$ and the $x-y$ plane over the region. If $f(x,y)=1$ as above, then this is just finding the area of the region.

You can also think of these multiple integrals in terms of finding masses, with $f(x,y)$ being some kind of density function. $f(x,y)=1$ would then represent a constant density, just giving area in the 2D case.

44.4.2 Triple Integration

Example. Write down the volume contained within the tetrahedron with vertices (0,0,0), (1,0,0), (0,1,0) and (0,0,1) as a triple integral. △



Let's integrate with the equilateral face being a function of x and y over the triangular region in the $x-y$ plane bounded by (0,0,0), (1,0,0) and (0,1,0). (Looking down at the tetrahedron from above, everything is contained within the triangle bounded by those three points.)

The lines of interest are the x and y axes, and $x+y=1$.

Using our previous method, x varies from 0 to 1 in this region, so our outermost integral goes from 0 to 1. Now, how does y vary in terms of x? The line $x+y=1$ is above the x axis here, so we go from $y=0$ to $x+y=1$, but again, no x's in our bounds, so rearrange for $y=1-x$, and we have our two outermost integrals sorted.

Now, just look at the z-axis. How does z vary in terms of x and y? The $x-y$ plane is below the plane $x+y+z=1$, so we go from the $x-y$ plane, which is given by $z=0$, to the plane $x+y+z=1$, which is rearranged to $z=1-x-y$. Overall, we have,

$$\int_0^1 \int_0^{1-x} \int_0^{1-x-y} dz\,dy\,dx$$

44.4.3 Change of Coordinate System

44.4.3.1 Polar Coordinates

In $\mathbb{R}^2$, the *Cartesian coordinates*, (x,y), and *polar coordinates*, (r,θ) are related by,

$$x = r\cos\theta$$
$$y = r\sin\theta$$

We may convert the double integral,

$$\iint\limits_R f(x,y)\,dA = \int\int f(r\cos\theta, r\sin\theta) r\,dr\,d\theta$$

Remember to multiply by an extra r.

In general, if you see many (x^2+y^2)'s in $f(x,y)$, converting to polar coordinates might be a good idea.

Example. Find the area in the first quadrant bounded by the polar curve $r = 1 + \cos\theta$. △

For polar double integrals, it is typically easier to use the angle in the outermost integral. So, using our formula from before, $f(x,y) = 1$, so we have,

$$\int\int r\,dr\,d\theta$$

But what are our bounds? We're looking for bounds in terms of angles, so clearly the first quadrant is bounded by $\theta = 0$ and $\theta = \frac{\pi}{2}$. Now, how does r vary? In this case, it's fairly simple, as r just goes from 0 up to the given curve, so we have,

$$\int_0^{\frac{\pi}{2}}\int_0^{1-\cos\theta} r\,dr\,d\theta$$

44.4.3.2 Cylindrical Coordinates

In $\mathbb{R}^3$, the Cartesian coordinates, (x,y,z), and *cylindrical coordinates*, (r,θ,z) are related by,

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta \\ z &= z \end{aligned}$$

We may convert the triple integral,

$$\iiint\limits_{\Omega} f(x,y,z)\,dV = \int\int\int f(r\cos\theta,r\sin\theta,z)r\,dr\,d\theta\,dz$$

Remember to multiply by an extra r.

Again, if you see many $(x^2 + y^2)$'s in $f(x,y,z)$, converting to cylindrical coordinates might be a good idea.

Example. Find the value of

$$\iiint\limits_{\Omega} (x^2 + y^2)\,dV$$

where Ω is the region bounded by the surfaces $x^2 + y^2 = 1$, $z = 2 - x$, and $z = 0$. △

First, plug in our formula:

$$\iiint\limits_{\Omega} x^2 + y^2\,dV = \iiint\limits_{\Omega} (r\cos\theta)^2 + (r\sin\theta)^2 r\,dr\,d\theta\,dz$$

Simplify the sines and cosines to get,

$$\iiint\limits_{\Omega} r^3\,dr\,d\theta\,dz$$

which certainly looks easier to do. But what are our bounds of integration? As we said earlier, it's easier to work with the angles on the outside, so let's swap the order of integration to $dr\,dz\,d\theta$. The surfaces given are now $r = 1$, $z = 2 - r\cos\theta$ and $z = 0$. $r = 1$ takes all angles, so we go from 0 to 2π for our outermost integral. Next, how does r vary? Well, r just goes from 0 (the z-axis) out to the curve $r = 1$, so we go from 0 to 1 for the middle integral. Finally, we look at how z varies. We are given two surfaces for z: $z = 2 - r\cos\theta$ and $z = 0$. As r is at most 1 over this region, $2 - r\cos\theta > 0$, so we know that z goes from 0 to $2 - r\cos\theta$. Overall, we have,

$$\int_0^{2\pi}\int_0^{1}\int_0^{2-r\cos\theta} r^3\,dr\,dz\,d\theta$$

Skipping over the actual integration, you find that this evaluates to π.

44.4.3.3 Spherical Coordinates

In $\mathbb{R}^3$, the Cartesian coordinates, (x,y,z), and *spherical coordinates*, (r,θ,ϕ) are related by,

$$\begin{aligned} x &= r\cos\theta\sin\phi \\ y &= r\sin\theta\sin\phi \\ z &= r\cos\phi \end{aligned}$$

(You can think of it like applying a cylindrical coordinate transformation to x, y and z, then polar coordinates to (x,y) and z)

Like in polar/cylindrical coordinates, θ is measured from the positive x-axis. ϕ is measured from the positive z-axis. Notice that, unlike θ, ϕ only ever varies from 0 to π. If ϕ goes past that, then the same point can be reached with a smaller angle, just by increasing θ by π radians.

We may convert the triple integral,

$$\iiint\limits_{\Omega} f(x,y,z)\,dV = \int\int\int f(r\cos\theta\sin\phi,r\sin\theta\sin\phi,r\cos\phi)r^2\sin\phi\,dr\,d\theta\,dz$$

Remember the extra factor of $r^2\sin\phi$.

Similarly, if you see many $(x^2+y^2+z^2)$'s in $f(x,y,z)$, converting to spherical coordinates might be a good idea.

Example. Find a triple integral expression for the volume of a sphere of radius R centred on the origin. △

Since we're finding a volume, $f(x,y,z)=1$, so $f(r\cos\theta\sin\phi,r\sin\theta\sin\phi,r\cos\phi)=1$. As before, it's easier to work with when angles are in the outermost integrals, let's do $dr\,d\theta\,d\phi$ for our order of integration. First, how does ϕ vary? Looking at the sphere, notice that it contains both the positive and negative z-axis, so ϕ varies from 0 to π (the sphere contains points directly above and below the origin). next, look straight down at the $x-y$ plane from the positive z-axis. The sphere lies in all 4 quadrants, so θ takes all values possible, varying from 0 to 2π. Finally, r goes from 0 up to the surface of the sphere which lies at a constant distance R away from the origin. Overall, we have,

$$\int_0^{\pi}\int_0^{2\pi}\int_0^{R} dr\,d\theta\,d\phi$$

Example. Find the value of

$$\iiint\limits_{\Omega} z^2\,dV$$

where Ω is the region bounded by two spheres of radius 1 and 2 centred at the origin. △

Given that we have two spheres, doing this in Cartesian coordinates is a horrible idea. So we convert our dV into $r^2\sin\phi\,dr\,d\theta\,d\phi$ and use spherical coordinates.

$$\iiint\limits_{\Omega}^{2} dV = \iiint\limits_{\Omega} (r\cos\phi)^2 r^2\sin\phi\,dr\,d\theta\,d\phi$$

We just worked out the bounds for a spherical region, so we can mostly just plug them in. However, we're looking for the volume between two spheres, so we look at how r varies in this situation. Thinking about the graph from the perspective of r, the sphere of radius 1 is "below" (closer to the origin) than the sphere of radius 2, so r varies from 1 to 2. So, overall, we have,

$$\int_0^{\pi}\int_0^{2\pi}\int_1^{2} r^4\cos^2\phi\sin\phi\,dr\,d\theta\,d\phi$$

Example. Find the value of

$$\iiint\limits_{\Omega} \sqrt{\exp(x^2+y^2+z^2)^3}\,dV$$

where Ω is the region bounded by $x^2+y^2+z^2=1$. △

We use the same conversion again, noting that the volume being integrated over is a sphere. Also note that $x^2+y^2+z^2=r^2$, so we can write,

$$\iiint\limits_{\Omega} \sqrt{\exp(x^2+y^2+z^2)^3}\,dV = \int_0^{\pi}\int_0^{2\pi}\int_0^1 sqrt\exp(r^2)^3\,dr\,d\theta\,d\phi$$

44.4.3.4 Arbitrary Change of Coordinates

The *Jacobian matrix* of a function, $\mathbf{f}:\mathbb{R}^n\to\mathbb{R}^n$, denoted $D\mathbf{f}$, is the matrix of partial derivatives,

$$\begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \dfrac{\partial f_1}{\partial x_3} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \dfrac{\partial f_2}{\partial x_3} & \cdots & \dfrac{\partial f_2}{\partial x_n} \\ \dfrac{\partial f_3}{\partial x_1} & \dfrac{\partial f_3}{\partial x_2} & \dfrac{\partial f_3}{\partial x_3} & \cdots & \dfrac{\partial f_3}{\partial x_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \dfrac{\partial f_n}{\partial x_3} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{bmatrix}$$

This can be more compactly written as,

$$\begin{bmatrix} \dfrac{\partial \mathbf{f}}{\partial x_1} & \cdots & \dfrac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix}$$

or,

$$\begin{bmatrix} \nabla^\top f_1 \\ \vdots \\ \nabla^\top f_n \end{bmatrix}$$

If $\mathbf{F}:\mathbb{R}^2\to\mathbb{R}^2$ is a bijection defined by $\mathbf{F}(u,v)=(x=a(u,v),y=b(u,v))$ for some functions, a and b, then the integral of $f:A\subseteq\mathbb{R}^2\to\mathbb{R}$ can be given by,

$$\iint\limits_{A} f(x,y)\,dx\,dy = \iint\limits_{B} f(u,v)|\det(\mathbf{J}(\mathbf{F}(u,v)))|\,du\,dv$$

where $\mathbf{J}$ is the Jacobian of the function $\mathbf{F}(u,v)$, and B is the same region in the new coordinate system.

From this point onwards, when the function is obvious and there is little room for confusion, we will write $\mathbf{J}$ to denote $\mathbf{J}(\mathbf{F}(\mathbf{x}))$.

Omitting the function $f(x,y)$, we can more compactly write $dA = dx\,dy = |\det(\mathbf{J})|\,du\,dv$.

For a triple integral, we similarly have, $dV = dx\,dy\,dz = |\det(\mathbf{J})|\,du\,dv\,dw$.

Example. Verify the area and volume element conversions previously found using the formula outlined above. △

For Cartesian $\mathbb{R}^2$ to polar, we have $\mathbf{F}(r,\theta) = (x = r\cos\theta, y = r\sin\theta)$.

$$\begin{aligned}\mathbf{J} &= \begin{bmatrix}\nabla x \\ \nabla y\end{bmatrix} \\ &= \begin{bmatrix}\dfrac{\partial x}{\partial r} & \dfrac{\partial x}{\partial \theta} \\ \dfrac{\partial y}{\partial r} & \dfrac{\partial y}{\partial \theta}\end{bmatrix} \\ &= \begin{bmatrix}\cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta\end{bmatrix}\end{aligned}$$

$\det(\mathbf{J}) = r(\cos\theta)^2 + r(\sin\theta)^2 = r((\cos\theta)^2 + (\sin\theta)^2) = r$, so $dA = r\,dr\,d\theta$, as we found before.

For Cartesian $\mathbb{R}^3$ to cylindrical, we have $\mathbf{F}(r,\theta,z) = (x = r\cos\theta, y = r\sin\theta, z = z)$.

$$\begin{aligned}\mathbf{J} &= \begin{bmatrix}\nabla x \\ \nabla y \\ \nabla z\end{bmatrix} \\ &= \begin{bmatrix}\dfrac{\partial x}{\partial r} & \dfrac{\partial x}{\partial \theta} & \dfrac{\partial x}{\partial z} \\ \dfrac{\partial y}{\partial r} & \dfrac{\partial y}{\partial \theta} & \dfrac{\partial y}{\partial z} \\ \dfrac{\partial z}{\partial r} & \dfrac{\partial z}{\partial \theta} & \dfrac{\partial z}{\partial z}\end{bmatrix} \\ &= \begin{bmatrix}\cos\theta & -r\sin\theta & 0 \\ \sin\theta & r\cos\theta & 0 \\ 0 & 0 & 1\end{bmatrix}\end{aligned}$$

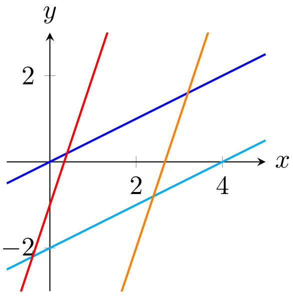
By Sarrus' rule, $\det(\mathbf{J}) = r(\cos\theta)^2 + r(\sin\theta)^2 = r((\cos\theta)^2 + (\sin\theta)^2) = r$, so $dV = r\,dr\,d\theta\,dz$ as before.

Spherical coordinates are similar, but algebraically involved and time consuming, so they will be omitted.

Example. Evaluate

$$\iint\limits_R \frac{x-2y}{3x-y}\,dA$$

where R is the region enclosed by the lines $y = \frac{x}{2}$, $y = \frac{x}{2} - 2$, $y = 3x - 1$ and $y = 3x - 8$. △



Let $u = x - 2y$ and $v = 3x - y$. Rearranging for x and y, we have $x = \frac{2v}{5} - \frac{u}{5}$ and $y = \frac{v}{5} - \frac{3u}{5}$.

Let $\mathbf{F}(x,y) = (x = \frac{2v}{5} - \frac{u}{5}, y = \frac{v}{5} - \frac{3u}{5})$.

$$\mathbf{J} = \begin{bmatrix} -\dfrac{1}{5} & \dfrac{2}{5} \\ -\dfrac{3}{5} & \dfrac{1}{5} \end{bmatrix}$$

$\det(\mathbf{J}) = \frac{1}{5}$, so we have,

$$\iint\limits_R \frac{x-2y}{3x-y}\,dA = \iint\limits_S \frac{u}{v}\frac{1}{5}\,du\,dv$$

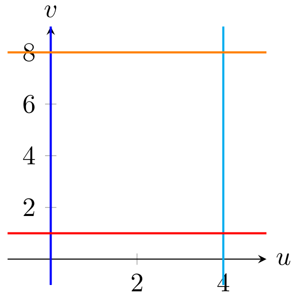
Now to find the bounds. Rearrange the lines defining the region:

$$y = \frac{x}{2} \Leftrightarrow x - 2y = 0 \Leftrightarrow u = 0$$

$$y = \frac{x}{2} - 2 \Leftrightarrow x - 2y = 4 \Leftrightarrow u = 4$$

$$y = 3x - 1 \Leftrightarrow 3x - y = 1 \Leftrightarrow v = 1$$

$$y = 3x - 8 \Leftrightarrow 3x - y = 8 \Leftrightarrow v = 8$$



So we have $u = 0$, $u = 4$, $v = 1$, and $v = 8$ as our bounds.

$$\int_1^8 \int_0^4 \frac{u}{5v}\,du\,dv$$

44.5 Vector Fields

Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is a scalar-valued function, assigning a scalar to every point in $\mathbb{R}^n$. Such a function may also be called a *scalar field*. Also recall that a function $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ is vector-valued, assigning a m-vector to every point in $\mathbb{R}^n$. Such a function may be called a *vector field*. We now consider the case for $n = m = 2$ and $n = m = 3$.

If a vector field, $\mathbf{F}(x,y)$ can be written as the gradient of some scalar valued function, $f(x,y)$, then $\mathbf{F}$ is a vector field whose flow is normal to the level sets of f. Such a field is called a *conservative* field.

For example, the vector field $\mathbf{F}(x,y) = (x,0)$ can also be written as the gradient of the function $f(x,y) = \frac{x^2}{2}$, so $\mathbf{F}$ is conservative.

44.5.1 Divergence & Curl

In this section, we will mostly be considering functions from $\mathbb{R}^3$ to $\mathbb{R}^3$. Recall the del operator, ∇, which we previously used to write the gradient function. We now use the same symbol to introduce some new operators:

$$\begin{aligned}\operatorname{grad} f &= \nabla f\\ \operatorname{div}\mathbf{F} &= \nabla\cdot\mathbf{F}\\ \operatorname{curl}\mathbf{F} &= \nabla\times\mathbf{F}\end{aligned}$$

These last two operators look rather strange, given that del is an operator and not a vector, but the way we've written them is a useful mnemonic as to how we actually calculate them: pretend the del is a vector full of partial differential operators, and perform the calculation as indicated. Whenever you multiply a function by a differential operator, instead apply the operator to the function.

Let $\mathbf{F}(x,y,z) = (f(x),g(y),h(z))$. Then,

$$\operatorname{div}\mathbf{F} = \nabla\cdot\mathbf{F} = \begin{bmatrix}\dfrac{\partial}{\partial x}\\ \dfrac{\partial}{\partial y}\\ \dfrac{\partial}{\partial z}\end{bmatrix} \text{“}\cdot\text{”} \begin{bmatrix}f\\ g\\ h\end{bmatrix} = \frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} + \frac{\partial h}{\partial z}$$

$$\operatorname{curl}\mathbf{F} = \nabla\times\mathbf{F} = \begin{bmatrix}\dfrac{\partial}{\partial x}\\ \dfrac{\partial}{\partial y}\\ \dfrac{\partial}{\partial z}\end{bmatrix} \text{“}\times\text{”} \begin{bmatrix}f\\ g\\ h\end{bmatrix} = \begin{bmatrix}\mathbf{i} & \mathbf{j} & \mathbf{k}\\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z}\\ f & g & h\end{bmatrix} = \begin{bmatrix}\dfrac{\partial h}{\partial y} - \dfrac{\partial g}{\partial z}\\ \dfrac{\partial h}{\partial z} - \dfrac{\partial g}{\partial x}\\ \dfrac{\partial h}{\partial x} - \dfrac{\partial g}{\partial y}\end{bmatrix}$$

The grad operator takes a scalar and returns a vector. The div operator takes a vector and returns a scalar. The curl operator takes a vector and returns a vector.

The curl of a conservative field is 0.

44.5.2 Parametric Surfaces

Recall that a curve in $\mathbb{R}^2$ or $\mathbb{R}^3$ can be parametrised with a one variable function. A surface in $\mathbb{R}^3$ can similarly be parametrised using a two variable function. You will likely have seen some of these previously, such as in the double vector parametrisation of a plane.

Example. Parametrise the surface in $\mathbb{R}^3$ given by $x^2 + y^2 + z^2 = A^2$. △

This is clearly a sphere, so try spherical coordinates with $r = A$. Then, $x = A\cos u \sin v$, $y = A\sin u \sin v$, $z = A\cos v$, with $u \in [0,2\pi), v \in [0,\pi]$. Remember to give intervals for your parameters.

Example. Find the Cartesian equation of the surface parametrised by,

$$\mathbf{r}(u,v) = (\sqrt{1-u}\cos v, \sqrt{1-u}\sin v, u), \quad u \in (-\infty, 1], v \in [0, 2\pi]$$

△

$x = \sqrt{1-u}\cos v$, $y = \sqrt{1-u}\sin v$ and $z = u$. Now look for ways to eliminate things. For example, $x^2 + y^2 = (1-u)\cos^2 v + (1-u)\sin^2 v = 1 - u = 1 - z$, so we have $x^2 + y^2 + z = 1$ for our Cartesian equation.

44.5.3 Surface Integrals

$$dS = \left|\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v}\right| du\, dv$$

Note: you only need the length of the cross product and not the vector itself, so to save time, you can check if $\frac{\partial \mathbf{r}}{\partial u} \cdot \frac{\partial \mathbf{r}}{\partial v} = 0$. If so, then you can instead use $\left|\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v}\right| = \left|\frac{\partial \mathbf{r}}{\partial u}\right|\left|\frac{\partial \mathbf{r}}{\partial v}\right|$.

Example. Find the formula for the surface area element for an arbitrary function $z = f(x,y)$. △

$$\begin{aligned}
\mathbf{r}(x,y) &= \begin{bmatrix} x \\ y \\ f(x,y) \end{bmatrix} \\
\frac{\partial \mathbf{r}}{\partial x} &= \begin{bmatrix} 1 \\ 0 \\ \dfrac{\partial f}{\partial x} \end{bmatrix} \\
\frac{\partial \mathbf{r}}{\partial y} &= \begin{bmatrix} 0 \\ 1 \\ \dfrac{\partial f}{\partial y} \end{bmatrix} \\
\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} &= \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & 0 & \dfrac{\partial f}{\partial x} \\ 0 & 1 & \dfrac{\partial f}{\partial y} \end{bmatrix} \\
&= \begin{bmatrix} -\frac{\partial f}{\partial x} \\ -\frac{\partial f}{\partial y} \\ 1 \end{bmatrix}
\end{aligned}$$

so $\left|\frac{\partial f}{\partial y} \times \frac{\partial f}{\partial v}\right| = \sqrt{\frac{\partial f}{\partial x}^2 + \frac{\partial f}{\partial y}^2 + 1}$, and,

$$\iint dS = \int\int \sqrt{\frac{\partial f}{\partial x}^2 + \frac{\partial f}{\partial y}^2 + 1}\, dx\, dy$$

44.5.4 Divergence Theorem

Flux is a vector quantity that describes how a fluid would flow through a surface, S, of a volume, V.

The *divergence theorem* says,

$$\iiint\limits_V \nabla \cdot \mathbf{F}\, dV = \iint\limits_S \mathbf{F} \cdot \mathbf{n}\, dS$$

where $\mathbf{n}$ is the outward-pointing unit normal to the surface S. Note, instead of S, sometimes, ∂V is written for the surface area element. Also, for the flux integral, sometimes $\mathbf{F} \cdot d\mathbf{S}$ is written instead of $\mathbf{F} \cdot \mathbf{n}\, dS$. $d\mathbf{S}$ is also equal to $\pm(\mathbf{r}_u \times \mathbf{r}_v)\, dy\, dv$. This form may be easier to caclulate, as you avoid finding the normal vector.

Example. Let V be the volume bounded by the unit sphere, and let $\mathbf{F}(x,y,z) = (z,y,x)$. Calculate the net flux over the surface of the sphere. △

First, write the equation for flux

$$\begin{aligned}
\iint\limits_S \mathbf{F}\cdot\mathbf{n}\,dS &= \iiint\limits_V \nabla\cdot\mathbf{F}\,dV \\
&= \iiint\limits_V \begin{bmatrix}\dfrac{\partial}{\partial x}\\ \dfrac{\partial}{\partial y}\\ \dfrac{\partial}{\partial z}\end{bmatrix}\cdot\begin{bmatrix}z\\ y\\ x\end{bmatrix} dV \\
&= \iiint\limits_V 0+1+0\,dV \\
&= \iiint\limits_V dV
\end{aligned}$$

which is just the volume of the unit sphere, which is $\frac{4\pi}{3}$.

Without the divergence theorem, the original integral is still doable with spherical coordinates, but is a lot more work to compute.

44.5.5 Line Integrals

Given a function, $\mathbf{F}:\mathbb{R}^n\to\mathbb{R}^n$, and a curve, $C=\mathbf{r}(t)$,$t\in[a$,$b]$, the integral of $\mathbf{F}\cdot d\mathbf{r}$ over C is called a *line integral.* In practice, we can calculate it as follows:

$$\oint\limits_C \mathbf{F}\cdot d\mathbf{r} = \int_a^b \mathbf{F}\cdot\frac{d\mathbf{r}}{dt}\,dt$$

The line integral of a curve across a force vector field represents the work done when moving an object along the curve C from where $t=a$ to $t=b$.

Example. Calculate the line integral of $\mathbf{F}(x$,$y)=(x^2,-xy)$ along the unit circle from (1,0) to (0,1) in the anticlockwise direction. △

First, parametrise the path being integrated along: $\mathbf{r}(t)=(\cos t,\sin t)$, $t\in[0,\frac{\pi}{2}]$. $\mathbf{r}'(t)=(-\sin t,\cos t)$. Now, write out the line integral, and replace the x and y with $\cos(t)$ and $\sin(t)$ in $\mathbf{F}$.

$$\begin{aligned}
\oint\limits_C \mathbf{F}\cdot d\mathbf{r} &= \int_0^{\frac{\pi}{2}} \mathbf{F}\cdot\frac{d\mathbf{r}}{dt}\,dt \\
&= \int_0^{\frac{\pi}{2}} \begin{bmatrix}\cos^2 t\\ -\cos t\sin t\end{bmatrix}\cdot\begin{bmatrix}-\sin t\\ \cos t\end{bmatrix} dt \\
&= \int_0^{\frac{\pi}{2}} -\cos^2 t\sin t-\cos^2 t\sin t\,dt \\
&= -2\int_0^{\frac{\pi}{2}} \cos^2 t\sin t\,dt \\
&= -\frac{2}{3}
\end{aligned}$$

The line integral between two given points of a conservative vector field is independent of the path taken.

For example, gravity is a conservative field; it doesn't matter how you climb up a mountain, you gain the same amount of gravitational potential energy regardless of choice of path. This also means that the line integral of a closed curve is zero. Going back to the gravitational example, if you end up back where you started, you'll have a net gain of 0 gravitational potential energy.

44.5.6 Circulation

We haven't talked much about curl yet, but its name gives a hint as to what physical characteristic of a field it may represent. Thinking of $\mathbf{F}(x,y,z)$ as the function that returns the velocity of a fluid, we can quantify the pointwise rotation of the fluid at any given point with the curl of the vector field evaluated at that point. The length of the curl is proportional to the speed of rotation, and its direction is normal to the plane of rotation.

If the point at which curl is being evaluated lies on a surface, S, with unit normal $\mathbf{n}$, we can define a quantity called pointwise circulation as, $\nabla \times \mathbf{F} \cdot \mathbf{n}$. Unlike curl, this is a scalar, whose sign indicates the direction of rotation around the point, relative to $\mathbf{n}$.

Similar to the corkscrew rule for the cross product, if you align your right hand thumb in the direction of $\mathbf{n}$, the direction of circulation flows along with your fingers if pointwise circulation is positive, and flows against your fingers if pointwise circulation is negative.

By integrating the circulation over the entire surface, we can find the net circulation over the surface.

44.5.6.1 Stokes' Theorem

Let $\mathbf{F}(x,y,z)$ be a vector field, S be a surface with unit normal $\mathbf{n}$ and boundary curve C, oriented according to the right-hand rule. Then, the net circulation of the surface S over the field $\mathbf{F}$ is equal to the line integral of $\mathbf{F}$ evaluated over C.

$$\iint\limits_S \nabla \times \mathbf{F} \cdot \mathbf{n}\, dS = \oint\limits_C \mathbf{F} \cdot d\mathbf{r}$$

Example. Calculate the integral of $\nabla \times \mathbf{F} \cdot dS$ over the surface $S : z = 4 - x^2 - y^2, z \in [0,4]$. △

The rim of the surface is a circle of radius 2 centred on the origin, which can be parametrised as, $\mathbf{r}(t) = (2\cos t, 2\sin t, 0), t \in [0,2\pi]$. $\mathbf{r}'(t) = (-2\sin t, 2\cos t)$. We then rewrite the vector field as $(2\sin t, 0, 4\cos^2 t)$.

$$\begin{aligned}\oint\limits_C \mathbf{F} \cdot \frac{d\mathbf{r}}{dt}\, dt &= \int\limits_C \begin{bmatrix} 2\sin t \\ 0 \\ 4\cos^2 t \end{bmatrix} \cdot \begin{bmatrix} -2\sin t \\ 2\cos t \\ 0 \end{bmatrix} dt \\ &= -4 \int_0^{2\pi} \sin^2 t\, dt \\ &= -4\pi\end{aligned}$$

Chapter 45

Multivariable Analysis

"Obvious is the most dangerous word in mathematics."

— Eric Temple Bell, *The Queen of the Sciences*

We begin by extending our work from analysis into higher dimensions, starting with exploring the notions of convergence and continuity for vector and matrix-valued functions before studying the Fréchet derivative. We then cover vector fields, line and surface integrals, and some fundamental integral theorems.

Some of the material here is repeated from previous chapters, but is reformulated more rigourously.

45.1 Notation

There is a wide variety of notation in this area, and many symbols are overloaded with several distinct meanings, so a short list has been included to disambiguate some of these.

$\|\mathbf{v}\|$	Euclidean norm of the vector $\mathbf{v}$. Also written as $\|\mathbf{v}\|_2$ when discussing other ℓ^p norms. Written as $|\mathbf{v}|$ instead, when matrix-norms are also in use.
$\|\mathbf{v}\|_\infty$	The infinity norm of the vector $\mathbf{v}$.
$\|\mathbf{v}\|_1$	The taxicab or Manhattan norm of the vector $\mathbf{v}$.
$\|\mathbf{A}\|_F$	Frobenius norm of the matrix $\mathbf{A}$. Treats the matrix like a vector, then computes the ordinary Euclidean norm.
$\|T\|$	Operator norm of the linear map $\mathbf{v}$. We will sometimes put a matrix into this norm, as they are isomorphic to linear transformations. Also just written as $\|\cdot\|_{\text{op}}$.
$C(U,\mathbb{R}^k)$	Space of continuous functions $f : U \to \mathbb{R}^k$. Also written as $C^0(U,\mathbb{R}^k)$ (see next entry), or as $C(U)$ when $k = 1$.

$C^n(U,\mathbb{R}^k)$ Space of functions $f : U \to \mathbb{R}^k$ continuously differentiable n times.

$\mathbb{B}_r(\mathbf{a})$ Open ball of radius r centred at a point $\mathbf{a}$. That is, the set $\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}-\mathbf{a}\| < r\}$. Also written as $\mathbb{B}(\mathbf{a},r)$ or $B(\mathbf{a},r)$.

$\mathbb{B}_r$ Open ball of radius r centred on the origin; $\mathbb{B}_r(\mathbf{0})$.

$\mathbb{B}$ Unit open ball centred on the origin; $\mathbb{B}_1(\mathbf{0})$

$\overline{\mathbb{B}}_r(\mathbf{a})$ Closed ball of radius r centred at a point $\mathbf{a}$. That is, the set $\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}-\mathbf{a}\| \leq r\}$. Also written as $\overline{\mathbb{B}}(\mathbf{a},r)$ or $\overline{B}(\mathbf{a},r)$.

$\overline{\mathbb{B}}_r$ Closed ball of radius r centred on the origin; $\overline{\mathbb{B}}_r(\mathbf{0})$

$\overline{\mathbb{B}}$ Unit closed ball centred on the origin; $\overline{\mathbb{B}}_1(\mathbf{0})$.

$S^n(r)$ The n-sphere of radius r; the boundary of $\mathbb{B}_r(\mathbf{0})$; the set $\{\mathbf{x} \in \mathbb{R}^{n+1} : |\mathbf{x}| = r\}$.

S^n The unit n-sphere; $S^n(1)$.

$L(\mathbb{R}^n,\mathbb{R}^k)$ The space of linear maps $T : \mathbb{R}^n \to \mathbb{R}^k$.

$L(\mathbb{R}^n)$ The space of linear maps $T : \mathbb{R}^n \to \mathbb{R}^n$; isomorphic to and hence interchangable with the space of $k \times n$ matrices with real entries.

$M(k \times n,\mathbb{R})$ The space of $k \times n$ matrices with real entries. Also abbreviated as $\mathbb{R}^{k \times n}$.

$M(n,\mathbb{R})$ The space of $n \times n$ matrices with real entries; $M(n \times n,\mathbb{R})$

$GL(n,\mathbb{R})$ The group of invertible linear maps $T : \mathbb{R}^n \to \mathbb{R}^k$; the group of nonsingular $n \times n$ matrices with real entries.

$\Delta(\mathbf{A})$ The multilinear function that sends a matrix $\mathbf{A}$ to its determinant. Not to be confused with the Laplacian.

M^* The Lipschitz constant of a function; the upper bound on how quickly a function can vary.

$\dfrac{\partial f}{\partial x_i}$ The partial derivative of a function f with respect to the variable x_i. Also written as $\partial_{x_i} f(x)$; just as $\partial_i f(x)$; or if f has few variables, as f_x, f_y, etc.

$\partial_{\mathbf{v}} f(\mathbf{x})$ The directional derivative of f at $\mathbf{x}$ in the direction of $\mathbf{v}$. Also written as $D_{\mathbf{v}} f(\mathbf{x})$. If $\mathbf{v}$ is one of the basis vectors, then this is the partial derivative.

∇ The del or nabla operator; can be thought of as a vector full of partial differential operators $\left[\frac{\partial}{\partial x_1},\frac{\partial}{\partial x_2},\ldots,\frac{\partial}{\partial x_n}\right]^\top$.

∇f The gradient of f; the vector $\left[\frac{\partial f}{\partial x_1},\frac{\partial f}{\partial x_2},\ldots,\frac{\partial f}{\partial x_n}\right]^\top$; also written as $\mathrm{grad}(f)$.

$\nabla \cdot \underline{v}$ The divergence of $\underline{v}$; calculated by "dotting" the del operator with the vector field $\underline{v}$; also written as $\mathrm{div}(f)$.

$\nabla \times \underline{v}$ The curl of $\underline{v}$; calculated by "crossing" the del operator with the vector field $\underline{v}$; also written as $\mathrm{curl}(f)$.

$Df(x)$ The Fréchet derivative of f at x.

∂f The Jacobian matrix of f. Also written as Df, because it's the same thing as the Fréchet derivative in finite dimensions.

$\mathcal{G}_f$ The graph of a function f; if f takes two variables, then $\mathcal{G}_f$ is the surface parametrised by $\mathbf{r}(x,y) = (x,y,f(x,y))$.

$\underline{v}$ A vector field; a function $\underline{v} : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^n$.

$\mathbf{v}^\perp$ The rotation of the vector $\mathbf{v} \in \mathbb{R}^2$ 90° clockwise; if $\mathbf{v} = (x,y)$, then $\mathbf{v}^\perp = (y, -x)$.

$\mathbf{r}(t)$ The parametrisation $r : [a,b] \to \mathbb{R}^2$ of a curve $C \subset \mathbb{R}^2$.

$\rho(s)$ The arclength or unit speed parametrisation $\rho : [0,L] \to \mathbb{R}^2$ of a curve C.

$\dot{\mathbf{r}}(t)$ The tangent to $\mathbf{r}(t)$; given by differentiating $\mathbf{r}$ componentwise.

$\dot{\rho}(s)$ The unit tangent to $\rho(s)$; given by differentiating ρ componentwise.

$\mathbf{N}(t)$ The normal to $\mathbf{r}(t)$; given by $\dot{\mathbf{r}}(t)^\perp$.

$\mathbf{n}(t)$ The unit normal to $\rho(s)$; given by $\dot{\rho}(s)^\perp$.

$\displaystyle\int_0^L \underline{v}\big(\rho(s)\big) \cdot \dot{\rho}(s) ds$ The tangential line integral of $\underline{v}$ along C. Calculated using $\int_a^b \underline{v}\big(\mathbf{r}(t)\big) \cdot \frac{d\mathbf{r}}{dt}\, dt$, where $\mathbf{r}$ is a parametrisation of C.

$\displaystyle\int_C \underline{v} \cdot d\mathbf{r}$ Alternative notation for the tangential line integral of $\underline{v}$ along C.

$\displaystyle\int_0^L \underline{v}\big(\rho(s)\big) \cdot \mathbf{n}(s)ds$ The flux integral of $\underline{v}$ along C; the normal line integral of $\underline{v}$ along C (compare with the tangential line integral above);. Calculated using $\int_a^b \underline{v}\big(\mathbf{r}(t)\big) \cdot \mathbf{N}(t)\,dt$.

$\mathbf{r}(u,v)$ The parametrisation of a surface $S \subset \mathbb{R}^3$.

$\displaystyle\iint_S \underline{v} \cdot \mathbf{n}\,dA$ The flux integral of $\underline{n}$ across a surface S. Calculated using $\iint_U \underline{v}\big(\mathbf{r}(u,v)\big) \cdot \big(\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v}\big)\,du\,dv$.

$\displaystyle\iint_S \underline{v} \cdot\, d\mathbf{S}$ Alternative notation for the flux integral of $\underline{v}$ across S. Also written as $\iint_S \underline{v} \cdot \mathbf{n}\,dS$, or $\iint_S \underline{v} \cdot\, d\mathbf{A}$.

Δf The Laplacian of f; calculated as $\nabla \cdot (\nabla f)$, or, the sum of the second derivatives of f; also written as $\nabla \cdot \nabla$ or ∇^2.

$D^2 f(x)$ The Hessian (transformation) of f.

$\partial^2 f(x)$ The Hessian matrix of f; also written as $\operatorname{Hess} f(x)$.

$\mathcal{N}_p$ An (open) neighbourhood of a point p.

Γ_c The level set of a function set equal to c; the set of inputs to a function such that the output is c.

45.2 Convergence and Continuity

45.2.1 Convergence in $\mathbb{R}^n$

A sequence $(\mathbf{x}_i)_{i=1}^{\infty}$ of vectors in $\mathbb{R}^n$ converges to $\mathbf{x} \in \mathbb{R}^n$ if,

$$\forall \varepsilon > 0, \exists N \in \mathbb{N} : i > N \to \|\mathbf{x}_i - \mathbf{x}\| < \varepsilon$$

where $\| \cdot \|$ is the euclidean norm.

Theorem 45.2.1 (Uniqueness of Limits)**.** *If $(\mathbf{x}_i)_{i=1}^{n}$ converges to both $\mathbf{x}$ and $\mathbf{y}$, then $\mathbf{x} = \mathbf{y}$.*

Theorem 45.2.2 (Componentwise Convergence)**.** *A sequence $(\mathbf{x}_i)_{i=1}^{n} \subseteq \mathbb{R}^n$ converges to $\mathbf{y}$ if and only if for each $i \in [1,n]$,*

$$\lim_{j\to\infty} \mathbf{x}_{i,j} = \mathbf{y}_i$$

That is, the real number sequences of components of $(\mathbf{x}_i)$ all individually converge to their corresponding component of $\mathbf{y}$.

The *uniform*, *max* or *infinity norm*, denoted by, $\| \cdot \|_\infty$ is defined by,

$$\|\mathbf{x}\|_\infty := \max\bigl(|x_1|, \ldots ,|x_n|\bigr), \qquad \mathbf{x} = (x_1, \ldots ,x_n)$$

The *taxicab* or *Manhattan norm*, denoted by $\| \cdot \|_1$ is defined by,

$$\|\mathbf{x}\|_1 := |x_1| + \cdots + |x_n|, \qquad \mathbf{x} = (x_1, \ldots ,x_n)$$

Theorem 45.2.3. *For all $\mathbf{x} \in \mathbb{R}^n$,*

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\| \leq \sqrt{n}\|x\|_\infty$$

and

$$\|\mathbf{x}\| \leq \|\mathbf{x}\|_1 \leq \sqrt{n}\|x\|$$

Theorem 45.2.4 (Algebra of Limits)**.** *If $(\mathbf{x}_i) \to \mathbf{x}$ and $(\mathbf{y}_i) \to \mathbf{y}$, then,*

$$\lim_{i\to\infty} (\alpha\mathbf{x}_i + \beta\mathbf{y}_i) = \alpha\mathbf{x} + \beta\mathbf{y}$$

for all $\alpha,\beta \in \mathbb{R}$;

$$\lim_{i\to\infty} \langle \mathbf{x}_i,\mathbf{y}_i \rangle = \langle \mathbf{x},\mathbf{y} \rangle$$

where $\langle -,- \rangle$ is any inner product, such as the scalar product;

$$\lim_{i\to\infty} \|\mathbf{x}_i\| = \|\mathbf{x}\|$$

where $\| \cdot \|$ is any norm.

A sequence $(\mathbf{x}_i)$ is *bounded* if there exists $M > 0$ such that $\|\mathbf{x}_i\| < M$ for all $i \in \mathbb{N}$.

Theorem 45.2.5 (Boundedness of Convergent Sequences)**.** *If $(\mathbf{x}_i)$ converges to some $\mathbf{x}$, then $(\mathbf{x}_i)$ is bounded.*

Theorem 45.2.6 (Bolzano-Weierstrass for Vectors)**.** *Any bounded sequence $(\mathbf{x}_i)_{i=1}^{\infty} \subseteq \mathbb{R}^n$ has a convergent subsequence $(\mathbf{x}_{i_j})$.*

45.2.2 Continuity

A function $\mathbf{f} : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^k$ is *continuous at a point* $\mathbf{p} \in U$ if,

$$\forall \varepsilon > 0, \exists \delta > 0 : \|\mathbf{x} - \mathbf{p}\| < \delta \to \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{p})\| < \varepsilon \qquad (\varepsilon - \delta \text{ Continuity})$$

or if, for all sequences $(\mathbf{x}_i) \to \mathbf{p}$,

$$\big(\mathbf{f}(\mathbf{x}_i)\big) \to \mathbf{p} \qquad \text{(Sequential Continuity)}$$

$\mathbf{f}$ is then said to be *continuous at* U if $\mathbf{f}$ is continuous at all points $\mathbf{p} \in U$.

We write $C(U,\mathbb{R}^k)$ to denote the space of continuous functions $f : U \to \mathbb{R}^k$.

A function $f : U \to \mathbb{R}^k$ has a *(continuous) limit* at $\mathbf{p} \in U$ if there exists a vector $\mathbf{q} \in \mathbb{R}^k$ such that,

$$\forall \varepsilon > 0, \exists \delta > 0 : (x \in \mathbf{U}) \wedge (0 < \|\mathbf{x} - \mathbf{p}\| < \delta) \to \|\mathbf{f}(\mathbf{x} - \mathbf{q})\| < \varepsilon$$

and we write $\lim_{\mathbf{x}\to\mathbf{p}} \mathbf{f}(\mathbf{x}) = \mathbf{q}$.

Just as for limits of sequences, continuous limits are unique. We also have that $\mathbf{f}$ is continuous at $\mathbf{p}$ if and only if $\lim_{\mathbf{x}\to\mathbf{p}} \mathbf{f}(\mathbf{x}) = \mathbf{p}$.

Given a real-valued function $f : \mathbb{R}^2 \to \mathbb{R}$, we define two families of functions g^y and h^x by

$$g^y(x) = f(x,y) = h^x(y)$$

In computer science terminology, g and h are the partial applications of f in the first and second arguments, respectively.

A function $f : \mathbb{R}^2 \to \mathbb{R}$ is *separately continuous* at (a,b) if g^b is continuous at a and h^a is continuous at b.

Continuity implies separate continuity, but not the converse.

Example. Define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x,y) = \begin{cases} 1 & xy \neq 0 \\ 0 & xy = 0 \end{cases}$$

Then, $g^0(x) = 0$ for all x and $h^0(y) = 0$ for all y, so f is separately continuous at $(0,0)$. But,

$$\lim_{(x,y)\to(0,0)} f(x,y) = \lim_{n\to\infty} f\left(0,\frac{1}{n}\right) = 0 \neq 1 = \lim_{n\to\infty} f\left(\frac{1}{n},\frac{1}{n}\right) = \lim_{(x,y)\to(0,0)} f(x,y)$$

so

$$\lim_{(x,y)\to(0,0)} f(x,y)$$

does not have a unique value and hence does not exist. △

Theorem 45.2.7 (Continuity of Sums). *If* $\mathbf{f},\mathbf{g} : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^k$ *are both continuous at* $\mathbf{p} \in U$*, then* $\alpha\mathbf{f} + \beta\mathbf{g}$ *is continuous at* $\mathbf{p}$ *for all* $\alpha,\beta \in \mathbb{R}$.

Theorem 45.2.8 (Continuity of Real-Valued Products). *If* $f,g : (U \subseteq \mathbb{R}^n) \to \mathbb{R}$ *are both continuous at* $\mathbf{p} \in U$*, then* $(fg)(x) \coloneqq f(x)g(x)$ *is continuous at* $\mathbf{p}$.

Theorem 45.2.9 (Continuity of Quotients). *If* $f,g : (U \subseteq \mathbb{R}^n) \to \mathbb{R}$ *are both continuous at* $\mathbf{p} \in U$*, and* $g(\mathbf{x}) \neq 0$ *for all* $\mathbf{x} \in \mathbb{U}$*, then* $(f/g)(\mathbf{x}) \coloneqq f(\mathbf{x})/g(\mathbf{x})$ *is continuous at* $\mathbf{p}$.

Theorem 45.2.10 (Continuity of Composition)**.** *If* $\mathbf{f} : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^k$ *is continuous at* $\mathbf{p} \in U$ *and* $\mathbf{g} : (V \subseteq \mathbb{R}^k) \to \mathbb{R}^m$ *is continuous at* $\mathbf{f}(\mathbf{p}) \in V$*, and* $f(U) \subseteq V$*, then* $g \circ f : \mathbf{U} \to \mathbb{R}^m$ *is continuous at* $\mathbf{p}$*.*

Theorem 45.2.11 (Componentwise Continuity)**.** *A function* $\mathbf{f} : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^k$ *defined by*

$$(x_1, \ldots ,x_n) \mapsto \big(f_1(x_1, \ldots ,x_n),f_2(x_1, \ldots ,x_n), \ldots ,f_k(x_1, \ldots ,x_n)\big)$$

where $(f_j)_{j=1}^k$ *are real-valued functions, is continuous at* $\mathbf{p} \in U$ *if and only if every* f_j *is continuous at* $\mathbf{p}$*.*

That is, $\mathbf{f}$ is continuous if and only if every component f_j is individually continuous.

Theorem 45.2.12. *If* $f : \mathbb{R} \to \mathbb{R}$ *is continuous at* $p \in \mathbb{R}$*, then any function* $\mathbb{R}^n \to \mathbb{R}$ *defined by*

$$(x_i)_{i=1}^n \mapsto f(x_j)$$

is continuous on $\{(x_i)_{i=1}^n : x_j = p\}$

That is, any function that is continuous as a function $\mathbb{R} \to \mathbb{R}$ is also continuous as a function $\mathbb{R}^n \to \mathbb{R}$.

A function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^k$ is *continuous along lines* or *linearly continuous* at a point $\mathbf{p} \in \mathbb{R}^k$ if the restriction $\mathbf{f}^L$ of $\mathbf{f}$ to the line L passing through $\mathbf{p}$ is continuous for every such line L.

Continuity implies linear continuity, but not the converse.

Example. Define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x,y) = \begin{cases} 1 & 0 < y < x^2 \\ 0 & \text{otherwise} \end{cases}$$

$f = 0$ over any sufficiently short line segment that passes through the point (0,0), so $\lim_{x\to 0} f(x,ax) = 0$ along any straight line path and f is linearly continuous at (0,0). But,

$$\lim_{n\to 0} f\left(n,\frac{1}{2}n^2\right) = 1 \neq 0 = f(0,0)$$

so f is discontinuous at (0,0). △

Linear continuity implies separate continuity, but not the converse.

Example. Define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x,y) = \begin{cases} 1 & xy = 0 \\ 0 & xy \neq 0 \end{cases}$$

$g^0(x) = 1$ for all x, and $h^0(y) = 1$ for all y, so f is separately continuous at (0,0). But,

$$\lim_{n\to 0} f(n,n) = 0 \neq 1 = f(0,0)$$

so f is not linearly continuous at (0,0). △

Continuity → Linear Continuity → Separate Continuity

45.3 Topology on $\mathbb{R}^n$

We define the *open ball* of radius $r > 0$ centred at a point $\mathbf{a} \in \mathbb{R}^n$, denoted by $\mathbb{B}_r(\mathbf{a})$ or $\mathbb{B}(\mathbf{a},r)$, to be the set $\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{a}\| < r\}$. We abbreviate $\mathbb{B}_r(\mathbf{0})$ to $\mathbb{B}(r)$, and $\mathbb{B}_1(\mathbf{0})$ (the *unit open ball*) to $\mathbb{B}$.

Similarly, the *closed ball* of radius $r > 0$ centred at a point $\mathbf{a} \in \mathbb{R}^n$, denoted by $\overline{\mathbb{B}}_r(\mathbf{a})$ or $\overline{\mathbb{B}}(\mathbf{a},r)$, to be the set $\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{a}\| \leq r\}$. We abbreviate $\overline{\mathbb{B}}_r(\mathbf{0})$ to $\overline{\mathbb{B}}(r)$, and $\overline{\mathbb{B}}_1(\mathbf{0})$ (the *unit closed ball*) to $\overline{\mathbb{B}}$.

A set $X \subseteq \mathbb{R}^n$ is *closed*, if for every sequence $(\mathbf{x}_i)_{i=1}^{\infty} \subseteq X$ of points in X that converges to a limit point $\mathbf{x} \in \mathbb{R}^n$, we also have $\mathbf{x} \in X$. That is, X is closed (in the algebraic sense) under sequential limits.

A set $U \subseteq \mathbb{R}^n$ is *open* if for all $\mathbf{x} \in U$, there exists $\varepsilon > 0$ such that $\mathbb{B}_\varepsilon(\mathbf{x}) \subset U$.

The empty set and $\mathbb{R}^n$ are both open and closed, or *clopen*.

Theorem 45.3.1. *A set is open if and only if its complement is closed.*

Theorem 45.3.2. *Open balls are open sets.*

Theorem 45.3.3. *Closed balls are closed sets.*

Theorem 45.3.4 (Arbitrary Union of Open Sets)**.** *If $(U_i)_{i \in I}$ is a (possibly uncountable) collection of open sets, then,*

$$\bigcup_{i \in I} U_i$$

is open.

Theorem 45.3.5 (Finite Intersection of Open Sets)**.** *If $(U_i)_{i=1}^{n}$ is a finite collection of open sets, then,*

$$\bigcap_{i=1}^{n} U_i$$

is open.

Corollary 45.3.5.1. *An arbitrary intersection or finite union of closed sets is closed.*

Let $E \subseteq \mathbb{R}^n$. Given $\varepsilon > 0$, the *ε-neighbourhood* $\mathcal{N}(E,\varepsilon)$ of E is defined by,

$$\mathcal{N}(E,\varepsilon) := \bigcup_{\mathbf{x} \in E} \mathbb{B}(\mathbf{x},\varepsilon)$$

The ε-neighbourhood of a set is always open.

45.3.1 Continuity and Topology

Rewriting the $\varepsilon - \delta$ definition of continuity in terms of open sets, a function $\mathbf{f} : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^k$ is continuous at a point $\mathbf{p} \in U$ if,

$$\forall \varepsilon > 0, \exists \delta > 0 : \mathbf{f}\big((\mathbb{B}(\mathbf{p},\delta) \cap U)\big) \subset \mathbb{B}(\mathbf{f}(\mathbf{p}),\varepsilon)$$

Applying the inverse to both sides of the inclusion, we have,

$$\forall \varepsilon > 0, \exists \delta > 0 : \mathbb{B}(\mathbf{p},\delta) \cap U \subset \mathbf{f}^{-1}\big(\mathbb{B}(\mathbf{f}(\mathbf{p}),\varepsilon)\big)$$

This suggests the following alternative characterisations of continuity:

Theorem 45.3.6. *For any function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^k$, the following statements are equivalent:*

- $\mathbf{f}$ *is continuous at all points of* $\mathbb{R}^n$.

- $\mathbf{f}^{-1}(V)$ *is open whenever* $V \subseteq \mathbb{R}^n$ *is open.*
- $\mathbf{f}^{-1}(\mathcal{F})$ *is closed whenever* $\mathcal{F} \subseteq \mathbb{R}^n$ *is closed.*

Note that this does not imply that the image of an open (closed) set under a continuous function is open (resp. closed): only inverse images preserve the topology of a set.

45.3.2 Compactness

A set $K \subseteq \mathbb{R}^n$ is *sequentially compact* if every sequence $(\mathbf{x}_i)_{i=1}^{\infty} \subset K$ has a convergent subsequence $(\mathbf{x}_{i_j})$ whose limit is in K.

A set $K \subseteq \mathbb{R}^n$ is *bounded* if there exists some $M > 0$ such that $\|\mathbf{x}\| \leq M$ for all $\mathbf{x} \in K$.

Theorem 45.3.7. *A set $K \subseteq \mathbb{R}^n$ is sequentially compact if and only if it is closed and bounded.*

Theorem 45.3.8 (Continuity Preserves Sequential Compactness)**.** *If $\mathbf{f} : K \to \mathbb{R}^k$ is continuous and K is sequentially compact, then $\mathbf{f}(K)$ is also sequentially compact.*

Theorem 45.3.9 (Extreme Value Theorem)**.** *Let $K \subset \mathbb{R}^n$ be sequentially compact, and let $f : K \to \mathbb{R}$ be continuous. Then, there exist $\mathbf{x}_*,\mathbf{x}^* \in K$ such that*

$$f(\mathbf{x}_*) \leq f(\mathbf{x}) \leq f(\mathbf{x}^*)$$

for all $x \in K$.

That is, a continuous real-valued function defined over a sequentially compact space attains its extreme values within that space.

Proof. Because f is continuous and K is sequentially compact, $f(K)$ is also sequentially compact, and is hence closed and bounded. Then, the values

$$U := \sup f(K), \quad L := \inf f(K)$$

are both finite and there exists sequences (a_i),$(b_i) \subset K$ such that $(a_i) \to L$ and $(b_i) \to U$. As $f(K)$ is closed, we have L,$U \in f(K)$, so $\mathbf{x}_* := f^{-1}(L)$ and $\mathbf{x}^* := f^{-1}(U)$ exist, and satisfy,

$$f(\mathbf{x}_*) = L \leq f(\mathbf{x}) \leq U = f(\mathbf{x}^*)$$

for all x in K, as required. ■

Corollary 45.3.9.1. *Let $K \subset \mathbb{R}^n$ be sequentially compact and let $\mathbf{f} : K \to \mathbb{R}^k$ be continuous. Then, there exists $\mathbf{x}_*,\mathbf{x}^* \in K$ in K such that*

$$\|f(\mathbf{x}_*)\| \leq \|f(\mathbf{x})\| \leq \|f(\mathbf{x}^*)\|$$

for all x in K.

Proof. The map $\| \cdot \| : \mathbb{R}^k \to \mathbb{R}$ is continuous, so $\mathbf{x} \mapsto \|f(\mathbf{x})\|$ is a continuous map $K \to \mathbb{R}$. ■

45.4 The Space $L(\mathbb{R}^n,\mathbb{R}^k)$ of Linear Maps

We denote the space of linear maps $T : \mathbb{R}^n \to \mathbb{R}^k$ by $L(\mathbb{R}^n,\mathbb{R}^k)$. If $n = k$, this is abbreviated to $L(\mathbb{R}^n)$. We denote the space of $k \times n$ matrices with real entries by $M(k \times n,\mathbb{R})$, also abbreviated to $\mathbb{R}^{k\times n}$.

We associate every matrix $\mathbf{A} \in \mathbb{R}^{k\times n}$ with a linear map $T \in L(\mathbb{R}^n,\mathbb{R}^k)$ defined by

$$T(\mathbf{x}) = \mathbf{A}\mathbf{x}$$

and we can write this association as a map $\mu : \mathbb{R}^{k\times n} \to L(\mathbb{R}^n,\mathbb{R}^k)$ that sends a matrix to the linear map it represents under the standard bases of $\mathbb{R}^n$ and $\mathbb{R}^k$. Moreover, μ is a linear isomorphism.

We also have that

$$\begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & & \vdots \\ a_{k,1} & \cdots & a_{k,n} \end{bmatrix} \cong [a_{1,1},\dots,a_{1,n},a_{2,1},\dots,a_{2,n},a_{3,1},\dots,a_{k,1},\dots,a_{k,n}]^\top$$

is a linear isomorphism, so,

$$\dim(L(\mathbb{R}^n,\mathbb{R}^k)) = \dim(\mathbb{R}^{k\times n}) = \dim(\mathbb{R}^{nk}) = nk$$

45.4.1 Matrix Norms

To discuss continuity of functions with matrix-valued inputs or outputs, we need to define a norm on $L(\mathbb{R}^n,\mathbb{R}^k)$, or equivalently, on $\mathbb{R}^{k\times n}$. In this section, we will write vector norms as $|\cdot|$, while matrix/linear map norms will be written as $\|\cdot\|$ for contrast.

The first such norm we might think of is to use the matrix-vector isomorphism above, and define the *Frobenius norm* $\|\cdot\|_F : \mathbb{R}^{k\times n} \to \mathbb{R}$ by,

$$\|(a_{i,j})\|_F := \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{n} a_{i,j}^2}$$

That is, treat the matrix as a vector, then calculate the ordinary Euclidean norm.

The *operator norm* $\|\cdot\|_{\mathrm{op}} : L(\mathbb{R}^n,\mathbb{R}^k) \to \mathbb{R}$, also denoted by just $\|\cdot\|$, is defined by,

$$\|T\| := \sup_{\mathbf{x}\in\mathbb{R}^n\setminus\{0\}} \frac{|T(\mathbf{x})|}{|\mathbf{x}|}$$

Informally, the operator norm is the maximum factor by which the transformation lengthens vectors. That is, the operator norm satisfies,

$$|T(\mathbf{x})| \le \|T\||\mathbf{x}|$$

for all $\mathbf{x} \in \mathbb{R}^n$.

Writing T as a matrix multiplication, by linearity, we have,

$$\begin{aligned} \frac{|\mathbf{Ax}|}{|\mathbf{x}|} &= \frac{1}{|\mathbf{x}|}\mathbf{Ax} \\ &= \left|\frac{1}{|\mathbf{x}|}\mathbf{Ax}\right| \\ &= \left|\mathbf{A}\left(\frac{\mathbf{x}}{|\mathbf{x}|}\right)\right| \end{aligned}$$

Because $\left|\frac{\mathbf{x}}{|\mathbf{x}|}\right| = 1$, this gives,

$$\|\mathbf{A}\| = \sup_{|\mathbf{x}|=1} |\mathbf{Ax}|$$

There are some more alternative characterisations of the operator norm for more general normed spaces. For instance, note that the above definitions are not well-defined if the codomain of the linear operator is the trivial space. Let $T : V \to W$ be a linear transformation with matrix $\mathbf{A}$. Then,

$$\|T\| = \inf\{M \ge 0 : |\mathbf{Av}| \le M|\mathbf{v}|,\mathbf{v} \in V\}$$

$$\begin{aligned}&= \sup\{|\mathbf{A}\mathbf{v}| : |\mathbf{v}| \leq 1, \mathbf{v} \in V\}\\ &= \sup\{|\mathbf{A}\mathbf{v}| : |\mathbf{v}| < 1, \mathbf{v} \in V\}\\ &= \sup\{|\mathbf{A}\mathbf{v}| : |\mathbf{v}| \in \{0,1\}, \mathbf{v} \in V\}\end{aligned}$$

Theorem 45.4.1. *The operator norm is a norm. That is, it satisfies,*

- $\|T\| = 0 \leftrightarrow T = 0$ *(Point separating)*
- $\|\alpha T\| = |\alpha|\|T\|$ *(Absolute homogeneity)*
- $\|T + U\| \leq \|T\| + \|U\|$ *(Triangle inequality)*

Theorem 45.4.2. *For any linear transformation T with matrix $\mathbf{A}$, we have,*

$$\frac{1}{\sqrt{n}}\|\mathbf{A}\|_F \leq \|T\| \leq \|\mathbf{A}\|_F$$

Theorem 45.4.3 (Composition Bound)**.** *For any $A \in L(\mathbb{R}^n,\mathbb{R}^k)$ and $B \in L(\mathbb{R}^k,\mathbb{R}^m)$, the map $B \circ A \in L(\mathbb{R}^n,\mathbb{R}^m)$ satisfies,*

$$\|B \circ A\| \leq \|A\|\|B\|$$

Proof.

$$\begin{aligned}|(B \circ A)(\mathbf{x})| &= |B(A(\mathbf{x}))|\\ &\leq \|B\||A(\mathbf{x})|\\ &\leq \|B\|\|A\||\mathbf{x}|\end{aligned}$$

for all $\mathbf{x} \in \mathbb{R}^n$, so,

$$\begin{aligned}\frac{|(B \circ A)(\mathbf{x})|}{|\mathbf{x}|} &\leq \sup \frac{|(B \circ A)(\mathbf{x})|}{|\mathbf{x}|}\\ &\leq \|B\|\|A\||\mathbf{x}|\end{aligned}$$

■

45.4.2 Convergence and Continuity in $L(\mathbb{R}^n,\mathbb{R}^k)$

These are defined identically to sequences $(\mathbf{x}_i)_{i=1}^{\infty} \subset \mathbb{R}^n$ and functions $\mathbf{f} : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^k$.

That is, a sequence $(T_i)_{i=1}^{\infty} \subset L(\mathbb{R}^n,\mathbb{R}^k)$ of linear transformations converges to $T \in L(\mathbb{R}^n,\mathbb{R}^k)$ is,

$$\forall \varepsilon > 0, \exists N \in \mathbb{N} : i > N \to \|T_i - T\| < \varepsilon$$

We can also use the Frobenius norm in place of the operator norm here to similarly define convergence of sequences of matrices, and because $\mathbb{R}^{k \times n} \cong \mathbb{R}^{kn}$, this implies that both $\mathbb{R}^{k \times n}$ and $L(\mathbb{R}^n,\mathbb{R}^k)$ are both complete spaces, so every convergent sequence of linear transformations or matrices is also Cauchy.

45.4.3 Matrix-Valued Functions

A function $f : U \to \mathbb{R}^{k \times n}$ is continuous at $x \in U$ if,

$$\forall \varepsilon > 0 \exists \delta > 0 : |y - x| < \delta \to \|f(y) - f(x)\|_F < \varepsilon$$

Because the Frobenius norm on matrices in $\mathbb{R}^{k \times n}$ is equivalent to the Euclidean norm on vectors in $\mathbb{R}^{nk}$, we also have that a matrix-valued function is continuous if and only if it is componentwise continuous.

This also provides an easy way to check if a linear-transformation-valued function $f : U \to L(\mathbb{R}^n,\mathbb{R}^k)$ is continuous: check if every entry of the matrix representing the linear transformation output is continuous.

Theorem 45.4.4. *The map $\Delta : \mathbb{R}^{n\times n} \to \mathbb{R}$ that sends a matrix to its determinant is continuous with respect to the Frobenius norm on $\mathbb{R}^{n\times n}$.*

Proof. The determinant is a polynomial of degree n in its n^2 variables, and polynomials are continuous on $(\mathbb{R}^{n^2},|\cdot|) \cong (\mathbb{R}^{n\times n},\|\cdot\|_F)$. ■

45.4.4 The Space $GL(n,\mathbb{R}) \subset L(\mathbb{R}^n)$ of Invertible Linear Maps

It is clear that if a linear map $T : \mathbb{R}^n \to \mathbb{R}^k$ is a bijection, then $n = k$ and $\ker(T) = \{\mathbf{0}\}$. But by the rank-nullity theorem, the converse also holds. That is, a linear map $T : \mathbb{R}^n \to \mathbb{R}^k$ is a bijection if and only if $k = n$ and $\ker(T) = \{\mathbf{0}\}$.

The *general linear group* over the real numbers, denoted by $GL(n,\mathbb{R})$, is defined by,

$$GL(n,\mathbb{R}) := \{T \in L(\mathbb{R}^n) : T \text{ is invertible}\}$$

with the group operation given by composition. In terms of matrices, this is equivalent to,

$$GL(n,\mathbb{R}) := \{\mathbf{A} \in \mathbb{R}^{n\times n} : \det(\mathbf{A}) \neq 0\}$$

Note that $GL(1,\mathbb{R}) \cong (\mathbb{R}^*,\times)$.

Theorem 45.4.5. *The space $GL(n,\mathbb{R})$ is an open subset of $\mathbb{R}^{n\times n}$*

Proof. $GL(n,\mathbb{R}) = \Delta^{-1}(\mathbb{R} \setminus \{0\})$, and $\mathbb{R} \setminus \{0\}$ is open, so $GL(n,\mathbb{R})$ is open by the continuity of Δ. ■

$GL(n,\mathbb{R})$ being open means that invertibility of a linear map in $L(\mathbb{R}^n)$ is a stable property: a linear map can be perturbed somewhat, and remain invertible. The next theorem quantifies exactly how much a linear map A can be perturbed, in terms of $\|A^{-1}\|$.

Theorem 45.4.6. *Given $A \in GL(n,\mathbb{R})$, let $\alpha := \frac{1}{\|A^{-1}\|}$. If $B \in L(\mathbb{R}^n)$ and $\|B - A\| < \alpha$, then B is invertible. That is, $\mathbb{B}_\alpha(A) \subset GL(n,\mathbb{R})$. Furthermore,*

$$\|B - A\| < \alpha \to \|B^{-1}\| \leq \frac{1}{\alpha - \|B - A\|}$$

Theorem 45.4.7. *The map $(\,\cdot\,)^{-1} : GL(n,\mathbb{R}) \to GL(n,\mathbb{R})$ defined by $A \mapsto A^{-1}$ is continuous.*

45.4.5 Lipschitz Continuity

A map $f : U \to \mathbb{R}^k$ is *Lipschitz continuous* on U if there exists an $M > 0$ such that,

$$|f(x) - f(y)| \leq M|x - y|$$

for all $x,y \in U$.

The *Lipschitz constant* or *modulus of (uniform) continuity* M^* of f is then defined by,

$$M^* := \sup_{\substack{x\neq y\\ x,y\in U}} \frac{|f(x) - f(y)|}{|x - y|}$$

Intuitively, a Lipschitz continuous function is limited in how fast it can change: for any pair of distinct points, the absolute value of the gradient of the line connecting them is bounded by this Lipschitz constant.

Note that Lipschitz continuity of a function is a very strong form of continuity, and it implies uniform (and hence regular) continuity of the function:

$$\forall\varepsilon > 0 : |x - y| < \frac{\varepsilon}{M} \to |f(x) - f(y)| < \varepsilon$$

Theorem 45.4.8. *Every linear map T is continuous.*

Proof. By linearity, $T(x) - T(y) = T(x - y)$, and so,

$$|T(x) - T(y)| = |T(x - y)| \leq \|T\||x - y|$$

■

Theorem 45.4.9. *The map $|\cdot| : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is Lipschitz continuous with Lipschitz constant $M^* = 1$.*

Proof. By the reverse triangle inequality, we have,

$$\big||x| - |y|\big| \leq |x - y|$$

■

The same holds for any norm, so the operator norm and Frobenius norm are both Lipschitz continuous with Lipschitz constant $M^* = 1$.

45.5 The Derivative

In this section, $U \subseteq \mathbb{R}^n$ will be an open subset of $\mathbb{R}^n$. This means that if $\mathbf{p} \in U$, then in any limit $\lim_{\mathbf{x}\to\mathbf{p}}$, $\mathbf{x}$ may approach $\mathbf{p}$ from any direction.

45.5.1 Partial Derivatives

A *partial derivative* of a multivariate function is its derivative with respect to one of its variables, with the other variables held constant.

Let $\{\mathbf{e}_i\}_{i=1}^n$ be the standard basis of $\mathbb{R}^n$. For any function $f : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^k$ with U open, a partial derivative of f at the point $\mathbf{x} \in U$ with respect to the i-th variable x_i is defined as,

$$\begin{aligned}\frac{\partial f}{\partial x_i} &= \lim_{h\to 0} \frac{f(x_1,x_2,\ldots,x_{i-1},x_i + h,x_{i+1},\ldots,x_n) - f(x_1,x_2,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n)}{h} \\ &= \lim_{t\to 0} \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x})}{h}\end{aligned}$$

Other notations include $\partial_{x_i} f(x)$ or $\partial_i f(x)$. If f is a function of only a few variables, then it is more common to write, say $f(x,y,z)$, rather than $f(x_1,x_2,x_3)$, and we write f_x for the partial derivative of f with respect to x.

Theorem 45.5.1 (Algebra of Partial Derivatives)**.** *If $f,g : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^k$, then,*

- $\partial_i(f + g) = \partial_i f + \partial_i g$;
- $\partial_i(fg) = (\partial_i f)g + f\partial_i g$.

45.5.2 Directional Derivatives

The rate of change of a multivariable function depends on the direction in which the change is measured.

Given a direction vector $\overrightarrow{\mathbf{v}} \in \mathbb{R}^n$ and a point $\mathbf{x} \in \mathbb{R}^n$, the line $L_{\mathbf{x},\overrightarrow{\mathbf{v}}}$ passing through $\mathbf{x}$ in the direction of $\overrightarrow{\mathbf{v}}$ is parametrised by $\mathbf{r}(t) = \mathbf{x} + t\overrightarrow{\mathbf{v}}$. Now, for any function $f : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^k$ with U open, there exists $\tau > 0$ such that the line segment $\mathbf{x} + t\overrightarrow{\mathbf{v}}$ is contained in U for $t \in (-\tau,\tau)$. The restricton $f_{\mathbf{x},\overrightarrow{\mathbf{v}}} : (-\tau,\tau) \to \mathbb{R}^k$ of f to this line segment is defined by,

$$f_{\mathbf{x},\overrightarrow{\mathbf{v}}}(t) \coloneqq f(\mathbf{x} + t\overrightarrow{\mathbf{v}})$$

This is now a function of a single real variable, so we can differentiate it componentwise.

The *directional derivative* of f in the direction of $\mathbf{v}$, denoted by $D_{\mathbf{v}}f(x)$ or $\partial_{\mathbf{v}}f(x)$, is defined by,

$$\begin{aligned}\partial_{\mathbf{v}}f(\mathbf{x}) &\coloneqq \frac{d}{dt}f_{\mathbf{x},\mathbf{v}}(t)\bigg|_{t=0}\\ &= \frac{d}{dt}f(\mathbf{x}+t\mathbf{v})\bigg|_{t=0}\\ &= \lim_{t\to 0}\frac{f(\mathbf{x}+t\overrightarrow{\mathbf{v}}) - f(\mathbf{x})}{t}\end{aligned}$$

In practice, you can calculate the directional derivative by multiplying the components of the normalised direction vector by the corresponding partial derivatives, or equivalently, by calculating the scalar product of the gradient and the direction vector: $\partial_{\mathbf{v}}f = \nabla f \cdot \mathbf{v}$ (where $\mathbf{v}$ is a unit vector).

Example. Find the directional derivative of $f(x,y) = x^2 - y^2$ in the direction of $\mathbf{v} = (a,b)$.

Since we are not given values for a and b, we do not modify $\mathbf{v}$, but in general, we would normalise $\mathbf{v}$ first.

We compute the directional derivative from the definition:

$$\begin{aligned}\frac{d}{dt}f\big((x,y)+t(a,b)\big)\bigg|_{t=0} &= \frac{d}{dt}f(x+ta,y+tb)\bigg|_{t=0}\\ &= \frac{d}{dt}\Big[(x+ta)^2 - (y+tb)^2\Big]\bigg|_{t=0}\\ &= 2a(x+ta) - 2b(y+tb)\bigg|_{t=0}\\ &= 2ax - 2by\end{aligned}$$

Alternatively, we can compute the partial derivatives (the components of ∇f);

$$\begin{aligned}\frac{\partial}{\partial x}f(x,y) &= 2x\\ \frac{\partial}{\partial y}f(x,y) &= -2y\end{aligned}$$

then multiply by the components of $\mathbf{v} = (a,b)$,

$$\nabla f \cdot \mathbf{v} = 2ax - 2by$$

△

For a function $f : \mathbb{R}^n \to \mathbb{R}$, the directional derivative existing for all $\mathbf{v} \in \mathbb{R}^n$ at a point $\mathbf{x}$ does *not* imply that f is continuous at $\mathbf{x}$, similarly to how linear continuity does not imply continuity.

45.5.3 The Fréchet Derivative

The derivative of a function $f : (a,b) \to \mathbb{R}$ at a point $x \in (a,b)$ is given by

$$\lim_{h\to 0} \frac{f(x+h) - f(x)}{h}$$

This definition cannot be easily extended to functions $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^k$, as there is no notion of division for vectors, unlike for real (or complex) numbers.

Instead, we rearrange the above to,

$$\lim_{h\to 0} \frac{|f(x+h) - (f(x) + f'(x)h)|}{|h|} = 0$$

That is, for a fixed x, the nonlinear mapping $h \mapsto f(x+h)$ is locally approximated by the affine linear map $h \mapsto f(x) + f'(x)h$.

Extending this idea to multivariate functions, a function $\mathbf{f} : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^k$ is differentiable at $\mathbf{x} \in U$ if there exists a linear map $T \in L(\mathbb{R}^n,\mathbb{R}^k)$ such that

$$\lim_{\mathbf{h}\to \mathbf{0}} \frac{\big|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \big(\mathbf{f}(\mathbf{x}) + T(\mathbf{h})\big)\big|}{|\mathbf{h}|} = 0$$

and we say that the linear map T is the *Fréchet derivative* of $\mathbf{f}$, also denoted by $D\mathbf{f}(\mathbf{x})$.

Expanding the $\varepsilon - \delta$ definition of the limit, we equivalently have: a function $\mathbf{f} : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^k$ is differentiable at $\mathbf{x} \in U$ if there exists a linear map $T \in L(\mathbb{R}^n,\mathbb{R}^k)$ such that,

$$\forall \varepsilon > 0, \exists \delta > 0 : |\mathbf{h}| < \delta \to \big|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \big(\mathbf{f}(\mathbf{x}) + T(\mathbf{h})\big)\big| < \varepsilon |\mathbf{h}|$$

Another characterisation is: a function $\mathbf{f} : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^k$ is differentiable at $\mathbf{x} \in U$ if there exists a linear map $T \in L(\mathbb{R}^n,\mathbb{R}^k)$ such that,

$$\mathbf{f}(\mathbf{x}+\mathbf{h}) = \mathbf{f}(\mathbf{x}) + T(\mathbf{h}) + \mathbf{E}(\mathbf{h})$$

where the error $\mathbf{E}(\mathbf{h}) \in o(\mathbf{h})$ grows asymptotically slower than linearly in $\mathbf{h}$.

Theorem 45.5.2. *If $\mathbf{f} : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^k$ is differentiable at $\mathbf{x} \in U$, then $\mathbf{f}$ is continuous at $\mathbf{x}$.*

Proof. As $\mathbf{f}$ is differentiable at $\mathbf{x}$, for all $\varepsilon > 0$, there exists $\delta > 0$ such that,

$$\begin{aligned} |\mathbf{h}| < \delta &\to |\mathbf{f}(\mathbf{x}+\mathbf{h}) - (\mathbf{f}(\mathbf{x}) + D\mathbf{f}(\mathbf{x})\mathbf{h})| \leq \varepsilon |\mathbf{h}| \\ &\to |\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x})| \leq |D\mathbf{f}(\mathbf{x})\mathbf{h}| + \varepsilon |\mathbf{h}| \\ &\to |\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x})| \leq (\|D\mathbf{f}(\mathbf{x})\| + \varepsilon)|\mathbf{h}| \end{aligned}$$

Set $\delta_* \coloneqq \min\big(\delta, \varepsilon/(\|D\mathbf{f}(\mathbf{x})\| + \varepsilon)\big)$. Then, if $|h| < \delta_*$, we have $|h| < \delta$, so,

$$|h| < \delta_* \to |\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x})| < (\|D\mathbf{f}(\mathbf{x})\| + \varepsilon)\delta_* < \varepsilon$$

■

Theorem 45.5.3 (Componentwise Differentiability)**.** *A function $\mathbf{f} : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^k$ defined by*

$$(x_1, \ldots, x_n) \mapsto \big(f_1(x_1, \ldots, x_n), f_2(x_1, \ldots, x_n), \ldots, f_k(x_1, \ldots, x_n)\big)$$

where $(f_j)_{j=1}^k$ are real-valued functions, is differentiable at $\mathbf{p} \in U$ if and only if every f_j is differentiable at $\mathbf{p}$.

That is, $\mathbf{f}$ is differentiable if and only if every component f_j is individually differentiable.

Theorem 45.5.4. *For a function $\mathbf{f} : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^k$, if $D\mathbf{f}(\mathbf{x})$ exists, then $\partial_{\mathbf{v}}\mathbf{f}(\mathbf{x})$ exists for all $\mathbf{v} \in \mathbb{R}^n$, and $\partial_{\mathbf{v}}\mathbf{f}(\mathbf{x}) = D\mathbf{f}(\mathbf{x})\mathbf{v}$.*

In particular, if $\mathbf{f}$ is differentiable at $\mathbf{x}$, then $\partial_{\mathbf{v}}\mathbf{f}(\mathbf{x})$ is linear in $\mathbf{v}$. That is,

$$\partial_{\alpha\mathbf{v}+\beta\mathbf{w}}\mathbf{f}(\mathbf{x}) = \alpha\partial_{\mathbf{v}}\mathbf{f}(\mathbf{x}) + \beta\partial_{\mathbf{w}}\mathbf{f}(\mathbf{x})$$

for all $\alpha,\beta \in \mathbb{R}$ and $\mathbf{v},\mathbf{w} \in \mathbb{R}^n$.

Note that the converse of this theorem does not hold – all directional derivatives existing does not guarantee that $\mathbf{f}$ is differentiable.

45.5.4 Gradient

Let $f : \mathbb{R}^n \to \mathbb{R}$. The *gradient* of f, denoted $\operatorname{grad} f$ or ∇f is the vector,

$$\nabla f = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} \\ \dfrac{\partial f}{\partial x_2} \\ \vdots \\ \dfrac{\partial f}{\partial x_n} \end{bmatrix}$$

$\nabla : \mathbb{R} \to \mathbb{R}^n$ by itself is the *grad operator*, and is effectively a vector full of partial derivative operators.

The *Jacobian matrix* of a function, $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^k$, denoted $\mathbf{J}$, $D\mathbf{f}$, or $\partial\mathbf{f}$, is the matrix of partial derivatives,

$$\partial\mathbf{f} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \dfrac{\partial f_1}{\partial x_3} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \dfrac{\partial f_2}{\partial x_3} & \cdots & \dfrac{\partial f_2}{\partial x_n} \\ \dfrac{\partial f_3}{\partial x_1} & \dfrac{\partial f_3}{\partial x_2} & \dfrac{\partial f_3}{\partial x_3} & \cdots & \dfrac{\partial f_3}{\partial x_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f_k}{\partial x_1} & \dfrac{\partial f_k}{\partial x_2} & \dfrac{\partial f_k}{\partial x_3} & \cdots & \dfrac{\partial f_k}{\partial x_n} \end{bmatrix}$$

This can be more compactly written as,

$$\partial\mathbf{f} = \begin{bmatrix} \dfrac{\partial \mathbf{f}}{\partial x_1} & \cdots & \dfrac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix}$$

or,

$$\partial\mathbf{f} = \begin{bmatrix} \nabla f_1 \\ \vdots \\ \nabla f_k \end{bmatrix}$$

Theorem 45.5.5. *If $\mathbf{f} : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^k$ is differentiable at $\mathbf{x} \in U$, and $\mathbf{h} \in \mathbb{R}^n$, then,*

$$D\mathbf{f}(\mathbf{x})(\mathbf{h}) = \partial\mathbf{f}(\mathbf{x})\mathbf{h}$$

On the left side, we have the linear map $D\mathbf{f}$ given by the Fréchet derivative, and on the right side, we have the Jacobian matrix, so this theorem just says that the Fréchet derivative is represented by the Jacobian matrix if $\mathbf{f}$ is already known to be differentiable at $\mathbf{x}$.

Theorem 45.5.6. *When f is differentiable at x, $Df(x)(h) = \partial_h f(x) = \partial f(x)h$.*

That is, whenever f is differentiable at x, the Fréchet derivative $Df(x)$ centred at x evaluated at h, the directional derivative $\partial_h f(x)$ of f at x in the direction of h, and the Jacobian matrix evaluated at x multiplied by h are all equal.

If f is not differentiable at x – that is, the Fréchet derivative $Df(x)$ does not exist – then directional derivative $\partial_h f(x)$ and the Jacobian $\partial f(x)$ may both still exist, but may not necessarily be equal.

However, if all partial derivatives are continuous at x (and hence the Jacobian is also continuous at x), then $Df(x)$ is guaranteed to exist:

Theorem 45.5.7. *Let $\mathbf{f} : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^k$ be a function and suppose there exists $\mathbb{B}_r(\mathbf{x}) \subset U$ such that the Jacobian matrix $\partial\mathbf{f}(\mathbf{y})$ exists at all points $\mathbf{y} \in \mathbb{B}_r(\mathbf{x})$ and that $\partial\mathbf{f}$ is continuous at $\mathbf{x}$. Then, $\mathbf{f}$ is differentiable at $\mathbf{x}$ and the Fréchet derivative is equal to the Jacobian matrix*

$$D\mathbf{f}(\mathbf{x})(\mathbf{h}) = \partial\mathbf{f}(\mathbf{x})\mathbf{h}$$

for all $\mathbf{h} \in \mathbb{R}^n$:

45.5.5 Geometric Approximation

Let $\mathbf{r} : [a,b] \to \mathbb{R}^k$ be a continuously differentiable parametrisation of a curve $C = r([a,b]) \subset \mathbb{R}^k$. Furthermore, suppose $\mathbf{r}$ is a *regular* parametrisation – that is, $\mathbf{r}'(t) \neq 0$ for all t. We can then interpret $\mathbf{r}'(t)$ to be the vector tangent to C at $\mathbf{r}(t)$, or alternatively, we can view $\mathbf{r}(t)$ to be the position of a particle tracing out C, and $\mathbf{r}'(t)$ is the velocity of the particle.

The line $L_{\mathbf{r}(t)}$ tangent to C at $\mathbf{r}(t)$ is then parametrised by,

$$\ell(h) = \mathbf{r}(t) + \mathbf{r}'(t)h$$

But, $\mathbf{r}'(t) = \partial\mathbf{r}(t)$, so the affine linear approximation of $h \mapsto \mathbf{r}(t+h)$ by $h \mapsto \mathbf{r}(t) + \big(\partial\mathbf{r}(t)\big)(h) = \ell(h)$ is a parametrisation of the tangent line $L_{\mathbf{r}(t)}$. That is, this approximation using Jacobian, for small h, corresponds to geometrically approximating the curve C by $L_{\mathbf{r}(t)}$ near $\mathbf{r}(t)$. This also holds true for more general parametrisations.

Let $U \subseteq \mathbb{R}^n$ be open, and let $\mathbf{r} : U \to \mathbb{R}^3$ be a continuously differentiable parametrisation of a surface $S = \mathbf{r}(U) \subset \mathbb{R}^3$. Furthermore, suppose $\mathbf{r}$ is a regular parametrisation – that is, $\partial\mathbf{r}(\mathbf{x})$ is of rank 2 for all $\mathbf{x} \in U$. If,

$$\mathbf{r}(u,v) = \begin{bmatrix} x(u,v) \\ y(u,v) \\ z(u,v) \end{bmatrix}$$

then,

$$\mathbf{r}_u = \begin{bmatrix} x_u \\ y_u \\ z_u \end{bmatrix}, \qquad \mathbf{r}_v = \begin{bmatrix} x_v \\ y_v \\ z_v \end{bmatrix}$$

(recall $\mathbf{r}_u = \frac{\partial\mathbf{r}}{\partial u}$, $x_u = \frac{\partial x}{\partial u}$, etc.) The Jacobian is given by,

$$\partial\mathbf{r} = \begin{bmatrix} x_u & x_v \\ y_u & y_v \\ z_u & z_v \end{bmatrix}$$

So, ∂ is of rank 2 if and only if $\mathbf{r}_u$ and $\mathbf{r}_v$ are linearly independent.

As in the 2-dimensional case, the affine linear approximation of $(h,k) \mapsto \mathbf{r}(u+h,v+k)$ by,

$$\begin{aligned}(h,k) &\mapsto \mathbf{r}(u,v) + \big(\partial \mathbf{r}(u,v)\big)(h,k)\\ &= \mathbf{r}(u,v) + h\mathbf{r}_u(u,v) + k\mathbf{r}_v(u,v)\end{aligned}$$

is then a parametrisation of the plane $\Pi_{\mathbf{r}(u,v)}$ tangent to S at $\mathbf{r}(u,v)$.

45.5.5.1 Graphs

Given a function $f : (U \subseteq \mathbb{R}^2) \to \mathbb{R}$, the *graph*, $\mathcal{G}_f$ of f is the surface parametrised by,

$$\mathbf{r}(x,y) = \big(x,y,f(x,y)\big)$$

That is, the height of the surface above the x-y plane is the value of $f(x,y)$, analogous to the 2-dimensional case where we plot the points given by $\big(x,f(x)\big)$.

Note that $\mathbf{r}_x = (1,0,f_x)$ and $\mathbf{r}_y = (0,1,f_y)$ are linearly independent for any function f.

A parametrisation of the plane tangent to the surface $\mathcal{G}_f$ at $\big(x,y,f(x,y)\big)$ is given by,

$$\begin{aligned}(h,k) &\mapsto \mathbf{r}(x,y) + \big(D\mathbf{r}(x,y)\big)(h,k)\\ &= \begin{bmatrix} x \\ y \\ f(x,y) \end{bmatrix} + h\begin{bmatrix} 1 \\ 0 \\ f_x \end{bmatrix} + k\begin{bmatrix} 0 \\ 1 \\ f_y \end{bmatrix}\\ &= \begin{bmatrix} x+h \\ y+k \\ f(x,y) + hf_x + kf_y \end{bmatrix}\\ &= \begin{bmatrix} x+h \\ y+k \\ f(x,y) + (h,k)\cdot \nabla f(x,y) \end{bmatrix}\end{aligned}$$

so f is not differentiable at $(x_0,y_0) \in U$ if and only if $\mathcal{G}_f$ does not have a tangent plane at $\big(x_0,y_0,f(x_0,y_0)\big)$.

45.5.6 Differentiation of Matrix-Valued Functions

$L(\mathbb{R}^n,\mathbb{R}^k) \cong \mathbb{R}^{k\times n} \cong \mathbb{R}^{nk}$, so the Fréchet derivative applies similarly to functions with domains and codomains in these spaces, the only difference being that the Euclidean norm $|\cdot|$ in the definition needs to be replaced by the operator norm $\|\cdot\|$ or Frobenius norm $\|\cdot\|_F$, respectively.

Example. Find the derivative of the map $f : L(\mathbb{R}^n) \to L(\mathbb{R}^n)$ defined by $f(T) = T \circ T = T^2$.

We consider $f(A+H) - f(A)$:

$$\begin{aligned}f(A+H) - f(A) &= (A+H)(A+H) - A^2\\ &= A^2 + AH + HA + H^2 - A^2\\ &= AH + HA + H^2\end{aligned}$$

The terms linear in H are $AH+HA$, so we should think that $\big(Df(A)\big)(H) = AH+HA$ is the derivative. However, we need to verify that it satisfies the required limit. First, rearrange to obtain,

$$f(A+H) - f(A) - (AH+HA) = H^2$$

Now verify the limit:

$$\begin{aligned}\lim_{H\to 0}\frac{\big\|f(A+H)-f(A)-\big(Df(A)\big)(H)\big\|}{\|H\|} &= \lim_{H\to 0}\frac{\|H^2\|}{\|H\|}\\ &\le \lim_{H\to 0}\frac{\|H\|^2}{\|H\|}\\ &= \lim_{H\to 0}\|H\|\\ &= 0\end{aligned}$$

so $\big(Df(A)\big)(H) = AH + HA$.

If we interpret f to act on matrices, we could also note that the entries of $f(\mathbf{A})$ are quadratic polynomials in the entries of $\mathbf{A}$, and are hence continuous. It then follows that f is differentiable, and $\big(Df(\mathbf{A})\big)(\mathbf{H}) = \partial_{\mathbf{H}}f(\mathbf{A})$, so we could calculate the directional derivative instead:

$$\begin{aligned}\partial_{\mathbf{H}}f(\mathbf{A}) &= \frac{d}{dt}f(\mathbf{A}+t\mathbf{H})\bigg|_{t=0}\\ &= \frac{d}{dt}(\mathbf{A}+t\mathbf{H})^2\bigg|_{t=0}\\ &= \frac{d}{dt}\mathbf{A}^2+t\mathbf{AH}+t\mathbf{HA}+t^2\mathbf{H}^2\bigg|_{t=0}\\ &= \mathbf{AH}+\mathbf{HA}+2t\mathbf{H}^2\Big|_{t=0}\\ &= \mathbf{AH}+\mathbf{HA}\end{aligned}$$

△

45.6 The Chain Rule

The following lemmata will be useful in the proof of the chain rule:

Lemma 45.6.1. *Given* $\mathbf{f} : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^k$, $\mathbf{x} \in U$, *and* $r > 0$ *such that* $\mathbb{B}_r(\mathbf{x}) \subset U$ *and* $T \in L(\mathbb{R}^n,\mathbb{R}^k)$, *we define* $\Delta_{\mathbf{x},T}\mathbf{f} : \mathbb{B}_r(\mathbf{0}) \to \mathbb{R}^k$ *by,*

$$\Delta_{\mathbf{x},T}\mathbf{f}(\mathbf{h}) = \begin{cases}\frac{\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})-T(\mathbf{h})}{|\mathbf{h}|} & \mathbf{h}\neq\mathbf{0}\\ 0 & \mathbf{h}=\mathbf{0}\end{cases}$$

Then, $\mathbf{f}$ *is differentiable at* $\mathbf{x}$ *with* $D\mathbf{f}(\mathbf{x}) = T$ *if and only if* $\Delta_{\mathbf{x},T}\mathbf{f}$ *is continuous at* $\mathbf{0}$.

Proof. If $\Delta_{\mathbf{x},T}\mathbf{f}$ is continuous at 0, then,

$$\begin{aligned}\lim_{\mathbf{h}\to\mathbf{0}}\left|\frac{\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})-T(\mathbf{h})}{|\mathbf{h}|}\right| &= \lim_{\mathbf{h}\to\mathbf{0}}\big|\Delta_{\mathbf{x},T}\mathbf{f}(\mathbf{h})\big|\\ &= \big|\lim_{\mathbf{h}\to\mathbf{0}}\Delta_{\mathbf{x},T}\mathbf{f}(\mathbf{h})\big|\\ &= \big|\Delta_{\mathbf{x},T}\mathbf{f}(\mathbf{0})\big|\\ &= \mathbf{0}\end{aligned}$$

so $\mathbf{f}$ is differentiable at $\mathbf{x}$ with $D\mathbf{f}(\mathbf{x}) = T$.

Conversely, if $\mathbf{f}$ is differentiable at $\mathbf{x}$, and we set $T = D\mathbf{f}(\mathbf{x})$, then we have $\lim_{\mathbf{h}\to\mathbf{0}}|\Delta_{x,T}\mathbf{f}(\mathbf{x})| = \mathbf{0}$. But then, $\lim_{\mathbf{h}\to\mathbf{0}}\Delta_{\mathbf{x},T}\mathbf{f}(\mathbf{h}) = \mathbf{0} = \Delta_{\mathbf{x},T}\mathbf{f}(\mathbf{0})$, so $\Delta_{\mathbf{x},T}\mathbf{f}$ is continuous at $\mathbf{0}$. ■

If $\mathbf{f}$ is differentiable at $\mathbf{x}$, we write $\Delta_{\mathbf{x}}\mathbf{f}(\mathbf{h})$ to abbreviate $\Delta_{\mathbf{x},D\mathbf{f}(\mathbf{x})}\mathbf{f}(\mathbf{h})$.

Lemma 45.6.2. *Let $\tau > 0$, and consider a function $\delta : (\mathbb{B}_\tau \subset \mathbb{R}^n) \to \mathbb{R}^k$ defined by*

$$\delta(\mathbf{h}) := \begin{cases} \xi(\mathbf{h})\eta(\mathbf{h}) & 0 < |\mathbf{h}| < \tau \\ \mathbf{0} & \mathbf{h} = \mathbf{0} \end{cases}$$

where $\xi : \big(\mathbb{B}_\tau \setminus \{\mathbf{0}\}\big) \to \mathbb{R}$ is bounded and $\eta : \mathbb{B}_\tau \to \mathbb{R}^k$ is continuous at $\mathbf{0} \in \mathbb{B}_\tau$ and $\eta(\mathbf{0}) = \mathbf{0}$. Then, δ is continuous at $\mathbf{0} \in \mathbb{B}_\tau$.

Proof. By continuity of η at $\mathbf{0}$, given $\varepsilon > 0$, there exists $\sigma \in (0,\tau)$ such that $|\eta(\mathbf{h})| < \varepsilon$ whenever $|\mathbf{h}| < \sigma$. By boundedness of ξ, there exists $M > 0$ such that $|\xi(\mathbf{h})| < M$ for all $\mathbf{h} \in \mathbb{B}_\tau \setminus \{0\}$.

Then, $|\delta(\mathbf{h})| < M\varepsilon$ whenever $0 < |\mathbf{h}| < \sigma$, which is to say, $\lim_{\mathbf{h}\to\mathbf{0}} \delta(\mathbf{h}) = \mathbf{0} = \delta(\mathbf{0})$, so δ is continuous at $\mathbf{0} \in \mathbb{B}_\tau$. ■

Theorem 45.6.3 (Chain Rule). *Let $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^k$ both be open. Suppose $\mathbf{f} : U \to \mathbb{R}^k$ is differentiable at $\mathbf{x} \in U$, and that $\mathbf{f}(\mathbf{x}) \in V$. If $\mathbf{g} : V \to \mathbb{R}^k$ is differentiable at $\mathbf{f}(\mathbf{x})$, then the composition $\mathbf{g} \circ \mathbf{f} : \mathbb{R}^n \to \mathbb{R}^k$ is differentiable at $\mathbf{x}$, and,*

$$D(\mathbf{g} \circ \mathbf{f})(\mathbf{x}) = D\mathbf{g}\big(\mathbf{f}(\mathbf{x})\big) \circ D\mathbf{f}(\mathbf{x})$$

Proof. We have

$$\mathbf{f}(\mathbf{x}+\mathbf{h}) = \mathbf{f}(\mathbf{x}) + D\mathbf{f}(\mathbf{x})\mathbf{h} + \Delta_{\mathbf{x}}\mathbf{f}(\mathbf{h})|\mathbf{h}|$$

and

$$\mathbf{g}\big(\mathbf{f}(\mathbf{x}) + \mathbf{k}\big) = \mathbf{g}\big(\mathbf{f}(\mathbf{x})\big) + D\mathbf{g}\big(\mathbf{f}(\mathbf{x})\big)\mathbf{k} + \Delta_{\mathbf{f}(\mathbf{x})}\mathbf{g}(\mathbf{k})|\mathbf{k}|$$

where

$$\Delta_{\mathbf{x}}\mathbf{f}(\mathbf{h}) := \begin{cases} \frac{\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})-D\mathbf{f}(\mathbf{x})\mathbf{h}}{|\mathbf{h}|} & \mathbf{h} \neq \mathbf{0} \\ \mathbf{0} & \mathbf{h} = \mathbf{0} \end{cases}$$

and

$$\Delta_{\mathbf{f}(\mathbf{x})}\mathbf{g}(\mathbf{k}) := \begin{cases} \frac{\mathbf{g}(\mathbf{f}(\mathbf{x})+\mathbf{k})-\mathbf{g}(\mathbf{f}(\mathbf{x}))-D\mathbf{g}(\mathbf{f}(\mathbf{x}))\mathbf{k}}{|\mathbf{k}|} & \mathbf{h} \neq \mathbf{0} \\ \mathbf{0} & \mathbf{h} = \mathbf{0} \end{cases}$$

Set $\mathbf{k}(\mathbf{h}) := D\mathbf{f}(\mathbf{x})\mathbf{h} + \Delta_{\mathbf{x}}\mathbf{f}(\mathbf{h})|\mathbf{h}|$ in the second equation. Then, by linearity of $D\mathbf{g}\big(\mathbf{f}(\mathbf{x})\big)$,

$$\mathbf{g}\big(\mathbf{f}(\mathbf{x}+\mathbf{h})\big) = \mathbf{g}\big(\mathbf{f}(\mathbf{x})\big) + D\mathbf{g}\big(\mathbf{f}(\mathbf{x})\big)\big(D\mathbf{f}(\mathbf{x})\mathbf{h}\big) + |\mathbf{h}|D\mathbf{g}\big(\mathbf{f}(\mathbf{x})\big)\big(\Delta_{\mathbf{x}}\mathbf{f}(\mathbf{h})\big) + |\mathbf{k}(\mathbf{h})|\Delta_{\mathbf{f}(\mathbf{x})}\mathbf{g}\big(\mathbf{k}(\mathbf{h})\big)$$

Therefore,

$$\mathbf{g}\big(\mathbf{f}(\mathbf{x}+\mathbf{h})\big) - \mathbf{g}\big(\mathbf{f}(\mathbf{x})\big) - D\mathbf{g}\big(\mathbf{f}(\mathbf{x})\big) \circ D\mathbf{f}(\mathbf{x})\mathbf{h} = |\mathbf{h}|(\delta_1\big(\mathbf{h}) + \delta_2(\mathbf{h})\big)$$

where

$$\delta_1(\mathbf{h}) := D\mathbf{g}\big(\mathbf{f}(\mathbf{x})\big)\big(\Delta_{\mathbf{x}}\mathbf{f}(\mathbf{h})\big)$$

and

$$\delta_2(\mathbf{h}) := \begin{cases} \frac{|\mathbf{k}(\mathbf{h})|}{|\mathbf{h}|}\Delta_{\mathbf{f}(\mathbf{x})}\mathbf{g}\big(\mathbf{k}(\mathbf{h})\big) & \mathbf{h} \neq \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{cases}$$

The proof will be complete once we show:

$$\lim_{\mathbf{h}\to\mathbf{0}}\big|\delta_1(\mathbf{h})\big| = 0 \qquad \text{and} \qquad \lim_{\mathbf{h}\to\mathbf{0}}\big|\delta_2(\mathbf{h})\big| = 0$$

We begin with δ_1.

$$\big|\delta_1(\mathbf{h})\big| \leq \big\|D\mathbf{g}\big(\mathbf{f}(\mathbf{x})\big)\big\|\big|\Delta_{\mathbf{x}}\mathbf{f}(\mathbf{h})\big|$$

and by differentiability of $\mathbf{f}$ at $\mathbf{x}$, we have

$$\lim_{\mathbf{h}\to\mathbf{0}}\big|\Delta_{\mathbf{x}}\mathbf{f}(\mathbf{h})\big| = 0$$

It follows immediately that $\lim_{\mathbf{h}\to\mathbf{0}}\big|\delta_1(\mathbf{h})\big| = 0$.

For δ_2, set

$$\xi(\eta) := \frac{\big|\mathbf{k}(\mathbf{h})\big|}{|\mathbf{h}|} \leq \frac{\big|D\mathbf{f}(\mathbf{x})\big|}{|\mathbf{h}|} + \big|\Delta_{\mathbf{x}}\mathbf{f}(\mathbf{h})\big| \leq \big\|D\mathbf{f}(\mathbf{x})\big\| + \big|\Delta_{\mathbf{x}}\mathbf{f}(\mathbf{h})\big|$$

for $\mathbf{h} \neq \mathbf{0}$. The continuity of $\Delta_{\mathbf{x}}\mathbf{f}$ at $\mathbf{0}$ implies that $\xi(\mathbf{h})$ is bounded on $\mathbb{B}_\tau \setminus \{0\}$ for some $\tau > 0$. Next, set

$$\eta(\mathbf{h}) := \Delta_{\mathbf{f}(\mathbf{x})}\mathbf{g}\big(\mathbf{k}(\mathbf{h})\big)$$

$\mathbf{k}(\mathbf{h})$ is a continuous function of $\mathbf{h}$, and $\mathbf{k}(\mathbf{0}) = \mathbf{0}$, so by differentiability of $\mathbf{g}$ at $\mathbf{f}(\mathbf{x})$, $\eta(\mathbf{h})$ is a continuous function of $\mathbf{h}$, and $\eta(\mathbf{0}) = \mathbf{0}$. We may then apply the previous lemma to $\delta_2(\mathbf{h}) = \xi(\mathbf{h})\eta(\mathbf{h})$ to obtain $\lim_{\mathbf{h}\to\mathbf{0}}\big|\delta_2(\mathbf{h})\big| = 0$. ■

Recall that the linear isomorphism $\mu : L(\mathbb{R}^n,\mathbb{R}^k) \to \mathbb{R}^{k\times n}$ maps linear transformations to their matrices. Applying this to the chain rule above gives a form of the chain rule with Jacobian matrices:

Theorem 45.6.4 (Chain Rule)**.** *Let $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^k$ both be open. Suppose $\mathbf{f} : U \to \mathbb{R}^k$ is differentiable at $\mathbf{x} \in U$, and that $\mathbf{f}(\mathbf{x}) \in V$. If $\mathbf{g} : V \to \mathbb{R}^k$ is differentiable at $\mathbf{f}(\mathbf{x})$, then the composition $\mathbf{g} \circ \mathbf{f} : \mathbb{R}^n \to \mathbb{R}^k$ is differentiable at $\mathbf{x}$, and,*

$$\partial(\mathbf{g} \circ \mathbf{f})(\mathbf{x}) = \partial\mathbf{g}\big(\mathbf{f}(\mathbf{x})\big)\partial\mathbf{f}(\mathbf{x})$$

Given functions $f : \mathbb{R}^m \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, the ith partial derivative of $g \circ f$ can be computed with the above chain rule as,

$$\partial_i(g \circ f)(\mathbf{x}) = g'(f(\mathbf{x}))\partial_i f(\mathbf{x})$$

so,

$$\nabla(g \circ f)(x) = g'(f(\mathbf{x}))\nabla f(\mathbf{x})$$

Example. Calculate $\nabla|\mathbf{x}|$, $\mathbf{x} \in \mathbb{R}^n \setminus \{0\}$.

$$|\mathbf{x}| = \sqrt{|\mathbf{x}|^2}$$

so we can apply the chain rule with $f(\mathbf{x}) = |\mathbf{x}|^2 = \sum_{i=1}^n x_i^2$ and $g(t) = \sqrt{t}$. First calculate ∇f and g':

$$\begin{aligned}\nabla f &= \nabla(x_1^2 + x_2^2 + \cdots + x_n^2) \\ &= \begin{bmatrix}\partial_1(x_1^2 + x_2^2 + \cdots + x_n^2) \\ \partial_2(x_1^2 + x_2^2 + \cdots + x_n^2) \\ \vdots \\ \partial_n(x_1^2 + x_2^2 + \cdots + x_n^2)\end{bmatrix}\end{aligned}$$

$$
\begin{aligned}
&= \begin{bmatrix} 2x_1 \\ 2x_2 \\ \vdots \\ 2x_n \end{bmatrix} \\
&= 2\mathbf{x} \\
g'(t) &= \frac{1}{2\sqrt{t}}
\end{aligned}
$$

$$
\begin{aligned}
\nabla|\mathbf{x}| &= \nabla(g \circ f)(\mathbf{x}) \\
&= g'(f(\mathbf{x}))\nabla f(\mathbf{x}) \\
&= \frac{1}{2\sqrt{|\mathbf{x}|^2}} 2\mathbf{x} \\
&= \frac{\mathbf{x}}{|\mathbf{x}|}
\end{aligned}
$$

with component form given by,

$$\frac{\partial}{\partial x_i}|\mathbf{x}| = \frac{x_i}{|\mathbf{x}|}$$

△

45.6.1 The Space $C^n(U,\mathbb{R}^k)$ of Continuously Differentiable Functions

Suppose $\mathbf{f} : (U \subseteq \mathbb{R}^n) \to \mathbb{R}^k$ is differentiable on U. Then, $\mathbf{f}$ is *continuously differentiable* at $\mathbf{p} \in U$ if the map $D\mathbf{f} : U \to L(\mathbb{R}^n,\mathbb{R}^k)$ defined by $\mathbf{x} \mapsto D\mathbf{f}(\mathbf{x})$ is continuous at $\mathbf{p}$. That is,

$$\forall \varepsilon - 0 \exists \delta > 0 : |\mathbf{x} - \mathbf{p}| < \delta \to \|D\mathbf{f}(\mathbf{x}) - D\mathbf{f}(\mathbf{p})\| < \varepsilon$$

Theorem 45.6.5. *A function $\mathbf{f} : U \to \mathbb{R}^k$ is continuously differentiable on U if and only if the Jacobian matrix $\partial\mathbf{f} : U \to \mathbb{R}^{k \times n}$ is continuous on U.*

This means that we can check if a function is continuously differentiable by computing all first-order partial derivatives $\partial_i f_j$ of $\mathbf{f} = (f_1, \ldots, f_k)$ and verifying that they are all continuous.

45.6.2 Mean Value Inequality

For any vector-valued function of a single real variable, $\mathbf{f} : [a,b] \to \mathbb{R}^k$, $\mathbf{f}(t) = (f_1(t), f_2(t), \ldots, f_k(t))$, we define the integral of $\mathbf{f}$ as,

$$\int_a^b \mathbf{f}(t)\, dt = \begin{bmatrix} \int_a^b f_1(t)\, dt \\ \int_a^b f_2(t)\, dt \\ \vdots \\ \int_a^b f_k(t)\, dt \end{bmatrix}$$

Lemma 45.6.6. *For any function $\mathbf{f} : [a,b] \to \mathbb{R}^k$,*

$$\left| \int_a^b \mathbf{f}(t)\, dt \right| \leq \int_a^b |\mathbf{f}(t)|\, dt$$

Proof. Let $\mathbf{I} \coloneqq \int_a^b \mathbf{f}(t)\,dt \in \mathbb{R}^k$. If $\mathbf{I} = \mathbf{0}$, then we have equality. Otherwise,

$$\begin{aligned}
|\mathbf{I}|\left|\int_a^b \mathbf{f}(t)\,dt\right| &= |\mathbf{I}|^2 \\
&= \mathbf{I}\cdot\mathbf{I} \\
&= \mathbf{I}\cdot\int_a^b \mathbf{f}(t)\,dt \\
&= \int_a^b \mathbf{I}\cdot\mathbf{f}(t)\,dt \\
&\leq \int_a^b |\mathbf{I}||\mathbf{f}(t)|\,dt \\
&= |\mathbf{I}|\int_a^b |\mathbf{f}(t)|\,dt
\end{aligned}$$

Dividing the first and last terms by $|\mathbf{I}|$ provides the result. ■

Theorem 45.6.7 (Generalised Mean Value Inequality)**.** *Suppose that* $\mathbf{x},\mathbf{y} \in U \subseteq \mathbb{R}^n$ *can be joined by a continuously differentiable path,* $\mathbf{r} : [a,b] \to U$*,* $\mathbf{r}(a) = \mathbf{x}$*,* $\mathbf{r}(b) = \mathbf{y}$*. Suppose that* $f \in C^1(U,\mathbb{R}^k)$*, and that there exists* $M \geq 0$ *such that the Jacobian satisfies* $\|\partial\mathbf{f}(\mathbf{x})\| \leq M$ *for all* $\mathbf{x} \in U$*. Then,*

$$|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x})| \leq M\,\mathrm{length}\big(\mathbf{r}([a,b])\big)$$

Proof.

$$\begin{aligned}
\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x}) &= \mathbf{f}\big(\mathbf{r}(b)\big) - \mathbf{f}\big(\mathbf{r}(a)\big) \\
&= \int_a^b \frac{d}{dt}\mathbf{f}\big(\mathbf{r}(t)\big)\,dt && \text{[Fundamental Theorem of Calculus II]} \\
&= \int_a^b \partial\mathbf{f}\big(\mathbf{r}(t)\big)\mathbf{r}'(t)\,dt && \text{[Chain Rule]}
\end{aligned}$$

so by the lemma above,

$$\begin{aligned}
|\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x})| &= \left|\int_a^b \partial\mathbf{f}\big(\mathbf{r}(t)\big)\mathbf{r}'(t)\,dt\right| \\
&\leq \int_a^b \left|\partial\mathbf{f}\big(\mathbf{r}(t)\big)\mathbf{r}'(t)\right|\,dt \\
&\leq \int_a^b \left\|\partial\mathbf{f}\big(\mathbf{r}(t)\big)\right\||\mathbf{r}'(t)|\,dt \\
&\leq \int_a^b M|\mathbf{r}'(t)|\,dt \\
&= M\,\mathrm{length}\big(\mathbf{r}([a,b])\big)
\end{aligned}$$

■

Corollary 45.6.7.1 (Vanishing Derivative)**.** *Suppose that* $U \subset \mathbb{R}^n$ *is differentiably path-connected and that* $\mathbf{f} : U \to \mathbb{R}^k$ *satisfies* $\partial\mathbf{f}(\mathbf{x}) = 0$ *for all* $\mathbf{x} \in U$*. Then,* $\mathbf{f}$ *is constant on* U*.*

Proof. Fix a point $\mathbf{y} \in U$. Then, by differentiable path-connectedness, given $\mathbf{x} \in U$, there exists a continuously differentiable path $\mathbf{r} : [a,b] \to U$ joining $\mathbf{x}$ to $\mathbf{y}$. So, by the generalised mean value inequality with $\partial\mathbf{f} = \mathbf{0}$, $\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{y})$ for all $\mathbf{x} \in U$. ■

This corollary does not hold if U is not path-connected, but the converse is true even if U is not path-connected.

For scalar valued functions $f : U \to \mathbb{R}$, this corollary can be stated as,

Corollary 45.6.7.2. *If $\nabla f(\mathbf{x}) = \mathbf{0}$ at all points $\mathbf{x}$ of a path-connected open set, then f is constant.*

A set $U \subseteq \mathbb{R}^n$ is *convex* if for all $\mathbf{x}$,$\mathbf{y} \in U$, the line,

$$L_{\mathbf{x},\mathbf{y}} = \{t\mathbf{x} + (1-t)\mathbf{y} : 0 \leq t \leq 1\}$$

is contained within U.

Corollary 45.6.7.3 (Mean Value Inequality)**.** *Let $U \subseteq \mathbb{R}^n$ be convex, and suppose that $\mathbf{f} \in C^1(U,\mathbb{R}^k)$ satisfies $\|\partial\mathbf{f}(\mathbf{x})\| \leq M$ for all $\mathbf{x} \in U$ for some $M \geq 0$. Then,*

$$|f(x) - f(y)| \leq M|x-y|$$

That is, $\mathbf{f}$ is Lipschitz continuous.

Proof. The result follows from the generalised mean value inequality with length$(L_{\mathbf{x},\mathbf{y}}) = |x-y|$. ■

The converse of the mean value inequality does not hold. That is, a function $\mathbf{f}$ being Lipschitz continuous does not imply that $\mathbf{f}$ is differentiable. For example, $(x,y) \mapsto \frac{x^3}{}$ is Lipschitz continuous on all of $\mathbb{R}^n$, but is not differentiable at $\mathbf{0}$ because none of the partial derivatives $\partial_i|\mathbf{x}|$ exist at $\mathbf{0}$. The function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by,

$$(x,y) \mapsto \frac{x^3}{x^2+y^2}$$

is also Lipschitz over all of $\mathbb{R}^2$, and, unlike $\mathbf{x} \mapsto |\mathbf{x}|$, has partial derivatives that exist everywhere, but is still not differentiable at (0,0).

45.7 Vector Fields

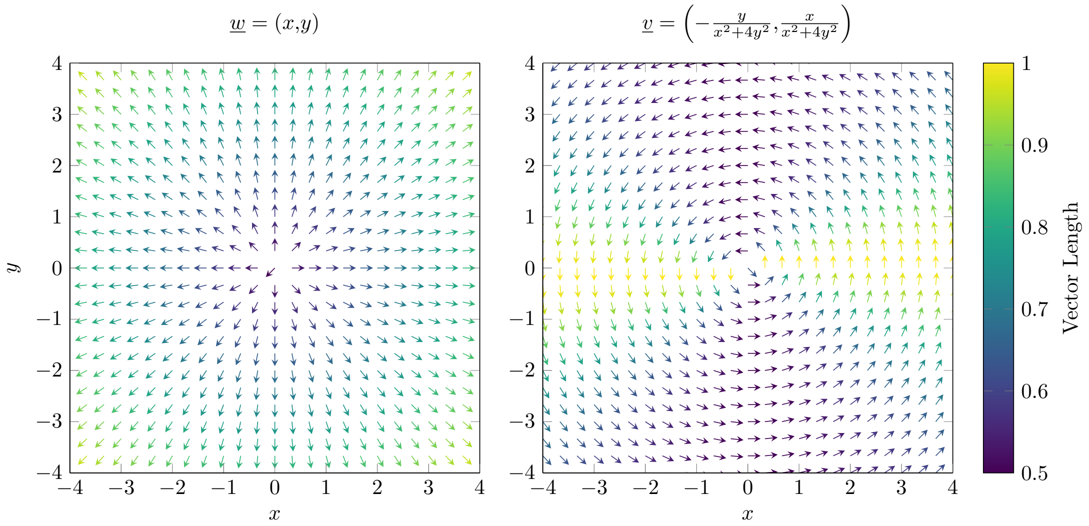
In this section, U will be a path-connected open subset of $\mathbb{R}^n$.

A *vector field* $\underline{v}$ on $U \subseteq \mathbb{R}^n$ is a function $\underline{v} : U \to \mathbb{R}^n$, so a vector field consists of n functions of n variables:

$$\underline{v}(\mathbf{x}) = \begin{bmatrix} v_1(x_1,x_2,\ldots,x_n) \\ v_2(x_1,x_2,\ldots,x_n) \\ \vdots \\ v_n(x_1,x_2,\ldots,x_n) \end{bmatrix}$$

We think of this function as associating a vector to every point in the input space.

Example.



△

A vector field will always be assumed to be at least continuous, and whenever it is differentiated, it will be assumed to be continuously differentiable.

45.7.1 Paths and Curves

A path $\mathbf{r} : [a,b] \to \mathbb{R}^n$ is said to be continuously differentiable on $[a,b]$ if,

- (i) $\mathbf{r}$ is continuous on $[a,b]$;
- (ii) $\mathbf{f}$ is continuously differentiable on (a,b);
- (iii) The limits $\lim_{t\to a^+} \mathbf{r}'(t)$ and $\lim_{t\to b^-} \mathbf{r}'(t)$ both exist so $\mathbf{r}'$ is a continuous function $[a,b] \to \mathbb{R}$.

We will always assume that a path $\mathbf{r} : [a,b] \to \mathbb{R}^n$ is continuous and piecewise continuously differentiable in the sense that there are a finite number of points $a_1,\ldots,a_\ell \in (a,b)$ with $a = a_0 < a_1 < a_2 < \cdots < a_\ell < a_{\ell+1} = b$ such that $\mathbf{r}$ is continosly differentiable on $[a_i,a_{i+1}]$ for all $0 \leq i \leq \ell$. If $\mathbf{r}'(t) \neq 0$ for all t, then $\mathbf{r}$ is *regular*.

Given $\mathbf{p},\mathbf{q} \in \mathbb{R}^n$, a curve $C_{\mathbf{p},\mathbf{q}}$ which goes from $\mathbf{p}$ to $\mathbf{q}$ is the image of some path $\mathbf{r} : [a,b] \to \mathbb{R}^n$ such that $\mathbf{r}(a) = \mathbf{p}$ and $\mathbf{r}(b) = \mathbf{q}$. The path $\mathbf{r}$ is then called a *parametrisation* of $C_{\mathbf{p},\mathbf{q}}$. If a curve C can be parametrised by a regular path, then the curve is also said to be regular.

Note that the parametrisation of a curve is not unique: If $\varphi : [\alpha,\beta] \to [a,b]$ is continuously differentiable, then $r \circ \varphi$ and r parametrise the same curve.

45.7.2 Tangential Line Integrals

The *component* of a vector $\mathbf{x} \in \mathbb{R}^n$ *in the direction of* a unit vector $\hat{\mathbf{v}}$ is defined as $\mathbf{x} \cdot \hat{\mathbf{v}}$. We also say that $\mathbf{x} \cdot \hat{\mathbf{v}}$ is the *component of* $\mathbf{x}$ *along* $\hat{\mathbf{v}}$.

If $\rho : [0,L] \to \mathbb{R}^n$ is the arclength or unit speed parametrisation of a regular curve $C_{pq} \subseteq \mathbb{R}^n$, then $\dot{\rho}(s) := \frac{d\rho}{ds}(s)$ is a unit vector called the *unit tangent* to C_{pq} at $\rho(s)$.

If $\underline{v}$ is a vector field, then $\underline{v}(\rho(s)) \cdot \dot{\rho}(s)$ is the *tangential component of* $\underline{v}$ *along* C_{pq}.

The *tangential line integral* of $\underline{v}$ along C_{pq} is defined as the integral of the tangential component of $\underline{v}$

along C_{pq}:

$$\int_0^L \underline{v}\big(\rho(s)\big) \cdot \dot{\rho}(s)\, ds$$

Because it is generally almost impossible to parametrise a curve by its arclength, we use a change of variable to actually compute these integrals in practice. Let $\mathbf{r} : [a,b] \to \mathbb{R}^n$ be a parametrisation of C_{pq}. Then, the mapping $\varphi : [0,L] \to [a,b]$ relates ρ and $\mathbf{r}$ by $\rho(s) = \mathbf{r}\big(\varphi(s)\big)$, so,

$$\int_0^L \underline{v}(\rho(s)) \cdot \dot{\rho}(s)\, ds = \int_a^b \underline{v}\big(\mathbf{r}(t)\big) \cdot \frac{d\mathbf{r}}{dt}\, dt$$

This line integral is also denoted by,

$$\int_{C_{pq}} \underline{v} \cdot d\mathbf{r}$$

obtained by cancelling the dt terms in the integral above.

Note that in the formula,

$$\int_0^L \underline{v}\big(\rho(s)\big) \cdot \dot{\rho}(s)\, ds$$

we have $\underline{v}(\rho(s)) \cdot \dot{\rho}(s) > 0$ whenever the angle between $\underline{v}$ and the unit tangent $\dot{\rho}$ is acute. Then,

$$\frac{1}{\text{length}(C_{pq})} \int_0^L \underline{v}\big(\rho(s)\big) \cdot \dot{\rho}(s)\, ds$$

represents the average value of $\underline{v} \cdot \dot{\rho}$ along C_{pq}, so the tangential line integral is a measure of the average rate at which the quantity described by the vector field $\underline{v}$ flows along C_{pq}.

If $\underline{v}$ represents a force, then $\int_C \underline{v} \cdot d\mathbf{r}$ represents the work done by $\underline{v}$ when moving an object along C. If C is a closed curve, then we write $\oint_C \underline{v} \cdot d\mathbf{r}$ instead, and the resulting value is sometimes called the *circulation* of $\underline{v}$ around C, as it measures the rate at which the quantity described by $\underline{v}$ circulates around C.

The value of the integral $\int_{C_{pq}} \underline{v} \cdot dr$ depends on the orientation of the path:

$$\int_{C_{pq}} \underline{v} \cdot d\mathbf{r} = -\int_{C_{qp}} \underline{v} \cdot d\mathbf{r}$$

When C is a closed curve, we write

$$\ointctrclockwise_C \underline{v} \cdot d\mathbf{r} = -\ointclockwise_C \underline{v} \cdot d\mathbf{r}$$

to indicate the orientation of the path.

45.7.3 Flux

45.7.3.1 Flux Across Curves in $\mathbb{R}^2$

Given a vector $\mathbf{v} = (x,y) \in \mathbb{R}^n$, we define $\mathbf{v}^\perp := (y, -x)$. That is, $\mathbf{v}^\perp$ is $\mathbf{v}$ rotated clockwise by 90°. In particular, $\mathbf{v} \cdot \mathbf{v}^\top = 0$, so $\mathbf{v}$ and $\mathbf{v}^\top$ are orthogonal.

The tangent $\dot{\mathbf{r}}(t)$ of a regular curve C with regular parametrisation $\mathbf{r}(t) = (x(t),y(t))$ is given by $\dot{\mathbf{r}}(t) = (\dot{x}(t),\dot{y}(t))$, so the *normal* to C is given by,

$$\mathbf{N}(t) := \dot{\mathbf{r}}(t)^\perp = \big(\dot{y}(t), -\dot{x}(t)\big)$$

If $\rho : [0,L] \to \mathbb{R}^n$ is the arclength parametrisation of C, then $\mathbf{n}(s) := \dot{\rho}(s)^{\perp}$ is the *unit normal* to C.

The *flux* of a vector field $\underline{v}(x,y) = \big(v_1(x,y),v_2(x,y)\big)$ across a curve C is defined as the integral of the normal component of $\underline{v}$,

$$\int_0^L \underline{v}\big(\rho(s)\big) \cdot \mathbf{n}(s)\, ds$$

Again, due to difficulties with parametrising a curve by its arclength, we compute this integral with another change of variable:

$$\int_0^L \underline{v}\big(\rho(s)\big) \cdot \mathbf{n}(s)\, ds = \int_a^b \underline{v}\big(\mathbf{r}(t)\big) \cdot \mathbf{N}(t)\, dt$$

45.7.3.2 Flux Across Surfaces in $\mathbb{R}^3$

A surface $S \subset \mathbb{R}^3$ is parametrised by the map $\mathbf{r} : (U \subseteq \mathbb{R}^2) \to \mathbb{R}^3$ with U open, defined by,

$$\mathbf{r}(u,v) = \big(x(u,v),y(u,v),z(u,v)\big)$$

The tangent plane $T_{\mathbf{r}(u,v)}S$ of S at $\mathbf{r}(u,v)$ is spanned by $\frac{\partial \mathbf{r}}{\partial u}(u,v)$ and $\frac{\partial \mathbf{r}}{\partial v}(u,v)$. It follows that the dimension of the tangent plane $T_{\mathbf{r}(u,v)}S$ is 2 if and only if the tangent vectors are linearly independent. If this is the case for all $(u,v) \in U$, then $\mathbf{r}$ is a *regular* parametrisation of S.

The tangent vectors $\frac{\partial \mathbf{r}}{\partial u}$ and $\frac{\partial \mathbf{r}}{\partial v}$ are linearly independent if and only if $\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} \neq \mathbf{0}$, in which case,

$$\mathbf{N}(u,v) := \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v}$$

is a *normal* to S at $\mathbf{r}(u,v)$.

Similarly to the definition of flux across a curve, the flux of a vector field $\underline{v}$ across a surface S is defined by,

$$\iint_S \underline{v} \cdot \hat{\mathbf{n}}\, dA$$

where $\mathbf{n}$ is a unit normal to S and dA is the area element on S. With respect to a parametrisation $\mathbf{r}$ of S, we have,

$$\mathbf{n}(u,v) := \frac{\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v}}{\left|\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v}\right|}, \qquad dA := \left|\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v}\right| du\, dv$$

or,

$$\mathbf{n}(u,v) := \frac{\mathbf{N}}{|\mathbf{N}|}, \qquad dA := |N|\; du\, dv$$

so the flux integral is given by,

$$\iint_S \underline{v} \cdot \hat{\mathbf{n}}\, dA = \iint_U \underline{v}\big(\mathbf{r}(u,v)\big) \cdot \left(\frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v}\right) du\, dv$$

The flux of $\underline{v}$ across S is also denoted by,

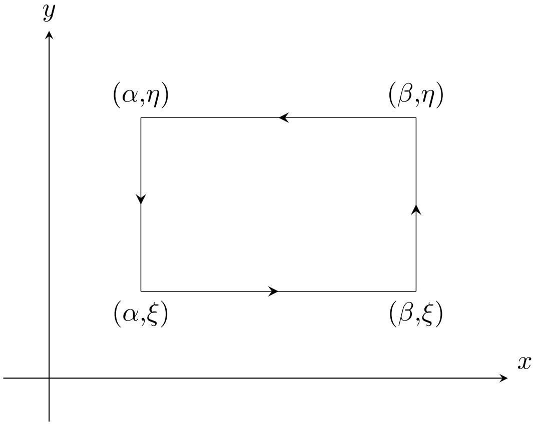
$$\iint_S \underline{v} \cdot d\mathbf{A}, \qquad \iint_S \underline{v} \cdot \mathbf{n}\, dS, \qquad \text{and} \qquad \iint_S \underline{v} \cdot d\mathbf{S}$$

45.8 The Integral Theorems of Vector Calculus

45.8.1 Green's Theorem for a Rectangle

Let $U \subseteq \mathbb{R}^2$ be open. A vector field $\underline{v} : U \to \mathbb{R}^2$ is called a *planar* vector field. As usual, we will assume $\underline{v}$ is continuously differentiable over U.

Let $\mathbb{R}$ denote the rectangle $[\alpha,\beta] \times [\xi,\eta]$ that is contained entirely (including the boundary ∂R) within U.



Consider the line integral of $\underline{v}$ around ∂R. If $\underline{v}(x,y) = \big(a(x,y),b(x,y)\big)$ for some scalar-valued functions $a,b : U \to \mathbb{R}$, then,

$$\oint_{\partial R} \underline{v} \cdot d\underline{r} = \int_\alpha^\beta a(x,\xi)\, dx + \int_\xi^\eta b(\beta,y)\, dy - \int_\alpha^\beta a(x,\eta)\, dx - \int_\xi^\eta b(\alpha,y)\, dy$$

By the fundamental theorem of calculus, we have,

$$\int_\alpha^\beta a(x,\xi)\, dx - \int_\alpha^\beta a(x,\eta)\, dx = \int_\alpha^\beta \int_\xi^\eta -\frac{\partial a(x,y)}{\partial y}\, dy\, dx$$

$$\int_\xi^\eta b(\beta,y)\, dy - \int_\xi^\eta b(\alpha,y)\, dy = \int_\xi^\eta \int_\alpha^\beta \frac{\partial b(x,y)}{\partial x}\, dx\, dy$$

So,

$$\oint_{\partial R} \underline{v} \cdot d\underline{r} = \iint_R \frac{\partial b(x,y)}{\partial x} - \frac{\partial a(x,y)}{\partial y}\, dx\, dy$$

obtaining the statement of Green's theorem for a rectangle.

45.8.1.1 Regions and Unit Normals

A *region* in $\mathbb{R}^n$ is a bounded open subset Ω of $\mathbb{R}^n$ for which there exists a function $f : \mathbb{R}^n \to \mathbb{R}$ such that,

- All partial derivatives of f are continuous;
- $\Omega = \{x \in \mathbb{R}^n : f(x) < 0\}$;
- $\nabla f(\mathbf{p}) \neq 0\, \forall \mathbf{p} \in f^{-1}\{0\} = \{\mathbf{p} \in \mathbb{R}^n : f(\mathbf{p}) = 0\}$.

The function f is then the *defining function* of Ω. The set $f^{-1}\{0\}$ is also the *boundary* of Ω, also denoted $\partial\Omega$. We also denote the *closure* $\Omega \cup \partial\Omega$ by $\overline{\Omega}$.

Example. Let $f(x,y,z) = x^2 + y^2 + z^2 - 1$. Then, the region defined by f is,

$$\begin{aligned}\{(x,y,z) : f(x,y,z) < 0\} &= \{(x,y,z) : x^2 + y^2 + z^2 < 1\}\\ &= \mathbb{B}_1(\mathbf{0})\end{aligned}$$

is the unit ball in $\mathbb{R}^3$, and its boundary is the unit 2-sphere $S^2(1) = \{\mathbf{x} \in \mathbb{R}^3 : |\mathbf{x}| = 1\}$. Note that

$$\begin{aligned}\nabla f(x,y,z) &= 2(x,y,z)\\ &\neq 0\end{aligned}$$

as required. △

The last requirement that $\nabla f(\mathbf{p}) \neq 0$ for all $\mathbf{p} \in \partial\Omega$ allows us to define the *outward unit normal* to Ω at $\mathbf{p}$:

$$\mathbf{n}_+(\mathbf{p}) := \frac{\nabla f(\mathbf{p})}{|\nabla f(\mathbf{p})|}$$

Unfortunately, this requirement also excludes some well-behaved subsets like polygons and polyhedra which do not have well defined normals at vertices and edges, so we also consider *piecewise regions.*

45.8.1.2 Boundary Orientation

We now focus on the 2-dimensional case. The boundary of a 2-dimensional, or *planar*, region Ω is a curve, or, if Ω is not simply connected as in the case of an annulus, a system of curves.

Let $\mathbf{n}_+(\mathbf{p}) = \big(h(\mathbf{p}),k(\mathbf{p})\big)$ be the outward unit normal to the region Ω at the point $\mathbf{p} \in \partial\Omega$. The *positively oriented unit tangent vector* $\mathbf{t}_+(p)$ at p is then the vector $\big(-k(\mathbf{p}),h(\mathbf{p})\big)$. That is, the outward unit normal rotated counterclockwise by 90°, or,

$$\mathbf{t}_+(\mathbf{p}) = -\mathbf{n}_+(\mathbf{p})^\perp$$

Informally, a tangent vector $\mathbf{t}$ to $\partial\Omega$ is positively oriented if, when facing in the direction of the vector, the interior of the region is to our left, and is negatively oriented otherwise. For example, if $\Omega = \mathbb{B}$, then a tangent vector that follows the unit circle in a counterclockwise manner is positively oriented.

However, take an annulus, for example. This region has two boundary curves; an *exterior* and *interior* boundary. A tangent vector on the exterior boundary is positively oriented if it follows the boundary counterclockwise, but a tangent vector on the interior boundary is positively oriented if it follows the boundary clockwise.

45.8.2 Green's Theorem for Planar Regions

In this section we will assume that all vector fields and functions considered are continuously differentiable on an open set $U \subseteq \mathbb{R}^2$, and that (the closure of) any region Ω lies entirely within U.

The *curl* of a planar vector field $\underline{v} : U \to \mathbb{R}^2$ defined by,

$$\underline{v}(x,y) = \big(a(x,y),b(x,y)\big)$$

is defined to be the function,

$$\operatorname{curl} \underline{v} = \frac{\partial b}{\partial x} - \frac{\partial a}{\partial y}$$

Theorem 45.8.1 (Green's Theorem for Planar Regions)**.** *Let Ω be a region in $\mathbb{R}^2$ and let $\underline{v} : U \to \mathbb{R}^2$ be a continuously differentiable planar vector field on U that contains $\overline{\Omega}$. Then,*

$$\iint_\Omega \operatorname{curl} \underline{v}(x,y)\, dA_{x,y} = \oint_{\partial\Omega} \underline{v} \cdot t_+ \, ds = \oint_{\partial\Omega} vli\ltimes \cdot d\mathbf{r}$$

where s is the arclength parameter along $\partial\Omega$, $\mathbf{r}$ is a positively oriented parametrisation of $\partial\Omega$, and the area element $dA_{x,y}$ is more often written as $dx\,dy$.

Recall that $\oint_{\partial\Omega} \underline{v} \cdot d\mathbf{r}$ is the circulation of $\underline{v}$ around $\partial\Omega$.

45.8.3 Flux and Divergence in the Plane

The *divergence* of a vector field $\underline{v}(x_1, \dots ,x_n) = (v_1(x_1, \dots ,x_n),v_2(x_1, \dots ,x_n), \dots ,v_n(x_1, \dots ,x_n))$, denoted by $\operatorname{div} \underline{v}$ and $\nabla \cdot \underline{v}$, is defined by,

$$\nabla \cdot \underline{v} := \sum_{i=1}^{n} \frac{\partial v_i}{\partial x_i}$$

Let $\mathbf{v},\mathbf{w} \in \mathbb{R}^2$. Then, $\mathbf{v} \cdot \mathbf{w} = \mathbf{v}^\perp \cdot \mathbf{w}^\perp$ and $(\mathbf{v}^\perp)^\perp = -\mathbf{v}$. So, if $\underline{v}$ is a planar vector field and Ω is a region in $\mathbb{R}^2$ that satisfy the hypotheses of Green's theorem, then,

$$\begin{aligned} \underline{v}^\perp \cdot \mathbf{t}_+ &= (\underline{v}^\perp)^\perp \cdot \mathbf{t}_+^\perp \\ &= -\underline{v} \cdot \mathbf{n}_+ \end{aligned}$$

The flux of $\underline{v}$ across $\partial\Omega$ is then given by,

$$\begin{aligned} \oint_{\partial\Omega} \underline{v} \cdot \mathbf{n}_+ \, ds &= -\oint \underline{v}^\perp \cdot \mathbf{t}_+ \, ds \\ &= -\iint_\Omega \operatorname{curl} \underline{v}^\perp \, dx\, dy \\ &= -\iint_\Omega \left(\frac{\partial(-a)}{\partial x} - \frac{\partial b}{\partial x} \right) dx\, dy \\ &= \iint_\Omega \left(\frac{\partial a}{\partial x} + \frac{\partial b}{\partial x} \right) dx\, dy \\ &= \iint_\Omega \nabla \cdot \underline{v}(x,y)\, dx\, dy \end{aligned}$$

Theorem 45.8.2 (Divergence Theorem for a Planar Region)**.** *Let Ω be a region in $\mathbb{R}^2$ and let $\underline{v} : U \to^2$ be a continuously differentiable planar vector field on U which contains $\overline{\Omega}$. Then,*

$$\iint_\Omega \nabla \cdot \underline{v}(x,y)\, dA_{x,y} = \oint_{\partial\Omega} \underline{v} \cdot \mathbf{n}_+ \, ds$$

where $\mathbf{n}_+$ is the outward unit normal to Ω.

Proof. Follows from Green's theorem as shown above. ■

45.8.4 Flux and Divergence in $\mathbb{R}^3$

Theorem 45.8.3 (Divergence Theorem)**.** *Let Ω be a region in $\mathbb{R}^3$ and let $\underline{v} : U \to \mathbb{R}^3$ be a continuously differentiable vector field on U which contains $\overline{\Omega}$. Then,*

$$\iiint_\Omega \nabla \cdot \underline{v}(x,y,z)\, dV_{x,y,z} = \iint_{\partial\Omega} \underline{v} \cdot \mathbf{n}_+ \, dA$$

where $\mathbf{n}_+$ is the outward unit normal to Ω, $dV_{x,y,z}$ is the volume element of Ω, more often written as $dx\,dy\,dz$.

45.8.5 Gradient Fields

If a vector field $\underline{v}$ is the gradient of a function $f : (U \subseteq \mathbb{R}^n) \to \mathbb{R}$, then $\underline{v}$ is called a *gradient field*, and the function f is called a *scalar potential* of $\underline{v}$.

How can we tell when a vector field $\underline{v}(x_1, \ldots ,x_n) = (v_1(x_1, \ldots ,x_n),v_2(x_1, \ldots ,x_n), \ldots ,v_n(x_1, \ldots ,x_n))$ is a gradient field? Or more generally, how can we recover the scalar potential of a gradient field?

This involves solving n simultaneous equations for one function f:

$$v_1(x_1, \ldots ,x_n) = \frac{\partial f}{\partial x_1},$$
$$\vdots$$
$$v_n(x_1, \ldots ,x_n) = \frac{\partial f}{\partial x_n},$$

For $n \geq 2$, we have more equations than unknowns, and so, we do not expect to be able to solve these equations for an arbitrary vector field $\underline{v}$.

We investigate under what conditions $\underline{v}$ has to satisfy in order for it to be a gradient field.

Theorem 45.8.4 (Fundamental Theorem of Calculus for Gradient Vector Fields)**.** *Given a continuously differentiable function $f : U \to \mathbb{R}$ and a curve $C_{\mathbf{pq}} \subset U$ from $\mathbf{p}$ to $\mathbf{q}$ parametrised by a continuously differentiable path $\mathbf{r} : [\mathbf{a},\mathbf{b}] \to \mathbf{U}$, we have,*

$$\int_{C_{pq}} \nabla f \cdot d\mathbf{r} = f(\mathbf{q}) - f(\mathbf{p})$$

Corollary 45.8.4.1. *This means that the value of a tangential line integral of a gradient field depends only on the orientation of C and the endpoints $\mathbf{p}$ and $\mathbf{q}$, and not on the shape of C itself.*

In particular, if the curve is closed, then the endpoints coincide, and we have:

Corollary 45.8.4.2. *For all closed curves C,*

$$\oint_C \nabla f \cdot d\mathbf{r} = 0$$

These two corollaries are equivalent in that any vector field that satisfies one will satisfy the other. Such a vector field is called a *conservative* vector field.

For example, gravity is a conservative field; it doesn't matter how you climb up a mountain, you gain the same amount of gravitational potential energy regardless of choice of path. Similarly, if you walk around but end up back where you started, you will have zero net gain of gravitational potential energy.

Theorem 45.8.5. *A continuous vector field $\underline{v} : U \to \mathbb{R}^n$ is a gradient field if and only if it is conservative.*

Proof. The forward direction follows from the above corollaries. For the reverse direction, pick $p \in U$, and define $f : U \to \mathbb{R}$ by:

$$f(x) = \int_{C_{px}} \underline{v} \cdot d\mathbf{r}$$

where $C_{\mathbf{px}}$ is any path in U from $\mathbf{p}$ to $\mathbf{x}$. (This is assuming that U is differentiably path-connected. If it is not, we pick a point and define this function for each path-connected component of U separately.) This integral does not depend on the choice of path from $\mathbf{p}$ to $\mathbf{x}$, so f is well-defined.

Let $(\mathbf{e}_i)_{i=1}^n$ be the standard basis of $\mathbb{R}^n$. Then,

$$\frac{\partial f}{\partial x_i} = \lim_{h\to 0} \frac{f(\mathbf{x}+h\mathbf{e}_i) - f(\mathbf{x})}{h}$$

Since U is open, there exists a $\delta > 0$ such that $\mathbb{B}_\delta(x) \subset U$. For $0 < h < \delta$, define $\mathbf{r}(t) = \mathbf{x} + t\mathbf{e}_i$ for $0 \le t \le h$. Thus, $\mathbf{r}$ parametrises the straight line from $\mathbf{x}$ to $\mathbf{x} + h\mathbf{e}_i$, and this line segment lies in U. Furthermore, we are free to choose the path in U from $\mathbf{p}$ to $\mathbf{x} + h\mathbf{e}_i$, so we can choose to go from $\mathbf{p}$ via $\mathbf{x}$; that is, we first go along $C_{\mathbf{px}}$, then go from $\mathbf{x}$ to $\mathbf{x} + h\mathbf{e}_i$ by means of $\mathbf{r}$. Thus,

$$\begin{aligned} f(\mathbf{x}+h\mathbf{e}_i) &= \int_{C_{\mathbf{px}}} \underline{v} \cdot d\mathbf{r} + \int_0^h \underline{v}\big(\mathbf{r}(t)\big) \cdot \frac{d\mathbf{r}}{dt}\,dt \\ &= f(\mathbf{x}) + \int_0^h \underline{v}\big(\mathbf{r}(t)\big) \cdot \mathbf{e}_i\,dt \end{aligned}$$

It follows that

$$\begin{aligned} \frac{\partial f}{\partial \mathbf{x}_i} &= \lim_{h\to 0} \frac{f(\mathbf{x}+h\mathbf{e}_i) - f(\mathbf{x})}{h} \\ &= \lim_{h\to 0} \left(\frac{1}{h} \int_0^h v_i\big(\mathbf{r}(t)\big)\,dt \right) \\ &= \frac{d}{dh} \left(\int_0^h v_i\big(\mathbf{r}(t)\big)\,dt \right) \Bigg|_{h=0} \\ &= v_i\big(\mathbf{r}(0)\big) \\ &= v_i(\mathbf{x}) \end{aligned}$$

■

Example. Given that

$$\underline{v}(x,y,z) = \big(xye^{xy} + e^{xy} + 1, x^2e^{xy}, 2z\big)$$

is conservative, find a scalar potential $f : \mathbb{R}^3 \to \mathbb{R}$ of $\underline{v}$.

We need to solve the three partial differential equations

$$\frac{\partial f}{\partial x} = xye^{xy} + e^{xy} + 1, \qquad \frac{\partial f}{\partial y} = x^2e^{xy}, \qquad \frac{\partial f}{\partial z} = 2z$$

for f.

Starting with the last equation, we have,

$$\begin{aligned} f(x,y,z) &= f(x,y,0) + \int_0^z \frac{\partial f}{\partial z}(x,y,t)\,dt \\ &= h(x,y) + \int_0^z 2t\,dt \\ &= h(x,y) + z^2 \end{aligned}$$

where $h(x,y) = f(x,y,0)$. Substituting into the second equation, we have $\frac{\partial h}{\partial y} = x^2e^{xy}$, which we similarly integrate into,

$$h(x,y) = h(x,0) + \int_0^y \frac{\partial h}{\partial y}(x,t)\,dt$$

$$\begin{aligned} &= g(x) + \int_0^y x^2 e^{xt}\,dt \\ &= g(x) + xe^{xy} - x \end{aligned}$$

where $g(x) = h(x,0)$. Again, we substitute into the first equation to obtain $g'(x) = 2$, so $g(x) = 2x + c$ for some constant $c \in \mathbb{R}$. Overall, this gives:

$$f(x,y,z) = xe^{xy} + x + z^2 + c$$

An alternative to this procedure is to integrate the last equation with respect to z directly, and write $f(x,y,z) = z^2 + h(x,y)$ where h is the "constant" of integration for the integration with respect to z, then repeating for the other variables. △

45.8.5.1 Incompressible and Irrotational Vector Fields

A vector field whose divergence is zero everywhere is called an *incompressible*, *solenoidal*, or *divergence-free* vector field.

Theorem 45.8.6 (Zero Flux Property)**.** *If $\underline{v} \in C^1(U \subseteq \mathbb{R}^3,\mathbb{R}^3)$ is incompressible, and $\overline{\Omega} \subset U$, then,*

$$\iint_{\partial\Omega} \underline{v} \cdot \mathbf{n}_+ \, dA = 0$$

Proof. $\underline{v}$ is incompressible, so $\nabla \cdot \underline{v} = 0$. By the divergence theorem,

$$\begin{aligned} \iint_{\partial\Omega} \underline{v} \cdot \mathbf{n}_+ \, dA &= \iiint_{\Omega} \nabla \cdot \underline{v}\, dV \\ &= \iiint_{\Omega} 0\, dV \\ &= 0 \end{aligned}$$

■

A vector field whose curl is zero everywhere is called an *irrotational* vector field.

Theorem 45.8.7. *Every conservative field is irrotational.*

Proof.

$$\begin{aligned} \operatorname{curl}(\nabla f) &= \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) - \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) \\ &= \frac{\partial f}{\partial x \partial y} - \frac{\partial f}{\partial y \partial x} \\ &= 0 \end{aligned}$$

■

45.8.5.2 Laplacian and Harmonic Functions

Let $\underline{v}$ be a incompressible conservative vector field with scalar potential f. Then, f satisfies the second order partial differential equation $\Delta f = 0$, where,

$$\begin{aligned} \Delta f &\coloneqq \nabla \cdot (\nabla f) \\ &= \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2} \end{aligned}$$

is the *Laplacian* of f.

For $f \in C^2(U)$, the equation $\Delta f = 0$ is called *Laplace's equation*, and its solutions are called *harmonic* functions or harmonic scalar fields.

45.9 Second Order Derivatives

45.9.1 Bilinear Forms

A linear map from a vector space to its field of scalars is called a *linear functional* or *covector*. The space $L(\mathbb{R}^n,\mathbb{R})$ of linear functionals on $\mathbb{R}^n$ is denoted by $(\mathbb{R}^n)^*$.

With respect to the standard basis of $\mathbb{R}^n$, $(\mathbb{R}^n)^*$ is identified with $\mathbb{R}^{1\times n}$. That is, every linear functional can be represented by a row vector.

A linear map $T : \mathbb{R}^n \to (\mathbb{R}^n)^*$ can be viewed as a bilinear form $\hat{T} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$:

$$\hat{T}(\mathbf{a},\mathbf{b}) := \mathbf{a}^\top T\mathbf{b}$$

45.9.2 The Hessian Matrix

Recall that if $f : (U \subseteq \mathbb{R}^n) \to \mathbb{R}$ is differentiable at $\mathbf{x}$, then there exists a linear transformation $Df \in L(\mathbb{R}^n,\mathbb{R}) = (\mathbb{R}^n)^*$ such that,

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\left|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \big(\mathbf{f}(\mathbf{x}) + Df(\mathbf{h})\big)\right|}{|\mathbf{h}|} = 0$$

(and further recall that $Df(\mathbf{x})$ is given by the Jacobian matrix, $\partial f(\mathbf{x})$).

Now, consider the case where $Df(\mathbf{x})$ itself is differentiable. Then, there exists some linear map $T \in L(\mathbb{R}^n,(\mathbb{R}^n)^*)$ such that,

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\left|Df(\mathbf{x}+\mathbf{h}) - \big(Df(\mathbf{x}) + T(\mathbf{h})\big)\right|}{|\mathbf{h}|} = 0$$

This map, if it exists, is called the *Hessian* of f, also denoted by $D^2f(\mathbf{x})$.

Suppose all second-order partial derivatives of $f : \mathbb{R}^n \to \mathbb{R}$ exist. Then, we define the *Hessian matrix*, denoted by $\mathbf{H}_f$ or $\partial^2 f(\mathbf{x})$, as,

$$\mathbf{H}_f = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n \partial x_1} & \dfrac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

That is, the (i,j)th entry is given by,

$$(\mathbf{H}_f)_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

This is the matrix that represents the Hessian transformation, if it exists. Note that the converse does *not* hold: even if all second order partial derivatives exist, and hence the Hessian matrix exists, Df may not necessarily be differentiable.

45.9.3 Non-Commutativity of Second Order Partial Derivatives

Second order partial derivatives do not, in general, commute. That is, for example,

$$\frac{\partial f}{\partial x \partial y} \neq \frac{\partial f}{\partial y \partial x}$$

However, these derivatives can commute under certain restrictions.

Theorem 45.9.1. *If the Hessian transformation $D^2 f(\mathbf{x})$ exists, then the second order partial derivatives at $\mathbf{x}$ commute. That is,*

$$\frac{\partial f}{\partial x \partial y}(\mathbf{x}) = \frac{\partial f}{\partial y \partial x}(\mathbf{x})$$

for all i,j; or, $\mathbf{H}_f$ is symmetric.

Corollary 45.9.1.1. *If all second order partial derivatives are continuous at $\mathbf{x}$, then the second order partial derivatives commute at $\mathbf{x}$.*

45.10 Inverse Function Theorem

45.10.1 Change of Variables and Inverse Functions

Let U and V be open subsets of $\mathbb{R}^n$. A change from variables $(x_1, \ldots, x_n) \in U$ to variables $(y_1, \ldots, y_n) \in V$ is achieved using a function $\Psi : U \to V$, with $\Psi = (\psi_1, \ldots, \psi_n)$such that,

$$\begin{aligned} y_1 &= (\psi_1(x_1, \ldots, x_n)) \\ y_2 &= (\psi_2(x_1, \ldots, x_n)) \\ &\vdots \\ y_n &= (\psi_n(x_1, \ldots, x_n)) \end{aligned}$$

If $\mathbf{\Psi}$ is bijective, then we can change back from y-variables to x-variables with the inverse map $\mathbf{\Psi}^{-1}$.

Theorem 45.10.1. *Suppose $\Psi : U \to V$ is a bijection differentiable at $\mathbf{x} \in U$, and suppose further that Ψ^{-1} is differentiable at $\mathbf{y} = \Psi(\mathbf{x}) \in V$. Then, $D\Psi(\mathbf{x})$ and $D\Psi^{-1}(\mathbf{y})$ are both invertible and,*

$$D\Psi^{-1}(\mathbf{y}) = \big(D\Psi(\Psi^{-1}(\mathbf{y}))\big)^{-1}$$

Proof. For all $\mathbf{y} \in V$,

$$\Psi\big(\Psi^{-1}(\mathbf{y})\big) = \mathbf{y}$$

Differentiating using the chain rule, we have,

$$D\Psi\big(\Psi^{-1}(\mathbf{y})\big) \circ D\Psi^{-1}(\mathbf{y}) = \mathrm{id}_{\mathbb{R}^n}$$

and the result follows. ■

In the 1-dimensional case, the Fréchet derivative is just the ordinary derivative, so this result is written as,

$$(\Psi^{-1})'(\mathbf{y}) = \frac{1}{\Psi'\big(\Psi^{-1}(\mathbf{y})\big)}$$

or more memorably as,

$$\frac{dx}{dy} = \frac{1}{\frac{dy}{dx}}$$

45.10.2 Local Inverses

Does the converse of the previous theorem hold? That is, if $\Psi : U \to V$ is differentiable at $\mathbf{x} \in U$, and $D\Psi(\mathbf{v})$ is invertible – does it then follow that Ψ^{-1} exists, and if it does, is it differentiable?

First, $D\Psi(\mathbf{x})$ depends only on the behaviour of Ψ near $\mathbf{x}$, so if $\big(D\Psi(\mathbf{x})\big)^{-1}$ exists, then Ψ^{-1} can exist at most "near" $\Psi(\mathbf{x})$, and not on all of $\Psi(U)$. We formalise what it means to be "near" a certain point.

Let $\mathbf{p} \in U$. If $\mathcal{N}_p \subseteq U$ is an open set containing $\mathbf{p}$, then we say that $\mathcal{N}_p$ is an (open) *neighbourhood* of $\mathbf{p}$.

Then, a function $\Psi : U \to V$ is a *local bijection* at $\mathbf{p} \in U$ if there is an open neighbourhood $\mathcal{N}_\mathbf{p}$ of $\mathbf{p}$ and an open neighbourhood $\mathcal{N}_\mathbf{q}$ of $\mathbf{q} = \Psi(\mathbf{p})$ such that the restriction $\Psi : \mathcal{N}_\mathbf{p} \to \mathcal{N}_\mathbf{q}$ is a bijection. We also say that Ψ is *locally invertible* at $\mathbf{p}$, and that inverse of the restricted function, $\Psi^{-1} : \mathcal{N}_\mathbf{q} \to \mathcal{N}_\mathbf{p}$ is the *local inverse* of Ψ. This local inverse is also called a *branch* of the *global* or *full* inverse Ψ^{-1} (if it exists).

Example. Consider $\Psi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ defined by $x \mapsto x^2$. Ψ is not injective, since $\Psi(x) = \Psi(-x)$.

But, take some $p > 0$, and the open neighbourhood $\mathcal{N}_p = (0,\infty)$. Then, $q = \Psi(p) = p^2 > 0$, and we can take the open neighbourhood $\mathcal{N}_q = (0,\infty)$, and indeed Ψ restricted to these neighbourhoods is bijective, with local inverse given by $\Psi^{-1}(y) = \sqrt{y}$. We can use the previous theorem to calculate the derivative of this inverse:

$$\begin{aligned} \Psi'(x) &= 2x \\ (\Psi^{-1})'(y) &= \frac{1}{\Psi'(\Psi^{-1}(y))} \\ &= \frac{1}{\Psi'(\sqrt{y})} \\ &= \frac{1}{2\sqrt{y}} \end{aligned}$$

If we instead take $p < 0$ with open neighbourhood $\mathcal{N}_p = (-\infty,0)$, and $q = p^2 > 0$ with open neighbourhood $\mathcal{N}_q = (0,\infty)$, Ψ is again bijective, with local inverse given by $\Psi^{-1}(x) = -\sqrt{x}$. This time, the derivative is given by,

$$\begin{aligned} \Psi'(x) &= 2x \\ (\Psi^{-1})'(y) &= \frac{1}{\Psi'(\Psi^{-1}(y))} \\ &= \frac{1}{\Psi'(-\sqrt{y})} \\ &= -\frac{1}{2\sqrt{y}} \end{aligned}$$

However, with our definition of a local inverse, there is no open neighbourhood around $p = 0$ such that Ψ is a bijection, so Ψ is not invertible on any open interval containing 0.

$\sqrt{y}$ and $-\sqrt{y}$ then form the two branches of the multivalued full inverse $\Psi^{-1}(y) = \pm\sqrt{y}$. △

Theorem 45.10.2 (Inverse Function Theorem)**.** *Let $U \subseteq \mathbb{R}^n$ be open, and suppose $\Psi : U \to \mathbb{R}^n$ is continuously differentiable. Suppose that the Fréchet derivative $D\Psi(\mathbf{p})$ is invertible at a point $\mathbf{p} \in U$ (that is, the Jacobian $\partial\Psi(\mathbf{p})$ has non-zero determinant), and define $\mathbf{q} = \Psi(\mathbf{p})$. Then,*

- *There exist neighbourhoods $\mathcal{N}_\mathbf{p} \subset U$ and $\mathcal{N}_\mathbf{q} \subset \Psi(U)$ of $\mathbf{p}$ and $\mathbf{q}$ respectively, such that the restriction $\Psi : \mathcal{N}_\mathbf{p} \to \mathcal{N}_\mathbf{q}$ is a bijection;*
- *The inverse of the restriction, $\Psi^{-1} : \mathcal{N}_\mathbf{q} \to \mathcal{N}_\mathbf{p}$, is continuously differentiable, and furthermore,*

$$D\Psi^{-1}(\mathbf{y}) = \big(D\Psi(\Psi^{-1}(\mathbf{y}))\big)^{-1}$$

for all $\mathbf{y} \in \mathcal{N}_{\mathbf{q}}$.

A map $\Psi : U \to V$ between two open subsets of $\mathbb{R}^n$ is called a *diffeomorphism* if it is bijective, continously differentiable on U, and its inverse is continuously differentiable on V.

Ψ is called a *local diffeomorphism near* $\mathbf{p} \in U$ if there exists a neighbourhood $\mathcal{N}_{\mathbf{p}} \subset U$ of $\mathbf{p}$ such that the restriction $\Psi : \mathcal{N}_{\mathbf{p}} \to \mathcal{N}_{\mathbf{q}}$ is a diffeomorphism, where $\mathbf{q} \coloneqq \Psi(\mathbf{p})$ and $\mathcal{N}_{\mathbf{q}} \coloneqq \Psi(\mathcal{N}_{\mathbf{p}}) \subset V$ is a neighbourhood of q.

45.11 Proof of the Implicit Function Theorem

Lemma 45.11.1. *A transformation* $T \in L(\mathbb{R}^n, \mathbb{R}^k)$ *is injective if and only if there exists* $\alpha > 0$ *such that* $|T(\mathbf{x})| \geq \alpha |\mathbf{x}|$ *for all* $\mathbf{x} \in \mathbb{R}^n$.

Proof. If $T(\mathbf{x}) = \mathbf{0}$ and $|T(\mathbf{x})| \geq \alpha|\mathbf{x}|$ for some $\alpha > 0$, then $\mathbf{x} = \mathbf{0}$, so T is injective.

Conversely, suppose there does not exist such an $\alpha > 0$, so there is a sequence $(\mathbf{x}_i)_{i=1}^{\infty} \subseteq \mathbb{R}^n \setminus \{\mathbf{0}\}$ such that $|T(\mathbf{x}_i)|/|\mathbf{x}_i| \to 0$ as $i \to \infty$.

Define $\mathbf{u}_i \coloneqq \mathbf{x}_i/|\mathbf{x}_i|$. Then, $|\mathbf{u}_i| = 1$ for all $i \in \mathbb{N}$, and $T(\mathbf{u}_j) \to \mathbf{0}$ as $i \to \infty$.

Since S^{n-1} is sequentially compact, there exists a subsequence $\mathbf{u}_{i_j}$ that converges to $u \in S^{n-1}$. However, $\mathbf{x} \mapsto |T(\mathbf{x})|$ is continuous, so

$$\begin{aligned} |T(\mathbf{x})| &= \lim_{i \to \infty} |T(\mathbf{x})| \\ &= 0 \end{aligned}$$

so $\mathbf{u} \in \ker(T)$, and T is not injective. ∎

Lemma 45.11.2. *Let* $U \subseteq \mathbb{R}^n$ *be open. Let* $\mathbf{f} \in C^1(U, \mathbb{R}^k)$ *and suppose that* $D\mathbf{f}(\mathbf{p})$ *is injective at some point* $\mathbf{p} \in U$. *Then, there exists* $\delta > 0$ *such that* $\mathbb{B}_\delta(\mathbf{p}) \subseteq U$ *and* $\mathbf{f}$ *is injective on* $\mathbb{B}_\delta(\mathbf{p})$.

Proof. By the previous lemma, there exists $\alpha > 0$ such that

$$\left| D\mathbf{f}(\mathbf{p})\mathbf{h} \right| \geq \alpha |\mathbf{h}| \tag{1}$$

for all $\mathbf{h} \in \mathbb{R}^n$. By continuity of $D\mathbf{f} : U \to L(\mathbb{R}^n, \mathbb{R}^k)$ at p, there exists $\delta > 0$ such that $\mathbb{B}_\delta(\mathbf{p}) \subseteq U$ and

$$\left\| D\mathbf{f}(\mathbf{p}) - D\mathbf{f}(\mathbf{x}) \right\| < \frac{1}{2}\alpha \tag{2}$$

for all $\mathbf{x} \in \mathbb{B}_\delta(p)$. Let $A \coloneqq D\mathbf{f}(\mathbf{p})$ and define $\mathbf{F} : U \to \mathbb{R}^k$ by $\mathbf{x} \mapsto \mathbf{f}(\mathbf{x}) - A(\mathbf{x})$. Then,

$$\begin{aligned} D\mathbf{F}(\mathbf{x}) &= D\mathbf{f}(\mathbf{x}) - A \\ &= D\mathbf{f}(\mathbf{x}) - D\mathbf{f}(p) \end{aligned}$$

so

$$\left\| D\mathbf{F}(\mathbf{x}) \right\| < \frac{1}{2}\alpha$$

for all $\mathbf{x} \in \mathbb{B}_\delta(\mathbf{p})$. Then, the mean value inequality yields

$$\left| \mathbf{F}(\mathbf{z}) - \mathbf{F}(\mathbf{x}) \right| \leq \frac{1}{2} |\mathbf{z} - \mathbf{x}|$$

for all $\mathbf{x}$,$\mathbf{z} \in \mathbb{B}_\delta(\mathbf{p})$. So,

$$\begin{aligned} \big|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{z})\big| &= \big|A(\mathbf{x}-\mathbf{z}) - \big(\mathbf{F}(\mathbf{z}) - \mathbf{F}(\mathbf{x})\big)\big| \\ &\geq \alpha|\mathbf{x}-\mathbf{z}| - \frac{1}{2}|\mathbf{x}-\mathbf{z}| \\ &= \frac{1}{2}\alpha|\mathbf{x}-\mathbf{z}| \end{aligned}$$

In particular, $\mathbf{f}$ is injective on $\mathbb{B}_\delta(\mathbf{p})$. ■

This lemma establishes the local injectivity of Φ in the statement of the inverse function theorem.

Note that this last inequality is a *quantitative* estimate of how strongly injective $\mathbf{f}$ is; the larger the value of α, the "more injective" $\mathbf{f}$ is. Similarly, regarding $\mathbf{F}$ as the difference between $\mathbf{f}$ and its affine linear approximation, the penultimate inequality gives a quantitative estimate of how close the linear approximation of $\mathbf{f}$ is to $\mathbf{f}$.

The next lemma helps prove that Φ is locally surjective; we have that there exists $\rho > 0$ such that all points within ρ of $\Psi(\mathbf{p})$ must lie in the image of Ψ.

The difficulty is in finding a preimage for each point $\mathbf{y} \in \mathbb{B}_\rho\big(\Psi(\mathbf{p})\big)$. The strategy is to look for $\mathbf{x}_*$ as a point which minimises the distance between $\mathbf{y}$ and $\Psi(\mathbf{x})$ as $\mathbf{x}$ moves around near $\mathbf{p}$.

Lemma 45.11.3. *Let $\mathbf{U} \subseteq \mathbb{R}^n$ be open. Let $\Psi \in C^1(U,\mathbb{R}^n)$, and suppose that $D\Psi(\mathbf{p})$ is surjective for some point $\mathbf{p} \in U$. Then, there exists $\rho > 0$ such that $\mathbb{B}_\rho\big(\Psi(\mathbf{p})\big) \subseteq \Psi(U)$.*

Proof. By the rank-nullity theorem, $D\Psi(\mathbf{p})$ is injective, so Lemma 45.11.2 applies. So, there exists $\alpha > 0$ such that

$$\big|D\Psi(\mathbf{p})\mathbf{h}\big| \geq \alpha|\mathbf{h}| \tag{1}$$

for all $\mathbf{h} \in \mathbb{R}^n$. As before, let $A := D\mathbf{f}(\mathbf{p})$ and define $\mathbf{F} : U \to \mathbb{R}^k$ by $\mathbf{x} \mapsto \Psi(\mathbf{x}) - A(\mathbf{x})$. Then, the following analogues of the last two inequalities in Lemma 45.11.2 hold: there exists $\delta > 0$ such that $\mathbb{B}_\delta(\mathbf{p}) \subseteq U$ and for all $\mathbf{x}$,$\mathbf{z} \in \mathbb{B}_\delta(\mathbf{p})$,

$$\big|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{z})\big| \leq \frac{1}{2}\alpha|\mathbf{x}-\mathbf{z}| \tag{2}$$

$$\big|\Psi(\mathbf{x}) - \Psi(\mathbf{z})\big| \geq \frac{1}{2}\alpha|\mathbf{x}-\mathbf{z}| \tag{3}$$

Set

$$\begin{aligned} K &:= \overline{\mathbb{B}}_{\frac{1}{2}\delta}(\mathbf{p}) \\ &= \big\{\mathbf{x} \in \mathbb{R}^n : |\mathbf{x}-\mathbf{p}| \leq \tfrac{1}{2}\delta\big\} \\ \partial K &= \big\{\mathbf{x} \in \mathbb{R}^n : |\mathbf{x}-\mathbf{p}| = \tfrac{1}{2}\delta\big\} \end{aligned}$$

Then, applying (3) with $\mathbf{z} = \mathbf{p}$ and $\mathbf{x} \in \partial K$ yields

$$\big|\Psi(\mathbf{x}) - \Psi(\mathbf{p})\big| \geq \frac{1}{4}\alpha\delta \tag{4}$$

for all $\mathbf{x} \in \partial K$. Set $\rho := \frac{1}{8}\alpha\delta$ and fix $\mathbf{y} \in \mathbb{B}_\rho\big(\Psi(\mathbf{p})\big)$.

We claim that $\mathbf{y} \in \Psi\big(\mathbb{B}_{\frac{1}{2}\delta}(\mathbf{p})\big)$. That is, if $\big|\mathbf{y} - \psi(\mathbf{p})\big| \leq \rho$, then the equation $\Psi(\mathbf{x}) = \mathbf{y}$ has a solution $\mathbf{x} \in \mathbb{B}_{\frac{1}{2}\delta}(\mathbf{p})$.

Define $\varphi : K \to \mathbb{R}$ by $\mathbf{x} \mapsto \big|\Psi(\mathbf{x}) - \mathbf{y}\big|$. Then, φ is continuous, and since K is sequentially compact, the extreme value theorem yields the existence of some $\mathbf{x}_* \in K$ such that

$$\varphi(\mathbf{x}_*) \leq \varphi(\mathbf{x})$$

for all $\mathbf{x} \in K$. We now establish that $\mathbf{x}_* \in \mathbb{B}_{\frac{1}{2}}(\mathbf{p})$ by showing that $\varphi(\mathbf{x}) > \varphi(\mathbf{p})$ for all $\mathbf{x} \in \partial K$, and hence $\mathbf{x}_* \notin \partial K$.

By (4), we have for all $\mathbf{x} \in \partial K$,

$$\begin{aligned}\varphi(\mathbf{x}) &= \big|\Psi(\mathbf{x}) - \mathbf{y}\big| \\ &\geq \big|\Psi(\mathbf{x} - \Psi(\mathbf{p})\big| - \big|\mathbf{y} - \Psi(\mathbf{p})\big| \\ &\geq \frac{1}{4}\alpha\delta - \frac{1}{8}\alpha\delta \\ &= \rho \\ &> \big|\mathbf{y} - \Psi(\mathbf{p})\big| \\ &= \varphi(\mathbf{p})\end{aligned}$$

so

$$|\mathbf{x}_* - \mathbf{p}| < \frac{1}{2}\delta \tag{5}$$

Since $D\Psi(\mathbf{p})$ is surjective, there exists $\mathbf{h} \in \mathbb{R}^n$ such that

$$\begin{aligned}A(\mathbf{h}) &= D\Psi(\mathbf{p})\mathbf{h} \\ &= \mathbf{y} - \Psi(\mathbf{x}_*)\end{aligned}$$

The idea is that if $\mathbf{y} \neq \Psi(\mathbf{x}_*)$, then we should be able to move $\mathbf{x}_*$ in a direction $\mathbf{h}$ such that $\Psi(\mathbf{x}_* + t\mathbf{h})$ would move closer to $\mathbf{y}$ for sufficiently small t, contradicting the construction of $\mathbf{x}_*$ as a point in K for which $\Psi(\mathbf{x}_*)$ is closest to $\mathbf{y}$. The affine linear approximation of $\Psi(\mathbf{x}_* + \mathbf{h})$ then asserts that $\Psi(\mathbf{x}_* + \mathbf{h}) - \Psi(\mathbf{x}_*) \approx D\Psi(\mathbf{x}_*)\mathbf{h} \approx D\Psi(\mathbf{p})\mathbf{h} = A(\mathbf{h})$. We wish to have $\Psi(\mathbf{x}_* + \mathbf{h}) = \mathbf{y}$, hence the requirement that $A(\mathbf{h}) = \mathbf{y} - \Psi(\mathbf{x}_*)$.

By (5), there exists $\eta > 0$ such that $\big|(\mathbf{x}_* + t\mathbf{h}) - \mathbf{p}\big| < \frac{1}{2}\delta$ whenever $|t| < \eta$. Invoking the affine linear approximation of $\Psi(\mathbf{x}_* + t\mathbf{h})$, we have for all $|t| < \eta$,

$$\begin{aligned}\Psi(\mathbf{x}_* + t\mathbf{h}) - \mathbf{y} &= \Psi(\mathbf{x} + t\mathbf{h}) - \Psi(\mathbf{x}_*) - tA(\mathbf{h}) + \big(\Psi(\mathbf{x}_*) - \mathbf{y} + tA(\mathbf{h})\big) \\ &= F(\mathbf{x} + t\mathbf{h}) - F(\mathbf{x}_*) + (1-t)\big(\Psi(\mathbf{x}_*) - \mathbf{y}\big)\end{aligned}$$

so applying (1) and (2), we have,

$$\begin{aligned}\varphi(\mathbf{x}_*) &\leq \varphi(\mathbf{x}_* + t\mathbf{h}) \\ &= \big|\Phi(\mathbf{x} + t\mathbf{h}) - \mathbf{y}\big| \\ &\leq \frac{1}{2}t\big|A(\mathbf{h})\big| + (1-t)\big|\Psi(\mathbf{x}_*) - \mathbf{y}\big| \\ &= \left(1 - \frac{1}{2}\right)\varphi(\mathbf{x}_*)\end{aligned}$$

Since this holds for all $t \in (0,\eta)$, we conclude that $\varphi(\mathbf{x}_*) = 0$, so $\mathbf{y} = \Psi(\mathbf{x}_*)$, and hence $\mathbb{B}_\rho\big(\Psi(\mathbf{p})\big) \subseteq \Psi\big(\mathbb{B}_{\frac{1}{2}\delta}(\mathbf{p})\big) \subseteq \Psi(U)$. ■

Corollary 45.11.3.1. *Let $U \subseteq \mathbb{R}^n$ be open. Let $\Psi \in C^1(U,\mathbb{R}^n)$, and suppose that $D\Psi(\mathbf{p})$ is invertible for all points $\mathbf{p} \in U$. Then Ψ is an open map. That is, the image of an open set in U under Ψ is open in $\mathbb{R}^n$, and in particular, if Ψ is injective on U, then $\Psi^{-1} : \Psi(U) \to U$ is continuous.*

Proof. Let $V \subseteq U$ be open. Applying the previous lemma to $\Psi\big|_V$, we have that for all $\mathbf{p} \in V$, there exists $\rho > 0$ such that $\mathbb{B}_\rho\big(\Psi(\mathbf{p})\big) \subseteq \Psi(V)$, i.e., $\Psi(V)$ is open. ■

Proof of Inverse Function Theorem. Ψ satisfies the hypotheses of Lemma 45.11.2 and Lemma 45.11.3. By inequalities (1) and (2) from Lemma 45.11.2 and the rank-nullity theorem, we deduce that $D\Psi(\mathbf{x})$ is invertible for all $\mathbf{x} \in \mathbb{B}_\delta(\mathbf{p})$. The previous corollary then implies that $\Phi\big(\mathbb{B}_\delta(\mathbf{p})\big)$ is open.

Set $\mathcal{N}_\mathbf{p} := \mathbb{B}_\delta(\mathbf{p})$ and $\mathcal{N}_\mathbf{q} := \Psi\big(\mathbb{B}_\delta(\mathbf{p})\big)$. Then, $\Psi : \mathcal{N}_\mathbf{p} \to \mathcal{N}_\mathbf{q}$ is a bijection.

It remains to prove that $\Psi^{-1} : \mathcal{N}_\mathbf{q} \to \mathcal{N}_\mathbf{p}$ is continuously differentiable.

Fix $\mathbf{y} \in \mathcal{N}_\mathbf{q}$ and choose $\rho > 0$ such that $\mathbb{B}_\rho(\mathbf{y}) \subseteq \mathcal{N}_\mathbf{q}$, and let $\mathbf{x} = \Psi^{-1}[y]$. It will be convenient to set

$$\Phi := \Psi^{-1}, \qquad A := D\Psi(\mathbf{x}), \qquad B := A^{-1}$$

For $|\mathbf{k}| < \rho$, define $\mathbf{h}(\mathbf{k}) := \Phi(\mathbf{y}+\mathbf{k}) - \Phi(\mathbf{y})$. Then, $\Phi(\mathbf{y}+\mathbf{k}) = \mathbf{x} + \mathbf{h}(\mathbf{k})$, so, using inequality (3) from Lemma 45.11.3, we have,

$$\begin{aligned}|\mathbf{k}| &= \big|(\mathbf{y}+\mathbf{k}) - \mathbf{y}\big| \\ &= \big|\Psi\big(\mathbf{x}+\mathbf{h}(\mathbf{k})\big) - \Psi(\mathbf{x})\big| \\ &\geq \frac{1}{2}\alpha\big|\mathbf{h}(\mathbf{k})\big|\end{aligned}$$

Since $\mathbf{h}(\mathbf{0}) = \mathbf{0}$, this implies that $\mathbf{h}(\mathbf{k})$ is continuous at $\mathbf{k} = \mathbf{0}$. Then, from Lemma 45.6.1, we have

$$\Psi\big(\mathbf{x}+\mathbf{h}(\mathbf{k})\big) = \Phi(\mathbf{x}) + D\Psi(\mathbf{x})\mathbf{h}(\mathbf{k}) + \Delta_\mathbf{x}\Psi\big(\mathbf{h}(\mathbf{k})\big)\big|\mathbf{h}(\mathbf{k})\big| \tag{1}$$

where

$$\Delta_{\mathbf{x},T}\mathbf{f}(\mathbf{h}) = \begin{cases} \frac{\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - T(\mathbf{h})}{|\mathbf{h}|} & \mathbf{h} \neq \mathbf{0} \\ 0 & \mathbf{h} = \mathbf{0} \end{cases}$$

and Δ_x abbreviates $\Delta_{\mathbf{x},D\mathbf{f}(\mathbf{x})}$.

Set $R(\mathbf{k}) := \Delta_\mathbf{x}\Psi\big(\mathbf{h}(\mathbf{k})\big)$. Then, the previous equation can be rewritten as

$$\mathbf{y}+\mathbf{k} = \mathbf{y} + A\big(\mathbf{h}(\mathbf{k})\big) + \big|\mathbf{h}(\mathbf{k})\big|R(\mathbf{k})$$

It follows that

$$B(\mathbf{k}) = \mathbf{h}(\mathbf{k}) + \big|\mathbf{h}(\mathbf{k})\big|B\big(R(\mathbf{k})\big)$$

and hence

$$\Phi(\mathbf{y}+\mathbf{k}) = \Phi(\mathbf{y}) + B(\mathbf{k}) - \big|\mathbf{h}(\mathbf{k})\big|B\big(R(\mathbf{l})\big) \tag{2}$$

Now, $\big|\mathbf{h}(\mathbf{k})\big| \leq \frac{2}{\alpha}|\mathbf{k}|$ and by continuity of composition and linearity of B, we have that $B\big(R(\mathbf{0})\big) = \mathbf{0}$. So, by Lemma 45.6.1, $\Psi^{-1} = \Phi$ is differentiable at $\mathbf{y}$ and

$$\begin{aligned}D\Psi^{-1}(\mathbf{y}) &= B \\ &= \big(D\Psi(\Psi^{-1}(\mathbf{y}))\big)^{-1}\end{aligned}$$

Finally, the continuity if $\mathbf{h}(\mathbf{k})$ at $\mathbf{k} = 0$ implies the continuity of Ψ^{-1} at $\mathbf{y}$ and therefore, since $D\Psi$ is continuous at $\mathbf{x} = \Psi^{-1}[y]$ and the inversion map $(-)^{-1} : GL(n,\mathbb{R}) \to GL(n,\mathbb{R})$ is continuous, we have that $D\Psi^{-1}$ is continuous at y. ■

45.12 Implicit Function Theorem

All functions in this section will be assumed to be continuously differentiable so we will write the Jacobian for the derivative instead of the Fréchet derivative.

Suppose we have a function $F : \mathbb{R}^2 \to \mathbb{R}$. If we set $F(x,y) = c$ for some $c \in \mathbb{R}$, then this equation defines y *implicitly* in terms of x, or x implicitly in terms of y.

For instance, if $F(x,y) = x^2 + y^2$ and $c > 0$, then $x^2 + y^2 = c$ describes a relation between x and y implicitly. If $c \neq \pm\sqrt{c}$, then the equation has two solutions for y in terms of x; namely, $y(x) = \sqrt{c - x^2}$ and $y(x) = -\sqrt{c - x^2}$, $-\sqrt{c} < x < \sqrt{c}$. Each of these solutions is called an *explict determination* of y in terms of x by the means of the functions $\sqrt{c - x^2}$ and $-\sqrt{c - x^2}$.

Let U be an open subset of $\mathbb{R}^{n+\ell} = \mathbb{R}^n \oplus \mathbb{R}^\ell$. We will write (x,y) for points in $\mathbb{R}^{n+\ell}$, where $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^\ell$.

Given a continuously differentiable function $\mathbf{F} : U \to \mathbb{R}^\ell$, we will write the Jacobian $\partial\mathbf{F}(x,y) \in \mathbb{R}^{\ell\times(n+\ell)}$ as $(\partial_x\mathbf{F}(x,y) \quad \partial_y\mathbf{F}(x,y))$ where $\partial_x\mathbf{F} \in \mathbb{R}^{\ell\times n}$ and $\partial_y\mathbf{F} \in \mathbb{R}^{\ell\times\ell}$

So, a matrix $\Lambda \in \mathbb{R}^{\ell\times(n+\ell)}$ can be written as $\Lambda = [\mathbf{A} \quad \mathbf{B}]$, where $\mathbf{A} \in \mathbb{R}^{\ell\times n}$ and $\mathbf{B} \in \mathbb{R}^{\ell\times\ell}$. If we then write a vector $\mathbf{z} \in \mathbb{R}^{n\times\ell}$ as $\mathbf{z} = (\mathbf{x},\mathbf{y})$ where $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^\ell$, then we can write a linear map $\mathbf{F} : \mathbb{R}^{n+\ell} \to \mathbb{R}^\ell$ defined by $\mathbf{F}(\mathbf{z}) = \Lambda\mathbf{z}$ as,

$$\mathbf{F}(\mathbf{x},\mathbf{y}) = [\mathbf{A} \quad \mathbf{B}] \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y}$$

Given some $\mathbf{c} \in \mathbb{R}^\ell$, we can then rewrite the equation $\mathbf{F}(\mathbf{x},\mathbf{y}) = \mathbf{c}$ as,

$$\mathbf{B}\mathbf{y} = \mathbf{c} - \mathbf{A}\mathbf{x}$$

This linear system of equations can be solved for the ℓ variables in $\mathbf{y}$ explicitly in terms of the n variables in $\mathbf{x}$ if $\mathbf{B}$ is invertible:

$$\mathbf{y} = \mathbf{B}^{-1}(\mathbf{c} - \mathbf{A}\mathbf{x})$$

If $\mathbf{B}$ is not invertible, then the system either has infinitely many solutions $\mathbf{y}$ if it is consistent, or no solutions if it is inconsistent. In either case, $\mathbf{y}$ cannot be written uniquely as a linear function of $\mathbf{x}$ if $\mathbf{B}$ is not invertible.

The implicit function theorem for a general continuously differentiable function $\mathbf{F} : U \to \mathbb{R}^\ell$ asserts that, if we have one solution $(\mathbf{x}_0,\mathbf{y}_0)$ of the equation $\mathbf{F}(\mathbf{x},\mathbf{y}) = \mathbf{c}$, and if $\partial_y\mathbf{F}(\mathbf{x}_0,\mathbf{y}_0) \in \mathbb{R}^{\ell\times\ell}$ is invertible, then we can solve for $\mathbf{y}$ in terms of $\mathbf{x}$ for $\mathbf{x}$ sufficiently near $\mathbf{x}_0$. The implicit function theorem is therefore concerned with converting an implicit relation $\mathbf{F}(\mathbf{x},\mathbf{y}) = \mathbf{c}$ to an explicit relation $\mathbf{y} = \mathbf{g}(\mathbf{x})$ such that the relation $\mathbf{F}(\mathbf{x},\mathbf{g}(\mathbf{x})) = \mathbf{c}$ holds for all $\mathbf{x}$ in some open neighbourhood $\mathcal{N}_{\mathbf{x}_0}$ containing $\mathbf{x}$.

Theorem 45.12.1 (Implicit Function Theorem)**.** *Let $U \subseteq \mathbb{R}^{n+\ell}$ be open and let $\mathbf{c} \in \mathbb{R}^\ell$. Suppose that $\mathbf{F} : U \to \mathbb{R}^\ell$ is continuously differentiable, and that the equation $\mathbf{F}(\mathbf{x},\mathbf{y}) = \mathbf{c}$ has a solution $(\mathbf{x}_0,\mathbf{y}_0) \in U$ such that $\det\big(\partial_y\mathbf{F}(\mathbf{x}_0,\mathbf{y}_0)\big) \neq 0$. Then, there exists an open neighbourhood $\mathcal{N}_{\mathbf{x}_0} \subseteq \mathbb{R}^n$ of $\mathbf{x}_0$ and a continuously differentiable function $\mathbf{g} : \mathcal{N}_{\mathbf{x}_0} \to \mathbb{R}^\ell$ such that,*

- $\mathbf{g}(\mathbf{x}_0) = \mathbf{y}_0$, $\Big\{\big(\mathbf{x},\mathbf{g}(\mathbf{x})\big) : \mathbf{x} \in \mathcal{N}_{\mathbf{x}_0}\Big\} \subset U$*, and* $\mathbf{F}\big(\mathbf{x},\mathbf{g}(\mathbf{x})\big) = \mathbf{c}$ *for all* $\mathbf{x} \in \mathcal{N}_{\mathbf{x}_0}$*;*
- *Furthermore,* $\partial_y\mathbf{F}(\mathbf{x},\mathbf{g}(\mathbf{x}))$ *is locally invertible over all* $\mathbf{x} \in \mathcal{N}_{\mathbf{x}_0}$*, and the derivative of* $\mathbf{g}$ *is given by,*

$$\partial\mathbf{g}(\mathbf{x}) = -\Big(\partial_y\mathbf{F}\big(\mathbf{x},\mathbf{g}(\mathbf{x})\big)\Big)^{-1} \cdot \partial_x\mathbf{F}\big(\mathbf{x},\mathbf{g}(\mathbf{x})\big)$$

for all $\mathbf{x} \in \mathcal{N}_{\mathbf{x}_0}$.

Example. In the previous example, we had $F(x,y) = x^2 + y^2$. Given a point (x_0,y_0) on the circle $F(x,y) = c > 0$ such that $y_0 > 0$, we have seen that $g(x) = \sqrt{c - x^2}$. The implicit function then gives the derivative of g to be,

$$\partial_x F(x,y) = [2x]$$

$$\begin{aligned}\partial_y F(x,y) &= [2y] \\ g'(x) &= -\big(\partial_y F(x,g(x))\big)^{-1} \cdot \partial_x F(x,g(x)) \\ &= -\frac{2x}{2g(x)} \\ &= -\frac{x}{\sqrt{c-x^2}}\end{aligned}$$

$\triangle$

Geometrically, the implicit function theorem asserts that if $\det\big(\partial_y \mathbf{F}(\mathbf{x}_0,\mathbf{y}_0)\big) \neq 0$, then near a point $(\mathbf{x}_0,\mathbf{y}_0)$, the level set $\Gamma_{\mathbf{c}} \coloneqq \{(\mathbf{x},\mathbf{y}) : \mathbf{F}(\mathbf{x},\mathbf{y}) = \mathbf{c}\}$ is the graph $\mathcal{G}_{\mathbf{g}}$ of a function $\mathbf{y} = \mathbf{g}(\mathbf{x})$. That is, near $(\mathbf{x}_0,\mathbf{y}_0)$,

$$\Gamma_{\mathbf{c}} = \mathcal{G}_{\mathbf{g}} \coloneqq \Big\{\big(\mathbf{x},\mathbf{g}(\mathbf{x})\big) : \mathbf{x} \in \mathcal{N}_{\mathbf{x}_0}\Big\} = \big\{(\mathbf{x},\mathbf{y}) : \mathbf{y} = \mathbf{g}(\mathbf{x}),\mathbf{x} \in \mathbf{x}_0\big\}$$

For instance, for $F(x,y) = x^2 + y^2 = c > 0$, $\partial_y F(x_0,y_0) \neq 0$ whenever $y_0 \neq 0$. If $y_0 > 0$, then (x_0,y_0) is contained in the upper semicircle which is the graph of $y = \sqrt{c - x^2}$, and if $y_0 < 0$, then (x_0,y_0) is contained in the lower semicircle which is the graph of $y = \sqrt{c - x^2}$.

The graph of a continuously differentiable function of one variable is a special case of a regular parametrisation of a curve, and the graph of a continuously differentiable function of two variables is a special case of a regular parametrisation of a surface. Regular parametrisations can be pieced together to form spaces that are known as *submanifolds* of Euclidean spaces. The circle in this case is a 1-dimensional submanifold of $\mathbb{R}^2$ that can be viewd as being pieced together along overlaps from the four semicircles which are the graphs of,

$$\begin{aligned}&y(x) = \sqrt{c-x^2}, x \in (-\sqrt{c},\sqrt{c}) \qquad && x(y) = \sqrt{c-y^2}, y \in (-\sqrt{c},\sqrt{c}) \\ &y(x) = -\sqrt{c-x^2}, x \in (-\sqrt{c},\sqrt{c}) \qquad && x(y) = -\sqrt{c-y^2}, y \in (-\sqrt{c},\sqrt{c})\end{aligned}$$

A set $M \subset \mathbb{R}^{n+\ell}$ is a *submanifold* (*without boundary*) of dimension n if, for each $\mathbf{p} \in M$, there exists an open neighbourhood $\mathcal{N}_{\mathbf{p}} \subset \mathbb{R}^{n+\ell}$ of $\mathbf{p}$, an open set $U \subset \mathbb{R}^n$ and a continuously differentiable function $\mathbf{r} : U \to \mathbb{R}^{n+\ell}$, such that $\mathbf{r}(\mathbf{x}_{\mathbf{p}}) = \mathbf{p}$ for some $\mathbf{x}_{\mathbf{p}} \in U$, $\mathbf{r} : U \to M \cap \mathcal{N}_{\mathbf{p}}$ is a bijection, and $\operatorname{rank}(\partial \mathbf{r}(\mathbf{x})) = n$ for all $\mathbf{x} \in U$.

The function $\mathbf{r}$ is then called a (regular) parametrisation of $M \cap \mathcal{N}_{\mathbf{p}}$. The *tangent space* $T_{\mathbf{r}(\mathbf{x})}M$ of M at $\mathbf{r}(\mathbf{x})$ is the image of $\partial \mathbf{r}(\mathbf{x})$ shifted by $\mathbf{r}(\mathbf{x})$; that is, $\mathbf{r}(\mathbf{x}) + \operatorname{span}\big(\partial_1 \mathbf{r}(\mathbf{x}), \dots \partial_n \mathbf{r}(\mathbf{x})\big)$, or,

$$T_{\mathbf{r}(\mathbf{x})}M = \Big\{\mathbf{r}(\mathbf{x}) + \big(\partial \mathbf{r}(\mathbf{x})\big)\mathbf{h} : \mathbf{h} \in \mathbb{R}^n\Big\}$$

Thus, the tangent space is identified with the image of the affine linear approximation of $\mathbf{r}$.

Theorem 45.12.2. *Given a continuously differentiable function $\mathbf{F} : U \subseteq \mathbb{R}^{n+\ell} \to \mathbb{R}^\ell$ with U open, and some fixed $\mathbf{c} \in \mathbb{R}^\ell$, define the level set $\Gamma_{\mathbf{c}} \coloneqq \{\mathbf{z} \in U : \mathbf{F}(\mathbf{z}) = \mathbf{c}\}$.*

Suppose that $\operatorname{rank}\big(\partial \mathbf{F}(\mathbf{z})\big) = \ell$ for all $\mathbf{z} \in \Gamma_{\mathbf{c}}$. Then, $\Gamma_{\mathbf{c}}$ is a submanifold (without boundary) of dimension n in $\mathbb{R}^{n+\ell}$. Furthermore, $T_z\Gamma_{\mathbf{c}} = \mathbf{z} + \ker\big(\partial \mathbf{F}(\mathbf{z})\big) = \big\{\mathbf{z} + \mathbf{v} : \partial \mathbf{F}(\mathbf{z})\mathbf{v} = \mathbf{0}\big\}$.

In the special case that $\ell = 1$, then $\Gamma_{\mathbf{c}}$ is called a *hypersurface* and,

$$\partial \mathbf{F}(\mathbf{z}) = (\partial_1 \mathbf{F}(\mathbf{z}), \dots \partial_{n+1} \mathbf{F}(\mathbf{z})), \qquad \nabla \mathbf{F}(\mathbf{z}) = \begin{bmatrix} \partial_1 \mathbf{F}(\mathbf{z}) \\ \vdots \\ \partial_{n+1} \mathbf{F}(\mathbf{z}) \end{bmatrix}$$

so,

$$\begin{aligned}\mathbf{v} \in \ker\bigl(\partial \mathbf{F}(\mathbf{z})\bigr) &\longleftrightarrow \bigl(\nabla \mathbf{F}(\mathbf{z})\bigr) \cdot \mathbf{v} = 0 \\ &\longleftrightarrow \nabla \mathbf{F}(\mathbf{z}) \perp T_{\mathbf{z}}\Gamma_{\mathbf{c}}\end{aligned}$$

so $\nabla \mathbf{F}$ is orthogonal to the level set $\Gamma_{\mathbf{c}}$, so the gradient of a function is the normal to the hypersurface it describes.

If we also have $n = 1$, then Γ_c is called a *level curve* in $\mathbb{R}^2$, and if $n = 1$, then $\Gamma_{\mathbf{c}}$ is called a *level surface* in $\mathbb{R}^3$.

Chapter 46

Differential Equations

> *"Since Newton, mankind has come to realise that the laws of physics are always expressed in the language of differential equations."*
>
> — Steven Strogatz

Differential equations are equations that relate functions with their derivatives. Here, we will mainly work with functions of single variables, with vector calculus being the focus of separate chapters, §44 and §45. For a foundational overview of calculus, see §34.

Newton's notation will not be used in this document. Lagrange's notation will be preferred, with Leibniz's notation used wherever differentials are more helpful (i.e. separable equations).

46.1 Functions and Variables

46.1.1 Terminology & Notation

46.1.1.1 Variables

Variables measure things. We can classify them into *independent* and *dependent* variables.

If a variable, for example, x, is a function of another variable, say, t, then x would be the dependent variable as its value is dependent on t, the independent variable. Usually, we see this written as $x(t)$.

There doesn't have to be a one-to-one correspondence between dependent and independent variables either: for example, you could have temperature as a function of position in 3D, $f(x,y,z)$, where f is the dependent variable, and x,y, and z are independent variables.

Dependent variables can usually be differentiated with respect to the independent variable(s).

46.1.1.2 Derivative Notation

When there is only one independent variable, we may save space and use Lagrange's (prime) notation over Leibniz's (quotient) notation:

$$\frac{dy}{dx} = y'$$

$$\frac{d^2y}{dx^2} = y''$$

$$\frac{d^ny}{dx^n} = y^{(n)}(x)$$

For nth derivatives in Lagrange's notation, do not omit the independent variable as to avoid confusion with exponents.

If the independent variable is time, we may also use Newton's (dot) notation:

$$\frac{dx}{dt} = \dot{x}$$
$$\frac{d^2x}{dt^2} = \ddot{x}$$

Newton's notation becomes rather unwieldy for derivatives of order higher than 2 or 3.

The partial derivative of a function $f(x,y,z)$ with respect to x, is variously written as,

$$\frac{\partial f}{\partial x}, f_x, \partial_x f$$

Other notations exist, but these are the main ones we will use.

The second-order partial derivative of f with respect to x is written as,

$$\frac{\partial^2 f}{\partial x^2}, f_{xx}, \partial_{xx} f, \partial_x^2 f$$

and the second-order mixed derivative of f with respect to x, then y is given by,

$$\frac{\partial^2 f}{\partial y \partial x}, f_{xy}, \partial_{yx} f, \partial_y \partial_x f$$

46.1.1.3 Properties of Differential Equations

If a differential equation only has one independent variable, it is referred to as an *ordinary differential equation*, or an *ODE*. A differential equation involving several independent variables is referred to as a *partial differential equation* or a *PDE*.

The *order* of a differential equation is the order of highest derivative present in the equation.

A differential equation is;

- *autonomous* if the independent variable does not appear in the ODE;
- *linear* if the ODE can be written in the form, $a(t)x + b(t)x' + c(t)x'' + \cdots = f(t)$;
- *homogeneous* if $f(t) = 0$ in the expression above.

46.1.2 Existence and Uniqueness of Solutions

Consider the ODE,

$$x'(t) = f(x,t)$$

If both $f(x,t)$ and $\frac{\partial f}{\partial x}$ exist and are continuous for $x \in (a,b)$ and $t \in (c,d)$, then, for any $X \in (a,b)$ and $T \in (c,d)$, the ODE has a unique solution on some open interval containing T (the formal definition of continuity is covered in §34).

Split up more finely, the theorem says that, if $\frac{dx}{dt} = f(x,t)$ and $x(a) = b$, then, a solution exists if $f(x,t)$ is continuous near (a,b), and that the solution is unique if $\frac{\partial f}{\partial x}$ is continuous near (a,b).

Example.

$$x^2 + t^2 \frac{dx}{dt} = 0; \quad x(0) = c; \quad c \neq 0$$

△

At $t = 0$, the equation reduces to $x^2 = 0$, but we have $x(0) \neq 0$, so this differential equation does not have any solutions.

Example.

$$\frac{dx}{dt} = \sqrt{x}; \quad x(0) = 0$$

△

Clearly, the constant function $x(t) = 0$ is a solution, but we also have,

$$x(t) = \begin{cases} 0 & t \leq c \\ \frac{(t-c)^2}{4} & t > c, \end{cases} \qquad c > 0$$

valid for any positive c. So, this differential equation does not have a unique solution.

But it might not be easy to find multiple solutions, so we can check using the theorem above. This differential equation fails the requirements because $x'(0)$ is not well defined, and is hence not continuous.

46.1.3 Fundamental Theorem of Calculus

Suppose $f : [a,b] \to \mathbb{R}$ is continuous. Let $G(x) = \int_a^x f(z)\,dz$. Then, $\frac{d}{dx}G(x) = f(x)$ (i.e., G is an antiderivative of f) and furthermore, $\int_a^b f(x) = F(a) - F(b)$ for any F such that $F'(x) = f(x)$.

46.2 First-Order Differential Equations

46.2.1 Linear

For brevity, we will not notate the independent variable from this point onwards (i.e. whenever we have x or x', we mean $x(t)$ and $\frac{dx}{dt}$, etc.) unless relevant or helpful to the method (i.e., separable equations).

46.2.1.1 Homogeneous with Constant Coefficients

$$x' + ax = 0$$

If you have a coefficient on x', divide everything by that coefficient to get it into the form above before proceeding.

The solution is given by,

$$x = Ae^{-at}, A = x(0)$$

46.2.1.2 Separable

$$\frac{dx}{dt} = f(x)g(t)$$

$$\begin{aligned} \frac{dx}{dt} &= f(x)g(t) \\ \frac{1}{f(x)}\frac{dx}{dt} &= g(t) \\ \int \frac{1}{f(x)}\frac{dx}{dt}\,dt &= \int g(t)\,dt \\ \int \frac{1}{f(x)}\,dx &= \int g(t)\,dt \end{aligned}$$

After evaluating these integrals, simply rearrange for x.

46.2.1.3 Homogeneous with Non-Constant Coefficients

$$x' + f(t)x = 0$$

This is just a separable equation:

$$\begin{aligned} x' &= f(t)x \\ \frac{dx}{dt} &= f(t)x \\ \frac{1}{x}\frac{dx}{dt} &= f(t) \\ \ln x &= \int f(t)\,dt \\ \ln x &= F(t) + C \\ x &= Ae^{F(t)} \end{aligned}$$

46.2.1.4 Non-Homogeneous

$$x' + f(t)x = g(t)$$

First, solve the homogeneous version, $x' + f(t)x = 0$, to get the *complementary function.*

Next, we need to get the *particular integral.*

We need to multiply both sides by some function, $I(t)$, such that we can apply the product rule in reverse on the LHS;

i.e., we want,

$$I(t)x' + I(t)f(t)x = (I(t)x)' \tag{1}$$

but,

$$(I(t)x)' = I(t)x' + I(t)x \tag{2}$$

so by equating (1) and (2), we have $I'(t) = I(t)f(t)$, so $I(t) = e^{\int f(t)\,dt}$.

Now, we have,

$$\begin{aligned} I(t)x' + I(t)f(t)x &= I(t)g(t) \\ (I(t)x)' &= I(t)g(t) \\ I(t)x &= \int I(t)g(t)\,dt \\ x &= \frac{1}{I(t)}\int I(t)g(t)\,dt \end{aligned}$$

Adding this to the complementary function found earlier gives the *general solution.*

$I(t)$ is an *integrating factor* of the differential equation.

46.2.2 Substitutions for Non-Linear ODEs

46.2.2.1 Type I

$$x' = f\left(\frac{x}{t}\right)$$

Let $u = \frac{x}{t}$. Then, $x = tu$

$$\begin{aligned} x &= tu \\ x' &= (tu)' \end{aligned}$$

We use the product rule ("*left dee-right plus right dee-left*") here, remembering that the derivative of t with respect to t is 1.

$$\begin{aligned} x' &= tu' + u \\ f(u) &= tu' + u \\ f(u) - u &= tu' \end{aligned}$$

which is a separable differential equation.

46.2.2.2 Type II

$$x' + f(t)x = g(t)x^n$$

Let $u = x^{1-n}$, so,

$$\begin{aligned} u' &= (1-n)x^{-n}x' \\ u' &= (1-n)x^{-n}\left(g(t)x^n - f(t)x\right) \\ u' &= (1-n)\left(g(t) - f(t)x^{1-n}\right) \\ u' &= (1-n)\left(g(t) - f(t)u\right) \\ u' + (1-n)f(x)u &= (1-n)g(t) \end{aligned}$$

which allows the use of an integrating factor.

46.2.3 Phase Lines

A non-linear ODE will often not have an explicit solution, but we can still analyse them in a couple of ways. We can identify and classify *fixed points* of an autonomous ODE with *phase lines.*

Given an ODE,

$$x' = f(x)$$

draw a graph with x' on the vertical axis, and x on the horizontal axis.

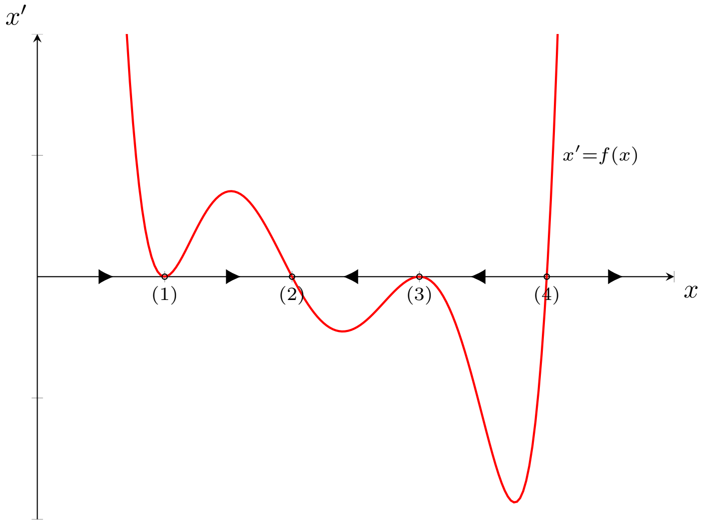
Wherever the graph lies above the line, a particle lying on the x-axis will have positive x', and will therefore move to the right. Similarly, wherever the graph lies below the line, the particle will have a negative x', and will move to the left.

You should indicate these directions with arrows on the x-axis.

If a point where the graph touches the x-axis has arrows pointing inwards, it is *stable.* If it has arrows pointing outwards, it is *unstable.* If arrows point inwards from one direction and outwards from another, it is *structurally unstable.*

The three prior cases are also known collectively as *fixed points*, *stationary points* or *equilibria.*

Example. △



In the diagram above, (1) is structurally unstable, (2) is stable, (3) is structurally unstable (but in a different manner than (1)), and (4) is stable.

We have not solved the ODE, but have still managed to determine some qualitative behaviours of the solutions. Notice that a particle that starts past point (4) will move to the right indefinitely, while a particle that starts anywhere to the left will eventually hit point (1). We call the behaviour of a solution as $t \to \infty$ the *asymptotic behaviour* or sometimes the *large time limit*, if the limit is well defined.

The stability of a fixed point clearly depends on how the line interacts with the x-axis. If the line has positive gradient when crossing the x-axis, the point is unstable, and if negative, stable. If the line touches the x-axis, but does not cross it, then the point is structurally unstable.

Note: having a gradient of zero is not sufficient (although necessary) to determine if a fixed point is structurally unstable. For example, the graph of $x' = x^3$ has zero gradient at $x = 0$, but still crosses the x-axis, causing it to be unstable. You should always draw a diagram.

46.2.4 Euler's Method

Consider the ODE,

$$x' = f(x,t), x(0) = X$$

and suppose we cannot find an analytic solution.

In *Euler's Method*, we find a numerical approximation to the solution.

First, we pick a small time step, h, and assume that x' is approximately constant over the small time step h. With that assumption, we use the Taylor expansion of $x(t + h)$,

$$\begin{aligned} x(t+h) &= x(t) + hx'(t) \\ &= x(t) + hf(x(t),t) \end{aligned}$$

so, the solution to the DE is approximated by the recurrence relation,

$$x(n+1) = x(n) + hf(x(n),nh)$$

Note that we only use the first two terms of the Taylor series, as any further derivatives are 0, since we assume $x'(t)$ is constant.

46.3 Second Order

46.3.1 Homogeneous

For a differential equation of the form

$$ax'' + bx' + cx = 0$$

we form and solve the *characteristic* or *auxiliary* equation,

$$a\lambda^2 + b\lambda + c = 0$$

There are three cases:

- *Two real roots*: If $\lambda = \alpha,\beta$, then $x = Ae^{\alpha t} + Be^{\beta t}$, where A and B are constant coefficients to be found.
- *Repeated real root*: If $\lambda = \alpha$ with multiplicity 2, then $x = (A + Bt)e^{\alpha t}$.
- *Complex roots*: If $\lambda = p \pm iq$, then $x = e^{pt}(A\cos(qt) + B\sin(qt))$

46.3.2 Damping

In the above equation, if,

$b = 0$, then the system is ***un***_damped_;
$b^2 - 4ac < 0$, then the system is ***under***_damped_;
$b^2 - 4ac = 0$, then the system is ***critically*** *damped*;
$b^2 - 4ac > 0$, then the system is ***over***_damped._

An undamped system represents a system without friction, and will oscillate regularly forever. Underdamped systems still oscillate, but a little bit of friction is present, causing the amplitude to decay over time. Critically damped systems generally do not oscillate, simply decaying to zero. Overdamped systems behave similarly, but with a slower decay.

46.3.3 Non-Homogeneous

For an equation of the form

$$ax'' + bx' + cx = f(t)$$

we first solve the homogenous version, $ax'' + bx' + cx = 0$, to get the *complementary function.*

Now, we want a *particular integral* to deal with the non-homogeneous part. We make an ansatz depending on the form of $f(t)$. If $f(t)$ is a polynomial, we set x equal to a general polynomial of the same degree. If $f(t)$ is exponential, we try the same. If $f(t)$ contains a sine, a cosine or both, we try a linear combination of **both**.

i.e., if $f(t) = 3\cos(5t)$, then we try $x = A\cos(5t) + B\sin(5t)$. Note that we keep the 5's intact, and that we use both sines and cosines, despite $f(t)$ only containing cosine.

Common oversight: Furthermore, if the complementary function matches $f(t)$ in any way, we must multiply our ansatz by t to avoid getting a solution we already have.

Find the first and second derivatives of your ansatz, and substitute into the original equation to solve for any unknown constants.

Remember to add the complementary function to your particular integral afterwards to get the general solution.

If you do not want to use the method above (*Undetermined Coefficients*), there is an alternative method: *Variation of Parameters.*

The method of undetermined coefficients only works when $f(t)$ is polynomial, exponential, (hyperbolic) trigonometric, or a linear combination of the previous.

Variation of parameters is a more powerful technique that works on a wider range of functions, but requires a little more work. See §46.6.2.

46.3.4 Resonance

We consider the ODE for a mass/spring system,

$$x'' + cx' + \omega^2 x = F\cos(\Omega t)$$

Where $F\cos(\Omega t)$ is some forcing term.

If the system is underdamped (§46.3.2), this ODE has the solution,

$$x(t) = A\cos(\Omega t - \phi) + Be^{-\frac{ct}{2}}\cos(\alpha t + \delta)$$

for very complicated and mostly irrelevant constants, α, A and ϕ.

But notice how as $t \to \infty$, the second term tends to 0 due to the negative exponential. This second term is the *transient behaviour* term, while the first term is the *steady state solution.*

If there is no forcing and no friction, i.e., $F = 0$ and $c = 0$, $\alpha = \omega$, and the system oscillates as a whole with *natural frequency* $\frac{\omega}{2\pi}$.

If forcing is present, then, as $\Omega \to \omega$, $A \to \infty$. This effect is *resonance.*

46.4 Recurrence Relations

46.4.1 First-Order

46.4.1.1 Homogeneous

Consider the recurrence relation,

$$x_n = ax_{n-1}$$

The solution can be found using *back substitution*:

$$\begin{aligned} x_n &= ax_{n-1} \\ &= a^2 x_{n-2} \\ &= a^3 x_{n-3} \\ &\vdots \\ &= a^n x_0 \end{aligned}$$

If initial conditions aren't given, then $x_n = Aa^n$ will suffice.

46.4.1.2 Non-Homogeneous

$$x_n = ax_{n-1} + f(n)$$

Solve the homogeneous version, $x_n = ax_{n-1}$, to get the complementary solution. Then choose an ansatz using the same procedure as outlined in §46.3.3 and substitute it into the non-homogeneous solution to solve for any unknowns.

Note, if $a = 1$ and $f(n)$ is a polynomial, you need to multiply your ansatz by n. But also, if $a = 1$, it may be easier to do back substitution anyway, so keep that in mind.

46.4.2 Second Order

46.4.2.1 Homogeneous

For a recurrence relation of the form

$$ax_{n+2} + bx_{n+1} + cx_n = 0$$

we solve the characteristic equation,

$$a\lambda^2 + b\lambda + c = 0$$

Again, there are three cases:

- *Two real roots*: If $\lambda = \alpha,\beta$, then $x = A\alpha^n + B\beta^n$, where A and B are constant real coefficients to be found.
- *Repeated real root*: If $\lambda = \alpha$ with multiplicity 2, then $x = (A + Bt)\alpha^n$, where A and B are constant real coefficients to be found.
- *Complex roots*: If $\lambda = p \pm iq$, then convert λ to polar form, $p \pm iq = re^{i\theta}$, and $x = r^n(A\cos(n\theta) + B\sin(n\theta))$. Or, if you hate working with trigonometry, use the same form as for two real roots, and solve for complex A and B.

46.4.2.2 Non-Homogeneous

See §46.4.1.2 and §46.3.3. These are done using the exact same procedure.

The only thing to note is, if you have an exponential in your ansatz, you have to be a little careful: if the base of the exponential is equal to one of the roots of the auxiliary equation, multiply by n, like with ODEs. But if the root is repeated, you need to multiply by n^2.

46.4.3 Other

A *fixed point* of a recurrence relation, $x_n = f(x_{n-1})$ is a value of $x = k$ such that $f(k) = k$. If $|f'(k)| < 1$, then k is a *stable* fixed point. If $|f'(k) > 1|$, then k is an *unstable* fixed point.

46.5 Systems of Linear First-Order ODEs

Much of the theory in this section depends on knowledge from §33. If you have not completed much linear algebra, many of the methods here may seem rather arbitrary and unexplained.

46.5.1 The Jacobian

The *Jacobian matrix* of a function, $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$, denoted $D\mathbf{f}$, is the matrix of partial derivatives,

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \frac{\partial f_1}{\partial x_3} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_2}{\partial x_3} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \frac{\partial f_3}{\partial x_1} & \frac{\partial f_3}{\partial x_2} & \frac{\partial f_3}{\partial x_3} & \cdots & \frac{\partial f_3}{\partial x_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \frac{\partial f_n}{\partial x_3} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

This can be more compactly written as,

$$\begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix}$$

or,

$$\begin{bmatrix} \nabla^\top f_1 \\ \vdots \\ \nabla^\top f_n \end{bmatrix}$$

If you are unfamiliar with the ∇ operator, see §44. You should familiarise yourself well with the Jacobian, as it appears everywhere in calculus. The Jacobian is effectively the higher dimensional version of the derivative – it encodes information about how a many-variable function changes.

The Jacobian matrix, interpreted as a linear transformation, represents how the function transforms locally around a point when evaluated at that point.

Using the Jacobian, we can now define:

46.5.2 Existence and Uniqueness 2: Electric Boogaloo

$$\frac{d}{dt}(\mathbf{x}(t)) = \mathbf{f}(\mathbf{x},t)$$

If $\mathbf{f}(\mathbf{x},t)$ and $D\mathbf{f}(\mathbf{x},t)$ exist and are continuous (§34) for $\mathbf{x} \in U \subseteq \mathbb{R}^n$ and $t \in (a,b)$, then for any $\mathbf{X} \in U$ and $T \in (a,b)$, there exists a unique solution to the equation above on some open interval containing T.

Now, with all the preamble done, we can move onto solving systems of ODEs.

46.5.3 Homogeneous 2×2 Systems with Constant Coefficients

The system of ODEs,

$$x' = ax + by$$
$$y' = cx + dy$$

can be written as a matrix equation,

$$\begin{bmatrix} x \\ y \end{bmatrix}' = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

or somewhat less descriptively as,

$$\mathbf{x}' = \mathbf{A}\mathbf{x}$$

Now, we find the eigenvalues and eigenvectors of the matrix equation (§33.5.2).

There is a method to solve matrix differential equations that avoids finding the eigenvalues and eigenvectors, but requires possibly more difficult calculations. This method, *matrix exponentiation*, is discussed in §33.9.3.2.

While more computationally difficult for small matrices, matrix exponentation generalises to larger systems of differential equations more easily, and can be more efficient for a computer to perform.

46.5.3.1 Distinct Real Eigenvalues

If u,v, $\mathbf{u}$, and $\mathbf{v}$ are distinct eigenvalues and their corresponding eigenvalues of $\mathbf{A}$, then the general solution is given by,

$$\mathbf{x} = Ae^{ut}\mathbf{u} + Be^{vt}\mathbf{v}$$

where A and B are constant coefficients to be determined.

46.5.3.2 Complex Eigenvalues

If $u = p + iq$ is a complex eigenvalue with corresponding eigenvector, $\mathbf{u} = \begin{bmatrix} a+ib \\ c+id \end{bmatrix}$, then we write,

$$\begin{aligned} \mathbf{x} &= Ae^{ut}\mathbf{u} \\ &= e^{(p+iq)t}\begin{bmatrix} a+ib \\ c+id \end{bmatrix} \end{aligned}$$

Using Euler's formula, we can rewrite this as,

$$\begin{aligned} &= e^{pt}(\cos(qt) + i\sin(qt))\begin{bmatrix} a+ib \\ c+id \end{bmatrix} \\ &= e^{pt}\left(\begin{bmatrix} a\cos(qt) + ib\cos(qt) \\ c\cos(qt) + id\cos(qt) \end{bmatrix} + \begin{bmatrix} ia\sin(qt) - b\sin(qt) \\ ic\sin(qt) - d\sin(qt) \end{bmatrix}\right) \\ &= e^{pt}\left(\underbrace{\begin{bmatrix} a\cos(qt) - b\sin(qt) \\ c\cos(qt) - d\sin(qt) \end{bmatrix}}_{\mathbf{v}_1(t)} + i\underbrace{\begin{bmatrix} a\sin(qt) + b\cos(qt) \\ c\sin(qt) + d\cos(qt) \end{bmatrix}}_{\mathbf{v}_2(t)}\right) \end{aligned}$$

so we have found two linearly independent solutions, so we can write the general solution as,

$$\mathbf{x} = e^{pt}\left(A\mathbf{v}_1(t) + B\mathbf{v}_2(t)\right)$$

where A and B are constant coefficients to be determined.

Note that you only have to do this process with one eigenvalue and eigenvector, as the other set will differ only by a minus sign, which eventually gets absorbed into the constant coefficients.

46.5.3.3 Repeated Real Eigenvalues

If λ is a eigenvalue with multiplicity 2, then find a vector, $\mathbf{v_1}$ that satisfies,

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_1 = \mathbf{0}$$

then, a second vector, $\mathbf{v}_2$ that satisfies,

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_2 = \mathbf{v}_1$$

The general solution is then given by,

$$e^{\lambda t}(A\mathbf{v}_1 + B(\mathbf{v}_2 + t\mathbf{v}_1))$$

46.5.4 Diagonalisation & Decoupling

If $\mathbf{Ax} = \mathbf{x}'$, and $\mathbf{A}$ has distinct eigenvalues, swapping to an eigenbasis will let you decouple a system of ODEs by defining a new variable in the eigenbasis.

i.e., Let $\mathbf{x}' = \mathbf{Ax}$, and suppose A has eigenvalues u and v and corresponding eigenvectors $\mathbf{u}$ and $\mathbf{v}$. Let P be the matrix with $\mathbf{u}$ and $\mathbf{v}$ as columns. P is a change of basis matrix from the eigenbasis, Y to the canonical basis, X:

$$\begin{tikzcd}
X \arrow[r, "A"] & X' \\
Y \arrow[u, "P"] \arrow[r, "B"] & Y' \arrow[u, "P"]
\end{tikzcd}$$

We see that $\mathbf{B} = \mathbf{P}^{-1}\mathbf{AP}$. Being in an eigenbasis, $\mathbf{B}$ will be a diagonal matrix with u and v along the diagonals.

Let $\mathbf{W} = \mathbf{P}^{-1}\mathbf{x}$, so,

$$\begin{aligned}
\mathbf{W}' &= \mathbf{P}^{-1}\mathbf{x}' \\
&= \mathbf{P}^{-1}\mathbf{Ax} \\
&= \mathbf{P}^{-1}\mathbf{APW} \\
&= \mathbf{P}^{-1}\mathbf{BW} \\
\begin{bmatrix} w' \\ z' \end{bmatrix} &= \begin{bmatrix} u & 0 \\ 0 & v \end{bmatrix} \begin{bmatrix} w \\ z \end{bmatrix}
\end{aligned}$$

So $w' = uw$ and $z' = vz$, so $w = Ce^{ut}$ and $z = De^{vt}$, where C and D are constant coefficients to be found.

Decoupling can also be done without transforming into an eigenbasis by defining new variables in the right way.

Example. Transform the third-order homogeneous differential equation,

$$\frac{d^3x}{dt^3} - 3\frac{dx}{dt} - 2x = 0$$

into a system of three first-order differential equations. △

Let $x = x$, $y = x'$ and $z = x''$.

$$\begin{aligned}
\begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} &= \begin{bmatrix} x' \\ x'' \\ x''' \end{bmatrix} \\
\frac{dx}{dt} &= x' \\
&= y \\
\frac{dy}{dt} &= x'' \\
&= z \\
\frac{dx}{dt} &= x'''
\end{aligned}$$

$$\begin{aligned}
\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} &= \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} \\
x''' - 3x' - 2x = 0 & \\
x''' &= 3x' + 2x \\
x''' &= 3y + 2x \\
\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 2 & 3 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} &= \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix}
\end{aligned}$$

46.5.5 Phase Portraits

Find all eigenvalues and eigenvectors of the system.

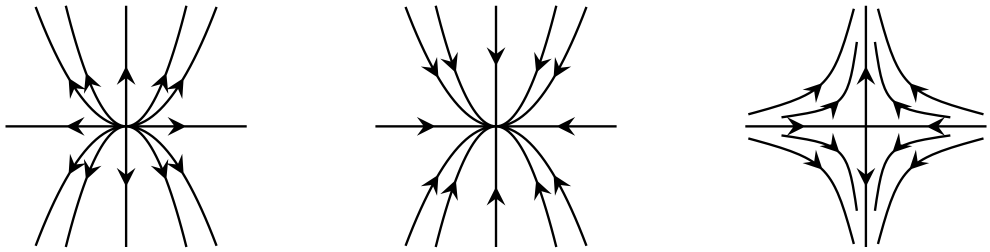
The following sections may be easier to remember if you recall the geometric interpretation of eigenvalues – the real part represents the local scaling, while the imaginary part represents the local rotation.

46.5.5.1 Distinct Real Eigenvalues

Draw the span of the eigenvectors, with arrows pointing outwards from the origin if the eigenvalue is positive, and inwards if negative.

If your eigenvalues are,

- both positive,
  - If you have eigenvalues, say, 3 and 2, then $e^{3t} \ggg e^{2t}$ as $t \to \infty$, so your trajectories should tend towards being parallel to the eigenvector with eigenvalue 3.
  - This is an *unstable node.*
- both negative,
  - With similar reasoning, your trajectories should tend towards being parallel to the eigenvector with the larger absolute value of eigenvalue.
  - This is a *stable node.*
- one positive, one negative,
  - One line should point inwards, and one points outwards.
  - Draw hyperbolae-esque trajectories between the lines as expected.
  - This is a *saddle point.*



$0 < \lambda_1 < \lambda_2$
Unstable Node

$\lambda_2 < \lambda_1 < 0$
Stable Node

$\lambda_1 < 0 < \lambda_2$
Saddle Point

All three figures have $\hat{\mathbf{\imath}}$ and $\hat{\mathbf{\jmath}}$ as eigenvectors.

46.5.5.2 Complex Eigenvalues

Say the system

$$\mathbf{Ax} = \mathbf{x}'$$

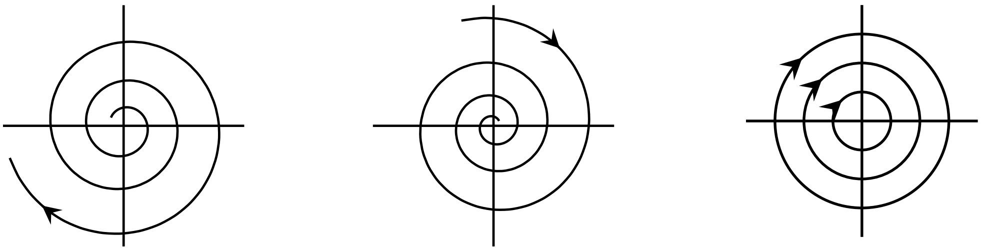
has matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

and you have eigenvalues $p + qi$, with $q \neq 0$. Then, if,

- $p > 0$, the trajectories will spiral outwards, an *unstable spiral* or *spiral source* ;
- $p < 0$, the trajectories will spiral inwards, a *stable spiral* or *spiral sink* ;
- $p = 0$, i.e., the eigenvalues are purely imaginary, the trajecctories will form circles or ellipses around the origin, a *centre.*

In all three cases, the motion is clockwise if $b - c > 0$, and anticlockwise if $b - c < 0$.



$\Re(\lambda) > 0$,$\Im(\lambda) \neq 0$
Unstable Spiral

$\Re(\lambda) < 0$,$\Im(\lambda) \neq 0$
Stable Spiral

$\Re(\lambda) = 0$,$\Im(\lambda) \neq 0$
Centre

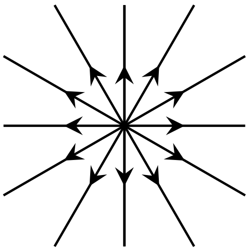
All three figures have $b - c < 0$.

46.5.5.3 Repeated Real Eigenvalues

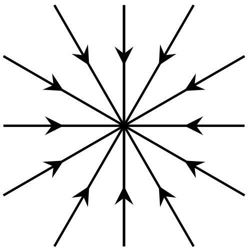
If the matrix is a multiple of the identity, then trajectories just point outwards/inwards evenly. This is a *star*, pointing outwards/is stable if the eigenvector is positive and pointing inwards/is unstable if negative. (Geometrically, if it's a multiple of the identity, then it's locally just a scaling transformation, so everything just moves directing in or out from the fixed point).

If the matrix is not a multiple of the identity, then sample some random points to get an idea of what the trajectories should look like. Easy points to sample are (1,0), (0,1), $(-1{,}0)$ and $(0,-1)$. You can get either an *improper node*, unstable if eigenvalue is positive, stable if negative, or a line of *(un)stable fixed points*. An unstable improper node can also be called a *degenerate sink*, and a stable improper node a *degenerate source.*

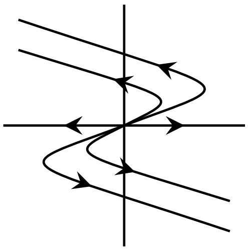
The trajectories in an improper node are parallel to the span of the eigenvector near the origin, then completely reverse in direction. For lines of (un)stable fixed points, the parallel sets of trajectories flow into or out from the line spanned by the eigenvector.



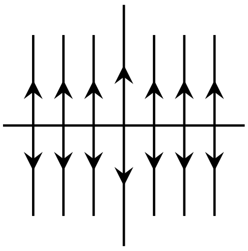
$\Re(\lambda) > 0$,$\Im(\lambda) \neq 0$
Unstable Star



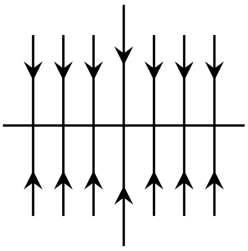
$\Re(\lambda) < 0$,$\Im(\lambda) \neq 0$
Stable Star



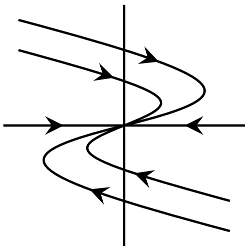
$\Re(\lambda) > 0$,$\Im(\lambda) \neq 0$
Unstable Improper Node



$\Re(\lambda) > 0$,$\Im(\lambda) \neq 0$
Line of Unstable Fixed Points



$\Re(\lambda) < 0$,$\Im(\lambda) \neq 0$
Line of Stable Fixed Points



$\Re(\lambda) < 0$,$\Im(\lambda) \neq 0$
Stable Improper Node

If you want a more general way to classify all these points, you can compute the trace and determinant of the matrix **A**.

Let $\Delta = (\operatorname{Tr} \mathbf{A})^2 - 4 \det \mathbf{A}$.

$\Delta, \operatorname{Tr} \mathbf{A}, \det \mathbf{A} = 0$, then the matrix is the zero matrix and every point is locally a fixed point;

$\det \mathbf{A} < 0$ – saddle;

$\Delta > 0, \operatorname{Tr} \mathbf{A} > 0, \det \mathbf{A} = 0$ – line of unstable fixed points;

$\Delta > 0, \operatorname{Tr} \mathbf{A} < 0, \det \mathbf{A} = 0$ – line of stable fixed points;

$\Delta > 0, \operatorname{Tr} \mathbf{A} > 0, \det \mathbf{A} > 0$ – unstable node;

$\Delta > 0, \operatorname{Tr} \mathbf{A} < 0, \det \mathbf{A} > 0$ – stable node;

$\Delta = 0, \operatorname{Tr} \mathbf{A} > 0, \det \mathbf{A} > 0$ – unstable improper node;

$\Delta = 0, \operatorname{Tr} \mathbf{A} < 0, \det \mathbf{A} > 0$ – stable improper node;

$\Delta < 0, \operatorname{Tr} \mathbf{A} > 0, \det \mathbf{A} > 0$ – unstable spiral;

$\Delta < 0, \operatorname{Tr} \mathbf{A} < 0, \det \mathbf{A} > 0$ – stable spiral;

$\mathbf{A} = k\mathbf{I}$,$k > 0$ – unstable star;

$\mathbf{A} = k\mathbf{I}$,$k < 0$ – stable star;

$\Delta < 0, \operatorname{Tr} \mathbf{A} = 0, \det \mathbf{A} > 0$ – centre.

46.5.6 Local Linearisation near Fixed Points

If we have the system,

$$x' = f(x,y)$$

$$y' = g(x,y)$$

where f and/or g are non-linear, and you are asked to draw a phase diagram of fixed points of this system, evaluate the Jacobian (§46.5.1) at each fixed point and use it as your matrix for determining eigenvalues/eigenvectors.

Example. Consider the system,

$$\begin{aligned} x' &= y \\ y' &= -x + y - x^2 y \end{aligned}$$

(This is the *Van der Pol oscillator*). Find and classify a fixed point of this system.

Clearly, (0,0) is a fixed point of this system. But what does the phase diagram look like?

First, compute the Jacobian:

$$D\mathbf{X} = \begin{bmatrix} 0 & 1 \\ -1 - 2xy & 1 - x^2 \end{bmatrix}$$

and evaluate it at our fixed point,

$$D\mathbf{X} = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}$$

which has eigenvalues $\frac{1}{2} \pm \frac{\sqrt{3}}{2}$, indicating that the phase portrait around the fixed point is locally an unstable spiral. △

46.6 Additional Techniques

This section will cover further techniques for integration that you may find faster and/or easier to perform. We recommend at least learning tabular integration by parts, as it streamlines the commonly taught formula to an extreme degree, particularly for repeated applications of integration by parts.

46.6.1 Tabular Integration by Parts

Say we want to integrate this function,

$$\int a(x)b(x)\,dx$$

Being a product of two functions, we use integration by parts.

Draw out a table, with D above the first column and I above the first, then put a column of alternating plusses and minuses, besides the first, starting with a plus. You will get a feel for how many rows is needed as you get more used to using this method, but for now, we will draw four.

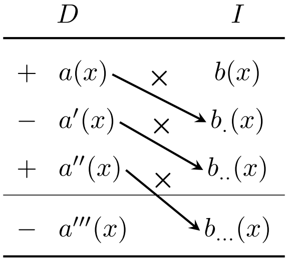
	D	I
$+$		
$-$		
$+$		
$-$		

Now, look at the integral, and decide which function is easier to differentiate. Or more usually, which function you don't want to integrate. Suppose we don't want to integrate $a(x)$, so we differentiate $a(x)$ and integrate $b(x)$.

Put $a(x)$ under D, and $b(x)$ under I, and differentiate and integrate them repeatedly, putting the result in the next row each time. For ease of reading, let $b_{.}(x)$ indicate the first integral of $b(x)$, $b_{..}(x)$ the second, and so on.

	D	I
$+$	$a(x)$	$b(x)$
$-$	$a'(x)$	$b_{.}(x)$
$+$	$a''(x)$	$b_{..}(x)$
$-$	$a'''(x)$	$b_{...}(x)$

When we decide to stop (we have three additional rows here), multiply diagonal elements, keeping the signs attached. Then, multiply the final row horizontally and throw it into an integral;



$$\int a(x)b(x)\,dx = [+a(x)b_{.}(x)] + [-a'(x)b_{..}(x)] + [a''(x)b_{...}(x)] + \int [-a'''(x)b_{...}(x)]\,dx$$

But when do we know when to stop?

There are three main stops:

- There is a 0 in the D column.
- You can integrate a row.
- A row appears more than once.

In the first case, when you multiply the last row together, the final integral just disappears. In the second case, if you can integrate a row, just stop the process and do the integral. In the third case, if a row appears more than once, that means you can rewrite the original integral in terms of itself, plus some extra stuff at the front, which you can rearrange for.

Example. Evaluate,

$$\int x^3 \sin(4x)\,dx$$

△

It's almost always ideal to differentiate the polynomial, as we know we can eventually get it to 0. The sine function is fine to integrate as well, so let's do that.

	D		I
$+$	x^3	$\searrow$ ×	$\sin(4x)$
$-$	$3x^2$	$\searrow$ ×	$-\frac{1}{4}\cos(4x)$
$+$	$6x$	$\searrow$ ×	$-\frac{1}{16}\sin(4x)$
$-$	6	$\searrow$ ×	$\frac{1}{64}\cos(4x)$
$+$	0		$\frac{1}{256}\sin(4x)$

$$\int x^3 \sin(4x)\,dx = -\frac{1}{4}x^3\cos(4x) + \frac{3}{16}x^2\sin(4x) + \frac{3}{32}x\cos(4x) - \frac{3}{128}\sin(4x)$$

Example. Evaluate,

$$\int x^3 \ln x\,dx$$

△

We like to differentiate polynomials, but integrating $\ln x$ requires integration by parts in the first place, which we would like to avoid, especially if we are repeatedly integrating it. So, we differentiate $\ln x$ and integrate x^3.

	D	I
$+$	$\ln x$	x^3
$-$	$\frac{1}{x}$	$\frac{1}{4}x^4$

If we look at the final row, we can already integrate its product, so we stop.

	D		I
$+$	$\ln x$	$\searrow$ ×	x^3
$-$	$\frac{1}{x}$		$\frac{1}{4}x^4$

$$\begin{aligned}\int x^3 \ln x\,dx &= \frac{1}{4}x^4\ln x - \frac{1}{4}\int x^3\,dx \\ &= \frac{1}{4}x^4\ln x - \frac{1}{16}x^4\end{aligned}$$

Example. Evaluate,

$$\int e^x \sin x\,dx$$

△

e^x and $\sin x$ are both easy to integrate and differentiate, so it doesn't really matter which way around we put them. Let's differentiate e^x and integrate $\sin x$.

	D		I
$+$	e^x	↘ ×	$\sin(x)$
$-$	e^x	↘ ×	$-\cos(x)$
$+$	e^x		$-\sin(x)$

We see that the final row is a copy of the first one (ignoring signs), so we can rewrite the integral as,

$$\begin{aligned}\int e^x \sin x\,dx &= -e^x\cos x + e^x \sin x - \int e^x \sin x\,dx\\ 2\int e^x \sin x\,dx &= -e^x\cos x + e^x \sin x\\ \int e^x \sin x\,dx &= \frac{1}{2}e^x \sin x - \frac{1}{2}e^x\cos x\end{aligned}$$

46.6.2 Variation of Parameters

Variation of parameters is a general method to solve non-homogeneous linear ODEs, though it can also be extended to solve PDEs as well.

Here, we will only consider second order ODEs,

$$x'' + bx' + cx = f(t)$$

(we divide through by the constant coefficient of x'' to simplify this method).

Consider the solution to the homogeneous case, which depends on the solutions to the auxiliary equation,

$$\begin{aligned}x &= Ae^{\alpha t} + Be^{\beta t}\\ x &= (A+Bt)e^{\alpha t}\\ x &= e^{pt}(A\cos(qt) + B\sin(qt))\end{aligned}$$

Notice how each solution can be split into two linearly independent parts (see §33.1.2 if you are unfamiliar with linear independence), x_1 and x_2, where,

$$\begin{aligned}x_1 &= Ae^{\alpha t}, & x_2 &= Be^{\beta t}\\ x_1 &= Ae^{\alpha t}, & x_2 &= Bte^{\beta t}\\ x_1 &= Ae^{pt}\cos(qt), & x_2 &= Be^{pt}\sin(qt)\end{aligned}$$

The functions x_1 and x_2 are the *fundamental solutions* of the equation.

We define the *Wronskian matrix* as,

$$\begin{bmatrix} x_1 & x_2 \\ x_1' & x_2' \end{bmatrix}$$

From linear algebra, we know that the *Wronskian determinant*, W, of this matrix cannot be 0. We use the Wronskian determinant to find the particular integral of the equation.

$$x = -x_1 \int \frac{x_2 f}{W}\,dx + x_2 \int \frac{x_1 f}{W}\,dx$$

Remember to add the complementary function to your particular integral afterwards to get the general solution.

46.6.3 Weierstrass Substitution

The *Weierstrass substitution* is a change of variable that transforms rational functions of trigonometric functions into an ordinary rational function of a parameter, t.

Letting $t = \tan \frac{x}{2}$, we can transform the integral,

$$\int f(\sin x, \cos x)\, dx = \int f\left(\frac{2t}{1+t^2}, \frac{1-t^2}{1+t^2}\right) \frac{2}{1+t^2}\, dt$$

Geometrically, as x varies, the point $(\cos x, \sin x)$ travels across the unit circle at unit speed. In other words, it is a *unit speed parametrisation* (see §44). The Weierstrass substitution is an alternative parametrisation of the unit circle such that the point $\left(\frac{1-t^2}{1+t^2}, \frac{2t}{1+t^2}\right)$ travels around the unit circle only once as t varies from $-\infty$ to ∞, starting and ending at $(-1{,}0)$. If you are familiar with projective geometry, this substitution can be viewed as the sterographic projection of the unit circle onto the y-axis from the point $(-1{,}0)$. This view can help you rederive various formulae on the fly, if required.

46.6.4 Reduction Formulae

A *reduction formula* allows you to write a recurrence relation for an integral in terms of related integrals with hopefully smaller exponents.

We do this by splitting up the exponent, substituting if needed, then integrating by parts.

Example.

$$\int \sin^n x\, dx$$

We wish to find a reduction formula for this integral. Start by setting,

$$\begin{aligned}
I_n &= \int \sin^n x\, dx \\
&= \int \sin^{n-1} x \sin x\, dx \\
&= -\sin^{n-1} x \cos x + \int (n-1) \sin^{n-2} x \cos^2 x\, dx \\
&= -\sin^{n-1} x \cos x + (n-1) \int \sin^{n-2} x(1 - \sin^2 x)\, dx \\
&= -\sin^{n-1} x \cos x + (n-1) \int \sin^{n-2} x\, dx - (n-1) \int \sin^2 x\, dx \\
&= -\sin^{n-1} x \cos x + (n-1) I_{n-2} - (n-1) I_n \\
I_n + (n-1)I_n &= -\sin^{n-1} x \cos x + (n-1) I_{n-2} \\
I_n &= -\frac{1}{n} \sin^{n-1} x \cos x + \frac{n-1}{n} I_{n-2}
\end{aligned}$$

So now, if we're given, for example, $\int \sin^{100} x\, dx$, we can repeatedly apply the reduction formula until the power is low enough for us to evaluate the integral by hand. △

46.6.5 Euler Substitution

If $f(a{,}b)$ is a rational function, then

$$\int f(x, \sqrt{ax^2 + bx + c})\, dx$$

can be changed into the integral of a rational function using *Euler substitutions.*

If $a > 0$, solve $\sqrt{ax^2+bx+c} = \pm x\sqrt{a} + t$ for x (the positive or negative sign can be chosen at will, depending on which is easier). The result will be a rational expression, that also allows us to write dx as a rational expression of t when we perform the substitution.

If $c > 0$, solve $\sqrt{ax^2+bc+c} = xt \pm \sqrt{c}$ for x, and use the result as your substitution. Again, the positive and negative sign can be chosen at will.

If ax^2+bx+c has real roots, α,β, then we solve $\sqrt{a(x-\alpha)(x-\beta)} = (x-\alpha)t$ for x, which will again result in a rational expression.

46.6.6 Laplace Transformations

The *Laplace transform* is an integral transform that converts a real-valued function (often, t) into a complex-valued function (often of a complex variable, s). This transform is useful because linear differential equations transform into simple algebraic equations.

The Laplace transform of a function, $f(t)$, is given by

$$\mathscr{L}(f(t)) = \int_0^\infty e^{-st} f(t)\, dt$$

While this might look complicated to calculate by hand, in practice, you just memorise the transforms of common functions and combine them from there. A short table of such transforms is included below. $\mathscr{L}(f(t))$ is also often written as $F(s)$.

$f(t)$	$\mathscr{L}(f(t))$
c	$\frac{c}{s}$
t	$\frac{1}{s^2}$
t^n	$\frac{n!}{s^{n+1}}$
$t^n e^{-\alpha t}$	$\frac{n!}{(s+\alpha)^{n+1}}$
$e^{-\alpha t}$	$\frac{1}{s+\alpha}$
$1-e^{-\alpha t}$	$\frac{\alpha}{s(s+\alpha)}$
$\sin \omega t$	$\frac{\omega}{s^2+\omega^2}$
$\cos \omega t$	$\frac{s}{s^2+\omega^2}$

In general, multiplying a function by $e^{-\alpha t}$ shifts the s along by α in the transform, i.e., $\mathscr{L}(e^{-\alpha t} f(t)) = F(s+\alpha)$.

If you intend on using Laplace transforms, you should commit this table, and more, to memory, as you will also need to be able to recognise them quickly in order to find the inverse Laplace transform of some given $F(s)$.

Example. Given,

$$F(s) = \frac{s+3}{s^2+6s+13}$$

what is $\mathscr{L}^{-1}(F(s))$?

Completing the square on the denominator, we have $\frac{s+3}{(s+3)^2+4}$, which matches the form for cosine. But the s is shifted along by 3, so we have $f(t) = e^{-3t}\cos 2t$. △

An important property of the Laplace transform, is that it is a linear operator (see §33). We should also look at the effect of taking the Laplace transform of a derivative:

$$\mathscr{L}(f'(t)) = \int_0^\infty e^{-st} f'(t)\, dt$$

$$\begin{aligned}
&= e^{-st}f(t)\Big|_0^\infty + \int_0^\infty se^{-st}f(t)\,dt \\
&= e^{-st}f(t)\Big|_0^\infty + s\int_0^\infty e^{-st}f(t)\,dt \\
&= e^{-st}f(t)\Big|_0^\infty + s\mathscr{L}f(t) \\
&= [0] - [f(0)] + s\mathscr{L}f(t) \\
&= s\mathscr{L}(f(t)) - f(0)
\end{aligned}$$

so we can rewrite the Laplace transform of a derivative as the Laplace transform of the original function, plus an initial condition. Similarly, we have,

$$\begin{aligned}
\mathscr{L}(f''(t)) &= s\mathscr{L}(f'(t)) - f'(0) \\
&= s^2\mathscr{L}(f(t)) - sf(0) - f'(0)
\end{aligned}$$

and this pattern continues for higher derivatives.

Now, let's use the Laplace transform to solve an initial value problem.

Example.

$$x'' + 5x' + 6x = 0, x(0) = 2, x'(0) = 3$$

$$\begin{aligned}
x'' + 5x' + 6x &= 0 \\
\mathscr{L}(x'' + 5x' + 6x) &= \mathscr{L}(0)
\end{aligned}$$

Recall that the Laplace transform is linear, and so,

$$\begin{aligned}
\mathscr{L}(x'') + 5\mathscr{L}(x') + 6\mathscr{L}(x) &= \mathscr{L}(0) \\
(s^2\mathscr{L}(x) - sx(0) - x'(0)) + 5(s\mathscr{L}(x) - x(0)) + 6\mathscr{L}(x) &= 0 \\
(s^2 + 5s + 6)\mathscr{L}(x) &= (s+5)x(0) + x'(0)
\end{aligned}$$

Use our initial conditions,

$$\begin{aligned}
(s^2 + 5s + 6)\mathscr{L}(x) &= 2s + 13 \\
\mathscr{L}(x) &= \frac{2s+13}{s^2+5s+6} \\
\mathscr{L}(x) &= \frac{2s+13}{(s+2)(s+3)}
\end{aligned}$$

Performing partial fraction decomposition,

$$\begin{aligned}
\mathscr{L}(x) &= 9\left(\frac{1}{s+2}\right) - 7\left(\frac{1}{s+3}\right) \\
\mathscr{L}(x) &= 9\mathscr{L}(e^{-2t}) - 7\mathscr{L}(e^{-3t}) \\
\mathscr{L}(x) &= \mathscr{L}(9e^{-2t} - 7e^{-3t}) \\
x &= 9e^{-2t} - 7e^{-3t}
\end{aligned}$$

△

It is important to note that you generally cannot find the Laplace transform of the product or composition of two functions. However, due to linearity, as long as your function can be written as the *sum* of known functions, you can work out its Laplace transform.

If you take probability or any kind of electrical engineering or signal/image processing, you may be familiar with *convolution*. You'll be happy to know that the Laplace transform of a convolution is simply the product of the Laplace transforms. That is, $\mathscr{L}((f * g)(t)) = \mathscr{L}(f(t)) \cdot \mathscr{L}(g(t)) = F(s) \cdot G(s)$.

For instance, in image processing, convolving an image with a kernel is required for a multitude of operations, including blurring, sharpening and edge detection. But convolving the naïve way can be an extremely slow process, especially for large kernels. Many modern convolution functions take an integral transform (often Fourier, rather than Laplace), allowing convolution to be applied as a multiplication, which is much faster to compute, before transforming back to the original image.

46.6.7 Leibniz Integration Rule

$$\frac{d}{dx}\left(\int_a^b f(x,t)\,dt\right) = \int_a^b \frac{\partial}{\partial x} f(x,t)\,dt$$

There is a longer form for non-constant bounds of integration, but we will focus on the special case of constant bounds.

This theorem allows us to integrate functions we otherwise wouldn't be able to.

Example. Evaluate

$$\int_0^\infty \frac{\sin t}{t}\,dt$$

(The integrand is also known as the (unnormalised) *sinc function*, a function occuring often in signal processing contexts. This particular definite integral is the *Dirichlet integral*, and cannot be evaluated using standard methods.)

We begin by defining a function,

$$f(s) = \int_0^\infty e^{-st}\frac{\sin t}{t}\,dt$$

(The similarity with the earlier Laplace transform is not a coincidence. There is a much faster way of doing this using the Laplace transform, combined with *Abel's theorem*, but that method will not be covered here, as it is beyond the scope of this document.)

We note that $f(0)$ is equal to the desired integral.

$$\begin{aligned}\frac{df}{ds} &= \frac{d}{ds}\int_0^\infty e^{-st}\frac{\sin t}{t}\,dt \\ &= \int_0^\infty \frac{\partial}{\partial s}e^{-st}\frac{\sin t}{t}\,dt \\ &= -\int_0^\infty e^{-st}\sin t\,dt\end{aligned}$$

You can alternatively use the complex definition of sine to perform this integral as an exercise.

$$\begin{aligned}&= -\left.\frac{e^{-st}(\cos t + s\sin t)}{s^2+1}\right|_{t=0}^{t=\infty} \\ &= -\frac{1}{s^2+1}\end{aligned}$$

Now, we integrate both sides with respect to s.

$$\begin{aligned}f(s) &= -\int \frac{1}{s^2+1}\,ds \\ &= -\arctan s + C\end{aligned}$$

$$-\arctan s + C = \int_0^\infty e^{-st}\frac{\sin t}{t}\,dt$$

Here, we can try some values of s to get some information about C. $s = 0$ doesn't work, because we just get the original problem back. Let's see what happens as $s \to \infty$.

$$\begin{aligned}
\lim_{s\to\infty} -\arctan s + C &= \lim_{s\to\infty}\int_0^\infty e^{-st}\frac{\sin t}{t}\,dt \\
\lim_{s\to\infty} -\arctan s + C &= \lim_{s\to\infty}\int_0^\infty \frac{\sin t}{te^{st}}\,dt \\
-\frac{\pi}{2} + C &= \int_0^\infty 0\,dt \\
-\frac{\pi}{2} + C &= 0 \\
C &= \frac{\pi}{2} \\
f(s) &= \frac{\pi}{2} - \arctan s \\
f(0) &= \frac{\pi}{2} - 0 \\
\int_0^\infty e^{-0t}\frac{\sin t}{t}\,dt &= \frac{\pi}{2} \\
\int_0^\infty \frac{\sin t}{t}\,dt &= \frac{\pi}{2}
\end{aligned}$$

△

46.6.8 Non-Elementary Integrals

If you somehow end up with one of these when constructing a differential equation for a question, you've probably done something wrong earlier.

The following is a non-exhaustive list of integrals that you will not be able to evaluate.

$$\int \sqrt{1+x^n}\,dx, \quad n\in\mathbb{N}, n\geq 3$$

$$\int \sqrt{1-x^n}\,dx, \quad n\in\mathbb{N}, n\geq 3$$

$$\int x^x\,dx$$

$$\int x^{-x}\,dx$$

$$\int \frac{1}{\ln x}\,dx$$

$$\int \frac{x^n}{e^x-1}\,dx \quad n\in\mathbb{N}$$

$$\int \sin(\sin x)\,dx$$

$$\int \arcsin(\arcsin x)\,dx$$

$$\int \sin(x^2)\,dx$$

$$\int \cos(x^2)\,dx$$

$$\int \frac{\sin x}{x}\,dx$$

$$\int \ln(\ln x)\,dx$$

$$\int e^{e^x}\,dx$$

$$\int e^{x^2}\,dx$$

$$\int e^{-x^2}\,dx$$

$$\int \frac{e^x}{x}\,dx$$

$$\int \frac{e^{-x}}{x}\,dx$$

$$\int x^{c-1}e^{-x}\,dx, \quad c\notin\mathbb{N}$$

While you don't have to memorise all of these, it's good to be able to recognise when you have an integral you can't evaluate, so you can go back and check your previous working, rather than wasting time on the integral.

Chapter 47

Probability

> *"The essential feature of statistics is a prudent and systematic ignoring of details."*
>
> — Erwin Schrödinger

It should be noted that there is a distinction between probability and statistics: probability is concerned with describing how likely events are to occur, or how likely a proposition is to be true. Statistics, on the other hand, is the branch of mathematics that concerns the analysis and interpretation of data.

This chapter is on probability. Here, we study the basics of probability, with a main focus on discrete probability distributions.

But first, there are some prerequisites that make discrete probability easier to work with.

47.1 Sample Spaces & Probabilities

A *probability space* consists of three elements:

- A *sample space*, Ω, the set of all possible outcomes;
- An *event space*, a family of sets $\mathcal{F} \subseteq \mathcal{P}(\Omega)$, with each set representing an *event*;
- A *probability measure*, $\mathbb{P} : \mathcal{F} \to [0,1]$, such that,
  - $\mathbb{P}(\Omega) = 1$;
  - $\mathbb{P}(\varnothing) = 0$;
  - If $\{A_i\}_{i=1}^{\infty} \subseteq \mathcal{F}$ are countably many disjoint events, then $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

and a probability space is *discrete* if Ω is at most countably infinite. An event is *elementary* if it is a set of size 1. $\mathcal{P}(\Omega)$ is also sometimes written as 2^{Ω}.

A set of events are *mutually exclusive* if they are pairwise disjoint – every pair of events is disjoint. Given the definition of a probability measure, this is equivalent to the intersection of any pair of events having probability 0. The empty set is disjoint with every set, including itself.

The complement of an event A, $\Omega \setminus A$, is also written as A' or A^c if the sample space is clear. Note that an event and its complement are mutually exclusive and partition the sample space.

Example. A D6 dice is rolled and a coin is thrown in an experiment. The sample space, Ω, is then $\{1,2,3,4,5,6\} \times \{H,T\}$ where $\times$ is the Cartesian product. $|\Omega| = 12$.

The event space is $\mathcal{P}(\Omega)$, and contains $2^{12} = 4096$ possible events. For example, $\{(1,H),(2,H),(3,H)\} = \{1,2,3\} \times \{H\}$ is the event of the die rolling a number less than or equal to 4 *and* the coin landing on heads. △

47.1.1 Algebra of Sets

The algebra of sets and boolean/logic statements are isomorphic algebraic structures.

You can transform equations about sets into boolean equations or logic statements by swapping,

$$\begin{array}{ccccc} A \cap B & \Leftrightarrow & a \wedge b & \Leftrightarrow & A\,\text{and}\,B \\ A \cup B & \Leftrightarrow & a \vee b & \Leftrightarrow & A\,\text{or}\,B \\ A^c & \Leftrightarrow & \neg a & \Leftrightarrow & \text{not}\,A \\ \varnothing & \Leftrightarrow & \bot & \Leftrightarrow & 0 \\ U & \Leftrightarrow & \top & \Leftrightarrow & 1 \end{array}$$

which might be helpful if you can still remember the first chapter on logic.

For those interested in abstract algebra, these are all complemented distributive lattice structures.

The binary operations of set union, $\cup$, and intersection, $\cap$ are in many ways analogous to the binary operations of addition and multiplication.

- Commutativity;
  - $A \cup B = B \cup A$
  - $A \cap B = B \cap A$
- Associativity;
  - $(A \cup B) \cup C = A \cup (B \cup C)$
  - $(A \cap B) \cap C = A \cap (B \cap C)$
- Distributive property;
  - $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
  - $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

However, unlike addition and multiplication, union and intersection distribute in both directions.

Two additional properties involve the set containing nothing, the *empty set*, $\varnothing$; and the set containing everything of interest, the *universe set*, U.

- Identity;
  - $A \cup \varnothing = A$
  - $A \cap U = A$
- Complement;
  - $A \cup A^C = U$
  - $A \cap A^C = \varnothing$

so $\varnothing$ and U are the identity elements for union and intersection, respectively. In probability, the universe set is often Ω.

- Idempotency;
  - $A \cup A = A$

  - $A \cap A = A$
- Domination;
  - $A \cup U = U$
  - $A \cap \varnothing = \varnothing$
- Absorption;
  - $A \cup (A \cap B) = A$
  - $A \cap (A \cup B) = A$
- De Morgan's Laws;
  - $(A \cup B)^C = A^C \cap B^C$
  - $(A \cap B)^C = A^C \cup B^C$
- Involution and Complement Laws;
  - $\varnothing^C = U$
  - $U^C = \varnothing$
  - $(A^C)^C = A$

You might notice that all the identities above are given in pairs that can be transformed into each other by interchanging $\cap$ and $\cup$, and $\varnothing$ and U.

These are examples of an extremely powerful property of Boolean algebras – the *principle of duality*, which asserts that the dual of a true statement obtained by interchanging unions/intersections, universes/empty sets and reversing inclusions (for computer scientists, this is the same as reversing a Hasse diagram to get another poset) is also true (note that the involution law is self-dual).

Duality is a concept with uses in a much broader range of applications, particularly in order and category theory.

47.1.2 Inclusion-Exclusion Principle

Theorem (Binary Inclusion-Exclusion). *Let A and B be sets. Then,*

$$|A \cup B| = |A| + |B| - |A \cap B|$$

Proof.

$$\begin{aligned} |A \cup B| &= |A \cup (B \setminus A)| \\ &= |A| + |B \setminus A| \qquad (1) \\ |B| &= |(B \setminus A) \cup (A \cap B)| \\ &= |B \setminus A| + |A \cap B| \qquad (2) \end{aligned}$$

Combining (1) and (2) gives the result. ■

Theorem 47.1.1. *Let A, B and C be sets. Then,*

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |B \cap C| - |C \cap A| + |A \cap B \cap C|$$

In general, to find the cardinality of the union of n sets, we include the cardinality of the sets, exclude the cardinalities of the pairwise intersections, include the cardinalities of the 3-wise intersections, exclude 4-wise, and continue up to n.

Symbolically,

Theorem (Inclusion-Exclusion). *For any collection of sets $(A_i)_{i=1}^n$,*

$$\left|\bigcup_{i=1}^n A_i\right| = \sum_{\varnothing\neq J\subseteq\{1,\ldots,n\}} (-1)^{|J|+1}\left|\bigcap_{j\in J} A_j\right|$$

Proof. Combinatorial proof given in §7.2.2. ■

If all sets are exchanged for events and cardinalities replaced with probability measures, all of the above equations still hold, i.e., $|A\cup B| = |A| + |B| - |A\cap B| \Leftrightarrow \mathbb{P}(A\cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A\cap B)$.

Example. Each square in a 3×3 grid is either shaded or unshaded with equal probability. What is the probability that a 2×2 square is shaded in?

Denote a square of dimensions $n\times n$ by S_n. There are $2^9 = 512$ possible ways to colour the entire S_3 square.

When one S_2 square is shaded in, there are $2^5 = 32$ ways to shade in the remaining 5 squares, and 4 possible positions to place S_2 within S_3, so there are $32\times 4 = 128$ possible ways to shade in S_3 with one S_2 shaded.

If there are two S_2 shaded, positioned in opposite corners, there are $2^2 = 4$ ways to shade in the remaining 2 squares, and 2 possible ways to arrange the two S_2 squares to be in opposite corners. If the S_2 squares are adjacent, there are 2^3 ways to shade in the remaining 3 squares, and 4 ways to arrange the S_2 squares such that they are adjacent. So, for two S_2 squares, there are a total of $2^2\times 2 + 2^3\times 4 = 40$ ways to shade in the S_3.

Three shaded S_2 squares leaves only 1 square remaining, and there are 4 ways to place three S_2 squares, so there are $2^1\times 4 = 8$ ways to shade three S_2 squares.

There is of course only 1 way to shade four S_2 squares at once.

To find the union of these sets, we use the inclusion-exclusion principle; the set of two S_2 squares is contained within certain shading patterns that include one S_2 square, but we then need to add back in any combinations that have three S_2 squares, then remove the combination which gives four S_2 squares again, as it is again contained within the cases for three S_2 squares. This gives $128 - 40 + 8 - 1 = 95$ ways to have at least one S_2 square shaded.

It follows that the probability of having at least one 2×2 square shaded is $\frac{95}{512}$.

△

47.2 Conditional Probability

Let A be an event, and let B be an event with non-zero probability. The probability of A occurring, given that B has occurred, is written as $\mathbb{P}(A|B)$, and can be calculated with,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}$$

It can be seen as the probability of A occurring within a new sample space over B.

47.2.1 Independence

Two events, A and B, are *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. A finite set of events is *pairwise independent* if every pair of events in the set is independent. A finite set of events is *mutually independent* if every event is independent from every other set and every intersection of every other event.

47.2.2 Law of Total Probability

Let S be a set. If $\{A_i\}_{i=1}^{\infty}$ are countably many disjoint non-empty sets such that $\bigcap_{i=1}^{\infty} A_i = S$, then $\{A_i\}_{i=1}^{\infty}$ are said to *partition* S.

If A is an event that can be written as a countable partition, $A = \{B_i\}_{i=1}^{\infty}$, then

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i)$$

or alternatively,

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A|B_i)\mathbb{P}(B_i)$$

47.2.3 Bayes' Theorem

For any events, A and B,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

If A and B are independent, this reduces to $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Extended form: Let $\{A_i\}_{i=0}^{n}$ partition the sample space. Then,

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{i=0}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}$$

Proof. By the definition of conditional probabilities,

$$\begin{aligned}
\mathbb{P}(A_i|B) &= \frac{\mathbb{P}(B \cap A_i)}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}(B \cap A_i)}{\mathbb{P}(B)} \cdot \frac{\mathbb{P}(A_i)}{\mathbb{P}(A_i)} \\
&= \frac{\mathbb{P}(B \cap A_i)}{\mathbb{P}(A_i)} \cdot \frac{\mathbb{P}(A_i)}{\mathbb{P}(B)} \\
&= \mathbb{P}(B|A_i) \cdot \frac{\mathbb{P}(A_i)}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{i=0}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}
\end{aligned}$$

■

Example. Three Prisoners Problem

Three prisoners, A, B, and C are in separate cells, supervised by a warden. Two of them have been sentenced to death and will be executed the following morning, but none of the prisoners know who is to be spared.

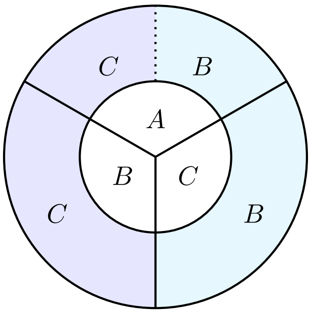
Prisoner A asks the warden what will happen tomorrow. The warden tells A that they won't say anything about A, nor anything about who will live. However, they do say that C is one of the prisoners to be executed.

Prisoner A is pleased as they believe that their own probability of surviving has gone up from 1 in 3, to 1 in 2, as it is now only between A and B who survives.

A secretly tells B the good news, who then reasons that A's chance of surviving is unchanged, while their own chances of survival has gone up to $\frac{2}{3}$.

Which prisoner is correct? △

In all cases, the warden will not tell A anything about A's fate. In the case that B is to live, the warden will also not say anything about who will live, so the warden can only say that C will be executed. In the case that C is to live, the warden will not say anything about who will live, so the warden can only say that B will be executed. However, if A is to live, then the warden has a choice. The warden can either say that B or C will be executed.



In the diagram above, the inner ring indicates who lives, with the outer ring indicating who the warden says will be executed. So, the warden says that B will die 50% of the time, or that C will die both of the time, corresponding to the right and left side of the circle, respectively. In both cases, A only lives $\frac{1}{3}$ of the time.

We can also write this using Bayes' theorem. Let A, B and C be the events that the corresponding prisoner is not executed, and let X be the event that the warden tells A that C is to be executed.

We see then that the probability that A survives is,

$$\begin{aligned}\mathbb{P}(A|X) &= \frac{\mathbb{P}(X|A)\mathbb{P}(A)}{\mathbb{P}(X)} \\ &= \frac{\mathbb{P}(X|A)\mathbb{P}(A)}{\mathbb{P}(X|A)\mathbb{P}(A)+\mathbb{P}(X|B)\mathbb{P}(B)+\mathbb{P}(X|C)\mathbb{P}(C)} \\ &= \frac{\frac{1}{2}\times\frac{1}{3}}{\frac{1}{2}\times\frac{1}{3}+1\times\frac{1}{3}+0\times\frac{1}{3}} \\ &= \frac{1}{3}\end{aligned}$$

and similarly,

$$\mathbb{P}(B|X) = \frac{\mathbb{P}(X|A)\mathbb{P}(A)}{\mathbb{P}(X)}$$

$$
\begin{aligned}
&= \frac{\mathbb{P}(X|A)\mathbb{P}(A)}{\mathbb{P}(X|A)\mathbb{P}(A) + \mathbb{P}(X|B)\mathbb{P}(B) + \mathbb{P}(X|C)\mathbb{P}(C)} \\
&= \frac{1 \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{3} + 1 \times \frac{1}{3} + 0 \times \frac{1}{3}} \\
&= \frac{2}{3}
\end{aligned}
$$

The denominators are the same in both cases, the difference stemming from the fact that the warden will always state that C is to be executed if B is to live, so $P(X|C) = 1$, but will only do so 50% of the time when A is to live, so $P(X|A) = \frac{1}{2}$.

47.2.4 Expected Value

The *expected value* of a random variable, X, is the weighted average of all possible values of X.

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} x_i p_i$$

where x_i are the possible values of X, and p_i are their corresponding probabilities of occurrence. The expected value is also sometimes denoted μ, particularly when working with normal distributions.

Expectation is linear, so,

$$\mathbb{E}\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1} c_i \mathbb{E}(X_i)$$

47.2.5 Variance

Variance is a measure of dispersion, representing how far a set of numbers is from their mean. Variance is the square of the standard deviation. It is often denoted as $\mathrm{Var}(X)$ or σ^2, and can be calculated from the expected value; $\mathrm{Var}(X) = \mathbb{E}(X^2) - E(X)^2$.

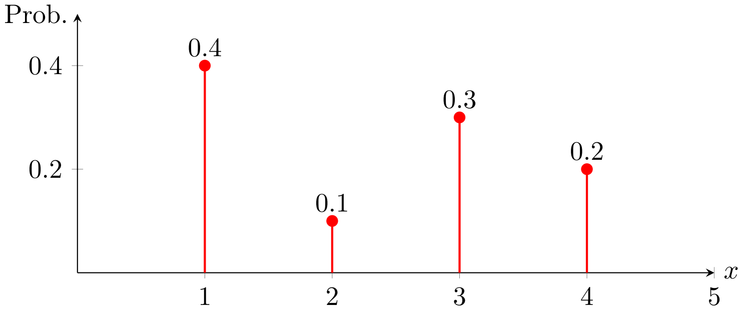
While you will not need to often calculate it by hand, the variance is an important summary statistic, and is frequently used as a parameter in various probability distributions.

47.3 Probability Distributions

A *random variable* is a quantity whose value depends on the outcome of a random event. Random variables are written in uppercase, with lowercase used to denote specific values the random variables can take.

A *probability mass function* or *discrete density function* is a function that gives the probability that a discrete random variable is equal to some given value. We write $\mathbb{P}(X = x)$ to denote the probability that the random variable X takes the particular value x. Then, the probability mass function, $p_X : \mathbb{R} \to [0,1]$ would be $p_X(x) = \mathbb{P}(X = x)$.

A probability mass function can be drawn on a plot,



Note that all the heights sum to 1, and that the probability mass function is zero at all the real numbers between valid outcomes.

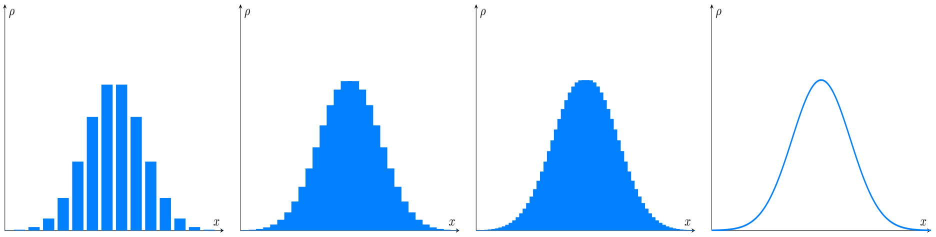
A defining feature of continuous probability distributions is that the probability for a random variable to take any specific value is 0, as there are infinitely many possible values the variable could take.

If every uncountably many particular values in some region all have non-zero probability, then the sum of all those probabilities goes to infinity. If all the probabilities are zero, then the whole sum is again zero, giving no meaningful information about the distribution.

To resolve this problem, we focus not on individual values, but ranges of values that the variable can take instead. For example, for a random variable that takes real values over [0,1], we might divide the interval into 10 parts and ask what the probability of falling into each region is.

When plotting this, rather than using the height of each bar to represent a probability, we use the areas.

As we make the intervals finer and finer, the smaller probability of falling into each interval is captured by the thinner width of each bars, so the height of the bars stay roughly the same as the intervals get smaller. Note that this wouldn't be the case if we used the heights to represent probability – in that case, every bar would shrink, and eventually reach 0 height in the limit. However, using areas, this process approaches a smooth curve.



So although each individual probability goes to zero, the overall shape of the distribution is preserved. With probability being proportional to the area of the bars, the vertical axis needs different units. Calling the width Δx, the height represents some kind of probability per unit in the x direction: $\frac{\text{Prob.}}{\Delta x}$, and we call this a *probability density*.

The curve that this process approaches is the *probability density function*, and is the continuous analogue to the probability mass function. To get the probability that a random variable lands within an interval, $[a,b]$, we integrate the probability density function between a and b to find the area under the curve. That is, if X is a random variable distributed according to the probability density function, f, then $\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\,dx$. Note that if $a = b$, then the integral returns 0, and integrating the probability density function over all of space returns 1.

47.3.1 Finite Discrete Uniform Probability Measures

A finite discrete probability measure is *uniform* if every pair of elementary events are equally likely.

In a discrete uniform probability measure, the probability of an event, A, happening is,

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

The plot of the probability mass function of a discrete uniform probability measure is a line of points, all with the same height.

Example. A fair D6 dice is rolled and a fair coin is thrown in an experiment. The event, $A = \{(1,H),(2,H),(3,H)\}$ has probability $\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{3}{12} = \frac{1}{4}$. △

Example. At a dinner party, 6 guests are seated around a table. Three pairs of hats are randomly distributed to the guests. What is the probability that every guest is sitting next to another guest with the same hat?

If we had a valid arrangement of hats, there are 6 ways to rotate it around the table, and since there are 3 sets of hats, there are 2 possible cycles (i.e., 1-2-3 and 1-3-2 are the two distinct cycles of 3 elements). Each hat can also be swapped around within its own pair, so there are $6 \times 2 \times 2^3$ valid arrangements, and 6! total arrangements, so the probability is $\frac{6\times2\times2^3}{6!} = \frac{2}{15}$. △

Most of the questions pertaining to these distributions will effectively reduce down to basic combinatorics and trying to find the size of an event.

You'll notice that we use the cardinality of Ω in the definition of probability for discrete uniform probability measures, and that this doesn't really make sense for infinite discrete sample spaces.

For more general sample spaces, we use the measure of those sets. As will be shown later, the measure of a countable set is zero, so this quotient is still not useful for us.

We could still attempt to define a distribution that assigns the same probability to each elementary event. Let X be a discrete random variable that takes values in a countably infinite set Ω, and suppose such a uniform distribution exists, so there exists some non-negative probability, p such that $\mathbb{P}(X = n) = p$ for all $n \in \Omega$. Since all the n are elementary and Ω is countably infinite, they are disjoint, so we can use the additive property of probability measures.

$$\begin{aligned} 1 &= \mathbb{P}(\Omega) \\ &= \mathbb{P}(X \in \Omega) \\ &= \sum_{n\in\Omega} P(X = n) \\ &= \sum_{n\in\Omega} p \end{aligned}$$

If $p = 0$, then $\sum_{n\in\Omega} p = 0$. If $p > 0$, then $\sum_{n\in\Omega} p = \infty$. In either case, we have a contradiction.

It turns out that there is no way to define a uniform distribution on a countably infinite set. So when someone says "discrete uniform distribution", they mean a finite discrete uniform distribution.

47.3.2 Continuous Uniform Probability Measures

If Ω is uncountably infinite, then the quotient using cardinalities is not well-defined. We instead use the *measure* of the sets involved. For $\Omega \subseteq \mathbb{R}$, we use the lengths of the sets; for $\Omega \subseteq \mathbb{R}^2$, the areas; and for $\Omega \subseteq \mathbb{R}^3$, the volumes.

Similar to countably infinite sets, not all subsets of $\mathbb{R}^n$ can be assigned a valid probability measure. This generally isn't a problem though, as all the sets we use in this module are measurable. For further reading, search *Vitali sets*, *Hausdorff paradox*, and the *Banach-Tarski theorem.*

Due to additivity, a lot of continuous probability problems reduce down to questions about geometry.

Example. Darts are thrown uniformly at a square with sides 2 units long. A unit circle dartboard is set in the square. What is the probability that any given dart will hit the dartboard?

The area of the dartboard is π, and the area of the square is 4, so the probability that the dart hits is $\frac{\pi}{4}$. △

Example. A coin is thrown and lands uniformly on an infinitely large table covered with a regular square grid.

1. If the coin has unit diameter, what is the probability that the coin does not land on any lines if the squares have side lengths,
   (a) 1?
   (b) 2?
   (c) $n \geq 1$?

Suppose the coin now has radius r, and the squares have side lengths of $L > 2r$.

What is the probability that the coin intersects,

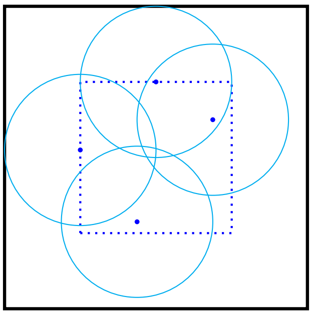
1. at most 1 line segment?
2. exactly 2 line segments?
3. exactly 3 line segments?

△

For all of these questions, we can just consider a single square, or a single intersection point, as the tiling is regular.

If the squares have side lengths 1, and the coin has unit dimeter, then the coin must land on the exact centre of the square, which is a point of zero area, so the probability that the coin does not land any any lines is 0.

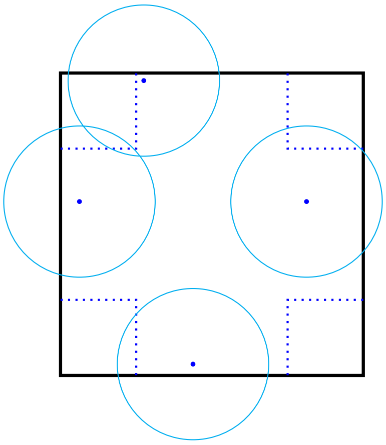
If the squares have side lengths 2, then the centre of the circle get to within radius of the outer square. This traces out another smaller square as our allowable area,



As long as the centre of the circle lands within the smaller square, the circle will not intersect any sides. The side length of the smaller square is the side length of the larger square, minus twice the radius (or minus the diameter) of the circle, so in this case, we have $2 - 1 = 1$, so the smaller square has unit length sides, and therefore has unit area. The larger square has area 4, so the probability that the circle does not intersect any lines is $\frac{1}{4}$.

Similarly, if the larger square has side lengths $n \geq 1$, then the smaller square will have side lengths $n-1$, so the probability is $\frac{(n-1)^2}{n^2}$. Or more generally, for a circle of radius r and a square of side length $L > 2r$, $\frac{(L-2r)^2}{L^2}$.

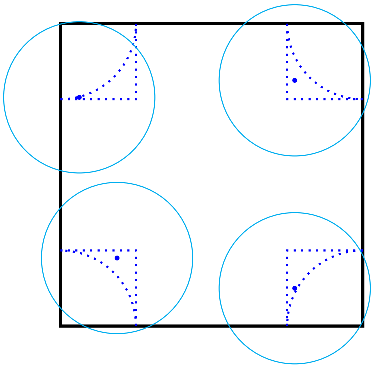
For the probability that the coin intersects at most one line segment, we look at how close the coin can get to the intersection points.



The smaller squares have side lengths equal to the radius of the coin, so the total area is $4r^2$ and the probability of intersecting at most 1 line segment is $\frac{4r^2}{L^2}$.

For intersecting exactly 1 line segment, we just subtract the probability of intersecting zero lines away, so we have $\frac{4r^2-(L-2r)^2}{L^2} = \frac{4r}{L} - 1$.

For intersecting exactly 2 line segments, the centre of the circle has to land in the complement of the region found before, but the circle cannot enclose the corner itself, so the centre cannot be within a radius distance from the corner. The required region is then,



which has area $4r^2 - \pi r^2 = (4-\pi)r^2$, so the probability that the coin intersects exactly two line segments is $\frac{(4-\pi)r^2}{L^2}$.

Given the setup of the grid, it is impossible for the coin to intersect exactly 3 lines, as intersecting more than 2 requires the circle to enclose a corner, which necessarily causes the coin to intersect 4 lines. It follows that the probability of the coin intersecting exactly 3 line segments is 0.

47.3.3 Measure Theory

If a real number in (0,1) is picked uniformly at random, what is the probability that the real number is rational?

The real numbers are uncountably infinite, so this is a continuous probability question, so we need to find the “length” of the rationals over (0,1).

The standard way to do this, is to cover the regions of interest with open intervals, then to add up the lengths of the intervals. One obvious way to do this is to just use (0,1), but we want to do this with the smallest total length possible. Can we do better than a length of 1?

We know that the rationals are a countable set, so there is a bijection between $\mathbb{Q}$ and $\mathbb{N}$. Cantor famously created one such bijection with his zig-zag argument, but we only need the rationals between 0 and 1.

There are many ways to do this, but one organised way is to start with $\frac{1}{2}$, then move onto $\frac{1}{3}$, $\frac{2}{3}$, then $\frac{1}{4}$ and $\frac{3}{4}$, then continuing with the reduced fractions with denominator 5, then 6, and so on. Doing this will list every rational in (0,1) exactly once, creating a bijection between the rationals and the naturals (the indexing set of the sequence).

$$\frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, \frac{1}{6}, \frac{5}{6}, \frac{1}{7}, \frac{2}{7}, \ldots$$

Now, we can assign an interval to each rational. Let $\varepsilon > 0$, and pick any convergent series, say $\sum_{i=1}^{\infty} \frac{1}{2^i} = 1$. Now, $\sum_{i=1}^{\infty} \frac{\varepsilon}{2^i} = \varepsilon$. Now, if we use the terms of this series as the lengths of intevals covering each rational, we can cover all the rationals in (0,1) using a total length of ε, so the length can be arbitrarily small.

We say that the rationals have a *Lebesgue measure* of 0. Doing the same process with the real numbers in (0,1), we find that this interval has a Lebesgue measure of 1. How this works in detail is somewhat involved, requiring more complicated topology techniques, and this question is just meant as a very brief introduction to measure theory, so the proof is omitted.

So, using the measures of the sets instead of cardinalities, we find that the probability that a real number randomly selected from (0,1) is rational is $\frac{0}{1} = 0$. This may be counterintuitive, given that the rationals are dense (§37.4.3) in the reals, and that it is certainly *possible* to select a rational from (0,1), but this kind of thing is very common in continuous probability.

We say that an event is said to happen *almost surely* if the set of possible exceptions has measure zero (and *almost never* is defined similarly). Note that this does not preclude the set of exceptions from being non-empty: rational numbers clearly exist between 0 and 1, but this set has measure 0, so we say that a rational is selected almost never, or equivalently, that an irrational is selected almost surely.

47.3.4 Binomial Distributions

A *Bernoulli trial* is an experiment with exactly two possible outcomes, often labelled “success” and “failure”, with the probabilities being the same every time the experiment is conducted.

If we define the random variable X to represent the number of successes in a fixed number of identical Bernoulli trials, then X is distributed *binomially*, and we write $X \sim B(n,p)$, where $n \in \mathbb{N}$ is the number of trials and $p \in [0,1]$ is the probability of success. Then, the following are equivalent notation for the probability mass function for the binomial distribution:

$$p_X(k) = f(k,n,p) = \mathbb{P}(k; n,p) = \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

A binomial distribution is a valid model for a random variable X if there are two possible outcomes, the number of trials and probability of success is fixed, and the trials are all independent from each other.

The expected value and variance of a random variable distributed binomially with parameters n,p, $X \sim B(n,p)$ is $\mathbb{E}(X) = np$ and $\mathrm{Var}(X) = np(1-p)$ (it is helpful to memorise these values, as they are used a lot, particularly in various approximations to the binomial distribution).

Situations where the binomial distribution could be used:

- Number of tails obtained on a (possibly biased) coin over 10 throws;
- Number of votes obtained by a candidate in a plurality voting election;
- Number of side effects of new medication experienced by 100 patients.

Situations where the binomial distribution is not valid:

- The colour of cards randomly removed from a deck without replacement (not independent – this is the *hypergeometric distribution*)
- The suit of cards randomly removed from a deck with replacement (not binary – this is the *multinomial distribution*)
- Number of times a die is rolled until a 6 is obtained (number of trials is not fixed – this is the *negative binomial* or *geometric distribution*);

47.3.5 Poisson Distribution

The series definition of the exponential function is,

$$e^x = \frac{x^0}{0!} + \frac{x^1}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

Multiplying both sides by e^{-x}, we have,

$$1 = \frac{x^0 e^{-x}}{0!} + \frac{x^1 e^{-x}}{1!} + \frac{x^2 e^{-x}}{2!} + \frac{x^3 e^{-x}}{3!} + \cdots$$

The right hand side sums to 1, so we can use these values as probabilities to define a probability distribution.

Using the probability mass function, $f(k;\lambda) = \mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$, this distribution is the *Poisson distribution*, taking a single parameter $\lambda > 0$.

A Poisson distribution is a valid model for a random variable X if events occur independently, singly in space or time, and at a constant average rate such that the mean number of occurrences over an interval is proportional to the length of the interval.

The expected value and variance of a random variable in a poisson distribution with parameter λ, $X \sim \text{Pois}(\lambda)$ is $\mathbb{E}(X) = \lambda = \text{Var}(X)$.

Situations where the Poisson distribution could be used:

- Number of alpha particles emitted by a radioactive source over a given time period;
- Number of patients arriving at an emergency room at a given hour of the day;
- Number of faulty parts manufactured at a factory in a day.

Situations where the Poission distribution is not valid:

- Number of students arriving at a lecture hall (not constant rate, and not independent);
- Number of earthquakes in a country per year (not independent);
- Number of articles published by tenured professors (to be tenured, a professor must have published at least once, so Poisson distribution is not a good fit due to the 0 output.)

47.3.6 Normal Distribution

The *normal* or *Gaussian distribution* has two parameters: μ, the population mean, and σ^2, the population variance. The distribution is symmetric about the mean, with mean=median=mode.

The probability density function of the normal distribution is,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The expected value and variance of a random variable distributed normally with parameters μ,σ^2, $X \sim N(\mu,\sigma^2)$ is $\mathbb{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

If some data is coded using the formula, $y = \frac{x-a}{b}$, then the mean and standard deviation of the coded data is given given by $\mu_y = \frac{\mu_x - a}{b}$ and $\sigma_y = \frac{\sigma_x}{b}$ (this is true of all random variables, not just normally distributed ones).

The *standard normal distribution* has mean 0 and standard devation 1. If $X \sim N(\mu,\sigma^2)$, then we can *standardise* X with the coding $Z = \frac{X-\mu}{\sigma}$. The resulting z-values are distributed according to the standard normal distribution, $Z \sim N(0,1)$. This works because every normal distribution is a version of the standard normal with the domain stretched by a factor of σ, and then translated by μ.

The probability density function of the standard normal distribution is often denoted $\phi(x)$, and is given by,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Instead of integrating the normal probability density function directly, we often standardise the given data and write the integral in terms of the standard normal density function. If $X \sim N(\mu,\sigma^2)$, then,

$$\begin{aligned}\mathbb{P}(a \leq X \leq b) &= \int_a^b f(x)\,dx \\ &= \frac{1}{\sigma}\int_a^b \phi\left(\frac{x-\mu}{\sigma}\right)\,dx\end{aligned}$$

47.4 Law of Large Numbers

Given a sequence of independent and identically distributed random variables $\{X_i\}_{i=1}^n$ with finite expected value $\mathbb{E}(X_1) = \mathbb{E}(X_2) = \cdots = \mathbb{E}(X_n) = \mu < \infty$, define a new random variable $\overline{X}_n = \sum_{i=1}^n \frac{X_i}{n}$. This variable is the *sample mean*.

As expectation is linear, $\mathbb{E}(\bar{X}_n) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i) = \frac{n\mu}{n} = \mu$, so the sample mean has the same mean as each of the individual variables, as we would expect.

Within statistics, there are various notions of convergence of random variables. These concepts are also called *stochastic convergence* in other areas of maths, and they formalise the idea that a sequence of random events can sometimes settle into some kind of stable behaviour with sufficiently large sample sizes.

We say that a sequence of random variables, $\{X_n\}$, *converges in distribution* or *converges weakly* towards a random variable X, if $\lim_{n\to\infty} F_n(x) = F(x)$ for all $x \in \mathbb{R}$ at which F is continuous, and where F_n and F are the cumulative distribution functions of X_n and X, respectively. This means that we increasingly expect that the next outcome in a sequence of random experiments is modelled better and better by the distribution of X. We use the notation $X_n \xrightarrow{\mathcal{D}} X$ or $X_n \rightsquigarrow X$ to represent this kind of convergence. This kind of convergence is used in the *weak law of large numbers.*

We say that a sequence of random variables, $\{X_n\}$, *converges in probability*, towards a random variable, X, if for all $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$. This means that the probability of an "unusual" outcome becomes smaller and smaller as the sequence progress. We use the notation $X_n \xrightarrow{P} X$ or $\operatorname{plim}(X_n) = X$ to represent this kind of convergence. This kind of convergence is used in the *central limit theorem.*

We say that a sequence of random variables, $\{X_n\}$ converges *almost surely*, *almost everywhere* or *strongly*, towards a random variable X, if $\mathbb{P}(\lim_{n\to\infty} X_n = x) = 1$. This type of convergence is very similar to pointwise convergence from analysis. This form of convergence means that the events for which X_n do not converge to X have probability 0 (the same as randomly selecting a rational from a reals interval; possible, but probability 0 – it has Lebesgue measure 0). We use the notation $X_n \xrightarrow{a.s.} X$ to represent this kind of convergence. This kind of convergence is used in the *strong law of large numbers.*

There is another stronger form of convergence analogous to uniform convergence from analysis called *sure convergence* or, but is rarely used in statistics as the only difference between sure and almost sure convergence in probability is in sets with Lebesgue measure 0.

The forms of convergence above are given in order of strength, with convergence in distribution being the weakest form. There are various other stronger forms of stochastic convergence not covered here.

There is a weak and a strong version of the law of large numbers. Both state that the sample average converges to the expected value;

$$\overline{X}_n \to \mu \text{ as } n \to \infty$$

The difference between the weak and strong versions is in the mode of convergence.

47.4.1 Weak Law of Large Numbers

The *weak law of large numbers* states that the sample mean converges in probability towards the expected value as the sample size increases;

$$\overline{X}_n \xrightarrow{P} \mu \text{ as } n \to \infty$$

That is, for any given error, $\varepsilon > 0$, there exists a sufficiently large sample size that will ensure that the average of the observations, $\overline{X_n}$ will almost always be within ε of the expected value, μ, which is the definition of a limit.

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| < \varepsilon\right) = 1$$

Equivalently, $\overline{X}_n$ will almost never be further than ε of the expected value, μ.

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) = 0$$

47.4.1.1 Bernoulli's Weak Law of Large Numbers

Suppose $X \sim B(n{,}p)$. Then, the expected value is $\mu = np$, so the weak law of large numbers says

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - np\right| > \varepsilon\right) = 0$$

However, for binary random variables, such as in the binomial distribution, we can also look at the mean of the proportion of successes, and not just the mean of the number of successes. Doing so, we have,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{X_n}{n} - p\right| > \varepsilon\right) = 0$$

47.4.2 Strong Law of Large Numbers

The *strong law of large numbers* states that the sample mean converges almost surely to the expected value;

$$\overline{X}_n \overset{a.s.}{\to} \mu \text{ as } n \to \infty$$

That is, $\mathbb{P}(\lim_{n\to\infty} \bar{X}_n = \mu) = 1$.

The weak law simply states that for some large n, $\bar{X}_n$ is likely to be close to μ, but does not preclude the possibility that $|\bar{X}_n - \mu| > \varepsilon$ happens infinitely many times (though, likely only at increasingly infrequent intervals for larger and larger n).

The strong law states that this almost surely does not occur (i.e., has Lebesgue measure 1). Note that this does not imply that for any $\varepsilon > 0$, there exists N such that $|\bar{X}_n - \mu| < \varepsilon$ holds for all $n > N$, since converging almost surely is not uniform convergence.

47.4.3 Central Limit Theorem

The *classical central limit theorem* states that if $\{X_i\}_{i=1}^n$ is an independent and identically distributed sequence of random samples drawn from a population with mean μ and variance σ^2, then the sample mean $\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$ converges in distribution to $N\left(\mu, \frac{\sigma^2}{n}\right)$, regardless of the distribution of the population.

47.5 Approximating the Binomial

With very large number of Bernoulli trials, it quickly becomes intractable to calculate factorials. For large n, we often approximate the binomial distribution with other, computationally easier distributions.

47.5.1 Poisson Limit Theorem

Let p_n be a sequence of real numbers in [0,1] such that the sequence np_n converges to some limit $\lambda < \infty$. Then,

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{\lambda^k e^{-\lambda}}{k!}$$

That is, if $X \sim B(n,p)$, n is large, and p is small, then X is approximately $\sim \text{Pois}(np)$ (recall that np is the expected value of X).

47.5.2 De Moivre–Laplace Theorem

The *De Moivre–Laplace theorem* is a special case of the central limit theorem. If $X \sim B(n,p)$, then, as $n \to \infty$, X converges in distribution to $N(\mu,\sigma^2)$, where μ is the expected value of X, which is np, and σ^2 is the variance of X, which is $np(1-p)$.

In other words, if $X \sim B(n,p)$, then for large n, X is approximately $\sim N\left(np, \sqrt{np(1-p)}^2\right)$.

Because the normal distribution is continuous, while the binomial is discrete, you need to apply a *continuity correction* when calculating probabilities. If $X \sim B(n,p)$ and $Y \sim N\left(np, \sqrt{np(1-p)}^2\right)$, then,

- $\mathbb{P}(X = a) \approx \mathbb{P}(a - 0.5 < Y < a + 0.5)$;
- $\mathbb{P}(X > a) \approx \mathbb{P}(Y > a + 0.5)$;
- $\mathbb{P}(X \geq a) \approx \mathbb{P}(Y > a - 0.5)$;
- $\mathbb{P}(X < a) \approx \mathbb{P}(Y > a - 0.5)$;
- $\mathbb{P}(X \leq a) \approx \mathbb{P}(Y < a + 0.5)$;

Chapter 48

Measure Theory

Chapter 49

Combinatorial Optimisation

> *"If you think it's simple, then you have misunderstood the problem."*
>
> — Bjarne Stroustrup

Combinatorial Optimisation is the study of finding optimal objects from finite sets, where the search space is discrete or discretisable. For instance, given a weighted finite graph, what is the optimal route to get from point A to point B? Combinatorial optimisation is closely related to complexity theory and theoretical computer science, as well as more practical applications in logistics and distribution optimisation problems.

Due to the nature of this chapter, we will unfortunately be mixing mathematical and programming notation, a *lot.* obj.flag represents an instance attribute, flag, attached to an object, obj. In algorithm blocks, single equality (=) or left arrow (←) represents variable assignment while double equality (==) represents an equality check. Setting a variable to "[ ]" indicates a *list* or an *array* being instantiated.

49.1 Complexity Analysis

49.1.1 Asymptotic Notation

Big O notation, also known as the *Landau symbols*, describes the limiting or *asymptotic* behaviour of a function as its argument tends towards some value, often infinity. Asymptotic descriptions like this allow us to quantify how good or bad an algorithm is mathematically, instead of running and testing several implementations of that algorithm on different machines, etc.

Let f be a real (or complex) valued function, and let g be a real valued function. Furthermore, let f and g be defined on some unbounded above subset of the positive real numbers, and let $g(x) > 0$ for all sufficiently large x.

Then, if there exists $M > 0$ and x_0 such that $|f(x)| \leq Mg(x)$ for all $x \geq x_0$, we write $f(x) \in O(g(x))$ as $x \to \infty$. The assumption that x is tending to infinity is often implicit, and we just write $f(x) \in O(g(x))$ alone. Essentially, $f(x) \in O(g(x))$ if $|f|$ is bounded above by g asymptotically, up to a constant factor.

$$f(x) \in O(g(x)) \coloneqq \exists M > 0 \exists x_0 \forall x > x_0 : |f(x)| \leq Mg(x)$$

There is also an analogous definition for x tending to a finite value a involving deltas, but it will not be discussed here. Instead, we can unify the two cases with the following alternative characterisation: if

$$f(x) \in O(g(x)) \coloneqq \limsup_{x \to a} \frac{|f(x)|}{g(x)} < \infty$$

then $f(x) \in O(g(x))$ as $x \to a$.

We sometimes use $=$ instead of $\in$, but this equality is not symmetric. $O(f(n)) = O(g(n))$ is not the same as $O(g(n)) = O(f(n))$. For example, $O(n) = O(n^2)$, but $O(n^2) \neq O(n)$. For this reason, $\in$ will be preferred in this document. You can also interpret $O(g(n))$ as a class of functions that don't grow faster than g, so the notation $\in$ also makes sense there (though, in this interpretation, you may be tempted to write $O(n) \subset O(n^2)$ when comparing these classes, which is not common notation).

Many of these classes have names, particularly in the context of analysing algorithm efficiency. Here are a few classes and algorithms, ordered by growth rate.

Class	Name	Example
$O(1)$	Constant	Returning the first element of a list, calculating $(-1)^n$
$O(\log\log n)$	Double Logarithmic	Implementing a van Emde Boas priority queue
$O(\log n)$	Logarithmic	Binary Search
$O(n)$	Linear	Searching through an unsorted list
$O(n \log n)$	Log-Linear or Linearithmic	Fast Fourier transform, merge sort, heapsort
$O(n^2)$	Quadratic	Naïve multiplication of n-digit numbers, bubble sort
$O(n^3)$	Cubic	Naïve matrix multiplication
$O(n^k)$, $k \in \mathbb{N}$	Polynomial	Determinant with LU decomposition, finding maximum matching in bipartite graph
$O(k^n)$	Exponential (linear exp)	Travelling salesman with dynamic programming, solving 3-SAT
$O(k^{n^m})$	Exponential	Decide a winning strategy for a game with polynomial turns and exponential moves, such as chess or go on arbitrary sized boards.
$O(n!)$	Factorial	Travelling salesman with brute force, determinant with Laplacian expansion
$O(k^{m^n})$	Double Exp	Deciding a FOL sentence over the naturals with the addition operation and equality predicate

Another notation is Ω or *big-Omega notation.* There are two incompatible definitions for this notation, but we will follow Knuth's convention: $f(x) \in \Omega(g(x))$ if and only if $g(x) \in O(f(x))$: f is bounded below by g asymptotically, up to a constant factor.

$$f(x) \in \Omega(g(x)) \coloneqq \exists M > 0 \exists x_0 \forall x > x_0 : |f(x)| \geq Mg(x)$$

or equivalently,

$$f(x) \in \Omega(g(x)) \coloneqq \liminf_{x\to\infty} \frac{f(x)}{g(x)} > 0$$

This is just the dual of big-O notation.

The last important notation we will cover is Θ, or *big-Theta notation.* $f(x) \in \Theta(g(x))$ if both $f(x) \in O(g(x))$ and $f(x) \in \Omega(g(x))$: f is bounded both above and below by g, up to a constant factor.

$$\exists M_1 \exists M_2 \exists x_0 \forall x > x_0 : M_1 g(x) \leq f(x) \leq M_2 g(x)$$

As we only care about the shape of growth as n becomes very large, when analysing runtime complexity of algorithms, we discard all coefficients, and keep only the term with the highest growth rate, as it will eventually dominate everything else. For instance, $2x^5 + 93x^2 + 50x + 12 \in O(x^5)$.

We also don't care about the base of logs in asymptotic notations:

$$\log_a(n) = \log_a(b)\log_b(n)$$

$$
\begin{aligned}
&= \frac{1}{\log_b(a)} \cdot \log_b(n) \\
&= k \log_b(n)
\end{aligned}
$$

so the base only affects the constant in the front which is discarded by the asymptotic notation.

Next, we give a sufficient (but not necessary) condition to test the asymptotic behaviour of a function: Consider two functions, $f(n)$ and $g(n)$.

Suppose

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} \to a$$

If,

- $a = 0$, then $f(n) \in O(g(n))$
- $a = \infty$, then $f(n) \in \Omega(g(n))$
- $a \in (0,\infty)$, then $f(n) \in \Theta(g(n))$

Applying a concave function, such as log, to both $f(n)$ and $g(n)$ does not change the asymptotic relationship between them.

Example. Is $2^{\log(\log(n))^3} \in \Omega\left(\sqrt{n}\right)$?

$$
\begin{aligned}
f(n) &= 2^{\log(\log(n))^3} \\
g(n) &= \sqrt{n} \\
f^*(n) &= \log(\log(f(n))) \\
&= \log\left(\log\left(2^{\log(\log(n))^3}\right)\right) \\
&= \log\left(\log(\log(n))^3 \log(2)\right) \\
&= \log\left(\log(\log(n))^3\right) + \log(\log(2)) \\
&= 3\log(\log(\log(n))) + \log(\log(2)) \\
g^*(n) &= \log\log(g(n)) \\
&= \log\left(\log\left(\sqrt{n}\right)\right) \\
&= \log\left(\frac{1}{2}\log(n)\right) \\
&= \log(\log(n)) + \log\left(\frac{1}{2}\right) \\
\lim_{n\to\infty} \frac{f^*(n)}{g^*(n)} &= \lim_{n\to\infty} \frac{3\log(\log(\log(n))) + \log(\log(2))}{\log(\log(n)) + \log\left(\frac{1}{2}\right)}
\end{aligned}
$$

Let $N = \log(\log(n))$. As $n \to \infty$, $N \to \infty$.

$$
\begin{aligned}
&= \lim_{N\to\infty} \frac{3\log(N) + \log(\log(2))}{N + \log\left(\frac{1}{2}\right)} \\
&= \lim_{N\to\infty} \frac{3\log(N)}{N + \log\left(\frac{1}{2}\right)} + \lim_{N\to\infty} \frac{\log(\log(2))}{N + \log\left(\frac{1}{2}\right)} \\
&= \lim_{N\to\infty} \frac{3\log(N)}{N + \log\left(\frac{1}{2}\right)}
\end{aligned}
$$

As $N \to \infty$, $3\log(N) \to \infty$ and $N + \log\left(\frac{1}{2}\right) \to \infty$, so apply L'Hôpital's rule.

$$\begin{aligned} &= \lim_{N\to\infty} \frac{3}{N} \\ &= 0 \end{aligned}$$

So $2^{\log(\log(n))^3} \in O\left(\sqrt{n}\right)$, and $2^{\log(\log(n))^3} \notin \Omega\left(\sqrt{n}\right)$. △

Example. Bubble sort.

To sort a list using *bubble sort*, we check the first two elements of the list, and swap them, if they are out of order. Then, we move along one, and check two elements again, then repeat until we reach the end of the list, where we start another pass. Once we pass through the list without performing any swaps, we know that the list is sorted.

Algorithm 13 Bubble Sort

```
procedure bubbleSort($A$)                              ▷ Input array
    $n \leftarrow$len($A$)
    repeat
        swapped = false
        for $i = 1$ to $n - 1$ do
            if $A[i-1] > A[i]$ then                    ▷ Check if the elements are out of order
                $(A[i-1], A[i]) \leftarrow (A[i], A[i-1])$         ▷ Swap elements
                swapped = True
            end if
        end for
    until swapped = False
    return $A$                                         ▷ Return the sorted list
end procedure
```

The comparison and swapping takes $\Theta(1)$ time, but runs $(n-1)$ times due to the for loop. The repeat statement will also run the for loop $(n-1)$ times, so overall, the algorithm takes $\Theta((n-1)^2) = \Theta(n^2)$.

One way to remember this result is to think about what happens if the smallest element is in the last place. Every time it is swapped, it only moves one place back, so $(n-1)$ passes are required, each one taking $\Theta(n)$ time, giving $\Theta(n^2)$. △

49.1.2 Master Theorem

Master Theorem: For an algorithm that has complexity that obeys the equation,

$$T(n) = aT\left(\frac{n}{b}\right) + \Theta(n^d), T(c) = \Theta(1)$$

we have,

$$T(n) \in \begin{cases} \Theta(n^d) & a < b^d \\ \Theta(n^d \cdot \log n) & a = b^d \\ \Theta(n^{\log_b a}) & a > b^d \end{cases}$$

Example. Merge sort.

In *merge sort*, we divide the list into halves, then run the algorithm again on each half, returning the list once the list is length 1. Then, we merge the sorted sublists together until only one list remains.

Algorithm 14 Merge Sort

```
procedure MERGESORT($A$)                                        ▷ Input array
    $n$ ←LEN($A$)
    if $n \leq 1$ then
        return $A$              ▷ If the list only contains one element, it is already sorted
    end if
    left ← [ ]
    right ← [ ]
    for $i = 1$ to $n$ do
        if $i < \frac{n}{2}$ then
            APPEND(left)                ▷ Split the list into two sublists, left and right
        else
            APPEND(right)
        end if
    end for
    left←MERGESORT(left)                                    ▷ Sort the two sublists
    right←MERGESORT(right)
    return MERGE(left,right)                ▷ Merge the two sorted sublists together
end procedure
```

where the merge subroutine combines two sorted lists into one sorted list in linear time.

The algorithm takes

$$T(n) = \underbrace{T\left(\left\lfloor \frac{n}{2} \right\rfloor\right)}_{\text{Sort left}} + \underbrace{T\left(n - \left\lfloor \frac{n}{2} \right\rfloor\right)}_{\text{Sort right}} + \underbrace{\Theta(n)}_{\text{Merge}}, n > 1$$

and we know $T(1) = \Theta(1)$, as the algorithm just returns the list for an array of length 1, taking constant time.

$$\sim 2T\left(\left\lfloor \frac{n}{2} \right\rfloor\right) + \Theta(n)$$

So, using the Master theorem, we have $a = 2$, $b = 2$, $d = 1$, so,

$$= \Theta(n \log n)$$

△

49.2 Graph Theory

49.2.1 Minimal & Maximal Elements

We briefly revisit the concept of minimal and maximal elements.

If for some x, $y \leq x$ only if $y = x$, then x is *minimal.* Or equivalently, x is minimal if there does not exist any y such that $y < x$. A partial order may have any number of minimal elements, including none. For example, the integers have no minimal element, the naturals have one minimal element, 0, and a set with k mutually incomparable elements has k minimal elements.

If an element x satisfies $x \leq y$ for all y, then x is a *minimum.* A partial order may have at most one minimum, such as 0 in the naturals, but can also have none at all, either because it contains an infinite descending chain like with the integers, or because it has more than one minimal element. Any minimum element is also minimal.

We define maximal and maximum elements similarly, as elements that are not less than any other element and elements that are greater than all other elements, respectively. Again, maximum elements are also maximal.

While these definitions seem similar, they are distinct, elements can be maximal, but not maximum. For example, consider the family of all subsets of $\mathbb{N}$ with at most three elements, ordered by $\subseteq$. Then, the set $\{0,1,2\}$ is a maximal element of this family, because it is not a subset of any larger set, but it is not a maximum, because it is not a superset of $\{3\}$ (and similarly for any other three-element set).

49.2.2 Basic Definitions & Theorems

A *graph* G is represented by V, a set of *vertices* or *nodes*, and E, a set of pairs of vertices, called *edges* or *arcs*, and we write $G = (V,E)$. If we are using multiple graphs at once, we can refer to the vertex (edge) set of a graph G by writing $V(G)$ $(E(G))$.

If the edge pairs are ordered, the graph is *directed* or *oriented*, and can also be referred to as a *digraph.*

A vertex and an edge are *incident* if the vertex is at either end of the edge. The *degree*, *valency* or *order* of a vertex is the number of edges incident to it. The *indegree* and *outdegree* of a vertex of a digraph is the number of edges pointing into and out from the vertex, respectively. A vertex of degree 1 is called a *leaf.* If every vertex of a graph has the same degree k, then the graph is said to be *k-regular.*

The *degree sequence* of a graph is a list of its vertex degrees. Conversely, if a sequence of numbers is the degree sequence of some graph, it is *graphical.* For example, 3, is not a graphical sequence, as there is no graph with a single node of degree 3, while 2,2,2 *is* a graphical sequence, because the triangle graph has 2,2,2 as a degree sequence.

For a vertex $x \in V$, we define $N(x) = \{y \in V : (x,y) \in E\}$ to be the *neighbourhood* of x. The degree of x can then also be written as $\deg(x) = |N(x)|$.

An edge that starts and ends at the same vertex is called a *loop.* If multiple copies of the same edge pair exists in the edge set, then the edges are called *parallel edges.*

A graph that does not contain loops nor parallel edges is called a *simple* graph. A graph that can contain parallel edges is a *multigraph.* A graph that can contain both loops and parallel edges is a *pseudograph.*

If each edge also has a number associated with it (the *weight* of the edge), the graph is a *weighted graph.* We write (G,w) for a weighted graph, where G is the underlying unweighted graph, and w is a function that maps edges to weights. When we write $w(S)$, where $S \subseteq G$ (or $S \subseteq E(G)$), we mean the sum of the weights of the edges of S, $\sum_{e \in V(S)} w(e)$ (or $\sum_{e \in S} w(e)$, respectively).

A *walk* is a route through a graph. A walk is *closed* if the first and last vertices are the same, and *open* otherwise. A *path* is a walk in which no vertex is visited more than once. A *trail* is a walk in which no edge is visited more than once. A *cycle* is a path in which the ending and starting vertex are the same. A *ray* is an infinite path that starts at a vertex, then travels through infinitely many other vertices.

A *Hamiltonian cycle* is a cycle that visits every vertex. An *Eulerian walk* is a trail which traverses every edge. An *Eulerian circuit* is both a trail and cycle which traverses every edge.

Theorem (Euler)**.** *An Eulerian circuit exists if and only if every vertex is of even degree.*

Corollary 49.2.0.1. *An Eulerian walk exists if and only if there are at most two vertices of odd degree.*

A graph that admits an Eulerian walk is *traversable* or *semi-Eulerian.* A graph that admits an Eulerian circuit is *Eulerian.*

Two vertices are *connected* if there is a path between them. Two vertices, u and v, are *adjacent* if they are connected by an edge, so $(u,v) \in E$. u and v are also called *neighbours.* In a directed graph, the *in-neighbours* of a vertex v, are all vertices u such that $(u,v) \in E$, and the *out-neighbours* are all vertices u such that $(v,u) \in E$.

$N(v)$ represents the set of neighbours of v, but does not include v itself. This notation can also be used on sets of vertices to represent the set of neighbours of that set of vertices.

A *path* graph, P_n, is a graph consisting of a sole path, without cycles. That is, a line of nodes, with a single path/trail running through it. Symbolically, the path graph is a graph on n nodes, $V = \{v_1, \ldots, v_n\}$ with $E = \{(v_i, v_{i+1}) : i \in [1, n-1]\}$.

A *cycle* graph, C_n, is a graph consisting of a sole cycle.

A *complete* graph, K_n, is a graph on n nodes with every possible edge included once.

A *tournament* is a directed complete graph. If an edge points from a vertex a to a vertex b, then a *dominates* b. If $D = (V, E)$ is a tournament, and $S \subset V$ with $|S| = k$, then S is a *k-strong set* if for every $v \in V \setminus S$, there exists a $u \in S$ such that $(u, v) \in E$. In other words, every vertex not in S is dominated by at least one vertex in S.

A *bipartite* graph is a graph that has a vertex set that can be partitioned into two subsets, commonly denoted L and R, such that for, every edge, $(u, v) \in E$ either $u \in L$ and $v \in R$ *or* $u \in R$ and $v \in L$. If a graph $G = (V, E)$ has partites L and R, we also write $G = (L \cup R, E)$ to represent this data.

The *complete bipartite graph* $K_{n,k}$ is the graph with two vertex partites of cardinality n and k with all possible edges between them. $K_{2,2} = C_4$. We also call a graph $K_{1,n}$ a *star graph*, and in particular, $K_{1,3}$ is the *claw graph.*

For all $k \in \mathbb{N}$, there exists a tournament on $n \in \mathbb{N}$ vertices without a k-strong set.

A graph is *connected* if every pair of vertices is connected.

A graph is a *tree* if it does not contain a cycle. An disconnected tree graph may also be called a *forest.* A directed forest is an *arborescence.* Every tree is bipartite.

A tree is *rooted* by distinguishing a vertex to be the *root.* From the root, a natural orientation of the edges can be assigned (i.e. pointing away or towards the root), forming a directed rooted tree. The maximum distance from the root to any leaves in the tree is called the *height* of the tree. If two nodes u and v are adjacent in a rooted tree, with u closer to the root, then we say that u is the *parent* of v, or that v is the *child* of u. If two vertices have the same parent node, then they are *sibling* nodes.

Given a graph G, we can *delete* a vertex by removing a vertex and removing all edges incident to it. We can similarly *delete* an edge by removing it. More interestingly, we can *contract* an edge by removing that edge, then combining the two incident vertices, such that every edge connecting to one of the original vertices connects to the new joined vertex (note that if one of several parallel edges are contracted, all the remaining parallel edges become loops on the joined vertex).

A *subgraph* of a graph G is a graph whose vertices and edges all belong to G. A subgraph is *induced* if every edge that can be included is included. In other words, an induced subgraph can be obtained by deleting vertices in G, but not edges. Given a graph G, and a subset $U \subseteq V(G)$, the subgraph of G induced by U is denoted $G[U]$.

A *spanning tree* is a subgraph that contains every vertex of the original graph, and is also a tree. A *connected component* of G is a maximal (with respect to inclusion) connected subgraph of G.

A subset of vertices $S \subseteq V$ is an *independent set* of the graph if there are no edges between any pair of vertices in S (this allows us to alternatively characterise trees as graphs whose vertex sets can be partitioned into two independent sets). Conversely, a *clique* is a subset of pairwise adjacent vertices. More generally, a *l-clique* is a subgraph that is a *complete* graph on l vertices. The *independence number* is the size of a maximum independent set, while the *clique number* is the size of a maximum clique.

Two graphs, $G = (V, E)$ and $H = (W, F)$ are *isomorphic* if there exists a bijective function, $\phi : V \to W$ such that if $(v_1, v_2) \in E$, then $\phi(v_1), \phi(v_2) \in F$, and vice versa. If such a function exists, we write $G \cong H$. The best known algorithm to determine whether two given graphs are isomorphic is $O(n^{\log n})$.

A graph G is called *H-free* if no induced subgraph of G is isomorphic to H.

The *complement* of a graph $G = (V,E)$, denoted $\bar{G}$ or G^c, is the graph (V,E'), where E' is the set of edges over V that are not in E. A graph is *self-complementary* if it is isomorphic to its complement.

A *matching* over a graph $G = (V,E)$ is a set of edges $M \subseteq E$ such that no vertex is incident to more than one edge. The *matching number* is the size of a maximum matching. A matching in which every vertex is incident to an edge is a *perfect matching.* A perfect matching is only possible on graphs with an even number of vertices.

An *alternating chain* with respect to a matching, M, is a path whose edges alternate between matched and unmatched edges. M admits an alternating chain if and only if M is not maximal.

A *vertex cover* is a subset, $S \subseteq V$ such that every edge in E is incident to at least one vertex in S.

For a graph $G = (V,E)$, if $M \subseteq E$ is a matching and $S \subseteq V$ is a vertex cover, then $|M| \leq |S|$. This also implies that the size of a maximal matching is at most the size of a minimal vertex cover.

Consider $G = (V,E)$ and let $S \subseteq V$. S is a vertex cover of G if and only if $V \setminus S$ is an independent set.

The distance between two vertices, u and v, written as $d(u,v)$, is the length of the shortest path from u to v. On an undirected graph,

- $d(u,v) = 0 \leftrightarrow u = v$ (Point separating);
- $d(u,v) = d(v,u)$ (Symmetry);
- $d(u,v) + d(v,w) \geq d(u,w)$ (Triangle inequality).

thus satisfying the requirements for a metric. A graph, along with this definition of a distance function, is a metric space.

Theorem 49.2.1. *A tree on n nodes has $n - 1$ edges.*

Proof. Let $P(n)$ be the statement that every tree on n nodes has $n - 1$ edges. $P(1)$ holds, as the trivial graph has $0 = 1 - 1$ edges. Assume that $P(n)$ holds for some fixed arbitrary value of $n \geq 1$.

Let T be a tree with $n+1$ nodes. As T is a tree, it cannot contain cycles, so at least one leaf node exists. Remove the leaf, and the edge incident to it. The new graph is a tree with n nodes. By the inductive hypothesis, this new graph has $n - 1$ edges, so it follows that T has n edges. Thus, $P(n) \to P(n + 1)$, completing the induction. ■

Corollary 49.2.1.1. *Every connected graph has a spanning tree. Every connected graph over n nodes has at least $(n - 1)$ edges, with exactly $(n - 1)$ edges if and only if the graph is a tree.*

A *cut* is a partition of the vertex set of a graph into two disjoint sets, L and R. An edge is *in* the cut (L,R) if it connects a vertex in L with a vertex in R. The set of edges in the cut C is denoted $\delta(C)$. The *value* of the cut C is the number of edges in the cut, $|\delta(C)|$.

If $G = (V,E)$ is a graph, then there exists a cut in G with value at least $\frac{|E|}{2}$.

The deletion of any edge from a tree partitions it into two connected components.

Lemma (Euler's Handshaking Lemma)**.** *In any undirected graph (V,E), the sum of the degrees of the vertices is equal to twice the number of edges.*

$$\sum_{v \in V} \deg(v) = 2|E|$$

Proof. Every edge connects two vertices, each contributing exactly 2 to the sum of the degrees. ■

Corollary 49.2.1.2. *The number of odd degree vertices is even.*

Corollary 49.2.1.3. *Every tree on $n \geq 2$ vertices has at least two leaves.*

Theorem 49.2.2. *The following statements are equivalent for any connected graph $G = (V,E)$:*

1. *G is a tree;*
2. *G has no cycles;*
3. *Any two vertices of G are connected by a unique path;*
4. *$G' = (V,E \setminus \{e\})$ is disconnected for any $e \in E$;*
5. *$|E| = |V| - 1$*

Proof. (1) $\leftrightarrow$ (2) by definition.

(2) $\leftrightarrow$ (3) because if the path is not unique, then the two paths together form a cycle.

(3) $\leftrightarrow$ (4) because if G' were connected, then the endpoints of e would be connected in G by two different paths.

(5) $\leftrightarrow$ (1) by Theorem 49.2.1. ■

49.2.3 Pigeonhole Principle

The *pigeonhole principle* states that if n elements are partitioned into m non-empty sets, with $n > m$, then at least set must contain more than one element.

Example. Let nine points be placed inside a square of side length 1, with no three points lying on the same line. Prove that it is always possible to select 3 points that form a triangle with an area of at most $\frac{1}{8}$.

Divide the unit square into 4 subregions of area $\frac{1}{4}$; for simplicity, and without loss of generality, let these regions be squares of side length $\frac{1}{2}$.

As there are 9 points, and 4 squares, there will always be at least one square containing at least 3 points by the pigeonhole principle (note: a point that lies on the edge of the square can be considered to be contained within that square). Selecting these three points within the square to be the vertices of a triangle, the entire triangle must be fully contained within that square.

The largest area it can be is half the area of the square. As the square has area $\frac{1}{4}$, it follows that the area of the triangle is at most $\frac{1}{8}$, as required. △

Example. Consider the complete graph on 6 vertices. Colour each edge either red, or blue. Prove that, no matter how the edges are coloured, the graph will always contain a triangle with all three sides the same colour.

Consider a particular vertex of the K_6 graph. There are 5 vertices adjacent to the selected vertex, and so, by the pigeonhole principle, at least three of the incident edges are of the same colour. Without loss of generality, assume that this colour is red. If any of the edges connecting those three vertices are red, a red monochromatic triangle including the original vertex is formed. If none are red, then all three must necessarily be blue, forming a blue monochromatic triangle.

Exercise. Can you extend this proof to show that a monochromatic triangle must always exist when colouring a complete graph on 17 vertices with 3 colours? △

This last example can also be stated in terms of cliques, where, instead of colouring edges red or blue, we include or exclude edges from a graph on 6 vertices, and want to prove that at least one 3-clique or 3-independent set exists. (The exercise, however, cannot. Cliques and independent sets are only equivalent for two-colour cases.)

49.2.4 Ramsey Numbers

For every $k,l \in \mathbb{N}$, $R(k,l)$ denotes the smallest positive integer such that every graph on $R(k,l)$ vertices contains a k-clique or an independent set of size l.

As shown in the last example of the previous section, we know that we always have a 3-clique or a 3-independent set on a graph with 6 vertices, so we have proved that $R(3,3) \leq 6$ (we haven't proved that smaller graphs don't have this property, so this is just an upper bound). However, we can easily prove $R(3,3) = 6$ by producing counterexamples for smaller graphs.

Clearly, $R(1,n) = R(n,1) = 1$ for all $n \in \mathbb{N}$, as the single vertex in the trivial graph simultaneously satisfies the condition for a 1-clique and a 1-independent set.

By symmetry, $R(k,l) = R(l,k)$ for any $k,l \in \mathbb{N}$. If every graph on n vertices satisfies $R(k,l)$, then every such graph contains an k-clique or a l-independent set. It follows that the complement of every graph then contains a l-clique or an k-independent set.

Theorem 49.2.3. *$R(k,l) \leq \binom{k+l-2}{k-1} \leq 2^{k+l}$ for all $k,l \in \mathbb{N}$.*

Theorem 49.2.4. *$R(k,k) \leq \binom{2k-2}{k-1} \leq 4^k$ for all $k \in \mathbb{N}$.*

Theorem 49.2.5. *If n,k are natural numbers satisfying $\binom{n}{m}2^{1-\binom{k}{2}} < 1$, then $R(k,k) > n$.*

49.2.5 Graph Traversal

Given a finite simple graph G and a vertex $v \in V$, also called the root, we wish to find a set $R \subseteq V$ of vertices reachable from v (i.e. for every $u \in R$, there exists a path from v to u), and a set $T \subseteq E$ of edges such that (S,T) is a tree.

The two classical algorithms for this are *depth first search* (DFS) and *breadth first search* (BFS).

DFS traverses the depth of any particular path as far as possible at each step. The algorithm starts from the root, moving from the current vertex to an adjacent unvisited vertex, continuing until there are no unvisited nodes left. Then, the algorithm backtracks along previously visited nodes in reverse order until one of these visited nodes has unvisited neighbours, at which point it proceeds down the new path as far as possible again. BFS starts at the root, and explores all nodes at a given depth, before moving on to nodes at the next depth level.

These algorithms perform the same task, but are suited to different applications. For instance, if you are building a chess AI, you might use a graph traversal algorithm to explore the possible graph of future game states. In this case, BFS would look at all possible first moves, before exploring all possible combinations of first and second moves. On the other hand, a naïve application of DFS would almost immediately get stuck in an infinite branch and never backtrack.

DFS can be given recursively, but we give a stack based implementation here:

Algorithm 15 Depth First Search

```
vertices = [ ]
edges = [ ]
procedure DFS($G$,$v$)
    $v$.visited = true
    $S$ = stack()
    $S$.push($v$)
    while $S$ is not empty do
        parent = $v$
        $v$ = $S$.pop()
        if $v$.visited == false then
            $v$.visited = true
            for $u \in N(v)$ do
                $S$.push($u$)
            end for
            edges.append((parent,$v$))
        end if
    end while
    for $v \in V$ do
        if $v$.visited == true then
            vertices.append($v$)
        end if
    end for
    return (vertices,edges)
end procedure
```

Note that the checking of whether a vertex has been visited is deferred until after the vertex is popped from the stack.

When giving a proof for an algorithm, we need to show termination, and correctness; that the algorithm will terminate within a finite number of steps, and that the algorithm works as intended, respectively.

We give a proof of DFS.

Proof. There are finitely many vertices, so the algorithm will terminate. Next, we prove correctness. At each pass of the while loop, the visited vertices form a tree by induction. Suppose there exists a vertex $t \in V \setminus \{R\}$ that is reachable from the root v, and let P be the path between v and t. Since $v \in R$ and $t \notin R$, there must exist two vertices x and y such that $x \in R$, $y \notin R$, and $(x,y) \in E$. Since $x \in R$, it must have been visited by the algorithm and hence have been in the stack. But then, all the neighbours of x, including y, would have been pushed onto the stack and hence marked visited (if not already visited), contradicting that $y \notin R$, and hence P and t do not exist. ■

BFS can similarly be given recursively, but we give a queue based implementation here:

Algorithm 16 Breadth First Search

```
vertices = [ ]
edges = [ ]
procedure BFS($G$,$v$)
    $v$.visited = true
    $S$ = queue()
    $S$.enqueue($v$)
    while $S$ is not empty do
        parent = $v$
        $v$ = $S$.dequeue()
        $v$.visited = true
        for $u \in N(v)$ do
            if $u$.visited == false then
                $S$.enqueue($u$)
            end if
        end for
        edges.append((parent,$v$))
    end while
    for $v \in V$ do
        if $v$.visited == true then
            vertices.append($v$)
        end if
    end for
    return (vertices,edges)
end procedure
```

The proof of correctness is similar to the proof for DFS.

Theorem 49.2.6. *A BFS-tree contains a path from the root v to every vertex reachable from v, which is shortest in G.*

Proof. Let $d_G(u,v)$ denote the distance between u and v in a graph G. Suppose (S,T) is the tree returned by the BFS algorithm.

Suppose that, when the algorithm ends, there are vertices $x \in S$ such that $d_G(v,x) < d_{(S,T)}(v,w)$. Without loss of generality, let w denote the vertex closest to v with this property.

Let P be a shortest path from v to w in G, and let (u,w) be the last edge in P. Then, by assumption, $d_G(v,u) = d_{(S,T)}(v,u)$, and hence $(u,v) \notin T$.

$$\begin{aligned} d_{(S,T)}(v,w) &> d_G(v,w) \\ &= d_G(v,u) + 1 \\ &= d_{(S,T)}(v,u) + 1 \end{aligned}$$

This implies that u was enqueued earlier than w, since vertices are enqueued according to their distance from the root in (S,T). In particular, w was not enqueued until after u was dequeued, since vertices are also dequeued in order with nondecreasing distance. But, w must have been enqueued via the edge (u,w) when u was enqueued, contradicting that $(u,w) \in T$. It follows that the assumption that there exists vertices $x \in S$ such that $d_G(v,x) < d_{(S,T)}(v,w)$ is false. ■

Theorem 49.2.7. *Graph traversal algorithms can be implemented in a graph G with $|V| = n$ vertices and $|E| = m$ edges to run in $O(n+m)$ time. Furthermore, the connected components of G can be detected in linear time.*

49.2.6 Minimum Cost Spanning Tree

Given a finite connected weighted simple graph G, we wish to find a spanning tree T of G such that the total weights of the edges in T is minimal. This is the minimum cost spanning tree (MST) problem.

Let (G,w) be a weighted graph.

Theorem 49.2.8. *The following statements are equivalent for any spanning tree T in G:*

1. *T is an optimum solution.*
2. *For every edge $f = (x,y) \notin E(T)$, no edge on the path from x to y in T has higher cost than f.*
3. *For every edge $e \in E(T)$, e is a minimum cost edge of $\delta(V(C))$, where C is a connected component of $T \setminus \{e\}$*

Proof. (1) → (2): Suppose T is optimum, but there is an edge $f = (x,y) \notin E(T)$, and an edge e on the path connecting x to y in T such that $w(f) < w(e)$. Then $(T \setminus \{e\}) \cup \{f\}$ is a spanning tree with lower cost.

(2) → (3): Suppose (3) does not hold, so there exists an edge $f = (x,y) \in \delta(V(C))$ such that $w(f) < w(e)$ Observe that e is the only edge in $\delta(V(C))$ in T, so $f \notin E(T)$, contradicting (2).

(3) → (1): Suppose T satisfies (3), but is not optimum. Let T' be an optimum tree maximising $|E(T) \cap E(T')|$, and suppose there exists $e \in E(T) \setminus E(T')$. Let C be a connected component of $T \setminus \{e\}$, so $e \in \delta(C)$. Clearly, $T' \cup \{e\}$ contains a cycle. Let $f \in \delta(C)$ be any other edge of the cycle. $(T' \setminus \{f\}) \cup \{e\}$ is a spanning tree, and since T' is optimum, $w(f) \leq w(e)$. However, we have $w(e) \leq w(f)$ from (3), so $w(f) = w(e)$, and hence $(T' \setminus \{f\}) \cup \{e\}$ is an optimum spanning tree. But this tree has more edges in common with T than T'. This contradiction shows that $E(T) \subseteq E(T')$, and hence $E(T) = E(T')$, so T is optimum. ■

The two classical algorithms here are *Kruskal's algorithm* and *Prim's algorithm.*

Algorithm 17 Kruskal's Algorithm

```
Sort the edges into ascending order of weight.
Select an edge of least weight to start the tree.
Consider the next edge of least weight. If it would form a cycle with the edges already included,
  move to the next edge. Otherwise, include the edge.
Repeat the previous step until all vertices are connected (or equivalently, if all edges remaining would
  form a cycle).
```

Proof. The algorithm terminates as G is finite. Correctness is proven in two parts: that the graph T produced is indeed a spanning tree, and that T is minimal.

T cannot have a cycle, as edges that would form a cycle are excluded by definition. T also cannot be disconnected, since the first encountered edge that joins disconnected components of T would have been added by the algorithm. Thus, T is a spanning tree of G.

Let P be the statement that if F is the set of edges chosen at any step of the algorithm, then there is some minimal spanning tree T that contains F and none of the edges rejected by the algorithm.

Clearly, P holds at the beginning when $F = \varnothing$ as any minimal spanning tree will suffice. Assume that P holds for some arbitrary non-final step of the algorithm.

If the next chosen edge e is in T, then P also holds for $F \cup \{e\}$. Otherwise, if $e \notin E(T)$, then $T \cup \{e\}$ has a cycle by construction, C. This cycle contains edges which are not in F, since e does not form a cycle when added to F, but does in T. Let $f \in C \setminus F$ be such an edge. Note that $f \in T$, and by P, has not been considered by the algorithm. f must therefore have a weight at least as large as e. Then, $(T \setminus \{f\}) \cup \{e\}$ is a tree with the same (or less) weight as T that contains $F \cup \{e\}$, so P also holds in this case.

By induction, P holds when F is itself a spanning tree, which is possible only if F also minimum. ■

Theorem 49.2.9. *For a graph $G = (V,E)$, a standard implementation of Kruskal's algorithm runs in $O(|E| \log |E|)$, or equivalently, $O(|E| \log |V|)$ time.*

Remark. These time classes are equivalent as $|E|$ is at most $|V|^2$, and $\log |V|^2 = 2 \log |V| \in O(\log |V|)$.

Proof. Sorting the edges takes $O(|E| \log |E|)$ time.

Selecting an edge of least weight is just returning the first element of a sorted list, which takes constant $O(1)$ time.

Checking if this edge creates a cycle is equivalent to checking if the edge connects two vertices that lie in different trees. We track which tree each vertex lies in using a union-find structure (similar to a disjoint union in set theory), which takes $O(|V|)$ operations to initialise. Then, during runtime, in the worst case, every edge needs to be iterated over, and for each edge, we perform two tree lookups, and possibly a union, which takes at most $O(|E| \log |V|)$ time.

Thus, the total time complexity is $O(|E| \log |E|) = O(|E| \log |V|)$. ■

Algorithm 18 Prim's Algorithm

```
Select any vertex to start the tree.
Select an edge of least weight that joins a vertex already in the tree to a vertex not yet in the tree.
Repeat the previous step until all vertices are connected (or equivalently, if all edges remaining would
  form a cycle).
```

Proof. The proof that the produced tree T is spanning is almost identical to that of Kruskal's algorithm.

Condition (3) of Theorem 49.2.8 guarantees that T is optimum. ■

Theorem 49.2.10. *Given the adjacency matrix of a graph $G = (V,E)$, a simple implementation of Prim's algorithm runs in $O\left(|V|^2\right)$ time.*

Proof. We can find the minimum weight edge to add by linearly searching the adjacency matrix, which has $|V|^2$ entries, giving $O\left(|V|^2\right)$ time complexity. ■

Remark. Using binary or Fibonacci heaps and adjacency lists, Prim's algorithm can be improved to run in $O((|V| + |E|) \log |V|) = O(|E| \log |V|)$ and $O(|E| + |V| \log |V|)$ time, respectively.

Kruskal's algorithm will find a spanning forest if G is disconnected, but Prim's algorithm will only find the tree spanning the connected component containing the starting vertex. Prim's algorithm can be extended to find spanning forests by iterating over the vertices.

49.2.6.1 Number of Spanning Trees

How many trees are there with n labeled vertices, up to isomorphism? Or equivalently, how many spanning trees does the complete graph K_n have?

On 3 vertices, the spanning tree is unique. On 3 vertices, the spanning tree is a path P_3 on the three nodes, and there are 3 ways to permute the order in which the path passes through the nodes, giving 3 spanning trees. On 4 vertices, there are $4!/2$ trees that are paths, for similar reasons, and 4 trees that are stars, giving 16 total. On 5 vertices, there are $5!/2$ copies of P_4, 5 stars with 4 leaves, and $5 \cdot 4 \cdot 3$ "stars" with 3 leaves, where one leaf is a chain of two vertices, giving 125 total.

Is there a pattern, or a general formula?

Theorem (Cayley)**.** *There are n^{n-2} trees on n labelled vertices.*

Proof. (Prüfer, 1918). For a tree T, consider its vertex set $V = \{1,2,\ldots,n\}$. Note that the number of sequences of length $n-2$ from $[n]$ is n^{n-2}. We will construct a bijection from the set of trees on n labelled vertices and the set of these sequences.

We convert a labelled tree into a sequence of length $n-2$ by removing the lowest labelled leaf until two vertices remained; each time a leaf is removed, its neighbour is added to the sequence.

To convert a sequence $S = (t_1,t_2,\ldots,t_{n-2})$ into a labelled tree T, let s_1 be the smaller vertex in $V \setminus S$, and we join s_1 to t_1 with an edge. Then, let s_2 be the smaller vertex in $V \setminus \{s_1\} \setminus S$, and join s_2 to t_2. Repeat this process until the elements of S have been exhausted, at which point n_2 edges have been added. Join the two vertices in $V \setminus \{s_1,s_2,\ldots,s_{n-2}\}$ to complete the tree. ■

49.2.7 Shortest Path Algorithm

Given a weighted digraph (G,w) and two vertices $s,t \in V(G)$, how can we find the shortest path from s to t (or decide that no such path exists).

If G is simple, unweighted and undirected, this can be solved using BFS by picking s to be the root node, as shown in Theorem 49.2.6.

Note that if a negative cycle exists, then there is no solution to this problem, as the path can be made arbitrarily negative by traversing the cycle arbitrarily many times.

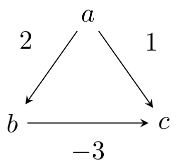
If we have an instance of this problem where the weights are non-negative, then we can solve this problem with *Dijkstra's algorithm.*

1. Mark all nodes as *unvisited.*
2. Assign to every node a *tentative distance*; set this value to 0 for the starting node, and infinity to all other nodes. As the algorithm progresses, this value represents the length of the shortest path from the starting node to the given node discovered. Since no path is known initially, (apart from the starting node, with path length 0), all other tentative distances are set to infinity.
3. Also assign each node a *previous node*, representing the vertex preceding it in the path. At the beginning, this value is undefined for each vertex.
4. Set the starting node as the *current node.*
5. For the current node, consider all of its unvisited neighbours, and calculate their tentative distances through the current node. That is, the add the tentative distance of the current node to the weight of the edge connecting the current node to that neighbour. If this tentative distance is lower than the one currently assigned to that neighbour, overwrite it, and also set the previous node of the neighbour to be the current node.
6. Once all neighbours have been visited, mark the current node as visited.

7. If the destination node has been marked visited or the smallest tentative distance among the unvisited nodes is infinity (this happens if the graph is disconnected and no path exists from the starting node to infinite tentative distance node), then terminate the algorithm.
8. Otherwise, select the unvisited node with the minimum tentative distance as the new current node, and return to step (5).

This method of approximating and updating tentative distances is known as a *relaxation* method.

Note that Dijkstra's algorithm does *not* work on digraphs with negative weights, as, once a vertex is marked as visited, it is never searched again, as it assumes that the path developed to this vertex is most efficient. However, this is not necessarily true with negative weights, as an overall more efficient longer path with negative weights may exist that is locally less efficient.



For instance, in this graph, starting at a, the algorithm would search first search c, then declare it visited. Then, it would search b, and discard the path to c, as it is already marked visited.

Algorithm 19 Dijkstra's Algorithm

```
procedure BFS($(G,w)$, start, end)
    for $v \in V(G)$ do
        $v$.visited = false
        $v$.distance = $\infty$
        $v$.previous = null
    end for
    start.distance = 0
    current = start
    while end.distance = $\infty$ do
        for $v \in N$(current) do
            if $v$.unvisited = false then
                newDist = current.distance + $w$(current,$v$)
                if newDist < $v$.distance then
                    $v$.distance = newDist
                    $v$.previous = current
                else
                    continue
                end if
            end if
        end for
        current.visited = true
        if $\min_{\{v:v.\text{visited}=\text{false}\}}(v.\text{distance}) = \infty$ then
            break
        end if
        current = $\min_{\{v:v.\text{visited}=\text{false}\}}(v.\text{distance})$
    end while
end procedure
```

We can also run Dijkstra's algorithm without giving a destination node, in which case, we change the termination condition to when all nodes have been visited, or if the smallest tentative distance among the unvisited nodes is infinity. This alternative version would find the shortest path from a source vertex to all other vertices.

Proof. G is finite, so the algorithm always terminates. Now, we show correctness.

To reduce the mixing of notation, for the purposes of this proof only, let $D(v) = v.\text{distance}$ and $p(v) = v.\text{previous}$. Also let R denote the set of *visited* vertices, and let $P_{[a,b]}$ denote the restriction of a path P to between vertices a and b in that path.

We will show that at any step of the algorithm, if a node v is the current node and s is the starting node, then,

1. $D(v) = d_{(G,w)}(s,v)$
2. If $D(v) < \infty$, then the path $v,p(v),p(p(v)),\dots$ contains s and is furthermore the shortest path from s to v.

(1): We induct on the number of while loop iterations. In the first iteration, the current node is s, and $D(d) = 0 = d(s,s)$.

Suppose a vertex v is selected, but the shortest path P from s to v has length $w(P) < D(v)$. If all vertices of P (except for v) have been visited, then $d(v) = w(P)$ by induction. Otherwise, let y be the first unvisited vertex of P, and let $x = p(y)$ be its predecessor.

$$s \longrightarrow \cdots\rightarrow x \longrightarrow y \cdots\rightarrow \longrightarrow v$$

$$\begin{aligned} D(y) &\leq D(x) + w\big((x,y)\big) \\ &= d_{(G,w)}(s,x) + w\big((x,y)\big) \\ &= w(P_{[s,x]}) + w\big((x,y)\big) \\ &\leq w(P) \\ &< D(v) \end{aligned}$$

contradicting the choice of v to be the current vertex.

(2): If $D(v) < \infty$, then $D(v)$ was decreased at some point, where $p(v)$ was also created.

The values of $D(v)$ and $p(v)$ can change several times before v is visited, but after the last change, $D(v) = d_{(G,w)}(s,v)$ by (1). Also, the sequence $p(v),p(p(v)),\dots$ contains s and defines a shortest path from s to $p(v)$ by induction, since $p(x)$ is visited for all visited x (with $p(v)$ being the base case). Thus, the sequence $v,p(v),p(p(v)),\dots$ contains s and defines a shortest path from s to v. ■

This implementation of Dijkstra's algorithm runs in $\Theta(|V|^2)$ time, where $|V|$ is the number of vertices in the graph. However, this can be optimised with the use of Fibonacci heap min-priority queues, running in $\Theta(|E| + |V|\log|V|)$ time. This variant is asymptotically the fastest known single-source shortest-path algorithm for arbitrary directed graphs with unbounded non-negative weights.

Another algorithm, is the *Bellman-Ford algorithm.* Bellman-Ford is slower than Dijkstra, but it works on a larger class of problems. Notably, it can handle graphs that contain negative weights, and can also detect negative cycles.

Like Dijkstra's algorithm, Bellman-Ford proceeds by relaxation. In Dijkstra's algorithm, this is done greedily by selecting the closest vertex that hasn't been searched yet in a priority queue. Bellman-Ford just relaxes all edges, and does so $|V| - 1$ times.

Bellman-Ford also requires a cycle detection subroutine, of which $O(|V|)$ solutions are known.

Algorithm 20 Bellman-Ford

```
 1: procedure BELLMANFORD(($G$,$w$), start)
 2:     for $v \in V(G)$ do
 3:         $v$.visited = false
 4:         $v$.distance = $\infty$
 5:         $v$.previous = null
 6:     end for
 7:     start.distance = 0
 8:     current = start
 9:     $i = 0$
10:     while $i \leq |V| - 1$ : do
11:         for $(u,v) \in E(V)$ do
12:             if $u$.distance + $w((u,v))$ < $v$.distance then
13:                 $v$.distance = $u$.distance + $w((u,v))$
14:                 $v$.previous = $u$
15:             end if
16:         end for
17:         $i = i + 1$
18:     end while
19:     for $(u,v) \in E(G)$ do
20:         if $u$.distance + $w((u,v))$ < $v$.distance then
21:             negativeLoop = [$v$,$u$]
22:             $i = 0$
23:             while $i \leq |V| - 1$ : do
24:                 $u$ = negativeLoop[0]
25:                 for $(u,v) \in E(G)$ do
26:                     if $u$.distance + $w((u,v))$ < $v$.distance then
27:                         negativeLoop = [$v$].append(negativeLoop)
28:                     end if
29:                 end for
30:                 $i = i + 1$
31:             end while
32:             return CYCLEDETECT(negativeLoop)
33:         end if
34:     end for
35: end procedure
```

Proof. G is finite, so the algorithm always terminates. Now, we show correctness.

To reduce the mixing of notation, for the purposes of this proof only, let $D(v) = v$.distance.

We induct on the number of iterations n of the for loop in line 11. In the zeroth iteration, the starting vertex has distance 0, which is correct. For other other vertices u, $D(u) = \infty$, which is also correct because there is no path from source to u with 0 edges.

For the base case, consider when a vertex's distance is updated by $D(v) = D(u) + w((u,v))$. By the induction hypothesis, $D(u)$ is the length of a path from the starting vertex to u. Then, $D(u) + w((u,v))$ is the length of the path from the starting vertex to v that follows the path from the starting vertex to u then to v.

Now, consider a shortest path P from the starting vertex to v with at most n edges. Let u be the last vertex before v on this path. Then, the section of the path from the start to u is a shortest path from the start to u with at most $n - 1$ edges, since if it were not, then there would be a path from the start to

u to which we could append the edge (u,v) to construct a path from the start to v strictly shorter than P, contradicting the choice of P. By the induction hypothesis, $D(u)$ after $n-1$ iterations is at most the length of this path from the start to u. It follows that $D(u) + w((u,v))$ is at most the length of P. In the nth iteration, $D(v)$ is compared to $D(u) + w((u,v))$, and is set to that, if it is shorter. So, after n iterations, $D(v)$ is at most the length of P, which is the length of a shortest path from the start to v with at most n edges, as required.

If there are no negative cycles, then every shortest path visits each vertex at most once, so in the for loop on line 19, no further improvements can be made. Now, suppose no improvements can be made. Then, for any cycle $v_1,v_2,\ldots,v_{k-1}$, we have,

$$D(v_i) \leq D(v_{i-1 \pmod k}) + w\big((v_{i-1 \pmod k},v_i)\big)$$

Summing over the cycle, the $D(v_i)$ and $D(v_{i-1 \pmod k})$ terms cancel, leaving,

$$0 \leq \sum_{i=1}^{k} w\big((v_{i-1 \pmod k},v_i)\big)$$

so the cycle is non-negative. It follows that the algorithm returns a cycle if and only if it is negative. ■

Bellman-Ford runs in $O(|V| \cdot |E|)$ time.

A graph is *locally finite* if every vertex in the graph has finite degree.

Lemma (Kőnig)**.** *Suppose G is connected, infinite, and locally finite. Then, G contains a ray.*

Proof. We give an inductive algorithm to generate such a ray.

Pick any vertex, $v_0 \in V(G)$. This vertex can be thought of a path of zero length, consisting of one vertex and no edges. By the assumptions of the lemma, each of the infinitely many vertices of G can be reached by a simple path that starts from v_0.

Next, as long as the current path ends at some vertex v_i, consider the infinitely many vertices that can be reached by paths that extend the current path, and for each of these vertices, construct a path to it that extends the current path. There are infinitely many of these extended paths, each of which connects from v_i to one of its neighbours, but v_i only has finitely many neighbours. It follows from the set-theoretic variant of the pigeonhole principle that at least one of these neighbours is used as the next step of infinitely many of these extended paths. Let v_{i+1} be such a neighbour, and extend the current path along the edge from v_i to v_{i+1}. By construction, this extension preserves the property that infinitely many vertices can be reached by paths that extend the current path.

Repeating this process for extending the path produces an infinite sequence of finite paths, each extending the previous path in the sequence by one edge. The union of these paths gives the required ray. ■

Corollary 49.2.10.1. *Every infinite tree contains either a vertex of infinite degree, or an infinite path.*

Proof. If the tree is locally finite, the lemma above applies, and thus contains a ray. Otherwise, it is not locally finite and contains a vertex of infinite degree. ■

49.2.8 Network Flow

A *network* (G,u,s,t) is a directed graph G with two distinguished nodes called the *source* node, s, and the *sink* node, t, along with a function $u : E(G) \to \mathbb{R}_{\geq 0}$ called the *edge capacity* function.

A *flow* in a network (G,u,s,t) is a function $f : E(G) \to \mathbb{R}_{\geq 0}$ such that $f(e) \leq u(e)$ for all $e \in E(G)$: that is, the flow over an edge cannot be higher than its capacity. This is called the *capacity constraint.*

A flow must also satisfy the *skew symmetry constraint*: $f\big((u,v)\big) = -f\big((v,u)\big)$. That is, the flow on an edge from a vertex u to a vertex v is equivalent to the negation of the flow from v to u.

A flow is *integral* if every edge is assigned an integer – that is, an integral flow is instead a function $f : E(G) \to \mathbb{Z}$.

A network equipped with a flow function is called a *flow network*.

Recall that a cut is a partition of a vertex set of a graph into two partites. If $X \subseteq V(G)$, then X and $V(G) \setminus X$ partition $V(G)$, thus defining a cut. Because this cut is determined entirely by X, we also denote it by X. Recall further that an edge is in the cut if it connects a vertex in one partite to a vertex in the other, and the set of edges in the cut X is denoted $\delta(X)$.

However, because G is directed, we can divide this set further. We let $\delta^+(X) \subseteq \delta(X)$ denote the set of edges leaving X, and $\delta^-(X) \subseteq \delta(X)$ denote the set of edges entering X.

If X is a singleton set containing the sole vertex v, we abbreviate $\delta(\{v\})$ as $\delta(v)$.

We define the *excess* function $x_f : V \to \mathbb{R}$ by,

$$x_f(v) \coloneqq \underbrace{\sum_{e\in\delta^-(v)} f(e)}_{\text{flow received by } v} - \underbrace{\sum_{e\in\delta^+(v)} f(e)}_{\text{flow sent by } v}$$

A node v is said to be *active* if $x_f(v) > 0$ (the node consumes flow), *deficient* if $x_f(v) < 0$ (the node produces flow), or *conserving* if $x_f(v) = 0$. A conserving node is said to satisfy the *flow conservation rule*. In flow networks, the source s is deficient and the sink t is active.

If every node apart from the source and sink is conserving, the flow is said to be a *feasible flow*. We only consider feasible flows, and shorten the name to just *flow*.

An *s-t-flow* in (G,u,s,t) is a flow f such that $x_f(s) < 0$, and $x_f(v) = 0$ for all $v \neq s,t$. The *value* of such a flow is the excess at the sink t:

$$\text{value}(f) \coloneqq x_f(t)$$

or equivalently, the negative of the excess at the source s:

$$\text{value}(f) \coloneqq -x_f(s)$$

Given a network (G,u,s,t), the *maximum flow problem* is to find an s-t-flow of maximum value.

An *s-t-cut* is a $\delta^+(A)$ for some $A \subseteq V(G)$ such that $s \in A$, $t \notin A$. A *minimum s-t*-cut is an s-t-cut of minimum total capacity.

49.2.8.1 Residual Networks

Let (G,u,s,t) be a network. For an edge $e = (x,y) \in E(G)$, let $\overleftarrow{e} = (y,x)$ denote the *reverse edge*.

Let $\overleftrightarrow{G}$ be the graph contained from G by adding the reverse edge for every edge of G. Note that $\overleftrightarrow{G}$ may be a multigraph, as there may now be parallel edges.

Given a network (G,u,s,t), and a flow f in it, the *residual network* (G_f,u_f,s,t) is defined by,

- $V(G_f) = V(G)$;
- $E(G_f) = \Big\{e \in E\big(\overleftrightarrow{G}\big) : u_f(e) > 0\Big\}$;
- $u_f(e) \coloneqq u(e) - f(e)$ for $e \in E(G)$;

- $u_f(\overleftarrow{e}) := f(e)$, where $\overleftarrow{e}$ is a reverse edge.

The residual network indicates the additional possible flow in the original network. If there is a path from source to sink in the residual network, then it is possible to add flow. The value of an edge in the residual graph is called the *residual capacity*, which is equal to the original capacity of the edge, minus the current flow given by f. An *f-augmenting path* is a path from s to t in the residual network.

Let f be a flow, P be an f-augmenting path, and let $0 < \gamma \leq \min\limits_{e \in E(P)} \big(u_f(e)\big)$.

We *augment f along P by γ*, by,

- Increasing $f(e)$ by γ for each $e \in E(P) \cap E(G)$;
- Decreasing $f(e)$ by γ for each $\overleftarrow{e} \in E(P)$.

We now have enough machinery to tackle the maximum flow problem.

Algorithm 21 Ford-Fulkerson Algorithm

```
1: Set $f(e) = 0$ for all $e \in E(G)$.
2: Find an $f$-augmenting path $P$. If none exist, then terminate the algorithm.
3: Compute $\gamma := \min_{e \in E(P)} \big(u_f(e)\big)$.
4: Augment $f$ along $P$ by $\gamma$, and return to line (2).
```

Lemma 49.2.11. *For any $A \subset V(G)$ such that $s \in A$ and $t \notin A$, and any s-t-flow f, we have,*

1.

$$\text{value}(f) = \sum_{e \in \delta^+(A)} f(e) - \sum_{e \in \delta^-(A)} f(e)$$

2.

$$\text{value}(f) \leq \sum_{e \in \delta^+(A)} u(e)$$

Note that (2) in the lemma above states that the value of a maximum s-t-flow cannot exceed the capacity of a minimum s-t-cut.

Proof. (1):

$$\begin{aligned} \text{value}(f) &= -x_f(s) \\ &= \sum_{e \in \delta^+(s)} f(e) - \sum_{e \in \delta^-(s)} f(e) \end{aligned}$$

Because $x_f(v) = 0$ for all $v \neq s$,

$$= \sum_{v \in A} \left(\sum_{e \in \delta^+(v)} f(e) - \sum_{e \in \delta^-(v)} f(e) \right)$$

For each $e = (x,y)$ with $x,y \in A$, $f(e)$ appears once positively and once negatively, so,

$$= \sum_{e \in \delta^+(A)} f(e) - \sum_{e \in \delta^-(A)} f(e)$$

(2) follows from (1) by using $0 \leq f(e) \leq u(e)$ for all $e \in E(G)$. ■

Theorem 49.2.12. *An s-t-flow is maximum if and only if there is no f-augmenting path.*

Proof. If there is no f-augmenting path, then t is not reachable from s in G_f. Let R be the set of vertices reachable from s in G_f. By the definition of G_f,

$$\forall e \in \delta_G^+(R), f(e) = u(e)$$

else $e \in G_f$, in which case, there is a vertex not in R reachable from s. We also have,

$$\forall e \in \delta_G^-(R), f(e) = 0$$

else $\overleftarrow{e} \in G_f$, in which case, there is a vertex not in R reachable from s. Then, by (1) of the above lemma, we have,

$$\text{value}(f) = \sum_{e\in\delta^+(A)} u(e)$$

and hence by (2) of the above lemma, f is maximum. ■

Remark.

1. If we allow irrational capacities, the Ford-Fulkerson algorithm may not terminate at all.
2. Even in the case of integer capacities, the number of augmentations can be exponential.
3. The maximum flow problem admits a polynomial-time implementation.

Theorem (Max-Flow Min-Cut). *In a network, the maximum value of an s-t-flow equals the minimum capacity of an s-t-cut.*

Theorem (Integral Flow). *If each edge in a flow network has integer capacity, then there exists an integral maximum flow.*

Note that this theorem does not say that the *value* of the flow is an integer (which follows directly from the max-flow min-cut theorem), but that the flow on *every edge* is an integer.

Theorem (Flow Decomposition). *Let (G,u,s,t) be a network, and let f be an s-t-flow in G. Then, there exists a family P^* of s-t-paths and a family C^* of cycles in G, along with a function $\omega : P^* \cup C^* \to \mathbb{R}_{\geq 0}$, such that,*

1.

$$f(e) = \sum_{\substack{K\in P^*\cup C^*\\ e\in K}} \omega(K)$$

2.

$$\text{value}(f) = \sum_{k\in P^*} \omega(K)$$

Moreover, if f is integral, then ω can be chosen to be integral.

Proof. We construct P^*, C^*, and ω by induction on the number of edges with non-zero flow. Let $e_0 = (v_0,w_0)$ be an edge with $f(e_0) > 0$. If $w_0 = t$, then we stop. Otherwise, there exists an edge $e_1 = (w_0,w_1)$ with $f(e_1) > 0$. If $w_1 = t$ r $w_1 = v_0$, then we stop. Otherwise, there exists an edge $e_2 = (w_1,w_2)$ with $f(e_2) > 0$. If $w_2 = t$ or $w_2 \in \{v_0,w_0\}$, then we stop. Continuing this process, in at most n steps, we either find a cycle, or reach vertex t. In the latter case, we repeat the procedure in the other direction and either find a cycle, or reach vertex s. In either case, we find either a cycle L, or a path L from s to t.

Set $\omega(L) = \min_{e\in L}\big(f(e)\big)$. For every $e \in L$, define $f'(e) := f(e)-\omega(L)$, and for all $e \notin L$, define $f'(e) := f(e)$.

There are strictly fewer edges of G with non-zero flow f', so, by the induction hypothesis, (1) and (2) holds for the flow f'.

We show that (1) also holds for f. If $e \notin L$, then (1) is valid for f, because in this case, $f(e) = f'(e)$. Let $e \in L$, and denote the members of $P^* \cup C^*$ containing e by $K_1,K_2,\dots,K_t$, where $K_t = L$. By induction,

$$f'(e) = \sum_{i=1}^{t-1} \omega(K_i)$$

and by definition, $f'(e) = f(e) - \omega(L)$. Therefore,

$$\begin{aligned} f(e) &= \omega(L) + f'(e) \\ &= \omega(K_t) + \sum_{i=1}^{t-1} \omega(K_i) \\ &= \sum_{i=1}^{t} \omega(K_i) \end{aligned}$$

so (1) holds for f.

Now, we show that (2) also holds for f. Suppose that L is a cycle. Then, G contains a cut $\delta(A)$ separating s from t which does not cross L, and hence value(f) = value(f') by claim (1) of the above lemma. Since (2) holds for f', we conclude it also holds for f in this case.

Now, suppose L is instead a path. Then,

$$\begin{aligned} \text{value}(f') &= \sum_{K \in P^* \setminus \{L\}} \omega(K) \\ &= \sum_{e \in \delta^+(A)} f'(e) - \sum_{e \in \delta^-(A)} f'(e) \\ &= \sum_{e \in \delta^+(A)} f(e) - \sum_{e \in \delta^-(A)} f(e) - \omega(L) \\ &= \text{value}(f) - \omega(L) \end{aligned}$$

and hence (2) holds for f. ■

Theorem (Menger (Edge Connectivity))**.** *Let G be a graph (directed or undirected), and let s,t be two distinct vertices of G. Let $k \in \mathbb{N}$. Then, these two statements are equivalent:*

1. *There are k edge-disjoint s-t-paths in G.*
2. *After deleting any $k-1$ edges from G, t is still reachable from s (e.g. G is connected).*

If the latter property holds in a graph for all s,t, then the graph is said to be *k-edge connected.*

Proof. (1) → (2) is trivial, because to destroy k edge-disjoint paths, at least k edges must be deleted (one per path).

(2) → (1): First, let G be directed. By assigning capacity $u(e) = 1$ to every edge $e \in E(G)$, we produce the network $G^* = (G,u,s,t)$. The capcity of a cut in this network is just the number of edges in the cut.

Assuming (2), the minimum capacity of a directed s-t-cut is at least k. By the max-flow min-cut theorem, G^* has an (integral) flow f of value at least k. Then, by the flow decomposition theorem,

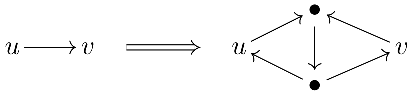
$$\text{value}(f) = \sum_{L \in P^*} \omega(L)$$

where P^* is a family of s-t-paths, and $\omega(L) = 1$ for all $L \in P^*$. The members of P^* are edge disjoint, because by the same theorem,

$$f(e) = \sum_{L \in P^* \cup C^*} \omega(L)$$

It follows that G contains k edge-disjoint paths.

Now, let G be undirected. Transform G into a directed graph $\overrightarrow{G}$ by replacing every edge as follows:



If (2) holds on G, then (2) also holds for $\overrightarrow{G}$. So, $\overrightarrow{G}$ has k edge-disjoint s-t-paths, and hence G has k edge-disjoint s-t-paths. ■

Corollary 49.2.12.1. *An undirected graph G on at least two vertices is k-edge connected if and only if for each pair of distinct vertices s and t, there are k edge-disjoint s-t-paths.*

Proof. Follows directly from the theorem. ■

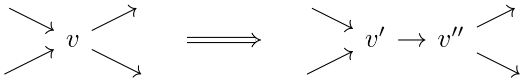
Theorem (Menger (Vertex Connectivity))**.** *Let G be a graph (directed or undirected), and let s,t be two non-adjacent vertices of G. Let $k \in \mathbb{N}$. Then, these two statements are equivalent:*

1. *There are k vertex-disjoint s-t-paths in G.*
2. *After deleting any $k-1$ vertices (distinct from s or t) from G, t is still reachable from s (e.g. G is connected).*

If the latter property holds in a graph for all non-adjacent s,t, then the graph is said to be *k-vertex connected.*

Proof. (1) → (2) is trivial, because to destroy k vertex-disjoint paths, at least k vertices must be deleted (one per path).

(2) → (1): First, let G be directed. Transform G into a new graph G' by replacing each vertex of G as follows:



Supose G' contains $k-1$ edges whose deletion makes t' unreachable from s''. Then, G has at most $k-1$ vertices whose deletion makes t unreachable from s. Because this contradicts (2), we conclude that after deleting any $k-1$ edges from G', t' is still reachable from s''. From the edge connectivity statement of Menger's theorem, G' has k edge-disjoint s''-t'-paths. It should be clear that these paths must also be vertex disjoint. It follows that G contains k vertex-disjoint s-t-paths.

The undirected version follows from the directed one in by the same construction as in the proof of the edge connectivity statement. ■

Corollary 49.2.12.2. *An undirected graph G on at least k vertices is k-vertex connected if and only if for each pair of distinct vertices s and t, there are k vertex-disjoint s-t-paths.*

Proof. Suppose G is k-vertex connected, but there exists vertices s and t in G such that there are not k vertex-disjoint s-t-paths. If s is not adjacent to t, then we apply the vertex connectivity statement of Menger's theorem to conclude that there is a set $U \subset V(G)$ of at most $k-1$ vertices such that $G \setminus U$ is disconnected, giving a contradiction.

Instead suppose that s and t are adjacent, and denote this edge e. By deleting e from G, we obtain a graph G' such that there are not $k-1$ vertex-disjoint s-t-paths. We again apply Menger's theorem to G', concluding there exists a set $X \subset V(G')$ of at most $k-2$ vertices such that $G' \setminus X$ is disconnected. Denote by S the connected component of $G' \setminus X$ containing s, and by T the connected component of $G' \setminus X$ containing T. At least one of them contains a vertex v different from s and t because $|V(G')| > k$. Without loss of generality, suppose that v is unreachable from s in $G' \setminus X$. Then, s and v are in different components of $G \setminus (X \cup \{t\})$, contradicting the assumption that G is k-vertex connected, as $|X \cup \{t\}| \leq k-1$. This completes the forward implication.

We prove the reverse implication by contraposition. Suppose G is not k-vertex connected, so there exists a set U of at most $k-1$ vertices such that $G \setminus U$ is disconnected. Take s from one connected component of $G \setminus U$, and t from another. Then, G has no k-vertex-disjoint s-t-paths. ■

The *vertex-connectivity* of a graph G is the maximum k such that G is k-(vertex)-connected. The *edge-connectivity* is similarly the maximum k such that G is k-edge connected.

49.2.9 Matchings

Let G be a simple graph. A *matching* in G is a subset $M \subseteq V(E)$ such that no two edges in M are incident to the same vertex. We say that a vertex v is *covered* by a matching M if v is incident to an edge $e \in M$. The *matching number* of a graph is the size of a maximum matching.

A matching is *perfect* if it covers all vertices of the graph. A perfect matching is only possible on graphs with an even number of vertices.

Given a simple graph G, the *maximum matching problem* is to find a matching of maximum cardinality in G.

Recall that a subset of vertices $S \subseteq V$ is an *independent set* of the graph if there are no edges between any pair of vertices in S, and that a graph is *bipartite* if its vertex set can be partitioned into two independent sets.

- The path graph P_n is bipartite for any value of n.
- The cycle graph C_n is bipartite for even values of n

Theorem (Characterisation of Bipartite Graphs)**.** *A graph is bipartite if and only if every closed walk in the graph is of even length.*

Proof. Suppose G is a bipartite graph with partites L and R, and let $C = (c_1,c_2,\ldots,c_k)$ be a closed walk in G. Without loss of generality, suppose we have $c_1 \in L$. Then, because G is bipartite, we have $c_2 \in R$, $c_3 \in L$, $c_{2n} \in R$, $C_{2n+1} \in L$. Because the walk is closed, it must be the case that $c_k \in R$, so k must be even. This completes the forward implication.

Now, suppose G is a simple graph with no closed walks of odd length. Without loss of generality, G is connected. Let $v,x,y \in G$.

Let P_x be a shortest path connecting v to x, and let P_y be a shortest path connecting v to y.

Let z be a vertex in both P_x and P_y closest to x and y. Then, $d(z,x)$ and $d(z,y)$ have the same parity. It follows that x and y are not adjacent, or else an odd cycle is created. This suggests the set $V_1 = \{u \in V(G) : d(u,v) \equiv 1 \pmod 2\}$ is an independent set. Through a similar argument, $V_2 = \{u \in V(G) : d(u,v) \equiv 0 \pmod 2\}$ is also an independent set. These two sets are clearly disjoint and together cover the vertex set of G, so they partition the vertex set of G and hence G is bipartite. ■

Corollary 49.2.12.3. *A graph is bipartite if and only if every cycle in the graph is of even length.*

Proof. The cycles in a graph are a strict subset of the closed walks. ■

In a graph G, a *vertex cover* is a subset $S \subseteq V(G)$ such that every edge in $E(G)$ is incident to at least one vertex in S. The *vertex cover number* is the size of a minimum vertex cover.

Lemma 49.2.13. *For all simple graphs G, the following statements hold:*

1. *A set $S \subseteq V(G)$ is a vertex cover if and only if $V(G) \setminus S$ is an independent set.*
2. *The sum of the independence number and the vertex cover number is equal to the number of vertices in G.*
3. *The vertex cover number is at most twice the matching number.*

Proof. (1) and (2) follow from the definitions of a vertex cover and an independent set.

(3): Let M be a maximal matching. Let V_m be the set of vertices incident with edges of M. Since M is maximal, every edge of the graph is incident to a vertex of V_m, and hence V_m is a vertex cover with $|V_m| = 2|M|$. Some of these vertices may not be required to form a vertex cover, so this gives an upper bound on the minimum vertex cover V, namely $|V| \leq 2|M|$. ■

If G is bipartite we can tighten this bound to an equality.

Theorem (Kőnig). *In any bipartite graph, $|M| = |S|$ for a maximum matching M and minimum vertex cover S.*

Proof. Let $G = (L \cup R, E)$ be a bipartite graph. Denote by G' the graph obtained from G by adding two vertices s and t, connecting s to every vertex of L, and connecting t to every vertex of R.

Then, the maximum number of vertex-disjoint s-t-paths in G' is equal to the matching number of G. The minimum number of vertices whose deletion makes t unreachable from s is also equal to the vertex cover number of G.

It follows from the vertex connectivity statement of Menger's theorem that these two values are equal. ■

49.2.9.1 Hall's Condition

Theorem (Hall). *Let $G = (L \cup R, E)$ be a bipartite graph. Then, G admits a matching covering L (an L-perfect matching) if and only if for all $X \subseteq L$, we have,*

$$|N(X)| \geq |X|$$

Proof. If G has an L-perfect matching, then $|N(X)| \geq |X|$ holds for all $X \subseteq L$ trivially.

Now, suppose $|N(X)| \geq |X|$ holds for all $X \subseteq L$, but there does not exist an L-perfect matching. Then by Kőnig's theorem, the vertex cover number is less than $|L|$.

Let $A \subseteq L$ and $B \subseteq R$ such that $A \cup B$ is a vertex cover of size $|A \cup B| \leq |L|$. Because G is bipartite, $N(L \setminus A) \subseteq B$, so it follows that,

$$\begin{aligned} |N(L \setminus A)| &\leq |B| \\ &< |L| - |A| \\ &= |L \setminus A| \end{aligned}$$

■

We can restate Hall's theorem in set-theoretic terms.

Consider a family of sets, S, with $A_1, A_2, \cdots A_n \subseteq S$. A *system of distinct representatives* (an *SDR*) is a set of distinct elements, $\{x_1, x_2, \ldots, x_n\} \subseteq S$, such that for all $i \in [1,n]$, $x_i \in A_i$.

A family of sets, S, satisfies *Hall's condition* if, for each subfamily $W \subset S$, we have,

$$|W| \leq \left| \bigcup_{A \in W} A \right|$$

A family of sets admits an SDR if and only if Hall's condition is satisfied. That is, there exists an SDR for a family of sets $A_1, A_2, \cdots A_n$ if the union of any k of these sets contains at least k elements for all $k \in [1,n]$.

We now give necessary and sufficient conditions for the existence of a perfect matching.

Theorem 49.2.14. *A bipartite graph $G = (L \cup R,E)$ admits a perfect matching if and only if $|A| = |B|$ and $|N(X)| \geq |X|$ for all $X \subseteq L$.*

Remark. If $G = (L \cup R,E)$ is k-regular, then $|E| = k|L| = k|R|$, and hence $|L| = |R|$. This allows us to rephrase the previous theorem.

Theorem 49.2.15. *Every regular bipartite graph has a perfect matching.*

Proof. Let $G = (L \cup R,E)$ be a k-regular bipartite graph. Let $X \subseteq L$. Because G is k-regular, there are $k|X|$ edges connected to $X \subseteq L$, and $k|N(X)|$ edges connected to $N(X) \subseteq R$. The former set of edges is contained within the latter, so $k|X| \leq k|N(X)|$, and hence $|X| \leq |N(X)|$, satisfying Hall's condition. By Hall's theorem, there exists a matching on L, so every vertex in L is paired with a vertex in R. But, $|L| = |R|$, so the matching is perfect. ■

Theorem 49.2.16. *The maximum matching problem can be solved for bipartite graphs with n vertices and m edges in $O(nm)$ time.*

Proof. Let $G = (L \cup R,E)$ be a bipartite graph. Construct a network G^* by:

- Adding a source s and connecting it to every vertex of L;
- Adding a sink t and connecting it to every vertex of R;
- Orienting all edges to point from s to A, from A to B, and from B to t;
- Defining the capacity function $u : E \to \mathbb{R}_{\geq 0}$ by $u(e) = 1$ for all edges $e \in E$.

Since all capacities are integers, there exists an integral maximum flow f. Because of flow conservation, the edges of G with a non-zero flow form a matching. Since the flow is maximum, the matching is maximum.

The maximum is attained after at most n augmentations in the Ford-Fulkerson algorithm. Since each augmentation takes $O(m)$ time, the total time complexity of finding a maximum matching in G is $O(nm)$. ■

49.2.9.2 Maximum Independent Set

Given a simple graph G, the *maximum independent set problem* is to find an independent set in G of maximum cardinality.

Recall:

- For any bipartite graph, the matching number is equal to the vertex cover number (Kőnig's theorem)
- For any graph, the sum of the independence number and the vertex cover number is equal to the number of vertices in the graph.

Theorem 49.2.17. *The maximum independent set problem can be solved for bipartite graphs with n vertices and m edges in $O(nm)$ time.*

49.2.9.3 Augmenting Paths

Let G be a graph, and M a matching in G. A path P is an *M-alternating chain* if $E(P) \setminus M$ is a matching. An M-alternating chain is additionally *M-augmenting* if its endpoints are not covered by M. That is, P is M-alternating if its edges alternate between being in and not in M. If *both* endpoints are not in M, then P is additionally M-augmenting.

Theorem (Berge)**.** *A matching M is maximum if and only if there are no M-augmenting chains.*

Proof. If an M-augmenting chain, P, exists, then $(M \setminus P) \cup (P \setminus M)$ is a matching of cardinality strictly greater than M, so M is not maximum. Intuitively, we simply flip the edges in the augmenting chain, and take the result to be the new matching.

Conversely, if M is not maximum, and M' is a matching such that $|M'| > |M|$, then $(M \setminus M') \cup (M' \setminus M)$ consists of vertex-disjoint alternating cycles and alternating paths, where at least one path has more edges in M' than in M. This path is M-augmenting. ■

49.2.9.4 Maximum Weight Matching

Given a simple weighted graph (G,w), the *maximum weight matching problem* is to find a matching in G of maximum total weight. Conversely, the *minimum weight perfect matching problem* is to find a matching in G of minimum total weight, or decide that G has no perfect matching.

Theorem 49.2.18. *The maximum weight matching problem is equivalent to the minimum weight perfect matching problem.*

Proof. Let (G,w) be an instance of minimum weight perfect matching, and let $K = 1 + \sum_{e \in E(G)} |w(e)|$. If $w'(e) = K - w(e)$ for each edge $e \in E(G)$, then any maximum weight matching in (G,w') gives a solution to the minimum weight perfect matching in (G,w).

Let (G,w) be an instance of maximum weight matching. Then, add $|V(G)|$ new vertices to G and all possible edges to create a complete graph G' on $2|V(G)|$ vertices. Define $w'(e) = -w(e)$ for the original edges of G and $w'(e) = 0$ for new edges. Then, a minimum weight perfect matching in (G',w') yields a maximum weight matching in G by deleting the edges not in G. ■

49.2.9.5 Maximum Independent Set

Given a simple graph G, the *conjugate*, *adjoint*, or *line graph* of G is a graph $\mathrm{L}(G)$ that represents the adjacencies between edges of G. We construct $\mathrm{L}(G)$ as follows:

- For each edge in G, we have a vertex in $\mathrm{L}(G)$;
- For every pair of edges in G that are incident to the same vertex, we include an edge between their corresponding vertices in $\mathrm{L}(G)$.

Remark. The claw graph $K_{1,3}$ is not a line graph, so any graph containing the claw as an induced subgraph is not a line graph.

A graph that does not contain the claw as an induced subgraph is called a *claw-free graph.*

Given a simple graph G, the *maximum independent set problem* is to find an independent set in G of maximum cardinality.

Theorem 49.2.19. *The maximum independent set problem restricted to the class of line graphs is equivalent to the maximum matching problem.*

Proof. $M \subseteq E(G)$ is a matching G if and only if M is an independent set in $\mathrm{L}(M)$. That is, finding a maximum matching in G is equivalent to finding a maximum independent set in $\mathrm{L}(G)$. ■

Let $G = (V,E)$ be a graph, and let $S \subseteq V$ be an independent set. Let H be a bipartite subgraph of G with partites A and B such that,

- $A \subseteq S$;
- $B \subseteq V \setminus S$;
- $\forall e \in B : N(e) \cap (S \setminus A) = \varnothing$ – the vertices of B do not have neighbours in $S \setminus A$;
- $|A| < |B|$.

Then, H is an *augmenting graph* for S.

Corollary (Characterisation of Maximum Independent Sets)**.** *An independent set S is maximum if and only if there is no augmenting graphs for S.*

Proof. If there is an augmenting graph for an independent set S, then S is not maximum, because $(S \setminus A) \cup B$ is a larger independent set. This proves the forward implication by contraposition.

If S is not maximum, and R is a larger independent set, then the bipartite graph with partites $S \setminus R$ and $R \setminus S$ is augmenting for S. This proves the reverse implication by contraposition. ■

The class of line graphs is a subclass of claw-free graphs.

Remark. Every bipartite claw-free graph has vertex degree at most 2. Every connected bipartite claw-free graph is either a path or a cycle. Every connected augmenting graph in the class of claw-free graphs is a path with odd number of vertices.

Theorem 49.2.20. *An independent set S in a claw-free graph is maximum if and only if there is no augmenting path for S.*

Theorem 49.2.21. *The problem of finding augmenting paths in claw-free graphs (and hence line graphs) is solvable in polynomial-time.*

49.2.10 Graph Transformations for Maximum Independent Sets

Lemma 49.2.22. *Let G be a graph and x,y be two adjacent vertices of G. If every vertex z adjacent to x is also adjacent to y, then the independence number of G is equal to the independence number of $G \setminus \{y\}$.*

Proof. Clearly, the independence number of G is at least the independence number of $G \setminus \{y\}$.

To prove the reverse inequality, let $S \subset V(G)$ be an independent set in G. If it does not contain y, then it is also an independent set in $G \setminus \{y\}$. Otherwise, if S contains y, then it contains neither x, nor any neighbour of x. But then, $(S \setminus \{y\}) \cup \{x\}$ is an independent set on $G \setminus \{y\}$ of size $|S|$. Then, the independence number of G is at least the independence number of $G \setminus \{y\}$. ■

Given a vertex x, suppose $N(x) = Y \cup Z$. We *vertex split* x by replacing it by three vertices, x', y, and z, such that $N(x') = \{y,z\}$, $N(y) = Y$, and $N(z) = Z$.

Lemma 49.2.23. *Let G' be the graph obtained by vertex splitting a vertex x in G. Then, the independence number of G' is one greater than the independence number of G.*

Proof. Let S be an independent set in G containing x. Then, $(S \setminus \{x\}) \cup \{y,z\}$ is an independent set in G' of size $|S| + 1$. If S does not contain x, then $S \cup \{x'\}$ is an independent set in G' of size $|S| + 1$. So, the independence number of G' is at least one greater than the independence number of G.

Conversely, let S be an independent set in G'. If it contains at most one vertex in $\{x',y,z\}$, then by deleting this vertex, we obtain an independent set in G of size $|S| - 1$. If S contains two vertices in

$\{x',y,z\}$, then these vertices must be y and z, and hence $(S \setminus \{y,z\} \cup \{x\})$ is an independent set on G of size $|S| - 1$. So, the independence number of G' is at most one greater than the independence number of G. ■

49.2.11 Stable Matching

Given two sets A and B of equal cardinality n, a *matching* is a bijection from the elements of one set to the other. Suppose further that each element $x \in A$ has an ordered list of preferences of elements in B, and similarly, each element in $y \in B$ has an ordered list of preferences of elements in A. If an element a prefers b to c, we write $a : b > c$.

A matching is *stable* if there does not exist elements $x \in A$ and $y \in B$ such that x prefers y over its assigned element and y also prefers x over its assigned element.

The *stable marriage problem* or *stable matching problem* (SMP) is to find a stable matching arrangement for two such sets A and B.

Example. $A = \{x,y\}$, $B = \{u,v\}$,

$x : u > v$

$y : v > u$

$u : x > y$

$v : x > y$

The matching $\{x,v\}$,$\{y,u\}$ is unstable, because x prefers u over v, and u also prefers x over y.

The matching $\{x,u\}$,$\{y,v\}$ is stable, because no pair prefers each other over their assigned elements. △

One algorithm to solve this problem is the *Gale-Shapley algorithm.*

1. At each point of the algorithm, each element is either *fixed* or *free*, with every element initially being free. Elements of A may alternate between being fixed and being free, but elements of B cannot be free after being fixed.
2. In each round of the algorithm, each element $x \in A$ interacts with its preferences in order, provided the preferences haven't been interacted with in previous rounds.
3. If the preference element y is free, the two are matched and both become fixed. Otherwise, y is fixed and already has a match, z. y then compares x to z. Whichever is preferred by y becomes the new match, becoming fixed, and the rejected element becomes free.
4. Repeat until every element is fixed.

Algorithm 22 Gale-Shapley Algorithm

```
procedure SMP($A$, $B$)
    matches = [ ]
    for $n \in A \cup B$ do
        $n$.free = true
    end for
    while $\exists x \in A : x$.free = true do
        $y = x$.preferences.pop()
        if $y$.free = true then
            matches.append(($x$,$y$))
            $x$.free = false
            $y$.free = false
        else if $\exists z : (z,y) \in$ matches then
            if $y : x > z$ then
                matches.remove(($z$,$y$))
                matches.append(($x$,$y$))
                $z$.free = true
            end if
        end if
    end while
end procedure
```

Proof. Each element in A interact at most n times, so the algorithm terminates after at most n^2 operations.

The algorithm stops when all elements are matched, and the two input sets are of equal cardinality, so the produced matching is perfect.

Now, suppose an element $x \in A$ prefers an element $y \in B$ to its assigned element. Then, x interacted with y and, either x was not preferred over the element assigned to y at that time, or, y preferred x over its assigned element, but later changed for a more preferable element. In both cases, y prefers its current assigned element over x, and hence the matching is stable. ■

A stable matching is *optimal* for an element x if there is no stable matching with an assignment x would prefer. Conversely, a stable matching is *pessimal* for x if there is no stable matching with a worse assignment for x. We say that x and y are a stable pair if there exists a stable matching where x and y are matched.

Theorem 49.2.24. *The stable matching produced by the Gale-Shapley algorithm is:*

- *Independent of the order of elements selected to interact;*
- *Optimal for elements of A;*
- *Pessimal for elements of B.*

Proof. Order the elements of A arbitrarily, and let x and y be matched by the algorithm in a stable matching M_1.

Suppose there exists y' such that $x : y > y'$, and suppose that (x,y') is a stable pair, so there exists a stable matching M_2 where x is matched with y'.

Then, x was rejected by y', and without loss of generality, suppose this was the first time a stable pair was rejected by the algorithm.

Now, suppose y' rejected x in favour of x', and let y'' be the match of x' in M_2. Then (x',y'') is also a stable pair. If $x' : y'' > y'$, then x' interacted with y'' before y', which means the stable pair (x',y'') was rejected before (x,y'), contradicting the assumption that (x,y') was the first stable pair rejected by the algorithm.

So, $x' : y' > y''$. But then, the matching M_2 is not stable, because x' and y' are not matched, and they both prefer each other over their assigned elements.

This contradiction shows that every element $x \in A$ is matched with its favourable stable partner; the matching is optimal for elements of A. Because the ordering was arbitrary, any ordering produces the same result.

Now, suppose $y : x > x'$, and suppose the algorithm matches y with x, but there is a stable matching M_3 where y is matched with x'. Let y' be the match of x in M_3. Since the algorithm produces an matching optimal for elements of A, it must be the case that $x : y > y'$. But then, x and y are not matched, and they both prefer each other over their assigned elements, contradicting the stability of M_3. ■

49.2.12 Eulerian Graphs

Recall that an *Eulerian walk* is a trail which traverses every edge. An *Eulerian circuit* is both a trail and cycle which traverses every edge.

A graph that admits an Eulerian walk is *traversable* or *semi-Eulerian.* A graph that admits an Eulerian circuit is *Eulerian.*

Theorem (Euler)**.** *A connected undirected graph admits an Eulerian circuit if and only if the degree of each vertex is even.*

A connected directed graph admits an Eulerian circuit if and only if the in-degree $|\delta^-(v)|$ is equal to the out-degree $|\delta^+(v)|$ for each vertex v.

Proof. The necessity of the degree conditions is obvious. Sufficiency is proved by the following algorithm. ■

Given a connected undirected graph G with even degree vertices, or a digraph with in-degree equal to out-degree for all vertices, *Fleury's algorithm* returns an Eulerian circuit.

1. Start at an arbitrary vertex, v_0.
2. At each step, choose an edge whose deletion would not disconnect the graph, unless no such edge exist, in which case, pick the remaining edge left at the current vertex.
3. Move to the other endpoint of the edge and delete the edge.
4. Now repeat until no edges remain.
5. The sequence from which the edges were chosen forms an Eulerian cycle.

Algorithm 23 Fleury's Algorithm (Undirected)

```
 1: Let $v_0 \in V(G)$ be arbitrary.
 2: procedure FLEURY($G$, $v_0$)
 3:     $W = [v_0]$
 4:     $x = v_0$
 5:     while $E(G) \neq \varnothing$ do
 6:         if $\delta(x) = \varnothing$ then                    ▷ For digraph $G$, check $\delta^+(x) = \varnothing$
 7:             $W = [v_0,e_0,v_1,e_1,\ldots,v_k,e_k,v_{k+1}]$
 8:             for $i = 0$ to $k$ do $W_i =$ FLEURY($G$, $v_i$)
 9:             end for
10:             $W = W_0,e_0,W_1,e_1,\ldots,W_k,e_k,v_{k+1}$
11:             return $W$
12:         else
13:             $e = (x,y),y \in \delta(x)$                       ▷ For digraph $G$, $y \in \delta^+(x)$
14:             $W = W,e,y$
15:             $x = y$
16:             $E(G) = E(G) \setminus \{e\}$
17:         end if
18:     end while
19: end procedure
```

Theorem 49.2.25. *Fleury's algorithm runs in $O(n+m)$ time for a graph with n vertices and m edges.*

Proof. We prove correctness by induction on m. The case $E(G) = \varnothing$ is trivial.

When line 7 is run, $v_{k+1} = v_1$ because of the degree conditions, so W is a closed walk at this stage. Let G' be the subgraph of G at this stage. Then, G' also satisfies the degree conditions.

Since G is connected, every connected component of G' contains at least one of v_i. Then, by the induction hypothesis, every edge of G' belongs to one of W_i, and hence the closed walk W composed in the last step is indeed Eulerian.

The runtime is linear because each edge is deleted immediately after being examined. ■

Corollary 49.2.25.1. *An Eulerian walk exists if and only if there are at most two vertices of odd degree.*

49.2.13 Chinese Postman

A postman must deliver mail along all streets of a town. How can he leave the post office, finish his job and return to the post office having traversed a minimum distance?

That is, given an weighted connected graph, the *Chinese postman problem* is to find a closed walk of minimum total weight visiting each edge at least once. More symbolically, the problem is, given a weighted connected graph (G,w), the task is to find a function $n : E(G) \to \mathbb{R}_{\geq 0}$ such that the graph G' constructed from G by taking $n(e)$ copies of each edge $e \in E(G)$ is Eulerian, and

$$\sum_{e \in E(G)} n(e)w(e)$$

is minimum.

If the graph is Eulerian, then the Eulerian walk provides an optimal solution. Otherwise, some edges must be visited more than once. It makes no sense to walk through an edge more than twice, so we we can restrict $n : E(G) \to \{1,2\}$. Therefore, the task simplifies to finding a subset $S \subseteq E(G)$ of minimum weight such that the graph obtained from G by doubling the edges in S is Eulerian.

As an aside, let us look at another problem.

Let G be an undirected graph, and let $T \subseteq V(G)$ be a subset of even cardinality. A subgraph J is a *T-join* if $|J \cap \delta(x)|$ is odd if and only if $x \in T$. In other words, a T-join is a spanning subgraph of G with the same vertex set as G, but only the edges that ensure that all the vertices in T have odd degree, and all the vertices not in T have even degree.

The fact that there are no T-joins for $|T|$ odd directly follows from the handshaking lemma.

Given an undirected weighted graph (G,w) and a set $T \subseteq V(G)$ of even cardinality, the *minimum weight T-join problem* is to find a minimum weight T-join in G, or decide that none exists.

Lemma 49.2.26. *Let G be a graph and let $T \subseteq V(G)$ be a subset of even cardinality. There exists a T-join in G if and only if $|V(C) \cap T|$ is even for each connected component C in G.*

Proof. If J is a T-join, then for each connected component C in G, we have,

$$\sum_{v \in V(C)} |J \cap \delta(v)| = 2|J \cap E(C)|$$

So, $|J \cap \delta(v)|$ is odd for an even number of vertices in $V(C)$. Since J is a T-join, this means that $|V(C) \cap T|$ is even. This completes the forward implication.

Conversely, let $|V(C) \cap T|$ be even for each connected component C of G. Then T can be partitioned into pairs $\{v_1,w_1\},\ldots,\{v_k,w_k\}$ with $k = |T|/2$ such that for each i, the pair $\{v_i,w_i\}$ belongs to the same connected component. Let P_i be an arbitrary v_i-w_i-path, and let,

$$J := \mathop{\triangle}_{i=1}^{k} E(P_i)$$

where $\triangle$ is the symmetric difference operation ($A \triangle B := (A \setminus B) \cup (B \setminus A) = \{x : (x \in A) \oplus (x \in B)\}$).

The symmetric difference of more than two sets consists of the elements that belong to an odd number of the sets. Observe that if the paths $P_1,\ldots,P_k$ are disjoint, then J is a T-join by definition, as it respects the degrees of vertices in T and not in T. If the paths are not disjoint, then the degree of each vertex has the same parity with respect to J as with respect to the disjoint union of the paths. In either case, J is a T-join, completing the reverse implication. ■

Lemma 49.2.27. *A T-join J in a weighted graph (G,w) has minimum weight if and only if for each cycle C in G, we have,*

$$w(J \cap E(C)) \leq w(E(C) \setminus J)$$

Proof. If $w(J \cap E(C)) > w(E(C) \setminus J)$, then $J \triangle E(C)$ is a T-join of lower weight than J.

Conversely, if J' is a T-join with $w(J) < w(J)$, then the subgraph of G formed by the edges of $J \triangle J'$ is Eulerian, as the degree of each vertex in this subgraph is even, in which case, it is the union of cycles. For at least one cycle C, we must have $w(J \cap E(C)) > w(J' \cap E(C)) = w(E(C) \setminus J)$. ■

Lemma 49.2.28. *Let (G,w) be a weighted graph, and let $T \subseteq V(G)$ be a subset of even cardinality. Every optimum T-join in G is the symmetric difference of $|T|/2$ paths whose endpoints are distinct and belong to T, and possibly some zero-weight cycles.*

Proof. We induct on T. The case $T = \varnothing$ holds trivially.

Let J be any optimum T-join in G. Without loss of generality, J contains no zero-weight cycle. By Lemma 49.2.27, J contains no cycle of positive weight. Since w is non-negative, J is a forest. Let x and y be two leaf nodes in the same connected component of this forest, and let P be the unique x-y-path in J. Then, by the definition of a T-join, $x,y \in T$. So, $J \setminus E(P)$ is an optimum T'-join, where $T' = T \setminus \{x,y\}$, so a cheaper T'-join would imply a cheaper T-join. The lemma follows by induction. ■

Theorem 49.2.29. *The minimum weight T-join problem with non-negative weights can be solved in polynomial time.*

Proof. For each pair $x,y \in T$, we find a shortest x-y-path $P_{x,y}$ and construct an auxiliary complete edge-weighted graph G^* with vertex set T, in which the weight of the edge (x,y) equals the length of the path $P_{x,y}$. Finding these paths for all possible pairs of vertices x and y can be done in $O(|V|^3)$ time using the Floyd-Warshall algorithm.

Then, we find in G^* a perfect matching M of minimum weight, which takes polynomial time.

Let J be the symmetric difference of the paths $P_{x,y}$ taken over all edges (x,y). Then J is a T-join, and is minimum because M has minimal weight. ■

Theorem 49.2.30. *If the weights are non-negative, then the minimum weight T-join problem coincides with the undirected Chinese postman problem.*

Proof. Otherwise, let T be the set of vertices of odd degree, noting that $|T|$ is even by the handshaking lemma, and set $w(e) = 1$ for all edges $e \in E(G)$. Now compute a minimum-cost T-join J with respect to w, and form the multigraph G^* by duplicating the edges in J. A Euerian cycle in G^* is now the desired Chinese postman tour in G. ■

49.2.14 Independence System

For a finite set S, we denote by $\mathcal{P}(S)$ or 2^S the power set of S (the set of all subsets of S).

A *set system* $(V,\mathcal{F})$ consists of a finite set V and a family $\mathcal{F} \subseteq \mathcal{P}(V)$ of subsets of V.

A set system $S = (V,\mathcal{I})$ is furthermore an *independence system* if,

$(M1)$ $\varnothing \in \mathcal{I}$

$(M2)$ For each $Y \subseteq X$, $Y \in F \to X \in \mathcal{I}$.

This latter property is also called the *hereditary property* or *downward-closedness.*

Elements of $\mathcal{I}$ are called *independent* or *feasible*, while elements of $\mathcal{I} \setminus V$ are *dependent* or *infeasible.*

Minimal dependent sets are called *circuits*, and maximal independent sets are called *bases.* For $X \subseteq V$, the maximal independent subsets of X are called *bases of X.*

Let $(V,\mathcal{I})$ be an independence system. For $X \subseteq V$, we define the rank $\text{rank}(X)$ of X as the size of a maximum subset of X that belongs to $\mathcal{I}$.

Example. The following are all independence systems:

1. $V = V(G)$ and $\mathcal{I}$ is the set of independent sets in a graph G.
2. $V = E(G)$ and $\mathcal{I}$ is the set of forests in G.
3. V is the set of columns of a matrix over some field and $\mathcal{I}$ is the power set of linearly independent columns.

4. V is any finite set, k is an integer, and $\mathcal{I}$ the subsets of V of cardinality at most k.

△

Given an independence system $(V,\mathcal{I})$ and a weight function $w : V \to \mathbb{R}$, a *minimisation problem* is to find a basis of minimum total weight, while a *maximisation problem* is to find an independent set of maximum total weight.

Many combinatorial optimisation problems can be formulated as minimisation and maximisation problems. For instance,

- MAXIMUM-WEIGHT-STABLE-SET
- TSP
- SHORTEST-PATH
- KNAPSACK
- MINIMUM-WEIGHT-SPANNING-TREE
- MAXIMUM-WEIGHT-FOREST
- STEINER-TREE
- MAXIMUM-WEIGHT-BRANCHING
- MINIMUM-WEIGHT-BRANCHING
- JSSP (JOB-SHOP-SCHEDULING-PROBLEM)

An independence system $(V,\mathcal{I})$ is a *matroid* if,

M3 $\forall X,Y \in \mathcal{I} : |X| > |Y| \to \Big(\exists x \in X \setminus Y : \big(Y \cup \{x\}\big) \in \mathcal{I}\Big)$ – if $X,Y \in \mathcal{I}$ and $|X| > |Y|$, then there exists an $x \in X \setminus Y$ such that $Y \cup \{x\} \in \mathcal{I}$.

Example.

- (Independent sets in a graph) is not a matroid.
- (Forests in a graph) is a matroid known as the *cycle (graphic) matroid.*
- (Linearly independent columns) is a matroid known as the *vector matroid.*
- (Subsets of size at most k) is a matroid known as the *uniform matroid.*

△

Theorem 49.2.31. *Let $(V,\mathcal{I})$ be an independence system. Then the following statements are equivalent:*

$(M3)$ $\forall X,Y \in \mathcal{I} : |X| > |Y| \to \Big(\exists x \in X \setminus Y : \big(Y \cup \{x\}\big) \in \mathcal{I}\Big)$

$(M3)'$ *For all $Z \subseteq V$, all bases of Z have the same cardinality.*

Proof. Suppose $(M3)$ is valid, but $(M3)'$ is not, and let X and Y be two bases if Z such that $|X| > |Y|$. Then, by $(M3)$, there is an $x \in X \setminus Y$ such that $Y \cup \{x\} \in \mathcal{I}$. Since $x \in X \setminus Y \subseteq X \subseteq Z$, $Y \cup \{x\} \subseteq Z$, contradicting that Y is a basis of Z.

Conversely, suppose $(M3)'$ is valid. If $|X| > |Y|$, the set Y cannot be a basis of $X \cup Y$ as Y is not maximal. Therefore, there exists at least one element $x \in (X \cup Y) \setminus Y = X \setminus Y$ such that $Y \cup \{x\} \in \mathcal{I}$. ■

Corollary 49.2.31.1. *Let $(V,\mathcal{I})$ be a matroid and let $X,Y \in \mathcal{I}$. If $|X| > |Y|$, then there exists a subset of $X \setminus Y$ of cardinality $|X| - |Y|$ such that $Y \cup Z \in \mathcal{I}$.*

Proof. By induction on $k = |X| - |Y|$. ■

($M3$) and this corollary are known as the *exchange*, *augmentation*, or *growth* property of matroids.

Algorithm 24 Greedy Algorithm for Matroid Minimisation

```
 1: procedure MINIMISE($(V,\mathcal{I}),w$)
 2:     SORT($V$, key = $\lambda t.w(t)$)      ▷ Sort elements by weight, so $w(e_1) \leq w(e_2) \leq \ldots \leq w(e_{|V|})$
 3:     $B = \varnothing$
 4:     for $i = 1$ to LEN(V) do
 5:         if $B \cup \{e_i\}$ then      ▷ Check the next cheapest edge is independent
 6:             $B = B \cup \{e_i\}$
 7:         end if
 8:     end for
 9:     return $B$
10: end procedure
```

Theorem 49.2.32. *The greedy algorithm solves the matroid minimisation problem optimally.*

Proof. Let $B = \{e_{j,1},e_{j,2}, \ldots ,e_{j,n}\}$ be the solution found by the algorithm. Suppose there is an element $e \in V \setminus B$. If $B \cup \{e\}$ were independent, then this element would have been added to to B in line 6. Since this element was rejected, $B \cup \{e\}$ is not independent, and hence B is a basis.

To prove optimality of B, let $B^* = \{e^*_{j,1},e^*_{j,2}, \ldots ,e^*_{j,n}\}$ be any optimal solution whose elements are sorted according by weight, as in the algorithm. Without loss of generality, B^* has the longest "prefix" coinciding with B.

Let j_k be the smallest index such that $e_{j,k} \neq e^*_{j,k}$. Since the set $\{e_{j,1},e_{j,2}, \ldots ,e^*_{j,k}\}$ is independent, $e_{j,k}$ appears before $e^*_{j,k}$ in the order, and hence $w(e_{j,k}) \leq w(e^*_{j,k})$.

If $e_{j,k}$ is the last element of B, then $wB \leq wB^*$, so B is optimal. Otherwise, $e_{j,k}$ is not the last element of B. Consider the set $B' = \{e_{j,1},e_{j,2}, \ldots ,e_{j,k}\}$. Since $|B'| < |B^*|$, there exists a set $Z \subseteq B^* \setminus B'$ of cardinality $|B^*| - |B'|$ such that the set $B'' = B' \cup Z$ is independent, and hence a basis. Then, $w(B'') \leq w(B^*)$, so B'' is an optimal basis, and this has a longer prefix coinciding with B, contradicting the choice of B^*. ■

Corollary 49.2.32.1. *An almost identical algorithm solves the matroid maximisation problem optimially.*

An *independence oracle* for an independence system $(V,\mathcal{I})$ is a function $\mathfrak{D} : \mathcal{P}(E) \to \{0,1\}$ defined by

$$\forall F \subseteq V, \mathfrak{D}(F) = \begin{cases} 1 & F \subseteq \mathcal{I} \\ 0 & \text{otherwise} \end{cases}$$

Remark. Because an independence system is determined entirely by V and $\mathcal{I}$, the independence oracle provides enough information to recover the independence system it describes, so we can flip the definition, and say that every independence oracle defines an independence system.

A *basis-superset oracle* is a function $\mathfrak{D} : \mathcal{P}(E) \to \{0,1\}$ defined by

$$\forall B \subseteq V, \mathfrak{B}(B) = \begin{cases} 1 & B \in \mathcal{I} \wedge \neg \exists x \in E : B \cup \{x\} \in \mathcal{I} \\ 0 & \text{otherwise} \end{cases}$$

The greedy algorithm requires sorting the elements of V, which takes $O(|V|\log|V|)$ time. However, more significantly, we need to consult with the basis-superset oracle at every step, so the complexity of the algorithm depends on the complexity of the oracle, given by $O(\mathfrak{D})$.

Theorem 49.2.33. *Let $(V,\mathcal{I})$ be an independence system. The greedy algorithm solves the maximisation problem optimally for any $w : E \to \mathbb{R}$ if and only if $(V,\mathcal{I})$ is a matroid.*

This theorem allows us to bound how well greedy algorithms can solve certain problems. For instance, the travelling salesman problem is an independence system, but *not* a matroid, so this theorem tells us that a greedy algorithm cannot optimally solve the travelling salesman problem.

Proof. Suppose $V,\mathcal{I}$ is not a matroid. That is, there exists $X,Y \in \mathcal{I}$ with $|X| < |Y|$ such that for all $e \in Y \setminus X$, $X \cup \{e\} \notin \mathcal{I}$.

Let $\varepsilon > 0$. Define the weight function by,

$$w(e) := \begin{cases} 1+\varepsilon & e \in X & \text{Choose first } |X| \text{ steps} \\ 1 & e \in Y \setminus X & \text{Can't choose} \\ 0 & e \in E \setminus \{X \cup Y\} & \text{Don't change weight} \end{cases}$$

So greedy outputs F with $w(F) = |x|(1+\varepsilon) + 0$. So, $w(F) = |X|(1+\varepsilon) < w(Y) = |Y|$ for $\varepsilon < |Y|/|X| - 1$, a contradiction to $w(F)$ being maximum for all weight functions w. This completees the forward implication.

Now, suppose $(V,\mathcal{I})$ is a matroid. This portion of the proof is similar to the proof of correctness for the greedy algorithm, so we give it more tersely. Let w be an arbitrary weight function, and let $F = \{f_1,f_2,\ldots,f_r\}$ be the output of the greedy algorithm. Without loss of generality, suppose $w(f_1) \geq w(f_2)\ldots \geq w(f_r)$. Suppose there exists $G \in \mathcal{I}$ such that $|F| < |G|$. By the augmentation property, there exists $g \in G \setminus F$ such that $F \cup \{g\}$ implies there exists t such that $\{f_1,\ldots,f_t,g,f_{t+1},\ldots,f_s\} = F \cup \{g\} \in \mathcal{I}$ with $w(f_t) \geq w(g) \geq w(f_{t+1})$. We also have $\{f_1,\ldots,f_t\} \subseteq \{f_1,\ldots,f_t,g\} \in \mathcal{I}$, so g should have been chosen in step $t+1$ of the greedy algorithm. So, F has maximum cardinality by contradiction.

Suppose there exists $G = \{g_1,g_2,\ldots,g_r\} \in \mathcal{I}$ such that $w(G) > w(F)$, and $w(g_i) \geq w(g_{i+1})$. So,

$$\sum_{g_i \in G} w(g_i) > \sum_{f_i \in F} w(f_i)$$

so, there exists k such that $w(g_k) > w(f_k)$ since $|G| \leq |F|$. Take $X = \{f_1,f_2,\ldots,f_{k-1}\}$ (= $\varnothing$if $k = 1$), and $Y = \{g_1,g_2,\ldots,g_k\}$. Clearly, $|X| < |Y|$, so by the augmentation property, there exists $g_t \in Y \setminus X$ with $t \leq k$ such that $\{f_1,f_2,\ldots,g_t\} = X \cup \{g_t\} \in \mathcal{I}$. Because $w(g_t) \geq w(g_k) > w(f_k)$, g_t should have been chosen before step k of the greedy algorithm, contradicting correctness, and so G does not exist and hence $w(f)$ is maximum.

This completes the reverse implication. ■

Given two matroids, $(V,\mathcal{I}_1)$ and $(V,\mathcal{I}_2)$, the *matroid intersection problem* is to find a set $X \in \mathcal{I}_1 \cap \mathcal{I}_2$ such that $|X|$ is maximum.

Theorem 49.2.34. *Edmonds' algorithm solves the matroid intersection problem. If the matroids are given by independence oracles with maximum complexity T, then the algorithm solves the problem in $O(|V|^3 T)$ time.*

The *partition matroid* is defined as follows. Let B_i be a collection of disjoint subsets of V, and let d_i be integers with $0 \leq d_i \leq |B_i|$. Define $I \subseteq V$ to be independent if $|I \cap B_i| \leq d_i$ for each i.

In particular, if $i = 1$ and $B_1 = V$, the partition matroid is the uniform matroid.

Given a bipartite graph $G = (A, \cup B, E)$, define two partition matroids M_A and M_B on E as follows:

M_A: for each vertex $i \in A$, let A_i be the set of edges incident to i and $d_i = 1$.

M_B: for each vertex $i \in B$, let B_i be the set of edges incident to i and $d_i = 1$.

Theorem 49.2.35. *The maximum matching problem for G coincides with the matroid intersection problem for M_A and M_B.*

Theorem 49.2.36. *The family of independent sets in a graph G forms a matroid if and only if every connected component of G is a clique.*

Proof. Let G be a graph, of which every connected component is a clique. For a subset $U \subseteq V(G)$, every basis in U contains exactly one vertex in each connected component in $G[U]$ (the subgraph induced by U). Therefore, all bases in U have the same size, and hence the family of independent sets in G forms a matroid, completing the forward implication.

If G contains a connected component with is not a clique, then it contains a subset $U \subseteq V(G)$ inducing a path on 3 nodes. But then U has two bases of size 1 and 2, so the family of independent sets in G do not form a matroid, completing the reverse implication. ■

Theorem 49.2.37. *Every independence system is the intersection of finitely many matroids.*

Proof. Let C be a circuit of $(V,\mathcal{I})$, and $\mathcal{I}_C$ the family of subsets $A \subseteq E$ such that C is not a subset of A. Then, (V,F_C) is a matroid because,

$(M1)$ $\varnothing \in \mathcal{I}_C$.

$(M2)$ For each $A \subseteq B$, $B \in F \to A \in \mathcal{I}$.

$(M3)'$ All bases of (V,F_C) have size $|V| - 1$.

and $(V,\mathscr{I})$ is the intersection $\bigcap(V,\mathcal{I}_C)$ taken over all circuits C of $(V,\mathscr{I})$. ■

Theorem 49.2.38. *The problem of finding a maximum independent set in the intersection of 3 matroids is NP-hard.*

49.3 Polynomial Time Solvability

For many combinatorial optimization problems, polynomial-time algorithms are known. However, there are also many important problems for which no polynomial-time algorithms are known to exist. Although we cannot prove that none exists, we can show that a polynomial-time algorithm for one "hard" problem would imply a polynomial-time algorithm for other "hard" problems.

49.3.1 Decision Problems

An *alphabet* in the context of formal languages is any set of symbols, often denoted by Σ. A *word* over an alphabet is any finite sequence of letters.

The *Kleene star*, also known as the *free monoid constructor*, is a unary operation, either on sets of strings, or sets of symbols or characters. The application of the Kleene star to a set V is written as V^*.

1. If V is a set of strings, then V^* is defined as the smallest superset of V that contains the empty string, ε, and is closed under string concatenation.
2. If V is a set of symbols or characters, then V^* is the set of all strings over symbols in V, including the empty string ε.

More formally, given a set V, we define the sets,

$$V^0 = \{\varepsilon\}$$
$$V^1 = V$$

and recursively define the set,

$$V^{i+1} = \{wv : w \in V^i, v \in V\} \text{ for each } i > 0$$

If V is a formal language, then V^i is a shorthand for the concatenation of V with itself i times. That is, V^i represents the set of all strings that can be represented as the concatenation of i strings in V. The Kleene star on V is then defined as:

$$V^* = \bigcup_{i \geq 0} V^i$$

The Kleene star is highly important in theoretical computer science, particularly in complexity and computability theory.

Note that if V is countable, then V^* is the countable union of countable sets, and is hence countable.

The set of all words over an alphabet Σ is then Σ^*. A *formal language* L over an alphabet Σ is a subset of Σ^*. See §2.3.4 for a more in-depth treatment on this topic.

Let $\{0,1\}^*$ be the set of all binary words, and let $L \subseteq \{0,1\}^*$ be a language. L can be interpreted as a decision problem as follows: given any binary string, decide whether it belongs to L.

Conversely, assuming a fixed efficient encoding, we can encode the input to any problem that can be answered positively or negatively as a binary string, in which case the set of all instances of the problem defines a language X, and the set of "yes" instances defines a subset $Y \subseteq X$.

A *decision problem* is a pair $P = (X,Y)$ where X is a language decidable in polynomial time, and $Y \subseteq X$. The elements of X are called *instances*, the elements of Y are *yes-instances*, and the elements of $X \setminus Y$ are *no-instances*. Decision problems in theoretical computer science are often written in (abbreviated) full capital letters.

An algorithm for a decision problem $P = (X,Y)$ is an algorithm computing the function $f : X \to \{0,1\}$ such that $f(x) = 1$ for $x \in Y$ and $f(x) = 0$ for $x \in X \setminus Y$. For instance, given an undirected graph, encoded as a binary string, we might ask, "Is there a Hamiltonian cycle in G?"

Theorem (Cantor's Diagonal Argument)**.** *There are functions $f : \mathbb{N} \to \{0,1\}$ that cannot be computed by any algorithm.*

Proof. Algorithms are finite sequences of a finite alphabet of possible instructions, so there are countably many possible algorithms, while the set of functions f has size $2^{\aleph_0} = \mathcal{P}(\mathbb{N}) = \mathfrak{c}$ which is uncountable, so no bijection can exist between the sets.

More specifically, by Cantor's Diagonal Argument, $\mathfrak{c}$ is strictly larger than $\aleph_0$, so there are more functions than there are algorithms, as required. ■

An *oracle* is an abstract machine (a generalisation of a function) that is assumed to be able to solve a specific problem (even non-decision problems) in a single operation. The problem is not assumed to even be computable – an oracle is simply a black box that is able to produce a solution for any instance of a given computation program.

A *certificate* or a *witness* is a string that certifies the membership of some string in a language. So, for the Hamiltonian cycle question, a certificate for a graph G would simply be a Hamiltonian cycle: clearly, if you have one, the graph G should be in X.

The class of all decision problems which admit a polynomial time algorithm is called *P* or *PTIME* (for *Polynomial Time*).

In contrast, *NP* (*Non-Deterministic Polynomial time*) is the class of decision problems that admit a polynomial-time certificate-checking algorithm. P is a subclass of NP, as every problem that is solvable in polynomial time can also be checked in polynomial time by just solving the problem. As shown above, Hamiltonian Cycle is NP, and, currently, there does not exist a polynomial time algorithm for Hamiltonial Cycle, so it is not P.

Many decision problems encountered in combinatorial optimisation belong to NP. For many of them, such as Hamiltonian Cycle, it is not known whether they admit polynomial time algorithms. However, we can say that certain problems are not easier than others. This can be formalised through the concept of *polynomial reduction.*

Let $P_1 = (X_1,Y_1)$ and $P_2 = (X_2,Y_2)$ be decision problems. Let $f : X_2 \to \{0,1\}$ with $f(x) = 1$ for $x \in Y_2$ and $f(x) = 0$ for $x \in X_2 \setminus Y_2$. We say that P_1 *polynomially reduces* to P_2 if there exists a polynomial-time algorithm for P_1 using f as an oracle.

Theorem 49.3.1. *If P_1 polynomially reduces to P_2, and there is a polynomial-time algorithm for P_2, then there is a polynomial-time algorithm for P_1.*

Proof. The oracle for P_2 is queried at most polynomially many times in the polynomial-time algorithm for P_1. If there is a polynomial-time algorithm for P_2, then it can be used as the oracle, so P_1 is the composition of two polynomial-time algorithms, and is hence polynomial-time. ■

Let $P_1 = (X_1,Y_1)$ and $P_2 = (X_2,Y_2)$ be decision problems. We say that P_1 *polynomially transforms* to P_2 if there exists a function $f : X_1 \to X_2$ computable in polynomial time such that $f(x_1) \in Y_2$ for all $x_1 \in Y_1$ and $f(x_1) \in X_2 \setminus Y_2$ for all $x_1 \in X_1 \setminus Y_1$.

A decision problem $\Pi \in$ NP is called *NP-complete* if all other problems in NP polynomially transform to Π. So, to prove a problem is NP-complete, we need to show it is in NP, and to polynomially transform a known NP-complete problem into it.

Conversely, a problem $\Pi \in$ NP is *NP-hard* if all problems in NP polynomially-*reduce* to Π.

49.3.2 Boolean Satisfiability

Revisit §2 if you have forgotten about predicate logic.

A *valuation* on a Boolean expression is an assignment of truth values to the literals in the expression.

A compound proposition is in *conjunctive normal form* or *CNF* if it is a conjunction of one or more *clauses*, where a clause is a disjunction of atoms; it is an AND of OR statements. A compound proposition is similarly in *disjunctive normal form* or *DNF* if it is the disjunction of one or more clauses, where a clause is a conjunction of atoms; it is an OR of AND statements.

Propositions in CNF:

- p
- $(p \lor \neg q) \land r$
- $(p \lor q) \land (\neg p \lor r) \land q \land (\neg q \lor \neg r)$
- $p \land \neg q \land r \land t \land \neg u \land v$

Propositions not in CNF:

- $(p \land q) \land (q \lor r)$
- $(p \lor q) \land (q \to \neg r) \land (\neg p \lor r)$

- $(p \lor (q \land r)) \land (p \lor \neg r)$

Interchanging $\land$ and $\lor$ above gives examples of clauses in and not in DNF.

Using the equivalence of material implication and disjunction, along with De Morgan's laws and the distributive laws, it is possible to rewrite any compound proposition in a normal form. However, applying these laws blindly does not necessarily produce the simplest normal form for a compound proposition.

For example,

$$\begin{aligned}(P \to Q) \land (\neg P \to Q) &\equiv (\neg P \lor Q) \land (P \lor Q)\\ &\equiv (\neg P \land P) \lor (\neg P \land Q) \lor (Q \land P) \lor (Q \land Q)\\ &\equiv 0 \lor (\neg P \land Q) \lor (Q \land P) \lor Q\\ &\equiv (\neg P \land Q) \lor (Q \land P) \lor Q\end{aligned}$$

Inspecting the clauses closer, we see that Q controls the value of the entire expression, so a simpler CNF for the proposition is just Q.

$$\equiv Q$$

We should really draw out a truth table to prove this formally, but it should be clear enough that this is true.

Theorem 49.3.2. *There is a polynomial time algorithm to reduce any Boolean expression to a DNF and CNF representation.*

Given a CNF C, the *satisfiability problem* or *SAT* is to determine if there is a valuation such that C evaluates to true.

Theorem (Cook–Levin)**.** *SAT is NP-complete.*

The satisfiability problem restricted to instances where each clause contains at most three literals is called *3-SAT*.

Theorem 49.3.3. *3-SAT is NP-complete.*

Proof. Clearly, 3-SAT belongs to NP. To prove completeness, we show that SAT polynomially transforms to 3-SAT.

Let $Z = (x_1 \lor x_2 \lor \ldots \lor x_k)$ be a clause containing $k > 3$ literals. Transform Z as follows:

$$(x_1 \lor x_2 \lor \ldots \lor x_k) \mapsto (x_1 \lor x_2 \lor \ldots \lor x_{k-1} \lor u) \land (\neg u \lor x_{k-1} \lor x_k)$$

where u is a new variable.

Suppose there is an assignment φ satisfying the original CNF. If Z is satisfied by one of the first $k-2$ literals, then by defining, $\varphi(u) = 0$, we extend φ to an assignment satisfying the transformed CNF. If Z is satisfied by x_{k-1} or x_k, we define $\varphi(u) = 1$ and obtain an assignment satisfying the transformed CNF.

Conversely, suppose there is an assignment φ satisfying the transformed CNF. If $\varphi(u) = 0$, then Z is satisfied by one of the first $k-2$ literals. If $\varphi(u) = 1$, then Z is satisfied by x_{k-1} or x_k.

So, an assignment φ satisfies Z if and only if it satisfies the transformed CNF.

Because a Boolean formula contains finitely many terms, this algorithm always terminates. Applying this transformation repeatedly, the original CNF can be transformed into an instance of 3-SAT which is satisfiable if and only if the original one is. ■

Theorem 49.3.4. *2-SAT is P.*

Proof. The implication $a \to b$ is logically equivalent to $\neg a \vee b$, so, in 2-SAT, the clause $x_1 \vee x_2$ is equivalent to the pair of implications $\neg x_1 \to x_2$ and $\neg x_2 \to x_1$. If x_1 is true, then x_2 must be true, and if x_2 is false, then x_1 must be false.

These implications are straightforward, so we just follow every possible implication chain and see if we ever derive both $\neg x$ from x or x from $\neg x$. If we do for some x, then the 2-SAT formula is unsatisfiable. Otherwise, it is satisfiable. The number of possible implication chains is polynomially bounded in the size of the input formula, so they are checkable in polynomial time. ■

Remark. With 3-SAT, we can express implications of the form $a \to b \vee c$, where a,b,c are literals. Now, if a is true, then one or both of b and c are true, but we don't know which. In this case, we have to do case analysis, and combinatorial explosion occurs.

Theorem 49.3.5. *HAMILTONIAN-CYCLE is NP-complete.*

Proof. Membership in NP is obvious. To prove completeness, we polynomially transform 3-SAT to HAMILTONIAN-CYCLE.

Let C be a CNF with clauses $Z_1,Z_2,\ldots,Z_m$ over the set of variables $X = \{x_1,x_2,\ldots,x_n\}$, with each clause containing three variables. We note that there are 2^n possible valuations on C. We model these 2^n possible valuations using a digraph with 2^n different Hamiltonian cycles.

Construct n paths $P_1,P_2,\ldots,P_n$ corresponding to the n variables, each consisting of $2k$ nodes, $P_i = (v_{i,1},v_{i_2},\ldots,v_{i,2k})$. We add edges from $v_{i,j-1}$ to $v_{i,j}$ on P_i corresponding to the assignment x_i = true (picturing the paths as lying left to right, these edges point left to right). We add edges from $v_{i,j}$ to $v_{i,j-1}$ on P_i corresponding to the assignment x_i = false (right to left). Next, add the edges connecting $v_{i,k}$ to $v_{i+1,k}$ for each k.

Next, add a source node s and target node node t, and connect s to $v_{1,1}$ and $v_{i,k}$, and connect t to $v_{n,1}$, $v_{n,k}$, and s.

Next, add a node $C_1,C_2,\ldots,C_m$ for each clause. If a clause C_j contains the variable x_i, connect C_j to $x_{i,2j-1}$ and $x_{i,2j}$, left to right (add edges $(x_{i,2j-1},C_j)$, and $(C_j,x_{i,2j})$) if C_j contains the positive literal x_i, and right to left (add edges $(x_{i,2j},C_j)$, and $(C_j,x_{i,2j-1})$) if C_j contains the negative literal $\neg x_i$.

Any Hamiltonian cycle in the graph traverses P_i either from right to left, or left to right, because any path entering a node $v_{i,j}$ has to exit from $v_{i,j+1}$ either immediately, or via one clause-node in between, in order to maintain the Hamiltonian property. Similarly, all paths entering at $v_{i,j-1}$ must exit from $v_{i,j}$.

Note that this graph can be constructed in polynomial time.

Since each path P_1 can be traversed in two possible ways, and we have n paths mapping to n variables, there can be 2^n Hamiltonian cycles in the graph $G \setminus \{C_1,C_2,\ldots,C_k\}$, each corresponding to a different valuation of $x_1,x_2,\ldots,x_n$.

If there exists a Hamiltonian cycle H in G

- If H traverses P_i from left to right, assign x_i = true;
- If H traverses P_i from right to left, assign x_i = false.

Since H visits each clause node C_j, at least one of the P_i was traversed in the correct direction relative to the node C_j, so the assignment obtained here satisfies the given 3-CNF.

Conversely, if there exists a satisfying assignment for the 3-CNF, select the path that traverses P_i from left to right if x_i = true, or right to left if x_i = false, including the clause nodes whenever possible. Connect the source to P_1, P_2 to t, and P_i to P_{i+1} appropriately so as to maintain the continuity of the path, then connect t to s to complete the cycle. Since the assignment is such that every clause is satisfied, the clause-nodes are included in the path. The P_i nodes, s and t are all included, and all the paths are traversed in one direction only so no node is repeated twice, so the path obtained is a Hamiltonian Cycle. ■

Theorem 49.3.6. *MAXIMUM-INDEPENDENT-SET is NP-complete.*

Proof. Membership in NP is obvious. To prove completeness, we polynomially transform SAT to MAXIMUM-INDEPENDENT-SET.

Let Z be a collection of clauses $Z_1,Z_2,\ldots,Z_m$ over the set of variables $X = \{x_1,x_2,\ldots,x_n\}$, with each Z_i containing k_i literals $(\beta_{i,1} \vee \beta_{i,2} \vee \ldots \vee \beta_{i,k})$. We construct a graph G such that G has an independent set of size m if and only if there is a truth assignment satisfying all m clauses.

For each clause Z_i, we introduce a clique of k_i vertices, one vertex per literal. Two vertices in different cliques (clauses) are connected by an edge if and only if they represent the same variable, but of different polarity.

Suppose G has an independent set S of size m. Then each of these cliques contains exactly one vertex. Setting each of these literals to be true, we obtain an assignment satisfying all m clauses since no two literals in S are in conflict.

Conversely, if there is a truth assignment satisfying all m clauses, then we choose a true literal out of each clause. The set of corresponding vertices defines an independent set in G of size m. ■

Theorem 49.3.7. *MAXIMUM-INDEPENDENT-SET is NP-complete for graphs of degree at most* 3.

Proof. If G has a vertex x of degree at least 4, we apply vertex splitting with $|Y| = 2$. In the transformed graph, x' has degree 2, y has degree 3, and z has degree $\deg(x) - 1$. Repeated applications of vertex splitting transforms G into a graph of vertex degree at most 3, and clearly this transformation is polynomial. ■

Theorem 49.3.8. *MINIMUM-VERTEX-COVER and MAXIMUM-CLIQUE are NP-complete.*

The *travelling salesman problem* (TSP) asks the following question: Given a list of cities and the distances between each pair of cities, what is the shortest possible route that visits each city exactly once and returns to the origin city?

That is, given a weighted complete graph (K_n,w), find a Hamiltonian cycle of minimum weight.

Theorem 49.3.9. *TSP is NP-hard.*

Proof. We give a reduction from HAMILTONIAN-CYCLE.

Let G be an instance of HAMILTONIAN-CYCLE. Construct an instance G' of TSP as follows: $V(G') = V(G)$ with every two vertices of G' being adjacent. Define $w((u,v)) = 1$ if $(u,v) \in E(G)$ and $w((u,v)) = 2$ otherwise. Then G has a Hamiltonian cycle if and only if the optimum tour in G' has length n. ■

49.3.3 Approximation Algorithms

An *absolute approximation algorithm* for an optimization problem P is a polynomial-time algorithm A for P for which there exists a constant k such that $|A(I) - \mathrm{Opt}(I)| \le k$ for all instances I of P, where $A(I)$ is the size of the solution found by the algorithm A and $\mathrm{Opt}(I)$ is the size of an optimal solution.

Let P be an optimization problem with non-negative weights and $k \ge 1$. A *k-factor approximation algorithm* for P is a polynomial-time algorithm A for P such that $1/k\,\mathrm{Opt}(I) \le A(I) \le k\,\mathrm{Opt}(I)$ for all instances I of P. We also say that A has *performance ratio k*.

The first inequality applies to maximization problems and the second one to minimization problems.

A 1-factor algorithm is an exact polynomial-time algorithm.

Theorem 49.3.10. *There is no k such that the greedy algorithm for VERTEX-COVER is a k-factor approximation algorithm.*

Proof. Let $p \in \mathbb{N}$ and G_p be a graph with vertex set $V(G) = A \cup B \cup C$, where $|A| = |B| = p$. For each $i \in [2,3,\ldots,p-1]$, split the vertices of B into $\lfloor p/i \rfloor$ groups and for each group introduce the vertex of A adjacent to the vertices of that group. The algorithm may first delete the vertices of A, in which case the size of the solution is $|A| + p$. On the other hand B is an optimal solution of size p, and the ratio $|A|/p + 1$. ■

Theorem 49.3.11. *There is a 2-factor approximation algorithm for VERTEX-COVER.*

Proof. Let M be a maximum matching in G. Then the set of vertices covered by M is a vertex cover containing $2|M|$ vertices. On the other hand, any vertex cover must contain at least $|M|$ vertices, so $|M| \leq \tau(G) \leq 2|M|$, where $\tau(G)$ is the size of a minimum vertex cover in G. Therefore, $2|M|/\tau(G) \leq 2$, so this algorithm is a 2-factor approximation. ■

49.3.4 Chromatic Numbers

A *vertex colouring* of G is a mapping $f : V(G) \to \mathbb{N}$ with $f(u) \neq f(v)$ whenever $(u,v) \in E(G)$.

In other words, in a vertex colouring, every colour class is a independent set, so vertex colouring is a partition of $V(G)$ into independent sets.

A *edge colouring* of G is a mapping $f : E(G) \to \mathbb{N}$ with $f(e) \neq f(e')$ for all edges e and e' incident to the same vertex.

Remark. An edge colouring of G is equivalent to a vertex colouring of the line graph of G.

Given an undirected graph G, the *vertex colouring problem* is to find a vertex colouring of G with minimum colours. The optimum value of the vertex colouring problem for G is called the *chromatic number* of G, denoted $\chi(G)$.

Given an undirected graph G, the *edge colouring problem* is to find an edge colouring with minimum colours. The optimum value of the edge colouring problem for G is called the *edge-chromatic number* or *chromatic index* of G, denoted $\chi'(G)$.

Theorem 49.3.12. *The following decision problems are NP-complete for any fixed value $k \geq 3$:*

(*i*) *Decide whether a given graph has a chromatic number at most k.*

(*ii*) *Decide whether a given graph has a chromatic index at most k.*

Moreover, (i) is NP-complete even for planar graphs of vertex degree at most 4, and (ii) is NP-complete for graphs of vertex degree at most 3.

Theorem 49.3.13. *Both problems can be solved in polynomial time for $k = 1,2$.*

Proof. $\chi(G) = 1$ if and only G has no edges. $\chi(G) = 2$ if and only if G is bipartite. In both caes, the problem can be solved in polynomial time. The chromatic index of G is at most 2 if and only if the chromatic number of $L(G)$ is at most 2. ■

Theorem 49.3.14. *For any graph G,*

$$\chi'(G) \geq \max_{v \in V(G)} \deg(v)$$

Proof. To reduce clutter, define $\Delta(G) \coloneqq \max_{v \in V(G)} \deg(v)$. We induct on $|E(G)|$.

Let $\Delta(G) = k$, and let $e = (u,v) \in E(G)$. By the induction hypothesis, $G \setminus \{e\}$ has an edge colouring f with k colours. Since the degree of u and v is strictly less than k in $G \setminus \{e\}$, there is a colour $i \in \{1, \ldots, k\}$ which is missing at u, and a colour $j \in \{1, \ldots, k\}$ which is missing at v. If $i = j$, we can assign this colour to the edge $e = (u,v)$ in G.

Otherwise, we consider the subgraph H of $G \setminus \{e\}$ formed by the edges of colour i and j. Every vertex of H has degree at most 2, and hence every connected component of H is either a path or a cycle. Each of u and v has degree 1 in H (degree 2 is not possible because each of them misses one of the two colours; degree 0 would allow to use the same argument as when both of them miss the same colour). Therefore, the connected component of H containing u is a path, and the connected component of H containing v is a path, and these two paths are different, otherwise we would have $i = j$. But now we can exchange the colours on the path containing u, and assign colour j to the edge e in G. ■

Theorem (Vizing)**.** *For any graph G,*

$$\Delta(G) \leq \chi'(G) \leq \Delta(G) + 1$$

Corollary 49.3.14.1. *The edge colouring problem admits an absolute approximation algorithm on simple graphs.*

Let $\omega(G)$ denote the size of a maximum clique in G.

Theorem 49.3.15. *For any graph G,*

$$\omega(G) \leq \chi(G) \leq \Delta(G) + 1$$

Proof. Since the vertices of any clique must have pairwise different colours in any proper colouring of G, we must have $\omega(G) \leq \chi(G)$.

For the second inequality, let $V(G) = \{v_1, \ldots, v_n\}$, and let S be a set of $\Delta(G) + 1$ colours. Assign any colour from S to v_1, and then proceed by induction as follows: for each i, vertex v_i has at most $\Delta(G)$ neighbours among $v_1, \ldots, v_{i-1}$, and hence at least one colour from S is missing among the neighbours of v_i. Assign this colour to v_i, and proceed to v_{i+1}. ■

Corollary 49.3.15.1. *A vertex colouring of G with $\Delta(G) + 1$ colours can be found in linear time.*

Theorem (Brooks)**.** *If G is a connected graph which is neither complete nor an odd cycle, then $\chi(G) \leq \Delta(G)$. Otherwise, $\chi(G) = \Delta(G) + 1$.*

Corollary 49.3.15.2. *Every connected graph of vertex degree at most 3 is 3-colourable, except for K_4.*

A graph is *planar* if it can be drawn on the plane in such a way that no edges cross each other. A *face* of a planar graph is a maximal section of the plane in which any two points can be joined by a curve that does not intersect any part of G. The *degree* of a face is the number of edges in the boundary surrounding the face.

Theorem (Euler)**.** *Let G be a connected planar graph with n vertices, m edges, and f faces. Then,*

$$n - m + f = 2$$

Proof. By induction on m. For $m = 0$, $G = K_1$, a graph with 1 vertex and 1 face. Suppose the formula is true for any connected planar graph with fewer than m edges, and let G have m edges. If G is a tree, then $m = n - 1$ and $f = 1$ and the formula holds. Otherwise, if G is not a tree, consider a cycle C, and an edge $e \in C$. The graph $G \setminus \{e\}$ is connected, has the same number of vertices, one edge fewer, and one face fewer. By the induction hypothesis, in $G \setminus \{e\}$, we have $n - (m - 1) + (f - 1) = 2$. Therefore in G, we have $n - m + f = 2$. ■

Corollary 49.3.15.3. *If G is a connected planar graph with $n \geq 3$ vertices and m edges, then $m \leq 3n-6$. If G is additionally triangle-free, then $m \leq 2n-4$.*

Proof. If we trace the boundary of all faces, we encounter each edge exactly twice. Denoting the number of faces of degree k by f_k, we conclude that,

$$\sum_k kf_k = 2m$$

Since the degree of any face in a simple planar graph is at least 3, we have,

$$\begin{aligned} 3f &= 3\sum_{k\geq 3} f_k \\ &= \sum_{k\geq 3} kf_k \\ &= 2m \end{aligned}$$

Together with Euler's formula, this proves $m \leq 3n-6$.

If G is additionally triangle-free, then,

$$\begin{aligned} 4f &= 4\sum_{k\geq 4} f_k \\ &= \sum_{k\geq 4} kf_k \\ &= 2m \end{aligned}$$

and therefore $m \leq 2n-4$. ■

Corollary 49.3.15.4. *K_5 and $K_{3,3}$ are not planar.*

Proof. For K_5, we have $m > 3n-6$, since $n = 5$ and $m = 10$, so K_5 is not planar. $K_{3,3}$ is triangle-free, and we have $m > 2n-4$, since $n = 6$ and $m = 9$, so $K_{3,3}$ is not planar. ■

Corollary 49.3.15.5. *Every planar graph has a vertex of degree at most 5.*

Proof. Let G be a connected planar graph with n vertices and m edges. Then,

$$\begin{aligned} \sum_{v\in V(G)} \deg(v) &= 2m \\ &\leq 2(3n-6) \\ &= 6n-12 \end{aligned}$$

and hence G has a vertex of degree at most 5. ■

A graph H is a *minor* of G if H can be obtained from G by vertex deletions, edge deletions, and edge contractions.

Theorem (Kuratowski–Wagner)**.** *A graph G is planar if and only if G does not contain K_5 nor $K_{3,3}$ as minors.*

Theorem (Six Colour)**.** *Every planar graph G can be vertex coloured with at most 6 colours.*

Proof. We induct on the number of vertices. Obviously, every graph with at most 6 vertices is 6-colourable.

If a planar graph has more than 6 vertices, then by Corollary 49.3.15.5, there exists a vertex v of degree at most 5. By the induction hypothesis, $G \setminus \{v\}$ is 6-colourable. Then, then neighbours of v use at most 5 different colours, so the unused colour can be used to colour v, and G is 6-colourable. ■

Theorem (Five Colour)**.** *Every planar graph G can be vertex coloured with at most* 5 *colours.*

Proof. We induct on the number of vertices. Obviously, every graph with at most 5 vertices is 5-colourable.

If a planar graph has more than 5 vertices, then by Corollary 49.3.15.5, there exists a vertex v of degree at most 5. Delete this vertex from G to form $G' = G \setminus \{v\}$. By the induction hypothesis, G' is 5-colourable. If the neighbours of v do not use all 5 different colours, the unused colour can be used to colour v, and G is 5-colourable. Otherwise, consider the vertices v_1,v_2,v_3,v_4,v_5 adjacent to v in cyclic order (which will depend on how we draw G), coloured with colours 1,2,3,4, and 5, respectively.

Consider the subgraph $G_{1,3}$ of G' consisting of the vertices coloured with colours 1 and 3 only, and the edges connecting them (this is called a *Kempe chain*). If v_1 and v_3 lie in different connected components of $G_{1,3}$, we can swap the 1 and 3 colours on the connected component containing v_1 without affecting the colouring of the rest of G'. This frees colour 1 for v, completing the task. If v_1 and v_3 lie in the same connected component of $G_{1,3}$, then we can find a path in $G_{1,3}$ consisting of only colour 1 and 3 vertices.

Now consider the subgraph $G_{2,4}$ of G' consisting of the vertices coloured with colours 2 and 4 only, and the edges connecting them, and appply the same arguments as before. We are then either able to reverse the 2-4 colouration on the subgraph of $G_{2,4}$ containing v_2 and colour v colour 2, or we can connect v_2 and v_4 with a path that consists of only colour 2 and 4 vertices. Such a path would necessarily intersect the 1-3 coloured path constructed before, since the vertices were given in cyclic order, contradicting the planarity of G. ■

Theorem (Four Colour)**.** *Every planar graph G can be vertex coloured with at most* 4 *colours.*

Remark. The Four Colour Theorem was one of the first major theorems to be proved with significant computer assistance.

If the Four Colour Theorem were false, then there would exist a minimal counterexample. After reducing the possibilities with various mathematical techniques, the remaining configurations were checked using a computer, taking over a thousand computer core hours to finish.

49.3.5 Bin Packing

The *knapsack problem* (KNAPSACK) is as follows: given a set of items, each with a weight and a value, determine which items to include in the collection so that the total weight is less than or equal to a given limit and the total value is as large as possible.

Theorem 49.3.16. *Deciding the knapsack problem ("Can a value of at least V be achieved without exceeding weight W?") is NP-complete.*

The *subset sum problem* (SSP) is as follows: given integers $C = (c_1,c_2,\dots,c_n)$ and a target number T, decide if there is a subset $S \subseteq \{1,\dots,n\}$ such that

$$\sum_{i \in S} c_j = T$$

SSP is a special case of KNAPSACK.

Theorem 49.3.17. *SSP is NP-hard. If $T = 0$, the problem is NP-complete. If the integers in C are all positive, then the problem is NP-complete.*

The *partition problem* is a variant of SSP where all inputs are positive, and the target sum is exactly half the inputs. Or equivalently,

$$\sum_{i \in S} c_j = \sum_{i \in \{1,\ldots,n\} \setminus S} c_j$$

Theorem 49.3.18. *PARTITION is NP-complete.*

Suppose we have n objects, each of a given size, and some bins of equal capacity. We want to assign the objects to the bins, using as few bins as possible. Of course, the total size of the objects assigned to one bin should not exceed its capacity. Without loss of generality, we assume that the capacity of each bin is 1.

Given a list of non-negative numbers $a_1,a_2,\ldots,a_n \leq 1$, the *bin packing problem* is to find a natural k and an assignment $f : \{1,\ldots,n\} \to \{1,\ldots,k\}$ with,

$$\sum_{\{i:f(i)=j\}} a_i \leq 1$$

for all $j \in \{1,\ldots,k\}$, such that k is minimum.

Theorem 49.3.19. *The following problem is NP-complete: given an instance I of Bin Packing, decide whether I has a solution with two bins.*

Proof. Membership in NP is obvious. To prove completeness, we transform PARTITION by choosing

$$a_i = \frac{2c_i}{\sum_{i=1}^{n} c_i}$$

■

Corollary 49.3.19.1. *It is NP-complete to distinguish whether the optimal solution is 2 or 3, and hence for any $\varepsilon > 0$, there is no $(3/2 - \varepsilon)$-factor approximation algorithm for Bin Packing.*

In the *online* variant of the problem, items arrive one after another, and the irreversible decision of where to place an item has to be made before knowing the next item, or even if there will be another one.

Most algorithms follow the same general pattern: if the next item fits in one of the currently open bins, put it in one of the bins. Otherwise, open a new bin and put the new item in it. These algorithms differ in the criterion by which they choose the open bin for the new item in the first step.

One algorithm for the online variant of BIN-PACKING is the *next fit algorithm.* In next fit, we always keep a single open bin. When the new item doesn't fit into it, it closes the current bin, and opens a new bin. Its advantage is that it is a bounded-space algorithm, since it only needs to keep a single open bin in memory.

Algorithm 25 Next Fit

```
procedure NF($C$)
    $k = 1$
    $S = 0$
    for $i = 1$ to $n$ do
        if $S + c_i > 1$ then
            $k = k + 1$
            $S = 0$
        end if
        $f(i) = k$
        $S = S + c_i$
    end for
    return $k$,$f$
end procedure
```

Let NF(I) denote the number of bins found by the next fit algorithm, OPT(I) be the minimum number of bins, and $\Sigma(I) = \sum_i c_i$

Theorem 49.3.20.

$$\mathrm{NF}(I) \leq 2\lceil\Sigma(I)\rceil - 1 \leq 2\,\mathrm{OPT}(I) - 1$$

Proof. Let $k = \mathrm{NF}(I)$, and let f be the assignment found by the next fit algorithm. For $j \in \{1, \ldots, \lfloor k/2 \rfloor\}$, we have,

$$\sum_{\left\{i:f(i)\in\{2j-1,2j\}\right\}} c_i > 1$$

Adding these inequalities, we obtain $\lfloor k/2 \rfloor < \Sigma(I)$. Since the left side is an integer, we conclude that $(k-1)/2 \leq \lfloor k/2 \rfloor \leq \lceil\Sigma(I)\rceil - 1$, proving $k \leq 2\lceil\Sigma(I)\rceil - 1$. The second inequality follows from the obvious fact that $\lceil\Sigma(I)\rceil \leq \mathrm{OPT}(I)$. ■

Corollary 49.3.20.1. *Next fit is a 2-factor approximation algorithm for BIN-PACKING.*

The *next-k-fit algorithm* (NkF) keeps the last k bins open, and chooses the first bin in which the item fits. It is therefore a k-bounded-space algorithm. For $k \geq 2$, NkF gives better results than NF, but increasing k to constant values large than 2 improves the alorithm no further in its worst-case behaviour.

The *first fit algorithm* keeps all bins open, in the order they were opened, attempting to place new items into the first bin in which it fits.

Algorithm 26 First Fit

```
procedure FF($C$)
    for $i = 1$ to $n$ do
        $f(i) = \min\left\{ j \in \mathbb{N} : \sum_{\{k<i:f(k)=j\}} a_k + a_i \leq 1 \right\}$
        $k = \max_{i\in\{1,\ldots,n\}} f(i)$
    end for
    return $k$,$f$
end procedure
```

Theorem 49.3.21. $\mathrm{FF}(I) \leq 7/4\,\mathrm{OPT}(I)$.

The *first fit decreasing algorithm* (FFD) sorts the items by descending size, then calls first fit. FFD requires being able to see the entire list first and thus solves the *offline* variant of Bin Packing.

Algorithm 27 First Fit Decreasing

```
procedure FFD($C$)
    SORT(C)
    return NF(C)
end procedure
```

Theorem 49.3.22. *FFD is a 3/2-factor approximation algorithm for Bin Packing.*

Proof. Let I be an instance of the problem and let $k = \mathrm{FFD}(I)$. Consider the jth bin for $j = \lceil 2k/3 \rceil$. If it contains an item of size greater than $1/2$, then each bin with smaller index did not have space for this item. Therefore, each of these bins has been assigned an item before. As the items are considered in non-increasing order, there are at least j items of size greater than $1/2$. Thus, $\mathrm{OPT}(I) \geq j \geq 2k/3$.

Otherwise, the jth bin and hence each bin with greater index contains no item of size greater than $1/2$. Therefore, the bins $j,j+1,\dots,k$ contain at least $2(k-j)+1$ items, none of which fit into bins $1,\dots,j-1$. Thus,

$$\begin{aligned}\Sigma(I) &> \min\{j-1, 2(k-j)+1\}\\ &\geq \min\left\{\left\lceil\frac{2k}{3}\right\rceil - 1, 2\left(k-\left(\frac{2k}{3}+\frac{2}{3}\right)\right)+1\right\}\\ &= \left\lceil\frac{2k}{3}\right\rceil - 1\end{aligned}$$

since $\mathrm{OPT}(I) \geq \Sigma(I) > \lceil 2k/3 \rceil - 1$, $\mathrm{OPT}(I) \geq \lceil 2k/3 \rceil \geq 2k/3$. ■

49.3.6 Steiner Trees

Let G be an undirected graph, and let $T \subseteq V(G)$. A *Steiner tree* for T is a set S such that $T \subseteq V(S) \subseteq V(G)$ and $E(S) \subseteq E(G)$. The elements of T are called *terminals*, and the elements of $V(G) \setminus T$ are the *Steiner points* of S.

Given an undirected weighted graph (G,w) and a set $T \subseteq V(G)$, the *Steiner tree problem* is to find a Steiner tree S for T of minimum weight.

MST ($T = V(G)$) and SHORTEST-PATH ($|T| = 2$) are special cases of STEINER-TREE solvable in polynomial-time.

Theorem 49.3.23. *STEINER-TREE is NP-hard, even for unit weights.*

Proof. We give a transformation from MINIMUM-VERTEX-COVER, which is known to be NP-complete.

Given a graph G, we transform it to a graph H by adding, for each edge $(u,v) \in E(G)$ a new vertex $x_{u,v}$ which is adjacent to both u and v, and by adding edges which are missing in G.

We set $w(e) = 1$ for all edges $e \in E(H)$, and set $T = \{x_{u,v} : u,v \in E(G)\}$. We will show that G has a vertex cover of size k if and only if H has a Steiner tree for T with $k + |E(G)| - 1$ edges.

Let $T \cup X$ be the set of vertices of a Steiner tree S in H, where $X \subseteq V(G) \subseteq V(H)$. The set T is independent in H as these vertices have neighbours only among the vertices of G. We also have that S is a connected graph, so every vertex of T has a neighbour in X. This means that X is a vertex cover in G, and $E(S) = |T| + |X| - 1 = |E(G)| + |X| - 1$, completing the forward implication.

Conversely, let X be a vertex cover in G. We can connect the vertices of X in the graph H by $|X| - 1$ edges. Since every edge of G is covered by a vertex of X, every vertex of T is connected by an edge to a vertex of X in H. The $|X| - 1$ edges connecting X and the $|T|$ edges incident to the vertices of T create a Steiner tree with $|T| + |X| - 1 = |E(G)| + |X| - 1$ edges, as required, thus completing the reverse implication. ■

Let (G,w) be a weighted graph with all weights positive. The *metric closure* of (G,w) is the pair (G^*,c^*), where G^* is the graph with $V(G^*) = V(G)$ in which two vertices x and y are adjacent if and only if they are connected in G by a path and $w^*(xy)$ equals the length of a shortest path between x and y in G.

Remark. w^* is symmetric, point separating, and satisfies the triangle inequality, thus defining a metric on G^*.

Theorem 49.3.24. *Let (G,w) be a weighted graph with all weights positive, let (G^*,w^*) be its metric closure, and let $T \subseteq V(G)$. If S is an optimum Steiner tree for T in G, and M is a minimum spanning tree in $G^*[T]$, then $w^*(E(M)) \leq 2w(E(S))$.*

Proof. Consider the graph H containing two copies of each edge of S. Then, H is Eulerian and hence contains an Eulerian walk W in H. This walk defines a Hamiltonian cycle W' in $G^*[T]$. Since w^* satisfies the triangle inequality,

$$\begin{aligned} w^*(W') &\leq w(W) \\ &= w(E(H)) \\ &= 2w(E(S)) \end{aligned}$$

However, we also have $w^*(E(M)) \leq w^*(W')$ since by deleting one edge of W' we obtain a spanning tree in $G^*[T]$. ■

This suggests the following 2-factor approximation algorithm for STEINER-TREE.

Algorithm 28 Steiner Tree

```
Compute the metric closure $(G^*,w^*)$
Compute the shortest path $P_{s,t}$ for all $s,t \in T$
Find a minimum spanning tree $M$ in $G^*[T]$
$E(S) = \bigcup_{(u,v) \in E(M)} P_{u,v}$
$V(S) = \{R \subseteq V(G) : (\forall v \in R : (\exists (u,v) \in E(S)) \vee (\exists (v,u) \in E(S)))\}$
return A minimal connected subgraph of $S$
```

Theorem 49.3.25. *This algorithm is a 2-factor approximation for STEINER-TREE and can be implemented in $O(|V(G)|^3)$ time.*

Recall that TSP is NP-hard.

Theorem 49.3.26. *Unless $P = NP$, there is no k-factor approximation algorithm for TSP for any $k \geq 1$.*

Proof. We will show that a k-factor approximation algorithm A for TSP implies that HAMILTONIAN-CYCLE (which is NP-complete) can be solved in polynomial time.

Given an instance G of HAMILTONIAN-CYCLE, we construct an instance G^* of TSP with $n = |V(G)|$ nodes and distances $w((i,j))$ as follows: if i is adjacent to j, then $w((i,j)) = 1$; otherwise, $w((i,j)) = 2 + (k - 1)n$.

Now, we apply A to the constructed instance of TSP. If the returned tour has length n, then this tour is a Hamiltonian cycle in G. Otherwise the returned tour has length at least $(n-1)+2+(k-1)n = kn+1$. Assuming that A is a k-factor approximation algorithm, we conclude that $(kn+1)/\operatorname{OPT}(G^*) \geq A(G^*)/\operatorname{OPT}(G^*) \geq k$, where $\operatorname{OPT}(G^*)$ is the length of the optimum tour. Hence, $\operatorname{OPT}(G^*) \geq n+1/k > n$, showing that G has no Hamiltonian cycle. ■

Metric TSP, also known as Δ-TSP, is TSP such that the underlying graph is its own metric closure. That is, given a positive weighted complete graph (K_n,w) with $w : (E(K_n)) \to \mathbb{R}_{\geq 0}$ satisfying $w(x,z) \leq w((x,y)) + w(y,z)$ for all $x,y,z \in V(K_n)$, find a Hamiltonian cycle of minimum weight.

Theorem 49.3.27. *METRIC-TSP is NP-hard.*

The greedy algorithm works badly in this problem and is not a k-factor approximation algorithm. However, the *double tree algorithm* has better performance on this problem.

Algorithm 29 Double Tree

```
1: Find a minimum weight spanning tree $T$ in $K_n$ with respect to $w$.
2: Walk around the tree, doubling each edge to create a Eulerian walk (circuit).
3: In the Eulerian walk, ignore all but the first occurence of each vertex.
4: **return** the tour constructed in line 3.
```

Theorem 49.3.28. *The double tree algorithm is a 2-factor approximation algorithm for METRIC-TSP.*

Proof. Clearly the algorithm is polynomial. Also, we have $w(E(T)) \leq \operatorname{OPT}(K_n,c)$, since by deleting an edge from any tour we obtain a spanning tree. Finally, the solution found by the algorithm is of weight at most $2w(E(T)) \leq 2\operatorname{OPT}(K_n,c)$. ■

49.4 Discrete Probability

A *probability space* consists of three elements:

- A *sample space*, Ω, which is the set of all possible outcomes;
- An *event space*, a family of sets $\mathcal{F} \subseteq \mathcal{P}(\Omega)$, with each set representing an *event*;
- A *probability function*, $\mathbb{P} : \mathcal{F} \to [0,1]$, such that
  - $\mathbb{P}(\Omega) = 1$
  - $\mathbb{P}(\varnothing) = 0$
  - If $\{A_i\}_{i=1}^{\infty} \subseteq \mathcal{F}$ are countably many disjoint events, then $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

and a probability space is *discrete* if Ω is at most countably infinite. An event is *elementary* if it is a set of size 1.

49.4.1 Boole's Inequality

If $\{A_i\}_{i=1}^{\infty} \subseteq \mathcal{F}$ are countably many events, then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

Two events, A and B, are *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. A finite set of events is *pairwise independent* if every pair of events in the set is independent. A finite set of events is *mutually independent* if every event is independent from every other set and every intersection of every other event.

49.4.2 Bayes' Theorem

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

for any events, A and B. If A and B are independent, then this reduces to $\mathbb{P}(A|B) = \mathbb{P}(A)$.

49.4.3 Law of Total Probability

If A is an event that can be written as a countable partition, $A = \{B_i\}_{i=1}^{\infty}$, then

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i)$$

or equivalently,

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A|B_i)\mathbb{P}(B_n)$$

49.4.4 Expected Value

The *expected value* of a random variable, X, is the weighted average of all possible values of X.

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} x_i p_i$$

where x_i are the possible values of X, and p_i are their corresponding probabilities of occurrence.

Expectation is linear, so,

$$\mathbb{E}\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1} c_i \mathbb{E}(X_i)$$

Markov's Inequality: If X is a random variable, and $a > 0$, then,

$$\mathbb{P}(x \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

49.4.5 The Probabilistic Method

49.4.5.1 First Moment Method

If X is a non-negative integer-valued random variable, we can find a lower bound for $\mathbb{P}(X = 0)$ using Markov's inequality. Since X takes integer values, $\mathbb{P}(X > 0) = \mathbb{P}(X \geq 1)$, so $\mathbb{P}(X \geq 1) \leq \mathbb{E}(X)$.

49.4.5.2 Second Moment Method

Similarly,

$$\mathbb{P}(X > 0) \geq \frac{(\mathbb{E}(X))^2}{\mathbb{E}(X^2)}$$

49.4.5.3 Lovász Local Lemma

We can (non-constructively) prove the existence of a structure with some desired property by proving that the probability of that property occuring in a random structure is greater than zero, or, equivalently, by proving that the probability of that property not occuring in a random structure is less than one.

Example. Let n,m,$d \in \mathbb{N}$. Suppose a town with n people contains m clubs, each of which contains exactly d members. Any person can be a member of multiple clubs, and there may also be people who are not members of any club.

Prove that, if $m < 2^{d-1}$, then there is always a way to partition the town into two sets in such a way that no club has all its members completely contained in either set.

Let Ω be the set of clubs, and suppose that $m < 2^{d-1}$.

Randomly assign each person to one of the two sets of the partition with equal probability $\frac{1}{2}$ of each.

For each club, $C \in \Omega$, let X_C be the event that C is contained entirely within one set of the partition. $\mathbb{P}(X_C)$ is the probability that every person $c \in C$ is assigned to the same set, multiplied by two, as there are two possible sets to be contained within. So,

$$\begin{aligned} \mathbb{P}(X_C) &= \frac{1}{2^d} \times 2 \\ &= \frac{1}{2^{d-1}} \end{aligned}$$

The probability that at least one club is a subset of one of the partition sets is therefore given by, $\mathbb{P}\left(\bigcup_{C\in\Omega} X_c\right)$ which can be bounded above by Boole's inequality.

$$\begin{aligned} \mathbb{P}\left(\bigcup_{C\in\Omega} X_c\right) &\leq \sum_{C\in\Omega} \mathbb{P}(X_c) \\ &\leq \frac{m}{2^{d-1}} \end{aligned}$$

As m is less than 2^{d-1}, $\frac{m}{2^{d-1}} < 1 \to \mathbb{P}\left(\bigcup_{C\in\Omega} X_c\right) < 1$, so the probability that a club is entirely contained within one set of the partition is less than 1 when $m < 2^{d-1}$. It follows that, if $m < 2^{d-1}$, there exists at least one partition such that no club has all its members completely contained within one set of the partition. △

49.5 Linear Programming

A *linear program* is a problem of the form,

"Minimise $\mathbf{c} \cdot \mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^n$, subject to the *constraint* $\mathbf{Ax} \leq \mathbf{b}$,
where $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, and the inequality is considered componentwise."

The vector $\mathbf{c}$ is then called the *cost vector* or the *objective function*.

The set,

$$\{\mathbf{x} \in \mathbb{R}^n : \mathbf{Ax} \leq \mathbf{b}\}$$

is called the *feasible region* of the linear program, as it contains every value of $\mathbf{x}$ that satisfies the constraint. The objective is then to find a value of $\mathbf{x}$ in the feasible region that yields a minimum value for $\mathbf{c} \cdot \mathbf{x}$.

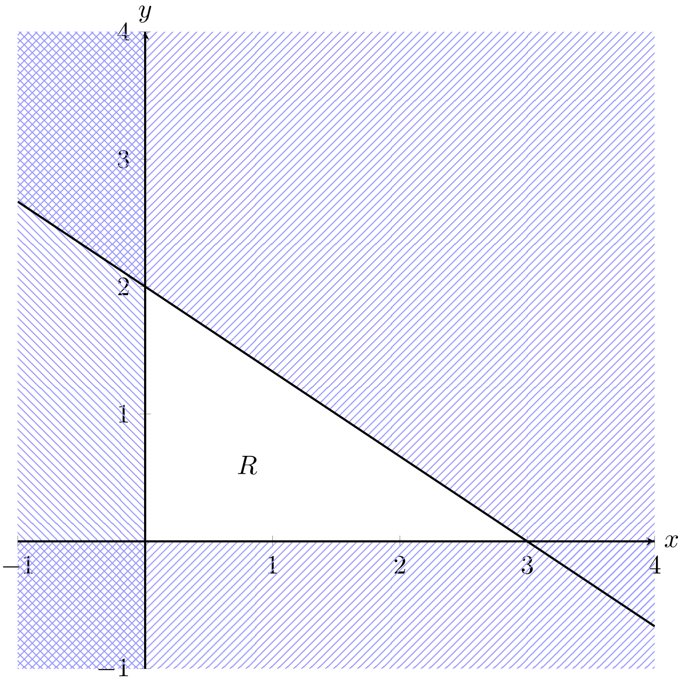
Example. Minimise $\begin{bmatrix}1\\2\end{bmatrix} \cdot \mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^2$ satisfiying,

$$\begin{bmatrix}-1 & 0\\0 & -1\\2 & 3\end{bmatrix}\mathbf{x} \leq \begin{bmatrix}0\\0\\6\end{bmatrix}$$

Writing $\mathbf{x} = [x,y]$, we can expand out the constraint into the system of equations,

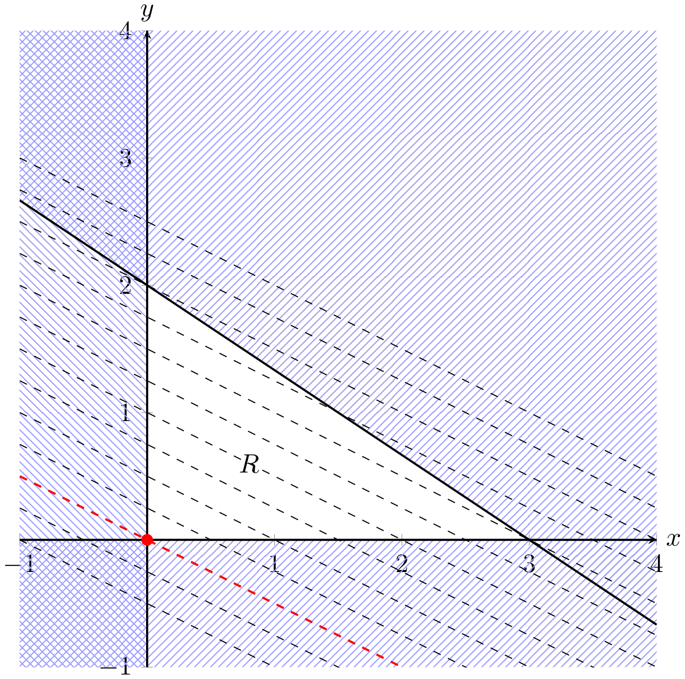
$$\begin{aligned}-x &\leq 0\\-y &\leq 0\\2x+3y &\leq 6\end{aligned}$$

which define the feasible region R,



It may be helpful to shade the unwanted region for each inequality so the feasible region is the only unshaded area left. Otherwise, you might have difficulty deciding which part is shaded by every inequality, especially for larger constraint matrices.

The objective function to minimise is then $x + 2y$, and we can picture various values of this function by looking at the lines $x + 2y = k$ for various values of k:



with k decreasing further down. Clearly, the objective function is minimised at the point $\mathbf{x} = (0{,}0)$, with value 0. △

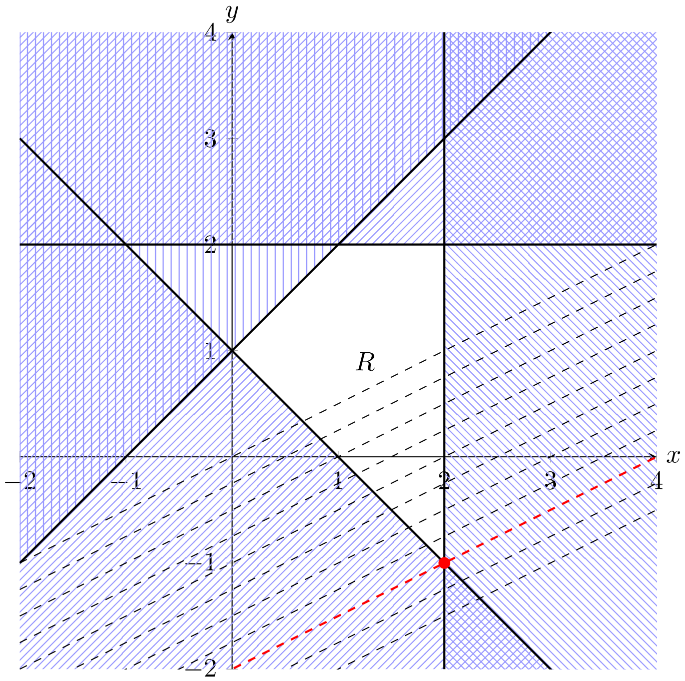
Example. Minimise, $\begin{bmatrix} -1 \\ 2 \end{bmatrix} \cdot \mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^2$ satisfiying,

$$\begin{bmatrix} -1 & -1 \\ -1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{x} \leq \begin{bmatrix} -1 \\ 1 \\ 2 \\ 2 \end{bmatrix}$$

Again, writing $\mathbf{x} = [x{,}y]$, we can expand out the constraint into the system,

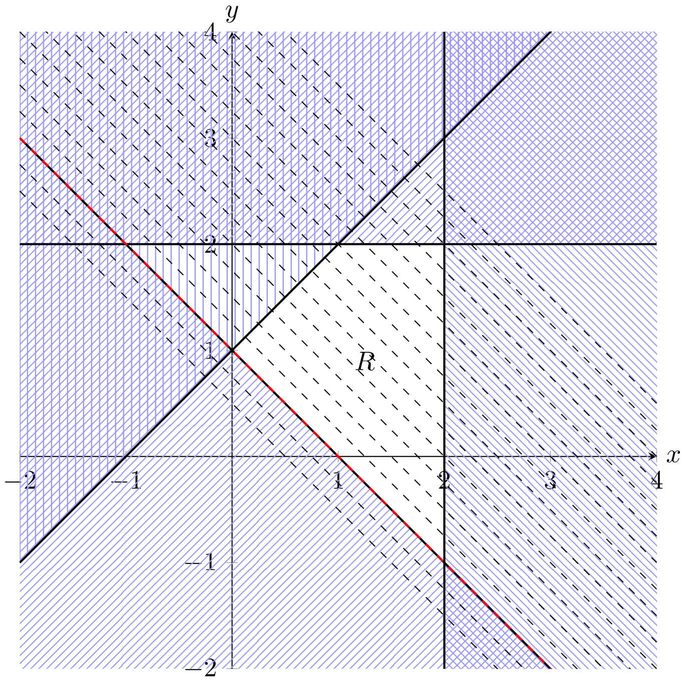
$$\begin{aligned} -x - y &\leq -1 \\ -x + y &\leq 1 \\ x &\leq 2 \\ y &\leq 2 \end{aligned}$$

with the objective function $-x + 2y = k$.



Again, the value of k decreases lower down, so the minimum is achieved at $(2, -1)$ with value -4. △

Now, suppose the objective function was instead $x + y$:



Now, we have a whole line of minimal solutions, and any point on the line $x + y = 1$ between $x = 0$ and $x = 2$ is a minimal solution.

49.5.1 Polyhedra

A *hyperplane* is a subset of $\mathbb{R}^n$ of the form,

$$\{\mathbf{x} \in \mathbb{R}^n : \mathbf{n} \cdot \mathbf{x} = \mathbf{b}\}$$

where $\mathbf{n}$ is the normal vector of the plane, while a *halfspace* is a subset of the form,

$$\{\mathbf{x} \in \mathbb{R}^n : \mathbf{n} \cdot \mathbf{x} \leq \mathbf{b}\}$$

so the hyperplane given by the equality is the boundary of this set. Notice that we can write a dot product as,

$$H = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{n}^\top \mathbf{x} = \mathbf{b}\}$$

and we can then consider $\mathbf{n}$ as a $1 \times n$ matrix.

A *polyhedron* is the intersection of finitely many halfspaces, and a *polytope* is a bounded polyhedron, or equivalently, a polytope is the convex hull of a finite set.

The *face* of a polyhedron $P = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$ that minimises some vector $\mathbf{c} \in \mathbb{R}^n$ is defined by,

$$\text{face}_{\mathbf{c}}(P) = \{\mathbf{x} \in P : \forall \mathbf{y} \in P, \mathbf{c} \cdot \mathbf{x} \leq \mathbf{c} \cdot \mathbf{y}\}$$

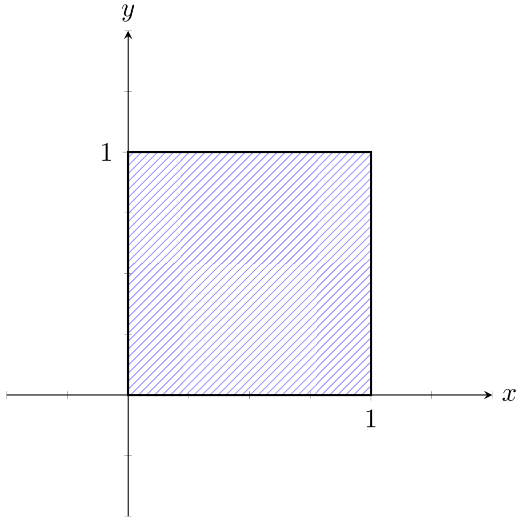
Example. Consider the polyhedron given by,

$$P = \big\{(x,y) \in \mathbb{R}^2 : 0 \leq x \leq 1, 0 \leq y \leq 1\big\}$$

What is the face minimising:

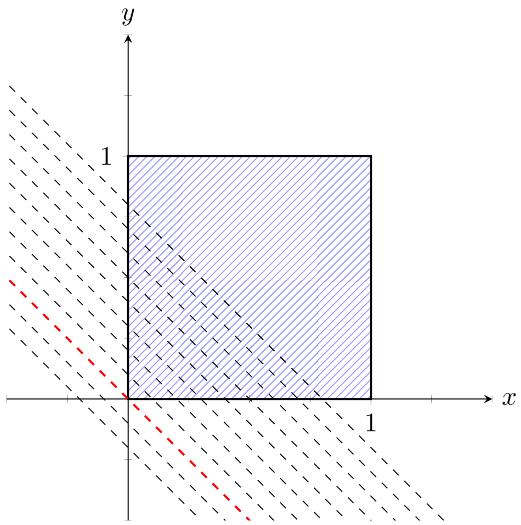
1. $\mathbf{c} = [1,1]$;
2. $\mathbf{c} = [-1,2]$;
3. $\mathbf{c} = [0,1]$?

P is given by,



$$\begin{aligned}\text{face}_{[1,1]}(P) &= \big\{(x,y) \in P : \forall (x',y') \in P, [1,1] \cdot [x,y] \leq [1,1] \cdot [x',y']\big\} \\ &= \big\{(x,y) \in P : \forall (x',y') \in P, x + y \leq x' + y'\big\}\end{aligned}$$

So, $\text{face}_{[1,1]}(P)$ is the set of points where the line $x + y = k$ intersects P for a minimal value of k:

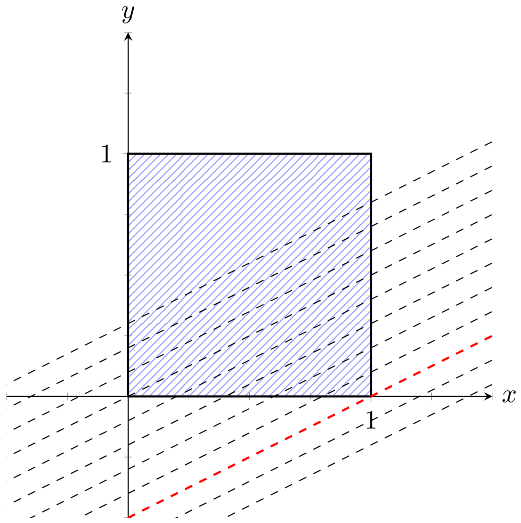


the minimal valued line intersects the polygon at (0,0), giving $\text{face}_{[1,1](P)} = \big\{(0{,}0)\big\}$.

Next, take $\mathbf{c} = [-1{,}2]$.

$$\begin{aligned}\text{face}_{[-1,2]}(P) &= \big\{(x{,}y) \in P : \forall (x'{,}y') \in P, [-1{,}2]\cdot[x{,}y] \le [-1{,}2]\cdot[x'{,}y']\big\}\\ &= \big\{(x{,}y) \in P : \forall (x'{,}y') \in P, -x + 2y \le -x' + 2y'\big\}\end{aligned}$$

This time, we look at the line $-x + 2y = k$,

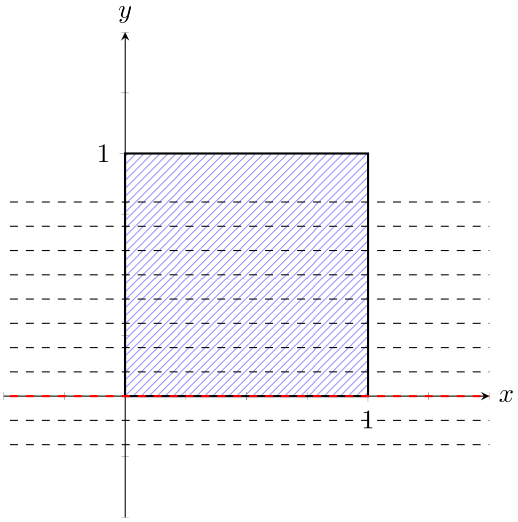


and now the minimal intersection point is at (1,0), so $\text{face}_{[-1,2]}(P) = \big\{(1{,}0)\big\}$

Next, take $\mathbf{c} = [0{,}1]$.

$$\begin{aligned}\text{face}_{[0,1]}(P) &= \big\{(x{,}y) \in P : \forall (x'{,}y') \in P, [0{,}1]\cdot[x{,}y] \le [0{,}1]\cdot[x'{,}y']\big\}\\ &= \big\{(x{,}y) \in P : \forall (x'{,}y') \in P, y \le y'\big\}\end{aligned}$$

So the line is now $y = k$,



Now, the minimal intersection points form an entire line segment, so $\text{face}_{[-1,2]}(P) = \big\{(x,0) : 0 \leq x \leq 1\big\}$. △

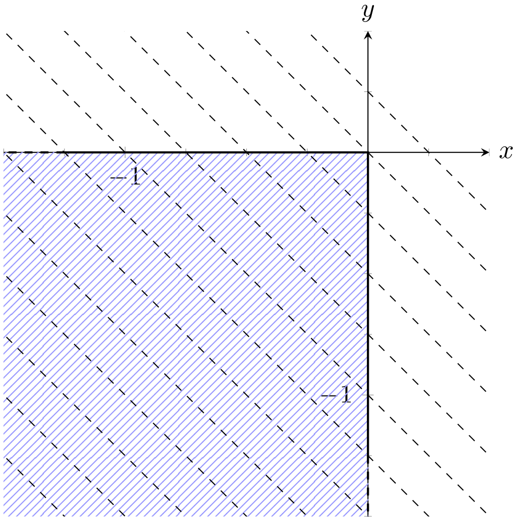
Example. Consider the polyhedron given by,

$$Q = \big\{(x,y) \in \mathbb{R}^2 : x \leq 0, y \leq 0\big\}$$

What is the face minimising $\mathbf{c} = [1,1]$?

$$\begin{aligned}\text{face}_{[1,1]}(Q) &= \big\{(x,y) \in Q : \forall (x',y') \in Q, [1,1] \cdot [x,y] \leq [1,1] \cdot [x',y']\big\} \\ &= \big\{(x,y) \in Q : \forall (x',y') \in Q, x + y \leq x' + y'\big\}\end{aligned}$$

We look at the line $x + y = k$:



so the minimum does not exist, and $\text{face}_{[1,1]}(Q) = \varnothing$. △

We have been solving these problems by finding minimal points, so we can also characterise faces as,

$$\text{face}_{\mathbf{c}}(P) = \{\mathbf{x} \in P : \forall \mathbf{y} \in P, \mathbf{c} \cdot \mathbf{x} \leq \mathbf{c} \cdot \mathbf{y}\}$$

$$= \left\{ \mathbf{Ax} \leq \mathbf{b} : \mathbf{c} \cdot \mathbf{x} = \min_{\mathbf{y} \in p} \mathbf{c} \cdot \mathbf{y} \right\}$$

and we can see that $\text{face}_{\mathbf{c}}(P)$ is a polyhedron if the minimum exists, and is otherwise empty.

The linear program,

“Minimise $\mathbf{c} \cdot \mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^n$ such that $\mathbf{Ax} = \mathbf{b}$.”

is equivalent to,

“Find $\mathbf{y} \in \text{face}_{\mathbf{c}}(P)$ where $P = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{Ax} \leq \mathbf{b}\}$ and compute $\mathbf{c} \cdot \mathbf{y}$.”

The *linear span* of a polyhedron P is the subspace,

$$\text{span}(\{\mathbf{x} - \mathbf{y} : \mathbf{x},\mathbf{y} \in P\})$$

of $\mathbb{R}^n$. The *dimension* of P is the dimension of its linear span.

Faces of dimension 0 are called *vertices*, and faces of dimension 1 are called *edges.*

49.5.2 Standard Form

Recall that a linear program is a problem of the form,

“Minimise $\mathbf{c} \cdot \mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^n$ such that $\mathbf{Ax} \leq \mathbf{b}$.”

where $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$.

A linear program is in *standard form* if it can be written as,

“Minimise $\mathbf{c} \cdot \mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^n$ such that $\mathbf{Ax} = \mathbf{b}$, and $\mathbf{x} \geq \mathbf{0}$.”

where $\mathbf{A} \in \mathbb{R}^{d \times n}$ and $\mathbf{b} \in \mathbb{R}^d$.

We can convert any linear program into standard form:

Algorithm 30 Linear Program Standard Form

```
1: Split each component $x_i$ of $\mathbf{x}$ into $x_i = x_i^+ - x_i^-$.
2: Add new variables to the inequality constraints to give $\mathbf{Ax} + \mathbf{s} = \mathbf{b}$.
3: Change the components $c_i$ of the cost vector $\mathbf{c}$ into $c_i = (c_i^+, - c_i^-)$, and set any components
   corresponding to slack variables to $c_s = 0$.
```

Example. Transform the following problem into standard form:

Minimise $\begin{bmatrix} 1 \\ 0 \end{bmatrix} \cdot \mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^2$ such that,

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \\ -1 & 1 \\ -1 & -1 \end{bmatrix} \mathbf{x} \leq \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

The constraint matrix gives,

$$\begin{aligned} x_1 + x_2 &\leq 1 \\ x_1 - x_2 &\leq 1 \\ -x_1 + x_2 &\leq 1 \\ -x_1 - x_2 &\leq 1 \end{aligned}$$

Perform the replacements on the components, and add slack variables to remove the inequalities to obtain,

$$\begin{aligned} x_1^+ - x_1^- + x_2^+ - x_2^- + s_1 &= 1 \\ x_1^+ - x_1^- - x_2^+ + x_2^- + s_2 &= 1 \\ -x_1^+ + x_1^- + x_2^+ - x_2^- + s_3 &= 1 \\ -x_1^+ + x_1^- - x_2^+ + x_2^- + s_5 &= 1 \end{aligned}$$

which can be written in matrix form as,

$$\begin{bmatrix} 1 & -1 & 1 & -1 & 1 & 0 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & 1 & 0 & 0 \\ -1 & 1 & 1 & -1 & 0 & 0 & 1 & 0 \\ -1 & 1 & -1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1^+ \\ x_1^- \\ x_2^+ \\ x_2^- \\ s_1 \\ s_2 \\ s_3 \\ s_4 \end{bmatrix} \leq \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Then, the original cost vector gives the constraint $\begin{bmatrix} 1 \\ 0 \end{bmatrix} \mathbf{x} = x_1$, so our new cost vector is,

$$\mathbf{c} = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

△

Example. Transform the following problem into standard form:

Minimise $\begin{bmatrix} 2 \\ 3 \end{bmatrix} \cdot \mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^2$ such that,

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix} \mathbf{x} \leq \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$$

The constraint matrix gives,

$$\begin{aligned} x_1 &\leq 1 \\ x_2 &\leq 1 \\ -x_1 - x_2 &\leq -1 \end{aligned}$$

so,

$$\begin{bmatrix} 1 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 1 & 0 \\ -1 & 1 & -1 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1^+ \\ x_1^- \\ x_2^+ \\ x_2^- \\ s_1 \\ s_2 \\ s_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$$

with cost vector given by,

$$\mathbf{c} = \begin{bmatrix} 2 \\ -2 \\ 3 \\ -3 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

△

To solve linear programs, we often look for faces of the feasible region. The standard form makes it easier to find vertices as solutions.

A vector $\mathbf{y} \in \mathbb{R}^n$ is a *basic solution* of a linear program if $\mathbf{Ay} = \mathbf{b}$ and the columns of $\mathbf{A}$ corresponding to non-zero entries of $\mathbf{y}$ are linearly independent. That is, if $\mathbf{A} = [\mathbf{A}_1|\mathbf{A}_2|\cdots|\mathbf{A}_n]$ and $\mathbf{y} = [y_1,y_2,\ldots,y_n]$, then $\{A_i : y_i \neq 0\}$ is a linearly independent set.

A *basic feasible solution* is a basic solution $\mathbf{y} \in \mathbb{R}^n$ such that $\mathbf{y} \geq \mathbf{0}$.

We will write $\mathbf{A}_i$ for the ith column of a matrix $\mathbf{A}$, and $\mathbf{a}_i$ for the ith row of $\mathbf{A}$. That is,

$$\mathbf{A} = [\mathbf{A}_1|\mathbf{A}_2|\cdots|\mathbf{A}_n] = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_d \end{bmatrix}$$

Example. Minimise $\begin{bmatrix} 1 \\ 2 \\ 0 \\ -1 \end{bmatrix} \cdot \mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^4$ such that $\mathbf{x} \geq \mathbf{0}$, and,

$$\underbrace{\begin{bmatrix} 1 & 0 & 1 & -2 \\ 0 & 1 & 1 & -2 \end{bmatrix}}_{\mathbf{A}} \mathbf{x} = \underbrace{\begin{bmatrix} 3 \\ -2 \end{bmatrix}}_{\mathbf{b}}$$

$\mathbf{y} = [5{,}0{,}-2{,}0]$ is a basic solution since $\mathbf{Ay} = \mathbf{b}$ and the first and third columns of $\mathbf{A}$ are linearly independent, but it is not a basic feasible solution since $\mathbf{y}$ contains a negative component.

$\mathbf{y} = [5{,}0{,}0{,}1]$ is a basic feasible solution since $\mathbf{Ay} = \mathbf{b}$ and the first and last columns of $\mathbf{A}$ are linearly independent, and $\mathbf{y} \geq \mathbf{0}$. △

To construct a basic solution, we choose d linearly independent columns $(\mathbf{A}_i)_{i\in\mathcal{I}}$ with $\mathcal{I} \subseteq [n]$ and $|\mathcal{I}| = d$, and set $y_i = 0$ for each $i \notin \mathcal{I}$. Augment these columns together into a $d\times d$ matrix $\mathbf{B} = [\mathbf{A}_{i_1}|\mathbf{A}_{i_2}|\cdots|\mathbf{A}_{i_d}]$.

Because the $(\mathbf{A}_i)$ are linearly independent, $\mathbf{B}$ is invertible, so $\mathbf{By} = \mathbf{b}$ has a unique solution given by

$$\mathbf{y_B} = \mathbf{B}^{-1}\mathbf{b} \in \mathbb{R}^d$$

where the coordinates in $\mathbb{R}^d$ are indexed by $\mathcal{I}$. Then, we set,

$$\mathbf{y} = \begin{cases} (\mathbf{y}_B)_i & i \in \mathcal{I} \\ 0 & i \notin \mathcal{I} \end{cases}$$

That is, we invert a matrix made of linearly independent columns of $\mathbf{A}$, then add in a 0 entry to the solution vector wherever we skipped a column from $\mathbf{A}$.

Example. Consider the polygon defined by,

$$P = \left\{(x,y) \in \mathbb{R}^n : -x+y \leq 1, x+y \leq 3, x \geq 0, y \geq 0\right\}$$

This can be written in matrix form as,

$$\begin{bmatrix} -1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ s_1 \\ s_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

noting that we did not have to split x and y as we are already given $x \geq 0$ and $y \geq 0$.

One linearly independent set is given by $\mathbf{A}_3$ and $\mathbf{A}_4$, so,

$$\mathbf{B} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and,

$$\begin{aligned} \mathbf{y_B} &= \mathbf{B}^{-1}\mathbf{b} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 3 \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ 3 \end{bmatrix} \end{aligned}$$

Since we skipped the first two columns, we have,

$$\mathbf{y} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 3 \end{bmatrix}$$

Because all the components are positive, this is a basic feasible solution.

Another linearly independent set is given by $\mathbf{A}_1$ and $\mathbf{A}_4$, so,

$$\mathbf{B} = \begin{bmatrix} -1 & 0 \\ 1 & 1 \end{bmatrix}$$

and,

$$\begin{aligned} \mathbf{y_B} &= \mathbf{B}^{-1}\mathbf{b} \\ &= \begin{bmatrix} -1 & 0 \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 3 \end{bmatrix} \\ &= \begin{bmatrix} -1 \\ 4 \end{bmatrix} \end{aligned}$$

Columns 2 and 3 were omitted from $\mathbf{B}$, so we have,

$$\mathbf{y} = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 4 \end{bmatrix}$$

This time, we have a negative component, so this basic solution is *not* feasible. △

Theorem 49.5.1. *A vector* $\mathbf{v} \in \mathbb{R}^n$ *is a basic feasible solution of a linear program if and only if it is a vertex of the corresponding polyhedron defined by* $P = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq 0\}$.

Theorem 49.5.2. *Either* $\min_{\mathbf{x} \in P} \mathbf{c} \cdot \mathbf{x} = -\infty$, *or there is a basic feasible solution* $\mathbf{y}$ *with* $\mathbf{c} \cdot \mathbf{y} \leq \mathbf{c} \cdot \mathbf{x}$ *for all* $\mathbf{x} \in P$.

49.6 The Simplex Algorithm

49.6.1 Geometric Simplex

Recall that an edge of a polyhedron is a one dimensional face. If a vector $\mathbf{y} \in P$ lies on an edge E, but does not lie on a vertex, then, up to scaling, there is a unique vector $\mathbf{d}$ such that $\mathbf{y} + \lambda\mathbf{d} \in E$ for sufficiently small λ. That is, $\mathbf{d}$ is the *direction vector* pointing along the edge.

Walking along an edge like this, we can eventually reach a vertex (assuming the edge is not a half-line that goes infinity in some direction). Because basic feasible solutions are vertices of polyhedra, this suggests an algorithm to find optimal basic feasible solutions:

1. Find a vertex of P.
2. Walk along edges of P, moving to vertices that reduce $\mathbf{c} \cdot \mathbf{x}$.
3. Stop when this isn't possible anymore.

Let $\mathbf{y} \in \mathbb{R}^n$ be a basic feasible solution obtained from the $d \times d$ matrix $\mathbf{B}_{\mathcal{I}} = [\mathbf{A}_{i_1}|\mathbf{A}_{i_2}|\cdots|\mathbf{A}_{i_d}]_{i\in\mathcal{I}}$ consisting of columns of $\mathbf{A}$, indexed by $\mathcal{I} \subseteq [n]$ with $|\mathcal{I}| = d$. Recall that $y_i = 0$ for all $i \notin \mathcal{I}$ by construction.

Denote by $\mathbf{y}_{\mathcal{I}}$ the vector obtained by restricting $\mathbf{y}$ to coordinates in $\mathbf{I}$. Then, $\mathbf{B}_{\mathcal{I}}\mathbf{y}_{\mathcal{I}} = \mathbf{b}$.

Let $k \notin \mathcal{I}$. We will look for a vector $\mathbf{d}$ with $\mathbf{d}_j = 0$ for all $\mathbf{j} \in \mathcal{I} \cup \{k\}$, $d_k = 1$, and $\mathbf{y} + \lambda\mathbf{d}$ is feasible for some $\lambda > 0$.

Then,

$$\begin{aligned}
\mathbf{A}(\mathbf{y} + \lambda\mathbf{d}) &= \mathbf{b} \\
\mathbf{A}\mathbf{y} + \lambda\mathbf{A}\mathbf{d} &= \mathbf{b} \\
\mathbf{b} + \lambda\mathbf{A}\mathbf{d} &= \mathbf{b} \\
\lambda\mathbf{A}\mathbf{d} &= \mathbf{0}
\end{aligned}$$

So $\mathbf{A}\mathbf{d} = \mathbf{0}$, and we can rewrite this as,

$$\mathbf{0} = \sum_{j=1}^{n} \mathbf{A}_j d_j$$

Recall that $d_j = 0$ for all $j \notin \mathcal{I} \cup \{k\}$, and $d_k = 1$, so,

$$\begin{aligned}
&= \sum_{j\in\mathcal{I}} A_j d_j + A_k \\
&= \mathbf{B}_{\mathcal{I}}\mathbf{d}_{\mathcal{I}} + \mathbf{A}_k
\end{aligned}$$

so $\mathbf{d}_{\mathcal{I}} = -\mathbf{B}_{\mathcal{I}}^{-1}\mathbf{A}_k$, giving,

$$d_j = \begin{cases} -\left(\mathbf{B}^{-1}\mathbf{A}_k\right)_j & j \in \mathcal{I} \\ 1 & j = k \\ 0 & j \notin \mathcal{I} \cup \{1\} \end{cases}$$

This vector $\mathbf{d}$ is called the *kth basic direction at* $\mathbf{y}$, and it depends on the choice of $\mathcal{I}$ and k.

Note that it is not always possible to find $\lambda > 0$ such that $\mathbf{y} + \lambda\mathbf{d}$ is feasible. That is, that $\mathbf{y} + \lambda\mathbf{d} \geq \mathbf{0}$.

A basic feasible solution $\mathbf{y}$ is *degenerate* if $|\{i : y_i \neq 0\}| < d$, and is *nondegenerate* otherwise.

Theorem 49.6.1. *If* $\mathbf{y}$ *is a nondegenerate basic feasible solution, then the kth basic direction is feasible.*

If this is the case, then $\mathbf{d}$ points along an edge, and we can move along it. Furthermore,

$$\mathbf{c} \cdot (\mathbf{y} + \lambda \mathbf{d}) = \mathbf{c} \cdot \mathbf{y} + \lambda \mathbf{c} \cdot \mathbf{d}$$

so the cost decreases if and only if $\mathbf{c} \cdot \mathbf{d} < 0$. We also have,

$$\begin{aligned}
\mathbf{c} \cdot \mathbf{d} &= \sum_{j=1}^{n} c_j d_j \\
&= \sum_{j \in \mathcal{I}} c_j d_j + c_k \\
&= \mathbf{c}_{\mathcal{I}} \cdot \mathbf{d}_{\mathcal{I}} + c_k \\
&= \mathbf{c}_{\mathcal{I}} \cdot (-\mathbf{B}_{\mathcal{I}}^{-1} \mathbf{A}_k) + c_k \\
&= c_k - \mathbf{c}_{\mathcal{I}}^\top \mathbf{B}_{\mathcal{I}}^{-1} \mathbf{A}_k
\end{aligned}$$

The *reduced cost* in direction k with respect to a basic feasible direction corresponding to $\mathcal{I}$ is given by $\overline{c}_i = c_i - \mathbf{c}_{\mathcal{I}}^\top \mathbf{B}_{\mathcal{I}}^{-1} \mathbf{A}_k$, and the *reduced cost vector* is given by $\overline{\mathbf{c}} = (\overline{c}_i)_{i=1}^n$

Note that if $\mathbf{y}$ is degenerate, then $\lambda^\bullet$ may be zero, and $\mathbf{y}' = \mathbf{y}$.

Theorem 49.6.2. *Let* $\mathbf{y}$ *be a basic feasible solution corresponding to* $\mathcal{I} \subseteq [n]$*, and let* $\overline{\mathbf{c}}$ *be the reduced cost vector. Then,*

- *If* $\overline{\mathbf{c}} \geq \mathbf{0}$*, then* $\mathbf{y}$ *is optimal.*
- *If* $\mathbf{y}$ *is optimal and nondegenerate, then* $\overline{\mathbf{c}} \geq 0$*.*

Algorithm 31 Geometric Simplex

```
1: Find a basic feasible solution $\mathbf{y}$ corresponding to $\mathcal{I} \subseteq [n]$.
2: Compute the reduced cost vector given by $\overline{c}_j = c_j - c_{\mathcal{I}}^\top \mathbf{B}_{\mathcal{I}}^{-1} \mathbf{A}_j$, where $c_j$ is the $j$th component of the cost vector $\mathbf{c}$. If $\overline{\mathbf{c}} \geq 0$, then $\mathbf{y}$ is optimal, and we may stop.
3: Choose $k \notin I$ with $\overline{c}_i < 0$, and compute the $k$th basic direction $\mathbf{d}$, given by
       $d_j = \begin{cases} -\left(\mathbf{B}^{-1}\mathbf{A}_k\right)_j & j \in \mathcal{I} \\ 1 & j = k \\ 0 & j \notin \mathcal{I} \cup \{1\} \end{cases}$
   If $\mathbf{d} \geq 0$, then the optimal cost is $-\infty$, and we may stop.
4: If $d_j < 0$ for some $j$, then let $\lambda^\bullet = \min_{d_j < 0} \frac{-y_j}{d_j}$, and let $\ell$ be the value of $j$ that achieves this minimum. That is, $\lambda^\bullet = \frac{-y_\ell}{d_\ell}$.
5: Set $\mathbf{y} = \mathbf{y} + \lambda^\bullet \mathbf{d}$ and $\mathcal{I} = (\mathcal{I} \setminus \{\ell\}) \cup \{k\}$. Go to step 2.
```

Example. TO DO △

49.6.2 Graph Optimisation Problems

The following problems on graphs can all be expressed as linear programs:

1. MST (Minimum Spanning Tree);
2. SHORTEST-PATH;

3. MAX-FLOW;
4. MAXIMUM-MATCHING;

TO DO.

Example. MST △

Example. SHORTEST-PATH △

Example. MAX-FLOW △

Example. MAXIMUM-MATCHING △

49.6.3 Simplex Tableau

This algorithm is rather involved, so we begin with an annotated worked example to give an overview of the method.

Example. Minimise $P = \begin{bmatrix} 3 \\ -1 \end{bmatrix} \mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^2$ subject to,

$$\begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & 4 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ r \\ s \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix}$$

and x,y,r,$s \geq 0$. (Here, r and s are slack variables.)

The *initial tableau* is written as follows:

Basic variable	x	y	r	s	Value
r	2	1	1	0	12
s	1	4	0	1	8
P	3	-1	0	0	0

The first row of the table shows the first constraint, the second row shows the second constraint, and the final *objective row* shows the objective function.

The "Basic variable" column indicates the variables that are not currently at zero. We start at the vertex (0,0), so $x = y = 0$.

Any variables in a simplex variable that are not basic variables have the value 0.

If $x = y = 0$, then $r = 12$ from the first row of the constraint matrix, and similarly, $s = 8$. We currently therefore have,

$$\mathbf{x} = \begin{bmatrix} 0 \\ 0 \\ 12 \\ 8 \end{bmatrix}$$

as our basic feasible solution with total value $P = 0$.

We scan the objective row of the tableau for the most negative number. This gives the *pivot column.* In this case, the pivot column is the y column.

For each other row, we then calculate a θ *value*, each given by dividing the value entry by the pivot entry.

Basic variable	x	y	r	s	Value	θ value
r	2	1	1	0	12	$12/1 = 12$
s	1	4	0	1	8	$8/4 = 2$
P	3	-1	0	0	0	

Next, we select the row containing the smallest positive θ value to be the *pivot row*.

Basic variable	x	y	r	s	Value	θ value
r	2	1	1	0	12	12
s	1	4	0	1	8	2
P	3	-1	0	0	0	

The entry at the intersection is then the *pivot*. We divide the values in the pivot row by the pivot, and replace the basic variable in the pivot row with the variable in the pivot column. In this case, s is replaced with y:

Basic variable	x	y	r	s	Value	Row operation
r	2	1	1	0	12	
y	$\frac{1}{4}$	1	0	$\frac{1}{4}$	2	$R2 \div 4$
P	3	-1	0	0	0	

Now use the pivot row to eliminate the pivot term from every other row:

Basic variable	x	y	r	s	Value	Row operation
r	$\frac{7}{4}$	0	1	$-\frac{1}{4}$	10	$R2 - R1$
y	$\frac{1}{4}$	1	0	$\frac{1}{4}$	2	
P	$\frac{13}{4}$	0	0	$\frac{1}{4}$	2	$R3 + R2$

There are no negative values in the objective row, so the solution is optimal. We read the entries in the value column for each variable, to obtain $x = 0$, $y = 2$, $r = 10$, and $s = 0$, recalling that any variable not listed in the first column is 0. This gives the vector,

$$\mathbf{x} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

We also have $P = -2$ (the simplex tableau is a maximisation method, so we've actually maximised the negative of P, so we need to negate the final value). △

Algorithm 32 Simplex Tableau

```
1: Draw the tableau(x) with a basic variable column on the left, one column for each variable (including slack variables), and a value column. Add a row for each constraint, and the bottom row for the objective function.
2: Enter the coefficients of the variables in the appropriate cells to form the initial tableau.
3: Find the most negative entry in the objective row to obtain the pivot column.
4: Calculate the $\theta$ values for each of the constraint rows, where $\theta$ is the value term divided by the pivot term.
5: Select the row with the smallest positive $\theta$ value to be the pivot row.
6: The element in the pivot row and pivot column is the pivot.
7: Divide the pivot row by the pivot, and change the basic variable in the first column to the variable at the top of the pivot column.
8: Use the pivot row to eliminate the pivot variable from other rows.
9: Repeat steps 3 to 8 until there are no negative values in the objective row.
10: The tableau is now optimal, and the non-zero values can be read off using the basic variable and value columns. If the objective function is to be minimised, take the negative of the objective value.
```

Chapter 50

Lambda Calculus

"The whole idea of Lambda Calculus really grew out of logic, and there's very beautiful dualities between programming on the one hand, and logic on the other. It's called the Curry-Howard Isomorphism, in which you can view a, let's say I have a function whose type is... it takes 2 integers and it produces an integer. Well, that type tells you something about the program. So, in a sense, it's a weak theorem about the program. It tells you something about the program, but not everything. And indeed, you could regard the program as a proof of that theorem."

— Simon Peyton Jones, *Functional Programming Languages and the Pursuit of Laziness*

50.1 Prefix Notation

For this section, it is helpful to be familiar with *prefix notation* (also known as *Polish notation*).

Many operators in mathematics, particularly arithmetic operations, relations, and orderings, are often written in *infix notation*, in which the operator is placed between the operands. For example, in the expression, $1 + 2$, the operator, $+$, is placed between its arguments, 1 and 2. In prefix notation, this would instead be written as $+\,1\,2$, more similarly to how we often write functions.

If the arity of operators are known, then, unlike in infix notation, brackets are unnecessary. That is, in an infix expression, brackets are required to override standard precedence rules, whereas in prefix expressions, ordering is sufficient.

For instance, the infix expression

$$(5 - 4) \times 3$$

is written as

$$\times(-\,5\,4)\,3$$

in prefix notation. But, since we know the subtraction operator takes two arguments, the brackets have no effect, so we may write,

$$\times - 5\,4\,3$$

instead.

$$5 - (4 \times 3)$$

is similarly written as

$$-5 \times 4\,3$$

Prefix notation is useful in that it can unambiguously express the order of operations like this without using brackets or other assumed precedence rules.

Prefix notation is also extremely easy to parse for computers: in a prefix string, push each symbol onto a stack. If a symbol is an operator of arity n, evaluate the operation from bottom up when there are exactly n non-operator symbols directly above it.

For example,

$$\times - 5\,4\,3$$

would be parsed as,

$$\begin{array}{ccccccccccccc}
 & & & & & & 4 & & & & & & \\
 & & & & 5 & & 5 & & & & 3 & & \\
 & & - & & - & & - & & 1 & & 1 & & \\
\times & \to & \times & \to & \times & \to & \times & \to & \times & \to & \times & \to & 3 \\
\hline
\end{array}$$

50.2 Motivation

A *computable function* is a formalisation of the intuitive notion of an algorithm. A function is computable if there exists an algorithm that can return the same outputs as that function, given the same inputs. We will be considering Turing-computable functions here. The lambda calculus provides a simple semantic system for handling these functions.

One simplification is the use of *anonymous functions* (§4.4.4), where we don't bind names to functions.

For example, the function,

$$\text{square\_sum}(x,y) = x^2 + y^2$$

can be written in anonymous form as,

$$(x,y) \mapsto x^2 + y^2$$

which we can read as, the tuple (x,y) is mapped to $x^2 + y^2$.

We can similarly write the identity map,

$$\text{id}(x) = x$$

as

$$x \mapsto x$$

In the lambda calculus, *all* functions are treated anonymously.

We also only use functions of a single input. Any ordinary function which takes multiple inputs, such as the square_sum function above, can be reformulated into an equivalent function that accepts a single input, and returns another function that takes a single input, and so on. For example,

$$(x,y) \mapsto x^2 + y^2$$

can be rewritten as

$$x \mapsto (y \mapsto x^2 + y^2)$$

This method is known as *currying*, and can be to convert any n-ary function into a chain of n unary functions.

We apply functions to arguments as usual, with the argument in brackets on the right of the function. So,

$$\begin{aligned}((x,y) \mapsto x^2 + y^2)(1,2) &= 1^2 + 2^2 \\ &= 5\end{aligned}$$

and

$$\Big((x \mapsto (y \mapsto x^2 + y^2))(1)\Big)(2) = (y \mapsto 1^2 + y^2)(2)$$
$$= 1^2 + 2^2$$
$$= 5$$

and we see that the curried function does indeed return the same output.

50.3 Lambda Terms

The lambda calculus consists of a formal language (§2.3.4) of sentences called *lambda terms* or *lambda expressions* defined by a formal syntax, and a set of transformation rules that allow us to manipulate these lambda terms.

The alphabet of the formal language underlying the lambda calculus consists of,

- variables, x,y,z, . . .;
- the *abstraction* symbols, λ (lambda) and . (dot);
- the *scoping* symbols, brackets ().

All syntactically valid lambda terms are then defined recursively (§5.1.2) as follows:

- Any variable, x, is a valid lambda term.
- If t is a lambda term, and x is a variable, then $\lambda x.t$ is a lamba term called an *abstraction.*
- If t and s are lambda terms, then (ts) is a lambda term called an *application.*
- No other string is a lambda term, unless implied by the previous rules.

Brackets can be used to disambiguate terms, but to reduce the number of necessary brackets and lambdas, we adopt the following conventions:

- Outermost brackets are dropped, so we write AB instead of (AB).
- Applications are left associative, so we write ABC for $((AB)C)$.
- The body of an abstraction extends as far right as possible, so $\lambda x.AB$ is read as $\lambda x.(AB)$ and not $(\lambda x.A)B$.
- Sequences of abstractions can be contracted, so $\lambda xyz.N$ is the same as $\lambda x.\lambda y.\lambda z.N$.

An abstraction $\lambda x.t$ defines* an anonymous function that binds a variable x as an input, and substitutes it into the expression, t. For example, $\lambda x.x + 1$ is an abstraction for the successor function, $f(x) = x + 1$, where the expression $x + 1$ is t.†

For the square_sum function, we would write $\lambda x.\lambda y.x^2 + y^2$, which can also be abbreviated as $\lambda xy.x^2 + y^2$ but it is important to remember that this latter notation represents a chain of unary functions, and is

[*] This is really just a model (see the addendum on formal languages and models) that is particularly human-friendly. The lambda calculus by itself is not a system for handling functions – the lambda calculus is a system for handling lambda expressions, which themselves are just lambda expressions, and don't inherently represent anything else. This is similar in flavour to groups in group theory. We may like to interpret the elements of the group S_n as permutations, but that's just one model of the group – the elements of the group aren't inherently permutations, just as lambda terms aren't inherently functions.

[†] Note that this first requires an implementation of "+" in the lambda calculus. Remember that the lambda calculus is a formal language, and any symbols not included in the definition above must first be defined in terms of more primitive objects.

Later on we will actually do this definition of addition in reverse order; defining the successor function first, then addition in terms of it.

not a single binary function. Also note that writing an abstraction merely defines the function, and does not invoke it.

An application ts is the application of an abstraction t, to an input term s. That is, it represents calling a function t on an input s to obtain $t(s)$.

There is no concept of variable declaration, however. In the lambda term $\lambda x.x + y$, y is treated as a variable, which just happens to be not yet defined. The expression is syntactically valid, and represents a function which adds its input to the unknown value y.

50.4 Function Functions

Unlike in most other situations, functions are *first class values* in the lambda calculus, meaning they can be the input or output of other functions. Note that this somewhat distinct from most other usages of functions – in common algebra, when we pass a function, $f(x)$, as the argument of another function, $g(x)$, as $g(f(x))$, we usually mean the composition of the two functions, $(g \circ f)(x)$. Here, the outer function can directly modify the inner function.

For example, $\lambda x.x$ represents the identity function, $x \mapsto x$, and $(\lambda x.x)y$ is an application representing the identity function being applied to y. The lambda term $\lambda x.y$ represents the constant function, $x \mapsto y$. We can apply these functions to each other: $(\lambda x.x)(\lambda x.y)$ is the identity function being applied to $\lambda x.y$, which just returns the abstraction $\lambda x.y$, which is the constant function, but note that this application does *not* return the constant, y, itself.

50.5 Free & Bound Variables

The *free variables* of a term are the variables not bound by an abstraction. They are similar to free and bound variables in predicate logic.

We define the set of free variables, FV, of a lambda expression recursively:

- $\text{FV}(x) = \{x\}$ – the set of free variables of a variable x contains just x alone.
- $\text{FV}(\lambda x.M) = \text{FV}(M) \setminus \{x\}$ – the set of free variables of the abstraction $\lambda x.M$ is the relative complement of the set of free variables of x relative to the set of free variables of M.
- $\text{FV}(MN) = \text{FV}(M) \cup \text{FV}(N)$ – the set of free variables of the application MN is the union of the set of free variables of M and the set of free variables of N.

For instance, in the lambda term $\lambda x.x + y$, x is bound by the abstraction, while y is free.

50.6 Reduction

The meaning of lambda expressions is defined by how the expressions can be reduced. There are three standard types of reduction:

- α-conversion: changing bound variables;
- β-reduction: applying functions to arguments;
- η-reduction: analogue for set extensionality.

We say two expressions are α-equivalent if they can be α-converted into the same expression, and similarly for β- and η-equivalence.

A term that is β-reducible is called a β-redex, short for *reducible expression.* η-redexes are defined similarly. The expression to which a redex reduces is called its *reduct.*

50.6.1 α-conversion

One basic form of equivalence on lambda terms is *alpha-equivalence* or *α-equivalence*. This form of equivalence captures the idea that the particular choice of bound variable in an abstraction does not matter. For instance, $\lambda x.x$ and $\lambda y.y$ both represent the identity function, and are therefore α-equivalent. The terms, x and y, however, are not α-equivalent, because they are not bound in an abstraction. We usually implicitly take α-equivalent terms to be identical.

Changing the bound variable of a lambda term is known as *α-conversion*, and transforms α-equivalent terms into each other.

During α-conversion, only variables bound by the same abstraction can be renamed – if the same variable is bound in several terms, this should be done with care. For example, a valid α-conversion of $\lambda x.\lambda x.x$ could be $\lambda y.\lambda x.x$, but not $\lambda y.\lambda x.y$. The original expression and the valid α-conversion both represent a function which returns the identity function, regardless of input. The invalid α-conversion instead represents a function which returns a constant function which returns the original input, which is clearly not equivalent.

Alpha-conversion is also not possible if the variable being changed *to* is already bound within the same expression. For example, in the lambda term, $\lambda x.\lambda y.x$, we cannot replace y with x.

50.6.2 Substitution

A *substitution*, $E[V := R]$ is a replacement of all free occurences of a variable V in the expression E with an expression R.

We define substitions on lambda expressions recursively. If x and y are variables, and M and N are any lambda expressions, then,

- $x[x := N] \equiv N$
- $y[x := N] \equiv y$, if $x \neq y$
- $(M_1 M_2)[x := N] \equiv (M_1[x := N])(M_2[x := N])$
- $(\lambda x.M)[x := N] \equiv \lambda x.M$
- $(\lambda y.M)[x := N] \equiv \lambda y.(M[x := N])$, if $x \neq y$, provided $y \notin \mathrm{FV}(N)$

Substituting into an abstraction may first require α-conversion. For instance, naïvely substituting $(\lambda x.y)[y := x]$ incorrectly gives $\lambda x.x$ because the substituted free variable, x, was bound into the abstraction. The correct substitution, up to α-equivalence, is instead $\lambda z.x$.

50.6.3 β-reduction

β-equivalence captures the idea that a function with a fixed argument is equivalent to that function being evaluated at that argument. That is, if $f(x) = x^2$, then we would like the expressions, "$f(3)$", and "3^2", to be equivalent.

β-reduction is defined in terms of substitution – the β-reduction of $(\lambda V.E)F$ is $E[V := F]$.

For example, suppose we have an encoding of squaring and of natural numbers in lambda calculus. Then, $((\lambda n.n^2)\,3)$ β-reduces to 3^2. In this case, we say that $((\lambda n.n^2)\,3)$ is a β-redex, and 3^2 is its reduct.

50.6.4 η-reduction

η-equivalence is similar to extensionality in set theory, where two sets are equal if and only if they contain the same elements, or more importantly, that a set is defined entirely by its contents. In the lambda calculus, two *functions* are equal if and only if they give the same outputs for the same inputs, for all possible inputs, or, a function is defined entirely by its output process. For instance, the functions, $f(x) = 2 \cdot x$ and $g(x) = x + x$, should intuitively be equivalent.

η-reduction converts between $\lambda x.(fx)$ and f whenever x is not free in f.

50.6.5 Normalisation

β-reduction allows us to calculate values from lambda terms. However, while values in set theory are sets, values in the lambda calculus are functions, which are represented by abstractions, so, to evaluate a lambda expression, we continue β-reducing the term until it looks like a function abstraction.

A lambda expression is in,

- *normal form* if no β- or η-reductions are possible. That is, it contains no β- or η-redex
- *head normal form* if it is in the form of a lambda abstraction whose body is not β-reducible.
- *weak head normal form* if it is in the form of a lambda abstraction.

50.7 Data Types

50.7.1 Variable Assignment

So far, we have a bunch of functions, and... well, just functions, really. Just as in set theory, to use other structures and *data types*, we need to encode them somehow. In set theory, we recursively encoded the naturals as a list of sets. In the lambda calculus, we pull a similar trick with functions, though in a different way than you'd might expect.

In the following section, whenever we write $A = B$, we do so in the programming sense, where the equality symbol is used for variable assignment. While the lambda calculus doesn't have variable assignment, we can wrap whatever we wanted to do with that variable into the body of an abstraction with the variable name as the bound variable, then apply that abstraction to the contents of the variable. That is, the code,

```
myvar = object
function(myvar),
```

is encoded in the lambda calculus as,

$$\lambda\texttt{myvar}.(\texttt{function(myvar)})(\texttt{object})$$

50.7.2 Boolean Variables & Logic Gates

With the Boolean values, true and false, they don't really mean anything by themselves, and are only useful in relation to other functions.[*] A function that returns a Boolean is called a *predicate.*

We can encode Boolean values using functions as follows:

$$\text{true} = \lambda x.\lambda y.x$$
$$\text{false} = \lambda x.\lambda y.y$$

So, given two values, true returns the first, while false returns the second. How do we actually use them?

If we know a variable b is a Boolean – that is, it is one of the above lambda expressions – we can test what value it is by passing it to this function:

$$\text{IsTrue} = \lambda b.b\,\text{true}\,\text{false}$$

[*] If you're thinking that we've had this exact discussion before, that's because we have: §11.5, §12.1.2, §33.6.

Intuitively, this works because if b is true, it returns the first value, true, and if b is false, it returns the second, false, but we can walk through this more formally by examining what happens when we evaluate this function on an input. Let's apply the function IsTrue to false, and we should expect the output to be false.

$$\text{IsTrue(false)} = (\lambda b.b\,\text{true}\,\text{false})\,\text{false}$$

Begin by expanding the function definitions. We're left with an application, so we perform β-reduction:

$$= \text{false}\,\text{true}\,\text{false}$$

Expand the definition of false:

$$= (\lambda x.\lambda y.y)\,\text{true}\,\text{false}$$

A few more β-reductions:

$$\begin{aligned} &= (\lambda y.y)\,\text{false} \\ &= \text{false} \end{aligned}$$

The working for true as an argument is almost identical. So, we can now check the contents of a Boolean variable. What about logic gates? Once we have those, we can basically go back to §2 and rebuild logic from there.

We first consider the simplest non-trivial Boolean function – the unary operator, NOT. We can make this one just by swapping the order of the inputs to the previous function:

$$\text{NOT} = \lambda b.b\,\text{false}\,\text{true}$$

Now, if b is true, it returns the first input, false, and if b is false, it returns the second, true.

Next, we can find the OR of two Booleans, a and b, by calling a with true and b.

$$\text{OR} = \lambda a.\lambda b.a\,\text{true}\,b$$

If a is true, it immediately returns true. Otherwise, it returns whatever value b is.

We similarly find the AND of a and b by calling a with b and false:

$$\text{AND} = \lambda a.\lambda b.a\,b\,\text{false}$$

If a is false, it immediately returns false. Otherwise, it returns whatever value b is.

NOT, OR and AND is a functionally complete set (§2.2.1), so we can now build up all of propositional logic by chaining these gates together.

Boolean also lends itself well to encoding binary. We first create a structure called a *pair*, which is, well, an ordered pair of objects. We only have Boolean objects so far, but that's enough for encoding binary.

This structure does exactly what we want:

$$\lambda x.\lambda y.\lambda b.b\,x\,y$$

We call it on two values, which are stored in x and y, say, X and Y, giving,

$$\lambda b.b\,X\,Y$$

Then, when we want to extract the data, we provide a third Boolean argument which selects x if true, and y if false.

If Y is another pair, we can create a linked list structure, where we can repeatedly enter false into the function to move to the next node, and a true to extract data from the current node.

If the data, X, is a Boolean, we can use this to encode binary numbers.

50.8 Church Numerals

We can represent natural numbers more efficiently, however.

We encode the natural number n as a function that maps any function, f, to its composition with itself n times. Writing $f^{\circ n} = \underbrace{f \circ f \circ \cdots \circ f}_{n}$ to denote this repeated composition, we have,

- $0 = \lambda f.\lambda x.x$;
- $1 = \lambda f.\lambda x.f\,x$;
- $2 = \lambda f.\lambda x.f(f\,x)$;
- $3 = \lambda f.\lambda x.f(f(f\,x))$;
- $4 = \lambda f.\lambda x.f(f(f(f\,x)))$;
- $n = \lambda f.\lambda x.f^{\circ n}\,x$;

Starting with 0 not applying the function at all, 1 applying the function once, 2 twice, 3 thrice, and so on, we construct the *Church numerals.*

It is important to note that it is this lambda term that composes functions itself, that is the value, and not the composite function for a particular choice of f, nor the end result of applying this function to some value. A Church numeral, n, really just represents the action of doing anything n times, and doesn't concern what the action itself is, nor what the action is acting on.

50.9 The Successor Function & Arithmetic

Many operations follow naturally from the definition of Church numerals, perhaps more so than they do from the von Neumann construction of the naturals in set theory.

We begin as usual with some version of the successor function, $\text{succ}(n) = n + 1$. With the above definition of natural numbers, we would intuitively think that succ would just take the input numeral, n, the function, f, and return $f(n)$,

$$\lambda n.\lambda f.f\,n$$

The problem is, when we actually perform an application and substitute in the lambda term for n, we get,

$$\lambda f.f(\lambda f.\lambda x.f\,x)$$

(for $n = 1$), leaving a floating lambda we don't want.

Instead, we also take x, then call n on f and x,

$$\lambda n.\lambda f.\lambda x.n\,f\,x = \lambda f.\lambda x.f\,x = n$$

which gives the actual value of n (again, the case $n = 1$ is shown above). Then, we wrap the entire body of the abstraction with another f to increase the value of n by 1:

$$\text{succ} = \lambda n.\lambda f.\lambda x.f\,(n\,f\,x)$$

Because a natural, n, is represented by the action of applying a function n times, we can take two naturals, a and b, and call b with succ and a. This will apply the successor function to a, b times, allowing us to perform addition.

$$\text{add} = \lambda a.\lambda b.b\,\text{succ}\,a$$

This entire definition of addition hinges on the fact that numerals are really just functions that repeatedly apply their first argument to the second.

If we call add with a, then apply it to 0, b times, we can perform multiplication.

$$\text{mult} = \lambda a.\lambda b.b\,(\text{add}\,a)\,0$$

We can also write these using identities to do with function composition. For example, the addition function, add$(a,b) = a + b$ uses the identity $f^{\circ(a+b)}(x) = f^{\circ a}(f^{\circ b}(x))$, and can also be written as,

$$\text{add} = \lambda a.\lambda b.\lambda f.\lambda x.a\,f\,(b\,f\,x)$$

which is somewhat less readable, but is still β-equivalent to the previous definition. We can also see that the successor function, succ n is β-equivalent to add $n\,1$ when using this definition of add.

Multiplication, mult$(a,b) = a \cdot b$ uses the identity $f^{\circ(a\cdot b)}(x) = (f^{\circ b})^{\circ a}(x)$.

$$\text{mult} = \lambda a.\lambda b.\lambda f.\lambda x.a\,(b\,f)\,x$$

Exponentiation is even easier. We can use the number of compositions in the definition of Church numerals, $n\,f\,x = f^{\circ n}\,x$, to encode exponents. Substituting a few variables, we have $a\,b\,f = a^{\circ b}f$ and,

$$\begin{aligned}\exp a\,b &= a^{\circ b}\\ &= b\,a\\ &= \lambda a.\lambda b.b\,a\end{aligned}$$

50.10 Predecessor

Subtraction and division are also definable, with a few caveats, but we first need a predecessor function, pred, which is the inverse of succ and returns the previous numeral. Because Church numerals only encode the naturals, we define the predecessor function to be,

$$\text{pred}(n) = \begin{cases} 0 & n = 0 \\ n-1 & n > 0 \end{cases}$$

To construct this function in the lambda calculus, we need to find a way to apply the function f in n one fewer time. That is, we need to construct $\lambda f.\lambda x.f^{\circ(n-1)}\,x$ from $\lambda f.\lambda x.f^{\circ n}\,x$.

To do this, we need to wrap the value of n in a container function such that f and x can be accessed from inside and outside the container. Call the container box, and define a new function, init to initialise an instance of box containing the value x.

$$\text{init} = \text{box}\,x$$

and a corresponding inverse operation to open a box to extract the value inside:

$$\text{unpack}\,(\text{box}\,v) = v$$

We now interface with the internals of the box by defining a new function, inc, that, when applied to the outside of a box, applies f to the contents of the box.

$$\text{inc}\,\text{init} = \text{box}(f\,x)$$

We can use these functions to redefine the identity function, id by calling inc on an instance of init n times, then unpacking the resulting box container.

$$\text{id} = \lambda n.\lambda f.\lambda x.\,\text{unpack}\,(n\,(\text{inc})\,(\text{init}))$$

$$
\begin{aligned}
&= \lambda n.\lambda f.\lambda x.\,\mathrm{unpack}\,(\mathrm{box}\,(n\,f\,x))\\
&= \lambda n.\lambda f.\lambda x.n\,f\,x\\
&= \lambda n.n
\end{aligned}
$$

Because inc defers the calling of f to the inside of a container, we can pass it a special container, const, that ignores the first application of f and just returns a box containing x.

$$\mathrm{inc}\,\mathrm{const} = \mathrm{box}\,x$$

Then, packing the resulting lambda expression inside the unpack clause of the identity function as defined above, we can extract this value with the ignored first application of f.

$$
\begin{aligned}
\mathrm{pred} &= \lambda n.\lambda f.\lambda x.\,\mathrm{unpack}\,(n\,(\mathrm{inc})\,(\mathrm{const}))\\
&= \lambda n.\lambda f.\lambda x.\,\mathrm{unpack}\,(\mathrm{box}\,((n-1)\,f\,x))\\
&= \lambda n.\lambda f.\lambda x.(n-1)\,f\,x\\
&= \lambda n.(n-1)
\end{aligned}
$$

While this works well enough, we still need to encode box, unpack, inc, init, and const as lambda expressions.

We want the box container to take two arguments, storing the first, v (the "contents" of the container), and, given a second argument, h, call h with v, so $\mathrm{box}\,v\,h = h\,v$, which is encoded as,

$$\mathrm{box} = \lambda v.(\lambda h.h\,v)$$

Once box is loaded with a value, k, we can easily extract by calling $\mathrm{box}\,k$ with the identity function as the input. As defined above, the box will then call the identity function with k as the argument, which just returns k, as desired.

$$
\begin{aligned}
\mathrm{unpack}(\mathrm{box}\,k) &= \mathrm{box}\ k\,(\lambda u.u)\\
&= (\lambda v.(\lambda h.h\,v)\,k)(\lambda u.u)\\
&= (\lambda h.h\,k)(\lambda u.u)\\
&= (\lambda u.u)\,k\\
&= k
\end{aligned}
$$

So box and unpack work as expected.

The term "container" is just syntactic sugar, helpful for thinking about how box separates f and x from the effect of inc, but this is more difficult to see in the lambda expressions. In this implementation, a boxed variable is really just a variable loaded into a function (the box) that takes another function and passes the variable into the received function. So, to actually pack a variable into a box, we just prime it to receive a function in this way.

$$\mathrm{init} = \lambda h.h\,x$$

Next, inc should take a loaded box containing a value, v, and return another instance of a box that contains $(f\,v)$.

So, we should begin by taking an input, $g = \mathrm{box}\,v$ and pack it into an application, $(g\,f)$ so that the box returns $(f\,v)$. Then, we finally pack this back into a box with a modified init, where, instead of returning x, we use the previous expression in its place.

$$
\begin{aligned}
\mathrm{inc} &= \lambda g.\,\mathrm{init'}\,(g\,f)\\
&= \lambda g.\lambda h.h\,(g\,f)
\end{aligned}
$$

Finally, we need a implementation of const to satisfy,

$$\mathrm{inc}\,\mathrm{const} = \mathrm{box}\,x$$

Rewriting this as a lambda expression, we have,

$$\lambda h.h(\mathrm{const}\ f) = \lambda h.h\,x$$

and we can easily see that $\mathrm{const}\,f = x$, which we can encode as,

$$\mathrm{const} = \lambda u.x$$

Packing these together, we find that the lambda expression for pred is,

$$\mathrm{pred} = \lambda n.\lambda f.\lambda x.n\,(\lambda g.\lambda h.h\,(g\,f))\,(\lambda u.x)\,(\lambda u.u)$$

50.11 Subtraction

To define subtraction, we can just replace the succ function in addition with our new pred function.

$$\mathrm{add} = \lambda a.\lambda b.b\,\mathrm{succ}\,a$$
$$\mathrm{sub} = \lambda a.\lambda b.b\,\mathrm{pred}\,a$$

Because Church numerals only encode natural numbers, if $a \leq b$, then $a - b$ returns 0 in this implementation.

Next, we will attempt to implement division. To handle the lack of negative numbers, we define division over Church numerals as,

$$b/a = \text{if } (a \geq b) \text{ then } 1 + (a - b)/a, \text{ else } 0$$

This definition of dividing through repeated subtractions is almost identical to our previous implementation in number theory (§10.1.1).

50.11.1 Comparison

The first problem is, how do we compare numbers in the lambda calculus? The simplest predicate for testing numbers is IsZero, which returns true if the argument is the Church numeral 0, and false otherwise.

We can implement this as follows:

$$\mathrm{IsZero} = \lambda n.n\,(\lambda x.\,\mathrm{false})\,\mathrm{true}$$

Because 0 represents not applying a function at all, when it gets passed to IsZero, the constant false function, $(\lambda x.\,\mathrm{false})$, is never applied, instead returning true. For any other Church numeral, the function is applied at least once, and hence returns false.

We can then compare numbers by using subtract, taking advantage of the fact that sub evaluates to 0 whenever the minuend is smaller than the subtrahend.

$$\mathrm{LEQ} = \lambda a.\lambda b.\,\mathrm{IsZero}\,(\mathrm{sub}\,a\,b)$$

Additionally, because $x = y \equiv (x \leq y \wedge y \leq x)$, we also get an implementation of the equality predicate for free:

$$\mathrm{EQL} = \lambda a.\lambda b.\,\mathrm{AND}\,(\mathrm{LEQ}\,a\,b)\,(\mathrm{LEQ}\,b\,a)$$

So, we can compare numbers, but the second problem is that this definition of division includes division within the definition itself. To handle this, we need to implement some *recursion.*

50.12 Recursion

Recursion is the idea of defining things in terms of themselves. One common example is the factorial function, where we have, $n! = \text{if } (n \leq 1) \text{ then } 1, \text{ else } n \cdot (n-1)!$, or,

```
def fac(n)
    if n <= 1:
        return 1
    else:
        return n*fac(n-1)
```

So, to find 3!, we expand the definition and see that $3 \not\leq 1$, so we have $3! = 3 \cdot 2!$. We again expand the definition, with $2 \not\leq 1$ now giving $3! = 3 \cdot 2 \cdot 1!$, and we find that $1 \leq 1$, so we return 1 to the call stack, finally giving $3! = 3 \cdot 2 \cdot 1$ which we can finally compute to be 6.

In the lambda calculus, there isn't a native way of performing recursion or looping, so we have to encode these things. Recursion in set theory is encoded fairly simply by using set comprehensions and including the set itself in the predicate clause, but how do we encode recursion in the lambda calculus, where we effectively just have functions?

In the lambda calculus, all functions are anonymous, so there is no way to refer to a value which has not yet been defined inside the lambda term defining the said value. We deal with this by arranging for a lambda expression to receive itself as its argument value.

To implement the factorial as a lambda expression, we need to create a function, f, that will receive the whole lambda expression representing the factorial itself as its value, so that calling the function as an application will instantiate another copy of the function itself, ready for the next invocation. So, the function, f, must always be passed to itself as an application within the body of the abstraction. We also pass everything into another function, F, to handle the self-application of the entire function. So, we want,

$$F = \lambda f.\lambda n.(\text{IsZero } n)\, 1\, (\text{mult } n\, (f\, f\, (\text{pred}\, \text{n})))$$

such that $f\, f\, x = \text{fact}\, x$ holds, which is needed for the self-application of f at the end. Because the body of the abstraction on the right is really the definition of a factorial, we also require $F\, f\, x = \text{fact}\, x$. It follows that $f = F$, so,

$$\text{fact} = F\, F = (\lambda x.x\, x)\, F$$

This self-application replicates F, passing the lambda expression representing the factorial into the next invocation. While this implementation works, it requires re-writing each recursive call as a self-application.

To avoid this, more sophisticated techniques are required. The factorial function is a bit complicated for this, so let's go back to a simpler recursion.

The simplest recursive definition is the *loop* function, which is defined as

`loop = loop`.

To find the value of `loop`, we expand its definition, which is `loop`, and we expand the definition again, and so on, giving a very simple loop which does nothing else. The way we encode this in the lambda calculus is by using *self application* – by applying something to itself. In this case, by calling a function* on itself.

$$\text{loop} = (\lambda x.x\, x)(\lambda x.x\, x)$$

We see a nested structure here, where we have two identical functions written next to each other, with the first being applied to the second – this is a self application. Within each function too, we have two x terms next to each other, with the first being applied to the second. Examining one of the two

* This function is more properly known as the Ω-Combinator.

main functions, we see that it takes an input, x, then applies it to itself, giving another instance of self application.

Let's see what happens when we actually "run" loop. That is, we will β-reduce the application. We call the first function by replacing all instances of x in the body of the abstraction with the second function. The first function just calls x on itself, written as a lambda application, so we get,

$$\begin{aligned}\text{loop} &= (\lambda x.x\,x)(\lambda x.x\,x)\\ &= (\lambda x.x\,x)(\lambda x.x\,x)\end{aligned}$$

and we see that this function β-reduces into itself. Or put another way, this function is its own β-redex. So, this definition just calls itself, as desired.

50.13 The Y Combinator

We can define a more general form of recursion as,

$$\text{fix}\, f = f(\text{fix}\, f)$$

so this function takes a function, then applies that function to itself. Expanding this definition, we find,

$$\text{fix}\, f = f(f(f(\cdots f(\text{fix}\, f)\cdots)))$$

so $\text{fix}\, f$ is a *fixed point* of f. In the lambda calculus, every function has a fixed point, and this function returns that fixed point.* More generally, any implementation of fix is called a *fixed-point combinator*, and these higher-order functions return the fixed point of any function passed to it.

In computer science, this type of function is also a form of *general recursion*, and it turns out that any arbitrary recursion can be defined in terms of these general ones. For instance, if $f = \lambda x.x$, that is, the identity function, then

$$\begin{aligned}\text{fix}\, f &= f(\text{fix}\, f)\\ &= \text{fix}\, f\end{aligned}$$

and we've just encoded loop in terms of fix.

The most common implementation of fix in the lambda calculus is the Y combinator.

$$\mathsf{Y} = \lambda f.(\lambda x.f\,(x\,x))(\lambda x.f\,(x\,x))$$

This is very similar to our implementation of loop, just with extra functions wrapped around everything.

We can verify that this implementation works:

$$\begin{aligned}\mathsf{Y}\, g &= (\lambda f.(\lambda x.f\,(x\,x))(\lambda x.f\,(x\,x)))\, g\\ &\equiv ((\lambda x.g\,(x\,x))(\lambda x.f\,(x\,x)))\\ &\equiv g((\lambda x.f\,(x\,x))(\lambda x.f\,(x\,x)))\\ &\equiv g(\mathsf{Y}\, g)\end{aligned}$$

When applied to a unary function, the Y combinator usually doesn't terminate. But with functions of two or more variables, the Y combinator becomes much more useful. The second variable can be used as a counter, so the resulting function is effectively an implementation of a `while` or `for` loop.

* The identity function trivially fixes every input, but the remarkable property of fixed-point combinators is that they construct a single value that is a fixed point of its input.

Now, back to the factorial, if we instead write,

$$F = \lambda f.\lambda n.(\mathrm{IsZero}\ n)\,1\,(\mathrm{mult}\ n\,(f\,(\mathrm{pred}\,\mathrm{n})))$$

then we would instead require that $f\,x = \mathrm{fact}\,x = F\,f\,x$ holds, so $F\,f = f$, and we deduce that $\mathrm{fact} = \mathrm{fix}\,F$, so,

$$\mathrm{fact} = \lambda f.(\lambda x.f\,(x\,x))(\lambda x.f\,(x\,x))\,\lambda f.\lambda n.(\mathrm{IsZero}\ n)\,1\,(\mathrm{mult}\ n\,(f\,(\mathrm{pred}\,\mathrm{n})))$$

50.14 The Z Combinator

50.15 Division

50.16 Logical Consistency

50.16.1 Kleene-Rosser Paradox

50.16.2 Curry's Paradox

50.16.3 Simply Typed Lambda Calculus

50.16.4 Combinatory Logic

Chapter 51

Category Theory I

> *"The difference between a regular mathematician and a category theorist is that, upon encountering a new theorem, the regular mathematician will ask 'What is an example of this theorem?', while the category theorist will ask 'What is this theorem an example of?'."*
>
> — Unattributed

The following chapters are taken from various essays of mine. There is some overlap in their content, since both define basic categorical terminology, but the first covers the very basics in far more detail than the second; it is recommended that you read them in order.

The first covers the basic definitions and categorical notions up to and including the celebrated *Yoneda lemma*, which details how we might examine an arbitrary object by looking at the maps into or out of it. The second covers structural axiomatisations of the foundations of mathematics, in particular, Lawvere's *Elementary Theory of the Category of Sets* (ETCS), before exploring more general logics internal to topoi.

51.1 Introduction

Many structures in mathematics come alongside some notion of maps, which are used to relate different objects with those structures. For instance, groups come alongside group homomorphisms, vector spaces with linear transformations, probability spaces with measurable functions, and topological spaces with continuous maps, to name a few.

A *category* is any collection of objects with maps between those objects that compose associatively. All of the previous examples are thus specific types of categories, and many common constructions in those areas do not actually rely on anything specific to that area and can be carried out and unified together when performed at the level of a category. For instance, the Cartesian product, direct product of groups (rings, monoids...), product topology, disjoint union, and graph tensor product are all instances of a categorical product. If we can prove something about the categorical product, we'll have proved a result about all of these different types of objects.

Just as many properties in metric spaces are actually topological in nature, many mathematical objects can be reduced to purely categorical constructions: direct sums, kernels, quotient objects, compactifications and completions are all also categorical in nature.

A common theme in category theory is that maps between objects are more important than the objects themselves, and it will often be the case that it is easier to describe an object by the properties it satisfies or what relations the object has, rather than what the object itself actually is. Even in abstract algebra, this is often the case – we care that the elements of a group have group structure, not what the elements themselves actually are or how we label them.

In mathematics, we often come across statements of the form, $\exists ! x : P(x)$, or, "There exists a unique x such that $P(x)$ holds." The property P is called a *universal property*, and it uniquely characterises the object x up to an isomorphism. For example,

Theorem 51.1.1. *Let* $\mathbf{1}$ *be a set with one element. Then, for all sets* X*, there exists a unique function from* X *to* $\mathbf{1}$.

Proof. For existence, we define a function that maps every element of X to the unique element of $\mathbf{1}$. Because every element of X only has one choice of destination, this function is unique. ■

So, the property "For all sets X, there exists a unique function from X to $\mathbf{1}$" uniquely characterises $\mathbf{1}$, up to relabelling of the element. Rather than describing an object itself, universal properties allow us to describe objects by how they relate to other objects in whatever universe we're working in.

The *Yoneda lemma* expands on this concept, suggesting that we may study a category $\mathscr{C}$ by examining the maps from $\mathscr{C}$ to the category of sets.

51.2 Categories

[Lei14] A *category* $\mathscr{C}$ consists of:

- A class $\mathrm{ob}(\mathscr{C})$ of *objects* in $\mathscr{C}$.
- For all (ordered) pairs of objects $A,B \in \mathrm{ob}(\mathscr{C})$, a class $\hom(A,B)$ of *maps* or *arrows* called *morphisms* from A to B, called the *hom-set* or *hom-class* of morphisms from A to B, also sometimes written $\mathscr{C}(A,B)$ or $\hom_{\mathscr{C}}(A,B)$ (particularly useful if multiple categories are in use). If $f \in \hom(A,B)$, we write $f : A \to B$ or $A \xrightarrow{f} B$. The collection of all of these classes is the hom-set of $\mathscr{C}$, and is written $\hom(\mathscr{C})$.
- For any three objects $A,B,C \in \mathrm{ob}(\mathscr{C})$, a binary operation, $\circ : \hom(A,B) \times \hom(B,C) \to \hom(A,C)$, $(g,f) \mapsto g \circ f$, called *composition*, such that,

- (*associativity*) if $f : A \to B$, $g : B \to C$, and $h : C \to D$, then $h \circ (g \circ f) = (h \circ g) \circ f$;
- (*identity*) for every object $X \in \mathrm{ob}(\mathscr{C})$, there exists a morphism $\mathrm{id}_X : X \to X$ called the *identity morphism* on X, such that every morphism $f : A \to X$ satisfies $\mathrm{id}_X \circ f = f$, and every morphism $g : X \to B$ satisfies $g \circ \mathrm{id}_X = g$.

In the above definitions, we use the term *class*. This is because these collections of data generally do not count as sets under ZFC or equivalent set theory axiomatisations (§51.6.2). For instance, the collection of all sets does not qualify as a set under ZFC, but such a collection of objects is highly useful in category theory, so we use classes instead. The notion of a class is informal in ZFC, since ZFC exclusively concerns things which are sets, but here, we will define* a class as a collection of sets that is unambiguously defined by a property all its members share, such as "being a set". Any class which is never a set is a *proper class*, while a class that is sometimes a set is a *small class.*

Throughout this document, we will largely ignore the distinction between the two, as the categories we will construct will generally be *locally small* – the class of morphisms between any pair of objects happens to be a set. That is, $\hom(A,B)$ is a set (in the sense that they can be constructed in ZFC) for all $A,B \in \mathrm{ob}(\mathscr{C})$. If we also have that $\mathrm{ob}(\mathscr{C})$ is a set, then $\mathscr{C}$ is furthermore a *small* category.

Let's go through these classes one by one.

An object can really be anything we want, but many of the simplest and most familiar examples will begin with sets, often with additional structure, such as groups or rings. For any two objects, A and B, the category has a set of morphisms from A to B, $\hom(A,B) = \{f,g,\ldots\}$. This doesn't really explain what a morphism actually *is*, but morphisms are so general that any more specificity is not particularly useful. Defining a morphism is somewhat like defining a vector – is a vector fundamentally an arrow in space which can be described with coordinates, or are they fundamentally ordered list of numbers? – the answer being, neither; a vector is anything that obeys the vector space axioms. It's more helpful to define them by the properties they satisfy, rather than what they themselves are, and as we will soon see, this viewpoint will become a recurring pattern. In fact, we should note that objects are in bijection with identity morphisms, so it is possible to define categories entirely in terms of morphisms, and ignore the objects entirely. We will not do so here, but it is yet another reminder that we will often care more about how an object interacts with other objects than about the object itself.

It might be helpful to view a morphism a type of (directional) relation, rather than a function. There is a morphism, f, from A to B if A is related to B, but B does not have to be related to A, and we write $f : A \to B$, or draw an arrow from A to B on a diagram to represent this.

$$\begin{tikzcd} A \arrow[r, "f"] & B \end{tikzcd}$$

It could be the case that A and B are not related at all, so the set of morphisms from A to B is empty. We can also have multiple morphisms from A to B if A is related to B in several ways.

$$\begin{tikzcd} A \arrow[r, bend left, "f"] \arrow[r, bend right, "g"'] & B \end{tikzcd}$$

An object can also be related to itself, and in multiple ways at once.

Morphisms must also have a binary operation defined on them, called *composition* that obey the composition law. If there are morphisms $A \xrightarrow{f} B \xrightarrow{g} C$, then the category must also contain a morphism $A \xrightarrow{h=g\circ f} C$. Furthermore, any three morphisms must compose associatively: that is, $(h \circ g) \circ f = h \circ (g \circ f)$ for all morphisms f, g and h (with the appropriate domains and codomains). Categories also require identity morphisms – for every object A, there must exist a morphism $\mathrm{id}_A : A \to A$ such that all morphisms $f : A \to B$ and $g : B \to A$ satisfy $\mathrm{id}_A \circ g = g$ and $f \circ \mathrm{id}_A = f$.

[*] A more formal way to handle these classes is through the introduction of *Grothendieck universes*. This is not of high importance to the main body of this document, and its discussion is relegated to §51.6.3 in the addendum.

Let $\text{ob}(\mathscr{C}) = \{A\}$, and $\hom(A,A) = \{\text{id}_A\}$. That is, $\mathscr{C}$ is a category containing only one object, and a single morphism from that object to itself. This morphism trivially satisfies the assocativity and identity requirements, so $\mathscr{C}$ is a category, called the *trivial category*, depicted below.

$$\begin{tikzcd} A \arrow[loop right, "\text{id}_A"] \end{tikzcd}$$

Apart from the trivial category, we usually omit the identity morphism from such diagrams. Conversely, a category which contains no morphisms apart from identity morphisms is called a *discrete category.*

Let $\text{ob}(\mathscr{C}) = \{A,B\}$, and the non-identity morphisms be $\hom(A,B) = \{f\}$ and $\hom(B,A) = \varnothing$. That is, $\mathscr{C}$ is a category containing two objects, and a single non-identity morphism connecting them in one direction only. This is the *arrow category.*

$$\begin{tikzcd} A \arrow[r, "f"] & B \end{tikzcd}$$

Now, let $(G, \cdot\,)$ be a group, $\text{ob}(\mathscr{C}) = \{*\}$, and $\hom(*,*) = G$. For any two morphisms, f and g, we define the composition $f \circ g$ to be $f \cdot g$, so the morphisms have group structure. Because groups require associativity and identities, the morphism axioms are satisfied, and we see that a group is really a one-object category. In fact, there's nothing specific to groups here – we could just have easily started with a monoid or any other algebraic structure with associativity and identities.

These categories are pretty simple, but they give us an idea of how basic categories can be. We will build a more complicated category next: **Set**. Unsurprisingly, the objects of **Set** are sets, while morphisms are ordinary set functions. Composition of morphisms is just regular function composition, and identity morphisms just identity functions which map elements of sets to themselves. The associativity and identity laws follow from elementary properties of function composition. So, **Set** is a category.

Many other commonly used categories follow this formula – that is, their objects are sets with additional structure, and their morphisms are functions that respect that structure. For example, in the category **Grp**, objects are groups, and morphisms are group homomorphisms. Composition and identities are inherited from **Set**, because everything in **Grp** is just a specialised version of something in **Set**. Similarly, **Ring** is the category of rings and ring homomorphisms; **Top**, topological spaces and continuous maps; $\mathbf{Vect}_K$, vector spaces over a field K and linear maps; etc.

However, this doesn't have to be the case, and in general, categories need not have sets as objects and structure-preserving maps as morphisms. We construct a basic example of such a category as follows: the objects in our category will be the real numbers, and for any real numbers, x and y, we define a unique morphism from x to y, if and only if $x \leq y$. The problem here, as opposed to in **Set** or **Grp**, is that we can't really say what a morphism really *is*. Here, it's not something that acts on any elements like a function in **Set** or a group homomorphism in **Grp**. In fact, it doesn't really seem to do anything at all, other than existing whenever x is less than or equal to y.

If we have morphisms $f : x \to y$ and $g : y \to z$, then we know $x \leq y$ and $y \leq z$. By transitivity of $\leq$, we have $x \leq z$, giving a unique morphism $h : x \to z$ by definition, which we can assign to the composition $g \circ f$. Because this morphism is unique, this assignment is well-defined and determines the composition of any pair of morphisms. Because $x \leq x$ holds for all x by reflexivity, there is a unique morphism from any element to themselves, which we can use as the identity morphism and the associativity and identity laws follow easily. So, $(\mathbb{R}, \leq)$ is a category. It may be noted that we used nothing specific to the real numbers, so any set equipped with a non-strict preorder is in fact a (small) category.

We can also construct a new category from a pair of existing categories. Given categories $\mathscr{C}$ and $\mathscr{D}$, the *product category* $\mathscr{C} \times \mathscr{D}$ is defined by $\text{ob}(\mathscr{C} \times \mathscr{D}) = \text{ob}(\mathscr{C}) \times \text{ob}(\mathscr{D})$, and $\hom_{\mathscr{C}\times\mathscr{D}}((A,B),(A',B')) = \hom_{\mathscr{C}}(A,A') \times \hom_{\mathscr{D}}(B,B')$, with compositions defined componentwise [Mac13]. That is, if $A \xrightarrow{f} A'$ and $B \xrightarrow{g} B'$ are objects and morphisms in categories $\mathscr{C}$ and $\mathscr{D}$ respectively, then we have the objects and morphism $(A,B) \xrightarrow{(f,g)} (A',B')$ in the product category $C \times D$. We just take pairs of objects in the constituent categories and pairs of corresponding morphisms between them.

The *principle of duality* states that every categorical definition and theorem has a *dual* definition and theorem, obtained by reversing the direction of all morphisms in the categories involved. We often prefix a dual notion with *co-*, such as in products and coproducts, or domains and codomains.

For instance, every category $\mathscr{C}$ has a *dual* or *opposite* category with the same class of objects, but with the domains and codomains of all morphisms interchanged, denoted $\mathscr{C}^{\mathrm{op}}$. That is, $\mathrm{ob}(\mathscr{C}) = \mathrm{ob}(\mathscr{C}^{\mathrm{op}})$, and $\hom_{\mathscr{C}}(A,B) = \hom_{\mathscr{C}^{\mathrm{op}}}(B,A)$ for all objects A and B. We note that this notion of duality for categories is involutive, so $(\mathscr{C}^{\mathrm{op}})^{\mathrm{op}} = \mathscr{C}$ for all categories $\mathscr{C}$.

Theorem 51.2.1 (Conceptual Duality)**.** *Let Σ be a statement that holds in all categories. Then the dual statement Σ^* holds for all categories.*

Proof. [Bor+94, adapted] If Σ holds in a category $\mathscr{C}$, then Σ^* holds in $\mathscr{C}^{\mathrm{op}}$. Every category is the dual of its dual, so Σ^* holds in all categories. ■

51.2.1 Commutative Diagrams

It is often helpful to depict categories visually. We have already been using arrows to show morphisms between objects, but we can do a lot more with these representations. If we take a selection of objects in a category and draw morphisms between them, we can compose morphisms by following a path through the diagram, and because of associativity, each path corresponds to a unique composition.

This is useful enough by itself, but certain diagrams have an additional helpful property. A diagram is *commutative* if, for every pair of objects in the diagram, all routes between them are equivalent. For example, this diagram is commutative if and only if $h = g \circ f$.

$$\begin{tikzcd}
X \arrow[r, "f"] \arrow[dr, "h"'] & Y \arrow[d, "g"] \\
 & Z
\end{tikzcd}$$

Suppose we have objects A and B in a category, and morphisms f from A to B and g from B to A such that the following diagram is commutative.

$$\begin{tikzcd}
A \arrow[loop left, "\mathrm{id}_A"] \arrow[r, shift left, "f"] & B \arrow[l, shift left, "g"] \arrow[loop right, "\mathrm{id}_B"]
\end{tikzcd}$$

That is, $f \circ g = \mathrm{id}_B$ and $g \circ f = \mathrm{id}_A$. f and g are then *isomorphisms* – morphisms with inverses – and we alternatively label g as f^{-1}. If an isomorphism between A and B exists, we say that A and B are *isomorphic*, and we denote this relation as $A \cong B$.

In the context of **Set**, isomorphisms are exactly the bijections, which is equivalent to the statement that a function has a two-sided inverse if and only if it is bijective. With this, we see that two sets are isomorphic if and only if they contain the same number of elements, possibly labelled in different ways – that is, if their cardinalities are equal. The actual contents of the set, and any extra structure the set has, aren't important in **Set**.

The isomorphisms in **Grp** are group isomorphisms, as you'd might expect, but this is non-trivial to prove, especially if the definition of a group isomorphism you use is a "bijective homomorphism".* To show that group isomorphisms are isomorphisms in **Grp**, we need to prove that the inverse of a bijective homomorphism is also a homomorphism. Similarly, the isomorphisms in **Ring** are exactly the ring isomorphisms.

An isomorphism is the mathematical way of saying that we only care about some specific property of an object. If we're working with the natural numbers, it doesn't matter if we're using the Peano

* Which is not true for say, topological spaces, or the category **Top**, where homomorphisms are continuous functions. However, the inverse of a bijective continuous function is not necessarily continuous, so bijective homomorphisms in **Top** are not necessarily isomorphisms. The isomorphisms in **Top** are instead *bicontinuous* maps, also called *homeomorphisms.*

construction or the von Neumann construction, because there are isomorphisms between them that preserve the behaviour of 0, 1, + and ·, which are the only things that matter for natural numbers (when considered as a semiring). If we're studying groups, then we don't really care about what elements are in each group, only that these elements have group structure. From the point of view of the category, isomorphic elements look the same because they share the only properties that the category cares about. You've probably heard that a topologist cannot tell the difference between a coffee mug and a doughnut. This is because in **Top**, these two objects have the same number of holes (a topological invariant that *does* matter in **Top**), and they can be bicontinuously and bijectively deformed into each other.

51.2.2 Functors

One central theme of category theory is the idea of mappings between objects. Whenever we encounter a new type of mathematical object, we should always ask if there is a sensible notion of a map between these objects. Of course, categories themselves are mathematical objects we can ask this question on.

Let $\mathscr{C}$ and $\mathscr{D}$ be categories. A *functor*, $F : \mathscr{C} \to \mathscr{D}$, consists of two parts: a mapping on objects, and a mapping on morphisms, that follow two constraints. $F : \mathrm{ob}(\mathscr{C}) \to \mathrm{ob}(\mathscr{D})$ associates each object X in $\mathscr{C}$ to an object, $F(X)$, in $\mathscr{D}$.

$$X \mapsto F(X)$$

Similarly, the map $F : \hom(\mathscr{C}) \to \hom(\mathscr{D})$ associates each morphism $f : X \to Y$ in $\mathscr{C}$ to a morphism $F(f) : F(X) \to F(Y)$ in $\mathscr{D}$ such that:

- $F(\mathrm{id}_X) = \mathrm{id}_{F(X)}$ for every object X in $\mathscr{C}$;
- $F(g \circ f) = F(g) \circ F(f)$ for all morphisms $f : X \to Y$ and $g : Y \to Z$ in $\hom(\mathscr{C})$.

That is, the functor preserves identity morphisms and composition of morphisms.

A more concise way to phrase this is, for every pair of objects $A{,}B \in \mathrm{ob}(\mathscr{C})$, the functor F induces a mapping $F_{A,B} : \hom_{\mathscr{C}}(A{,}B) \to \hom_{\mathscr{D}}(F(A){,}F(B))$ that respects the structure of the categories.

$$\begin{tikzcd}
A \arrow[r, "f", ""{name=U, below}] & B \\
F(A) \arrow[r, "F(f)", ""{name=D, above}] & F(B)
\arrow[from=U, to=D, mapsto, shorten=2mm]
\end{tikzcd}$$

Theorem 51.2.2. *Functors preserve commutativity of diagrams.*

Proof. Because functors preserve composition of morphisms, for any two paths $a_1 \circ a_2 \circ \ldots \circ a_n$ and $b_1 \circ b_2 \circ \cdots \circ b_m$ connecting two objects in a commutative diagram of $\mathscr{C}$, we have,

$$\begin{aligned} F(a_1) \circ F(a_2) \circ \cdots \circ F(a_n) &= F(a_1 \circ a_2 \circ \cdots \circ a_n) \\ &= F(b_1 \circ b_2 \circ \cdots \circ b_m) \\ &= F(b_1) \circ F(b_2) \circ \cdots \circ F(b_m) \end{aligned}$$

so the corresponding paths in $\mathscr{D}$ are also equal, and hence the diagram of $\mathscr{D}$ commutes. ■

Corollary 51.2.2.1. *In particular, functors preserve isomorphism diagrams, so if f is an isomorphism in $\mathscr{C}$, then $F(f)$ is an isomorphism in $\mathscr{D}$.*

One of the most basic examples of a functor is the *constant functor* ΔX which associates every object in $\mathscr{C}$ to a single object $X \in \mathrm{ob}(\mathscr{D})$, and every morphism to id_X. Because every morphism is transformed into the identity morphism on X, composition and identities are trivially preserved, satisfying functoriality.

For a possibly more familiar example, let $(G, \cdot\,)$ and $(H, *\,)$ be groups, interpreted as categories $\mathscr{G}$ and $\mathscr{H}$. Any functor $F : \mathscr{G} \to \mathscr{H}$ must associate the only object in $\mathscr{G}$ to the only object in $\mathscr{H}$, and is thus determined only by its action on the morphisms. The functor must satisfy $F(\mathrm{id}_{\mathscr{G}}) = \mathrm{id}_{\mathscr{H}}$, and

$F(g \cdot h) = F(g) * F(h)$ for all morphisms $g,h \in \hom(\mathscr{G})$. So, any functor $\mathscr{G} \to \mathscr{H}$ is just a group homomorphism $G \to H$ (and again, we haven't mentioned inverses, so this holds similarly for monoids).

One very important type of functor is the so-called *forgetful functor*. Forgetful functors do nothing to the objects and morphisms of a category apart from "forgetting" some additional structure that mattered in the original category. For instance, the forgetful functor $U : \mathbf{Grp} \to \mathbf{Set}$.

Every object in **Grp** is a group – a set G with some extra structure in the form of a binary operation and a set of axioms. The forgetful functor U "forgets" this extra structure on objects, and gives $(G, \cdot\,) \mapsto G$, which is just a set – or rather, an object in **Set**. Similarly, morphisms in **Grp** are group homomorphisms, which are just set functions that happen to respect this extra structure. Forgetting that additional structure still leaves a normal set function – that is, a morphism in **Set**. Since morphisms are effectively unchanged, identity and composition morphisms still exist, so U is a well-defined functor.

Let $\mathscr{C}$ and $\mathscr{D}$ be categories. A *contravariant* functor from $\mathscr{C}$ to $\mathscr{D}$ is a functor $\mathscr{C}^{\mathrm{op}} \to \mathscr{D}$ (or equivalently, a functor $\mathscr{C} \to \mathscr{D}^{\mathrm{op}}$). In contrast, a *covariant* functor from $\mathscr{C}$ to $\mathscr{D}$ is an ordinary functor $\mathscr{C} \to \mathscr{D}$. Informally, a contravariant functor from $\mathscr{C}$ to $\mathscr{D}$ is just an ordinary functor $\mathscr{C}$ to $\mathscr{D}$ that "reverses all morphisms and compositions". This terminology is often used when a named category is involved – it is more convenient to say that a functor is contravariant, than to start writing $\mathbf{Set}^{\mathrm{op}}$ everywhere. However, contravariance can also arise naturally in some constructions:

For instance, the function that sends a set X to its power set $\mathcal{P}(X)$ defines the object mapping of a functor from **Set** to **Set**. We can define its action on morphisms $f : X \to Y$ by mapping f to the *direct image* function $\mathcal{P}(f) : \mathcal{P}(X) \to \mathcal{P}(Y)$ defined by $A \mapsto f(A)$, thus defining the *covariant* power set functor $\mathcal{P}(-) : \mathbf{Set} \to \mathbf{Set}$. However, we could alternatively define the morphism mapping by mapping f to the *inverse image* function $\mathcal{P}(f) : \mathcal{P}(Y) \to \mathcal{P}(X)$ defined by $A \mapsto f^{-1}(A)$. The inverse image function naturally reverses the direction of morphisms, thus defining the *contravariant* power set functor $\mathcal{P}(-) : \mathbf{Set}^{\mathrm{op}} \to \mathbf{Set}$.

51.2.3 Full and Faithful Functors

For set functions, it is often helpful to consider properties the function may satisfy on the codomain, such as surjectivity and injectivity. There exists a similar notion for functors: let $\mathscr{C}$ and $\mathscr{D}$ be locally small* categories, and let $F : \mathscr{C} \to \mathscr{D}$ be a functor. If for every pair of objects, $A,B \in \mathrm{ob}(\mathscr{C})$, the induced function $F_{A,B} : \hom_{\mathscr{C}}(A,B) \to \hom_{\mathscr{D}}(F(A),F(B))$ is:

- surjective, then F is *full*;
- injective, then F is *faithful*;
- bijective, then F is *fully faithful*.

Note that faithfulness is distinct from injectivity, in that faithful functors are not necessarily injective on objects or morphisms. For instance, let $\mathscr{C}$ be the discrete category on two objects, A and B, and let $\mathscr{D}$ be the trivial category on an object X. Any functor $F : \mathscr{C} \to \mathscr{D}$ will map the two objects in $\mathscr{C}$ to the unique object of $\mathscr{D}$, and similarly, the identity morphisms on A and B are both mapped to the identity morphism of X, so F is not injective on objects or morphisms. However, the functions $F_{A,A}$ and $F_{B,B}$ each map one morphism to one morphism, and are hence injective (in fact, bijective), while $F_{A,B}$ and $F_{B,A}$ are empty functions, and are hence vacuously injective (but not surjective, as $F(A) = F(B) = X$, and $\hom_{\mathscr{D}}(X,X)$ is non-empty). It follows that F is a faithful functor, but is injective on neither objects nor morphisms. Similarly, full functors are also not necessarily surjective on objects or morphisms, which can be shown by constructing a functor $G : \mathscr{D} \to \mathscr{C}$ in the previous example.

* Because the following definitions are in terms of properties of functions on hom-sets, we require that the hom-sets are indeed sets as these notions are set-theoretic in nature and do not extend readily to proper classes. For large categories, we extend the definition of full and faithful functors to left and right cancellative, respectively, instead.

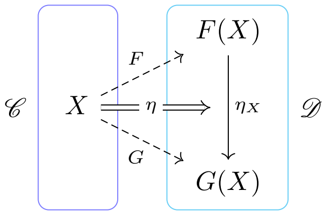
A fully faithful functor that is injective on isomorphism classes of objects is additionally said to be an *embedding*. If there exists an embedding F from $\mathscr{C}$ to $\mathscr{D}$, then F is said to *embed* $\mathscr{C}$ into $\mathscr{D}$.

51.3 Natural Transformations

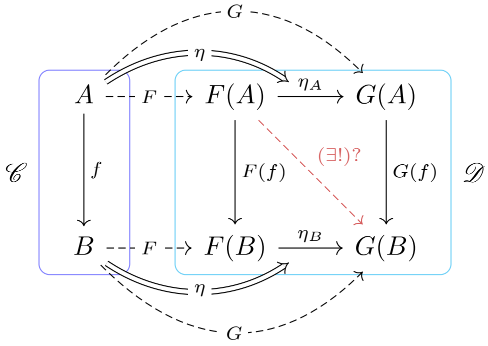
We have morphisms between categories in the form of functors, but the next obvious question to ask is, is there a notion of mappings between functors?

Fix categories $\mathscr{C}$ and $\mathscr{D}$, and let $\mathscr{C} \underset{G}{\overset{F}{\rightrightarrows}} \mathscr{D}$ be functors. A mapping $\begin{tikzcd} \mathscr{C} \arrow[r, bend left, "F", ""{name=U, below}] \arrow[r, bend right, "G"', ""{name=D}] & \mathscr{D} \arrow[Rightarrow, from=U, to=D, "\eta"] \end{tikzcd}$ or $\eta : F \Rightarrow G$ is then called a *natural transformation*.

F and G map objects and morphisms in $\mathscr{C}$ to objects and morphisms in $\mathscr{D}$, respectively. To define a mapping from F to G, we would like to associate objects and morphisms in $\mathscr{D}$ mapped by F to corresponding objects and morphisms in $\mathscr{D}$ mapped by G. For objects, this just means that if X is an object in $\mathscr{C}$, then we wish to associate $F(X)$ with $G(X)$. However, $F(X)$ and $G(X)$ are objects in $\mathscr{D}$, so a relation between them is just a morphism in $\hom_{\mathscr{D}}(F(X),G(X))$. So, η just maps each X in $\mathscr{C}$ to a morphism $F(X) \xrightarrow{\eta_X} G(X)$ called a *component* of η.



However, $\hom_{\mathscr{D}}(F(X),G(X))$ possibly contains many morphisms we could assign to η_X. To help us decide which one to use, consider a morphism $f : A \to B$ in $\mathscr{C}$. Under F and G, f gives the two morphisms $F(f) : F(A) \to F(B)$ and $G(f) : G(A) \to G(B)$. It would seem sensible for $F(f)$ to be related to $G(f)$ under η. From the mapping on objects, we also have $\eta_A : F(A) \to G(A)$ and $\eta_B : F(B) \to G(B)$, giving a square diagram of morphisms.



In this diagram, there are two paths from $F(A)$ to $G(B)$: $\eta_B \circ F(f)$, and $G(f) \circ \eta_A$. Because categories require compositions, these morphisms always exist, but if η_A and η_B were assigned without any other constraints, these compositions are not necessarily equal and there could be multiple distinct morphisms from $F(A)$ to $G(B)$. However, we can use this to relate $F(f)$ with $G(f)$ by enforcing that these compositions are equal, or equivalently, that the diagram commutes. This requirement is the *naturality* condition.

So, for categories $\mathscr{C}$ and $\mathscr{D}$, and functors $\mathscr{C} \underset{G}{\overset{F}{\rightrightarrows}} \mathscr{D}$, a natural transformation $\begin{tikzcd} \mathscr{C} \arrow[r, bend left, "F", ""{name=U, below}] \arrow[r, bend right, "G"', ""{name=D}] & \mathscr{D} \arrow[Rightarrow, from=U, to=D, "\eta"] \end{tikzcd}$ is a collection of morphisms $\left(F(X) \xrightarrow{\eta_X} G(X)\right)_{X \in \mathrm{ob}(\mathscr{C})}$ such that the following diagram commutes:

$$\begin{tikzcd}
A \arrow[d, "f"] & F(A) \arrow[r, "\eta_A"] \arrow[d, "F(f)"] & G(A) \arrow[d, "G(f)"] \\
B & F(B) \arrow[r, "\eta_B"] & G(B)
\end{tikzcd}$$

That is, $\eta_B \circ F(f) = G(f) \circ \eta_A$ for all $f : A \to B$ in $\hom(\mathscr{C})$.

We next need to verify that these natural transformations actually function as categorical morphisms. That is, that there always exists an identity, and that natural transformations compose associatively.

Following the component definition, the identity natural transformation on a functor $F : \mathscr{C} \to \mathscr{D}$ is a natural transformation $\mathrm{id}_F : F \Rightarrow F$ that maps each $X \in \mathrm{ob}(\mathscr{C})$ to a morphism $F(X) \xrightarrow{(\mathrm{id}_F)_X} F(X)$. This is just the identity morphism on $F(X)$, which always exists, so every component of id_F also always exists. The diagram, consisting of a single morphism and two identities, then trivially commutes, satisfying naturality, and hence id_F always exists. Identities, however, need to compose with other morphisms, and leave them unchanged. How do natural transformations compose?

51.3.1 Vertical Composition

Fix categories $\mathscr{C}$ and $\mathscr{D}$, and let F,G,H : $\mathscr{C} \to \mathscr{D}$ be functors. Consider the natural transformations $\alpha : F \Rightarrow G$ and $\beta : G \Rightarrow H$.

$$\begin{tikzcd}
\mathscr{C} \arrow[r, bend left=60, "F"{name=F}] \arrow[r, "G"{name=G, description}] \arrow[r, bend right=60, "H"'{name=H}] & \mathscr{D}
\arrow[Rightarrow, from=F, to=G, "\alpha"] \arrow[Rightarrow, from=G, to=H, "\beta"]
\end{tikzcd}$$

From the diagram, it would seem sensible to define the composition $\beta \circ \alpha$ to be a map from F to H. Such a composition of natural transformations is called a *vertical composition.*

Consider an object X in $\mathscr{C}$. The two components of α and β at X are then $\alpha_X : F(X) \to G(X)$ and $\beta_X : G(X) \to H(X)$. Because these are just morphisms in $\mathscr{D}$, they can be composed according to regular morphisms composition rules, and so, we can define the component $(\beta \circ \alpha)_X$ to be $\beta_X \circ \alpha_X : F(X) \to H(X)$. Because identity natural transformations map objects to identity morphisms, this also verifies that they do in fact function as identities with respect to vertical composition.

However, it remains to show that these components satisfy the naturality requirement.

$$\begin{tikzcd}
A \arrow[d, "f"] & F(A) \arrow[r, "\alpha_A"] \arrow[d, "F(f)"] & G(A) \arrow[r, "\beta_A"] \arrow[d, "G(f)"] & H(A) \arrow[d, "H(f)"] \\
B & F(B) \arrow[r, "\alpha_B"] & G(B) \arrow[r, "\beta_B"] & H(B)
\end{tikzcd}$$

Because α and β are natural transformations, they individually satisfy the naturality requirement, so each square commutes individually, and hence the diagram as a whole also commutes.

For any two categories, we can now define functors between them, and natural transformations between those functors that obey the morphism axioms. This suggests the construction of a new category, where the objects are functors, and the morphisms are natural transformations.

Let $\mathscr{C}$ and $\mathscr{D}$ be categories. We construct the *functor category* $[\mathscr{C},\mathscr{D}]$ by taking objects to be functors from $\mathscr{C}$ to $\mathscr{D}$, morphisms to be natural transformations, composition of morphisms to be vertical composition of natural transformations, and identity morphisms to be identity natural transformations.

Given the name of vertical composition, it is unsurprising that we have a notion of *horizontal composition*, but its discussion is relegated to §51.6.4 in the addendum.

51.3.2 Natural Isomorphisms

Fix categories $\mathscr{C}$ and $\mathscr{D}$. A *natural isomorphism* between functors from $\mathscr{C}$ to $\mathscr{D}$ is an isomorphism in the functor category $[\mathscr{C},\mathscr{D}]$.

That is, $\eta : F \Rightarrow G$ is an natural isomorphism if η is a natural transformation and there exists a natural transformation $\vartheta : G \Rightarrow F$ such that $\vartheta \circ \eta = \mathrm{id}_F$ and $\eta \circ \vartheta = \mathrm{id}_G$, and we write η^{-1} for ϑ.

$$\begin{tikzcd} A \arrow[loop left, "\mathrm{id}_A"] \arrow[r, shift left, "f"] & B \arrow[l, shift left, "f^{-1}"] \arrow[loop right, "\mathrm{id}_B"] & \leftarrow \text{cf.} \rightarrow & \mathscr{C} \arrow[d, bend right=60, "F"', "\mathrm{id}_F \circlearrowright"'{name=F}] \arrow[d, bend left=60, "G", "\circlearrowleft \mathrm{id}_G"{name=G}] \arrow[from=F, to=G, Rightarrow, shift left, "\eta"] \arrow[from=G, to=F, Rightarrow, shift left, "\eta^{-1}"] \\ & & & \mathscr{D} \end{tikzcd}$$

In this case, we say F and G are *naturally isomorphic*, and because natural isomorphisms are just isomorphisms in a specific type of category, we reuse notation and write $F \cong G$, or we say that $F(X) \cong G(X)$ *naturally in* X whenever we need to bind a variable.

The next theorem gives an alternative characterisation of natural isomorphisms.

Theorem 51.3.1. *Let* $\begin{tikzcd} \mathscr{C} \arrow[r, bend left=40, "F", ""{name=U, below}] \arrow[r, bend right=40, "G"', ""{name=D}] & \mathscr{D} \arrow[Rightarrow, from=U, to=D, "\eta"] \end{tikzcd}$ *be a natural transformation. Then, η is a natural isomorphism if and only if $\eta_X : F(X) \to G(X)$ is an isomorphism for all $X \in \mathrm{ob}(\mathscr{C})$.*

Proof. Suppose η is a natural isomorphism, so there exists ϑ such that $\vartheta \circ \eta = \mathrm{id}_F$. Then, $(\vartheta \circ \eta)_X = \vartheta_X \circ \eta_X = (\mathrm{id}_F)_X$ for all $X \in \mathrm{ob}(\mathscr{C})$, so every component is an isomorphism, completing the forward implication.

Now, suppose that $\eta_X : F(X) \to G(X)$ is an isomorphism for all $X \in \mathrm{ob}(\mathscr{C})$. Define $\vartheta : G \Rightarrow F$ by $\vartheta_X = (\eta_X)^{-1}$. Because η is a natural transformation, we have $\eta_B \circ F(f) = G(f) \circ \eta_A$. Left and right multiplying by ϑ_B and ϑ_A respectively, we have, $F(f) \circ \vartheta_A = \vartheta_B \circ G(f)$ which is exactly the naturality condition, and hence ϑ is a natural transformation. Then, $\vartheta \circ \eta = \Big(F(X) \xrightarrow{\vartheta_X \circ \eta_X} F(X)\Big)_{X \in \mathrm{ob}(\mathscr{C})} = \mathrm{id}_F$, and $\eta \circ \vartheta = \Big(G(X) \xrightarrow{\eta_X \circ \vartheta_X} G(X)\Big)_{X \in \mathrm{ob}(\mathscr{C})} = \mathrm{id}_G$, and hence η is a natural isomorphism, completing the backward implication. ■

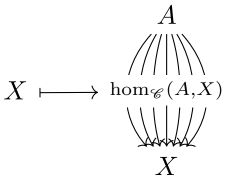
In the reverse direction, we used that η is a natural transformation to obtain naturality for ϑ. Without this, it could still be the case that $F(X) \cong G(X)$ for all X, but there does not exist a natural transformation from F to G at all, so "$F(X) \cong G(X)$ naturally in X" is a much stronger condition than just "$F(X) \cong G(X)$ for all X".

51.4 Hom-Functors

Suppose we wish to study the properties of an object A in a locally small category $\mathscr{C}$. One way to do so is to look at A from a different object, X. Then, look at A from another object, Y, and repeat. By looking at how A is seen by other objects, we can obtain a lot of information about A. The relationships an object X has with A are exactly the hom-sets $\hom(A,X)$ and $\hom(X,A)$, but these sets are different for each X. In fact, in locally small categories, this assignment of hom-sets with respect to a fixed A is functorial in X. That is, given a fixed A, every morphism $X \to Y$ induces a function $\hom_{\mathscr{C}}(A,X) \to \hom_{\mathscr{C}}(A,Y)$.

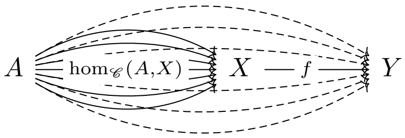
Let $\mathscr{C}$ be a locally small category, and fix an object $A \in \mathrm{ob}(\mathscr{C})$. We define the (covariant) *hom-functor*, $\hom(A,-) : \mathscr{C} \to \textbf{Set}$, also denoted h_A, as follows.

For each object $X \in \mathrm{ob}(\mathscr{C})$, we define $\hom(A,-)(X) = \hom_{\mathscr{C}}(A,X)$, so each object is mapped to the set of maps from A to that object.



We can interpret this as h_A mapping objects to how they are "seen" by A.

For each morphism $f : X \to Y$, we define $\hom(A,-)(f)$ to be the function $\hom(A,f) : \hom_{\mathscr{C}}(A,X) \to \hom_{\mathscr{C}}(A,Y)$, also denoted $h_A(f)$, defined by the postcomposition $g \mapsto f \circ g$.



[nLa23, adapted]

That is, we map each morphism $X \xrightarrow{f} Y$ in $\hom(\mathscr{C})$ to the function $h_A(f)$ that maps each morphism $A \xrightarrow{g} X$ to the composite morphism $A \xrightarrow{g} X \xrightarrow{f} Y$. In the above diagram, the morphisms on the left are "combed" through to Y through f, and we can interpret this as h_A mapping morphisms $X \xrightarrow{f} Y$ to how A "sees" the object Y through f.

The contravariant hom-functor $\hom(-,B)$, also denoted[*] h^B, is defined dually, with h^B mapping objects and morphisms to how they see B, rather than how they are seen from B.

$$\begin{tikzcd}
\mathscr{C} \arrow[rr, "h_A"] & & \textbf{Set} & \mathscr{C}^{\mathrm{op}} \arrow[rr, "h^B"] & & \textbf{Set} \\
X \arrow[d, "f"'] & \mapsto \arrow[d, phantom, "\mapsto"] & \hom_{\mathscr{C}}(A,X) \arrow[d, "h_A(f)"'] & X \arrow[d, "f"'] & \mapsto \arrow[d, phantom, "\mapsto"] & \hom_{\mathscr{C}}(X,B) \\
Y & \mapsto & \hom_{\mathscr{C}}(A,Y) & Y & \mapsto & \hom_{\mathscr{C}}(Y,B) \arrow[u, "h^B(f)"]
\end{tikzcd}$$

[Rie17, adapted]

Theorem 51.4.1. *h_A is a functor.*

Proof. We verify the functor axioms.

- Let $A \xrightarrow{f} X$ be a morphism. Then,

$$\begin{aligned}\big[h_A(\mathrm{id}_X)\big](f) &= \mathrm{id}_X \circ f \\ &= \mathrm{id}_{h_A(X)}(f)\end{aligned}$$

so h_A preserves identities.

[*] The usage of h_X for the covariant hom-functor and h^X for the contravariant hom-functor is not standardised. Some texts – notably [Lei14] – reverse the labelling, or use different notation entirely.

- Let $A \xrightarrow{h} X \xrightarrow{g} Y \xrightarrow{f} Z$ be morphisms. Then,

$$\begin{aligned}\big[h_A(g \circ f)\big](h) &= (g \circ f) \circ h \\ &= g \circ (f \circ h) \\ &= \big[h_A(g)\big](f \circ h) \\ &= \big[h_A(g)\big]\Big(\big[h_A(f)\big](h)\Big) \\ &= \big[h_A(g) \circ h_A(f)\big](h)\end{aligned}$$

  and h_A preserves compositions. ■

Corollary 51.4.1.1. *By duality, h^B is also a functor.*

For each object A, we have assigned a functor h_A, encapsulating how the category is seen from A, and as A varies, this view varies. However, it is the same category being seen from all objects, so it wouldn't be unusual for us to expect that this assignment has some internal consistency.

As it turns out, any morphism $f : A \to B$ induces a natural transformation $h_f : h_B \Rightarrow h_A$. Note the change in direction here! A collection of covariant functors come together to define a contravariant natural transformation. And, if we started with the contravariant hom-functors, they would all come together to define a covariant natural transformation.

Consider the component $h_B(X) \to h_A(X)$ of h_f at an object X. Recall that a map $h_B(X) \to h_A(X)$ just sends morphisms $B \to X$ to $A \to X$. We can interpret these hom-sets as contravariant hom-functors at a fixed X, so we're really just looking for a morphism $h^X(B) \to h^X(A)$, which is given exactly by precomposition by f. That is, each morphism $g : B \to X$ is mapped to the morphism $g \circ f : A \to X$.

In fact, there's no reason why we should have to fix one argument at a time. The notation hom(A,−) and hom(−,B) suggests that we may take both inputs to the hom-functor to be variable. Let $f : X \to Y$ and $h : B \to A$ be morphisms, and consider the following diagram:

$$\begin{tikzcd}\hom(A,X) \arrow[r, "\hom(h,X)"] \arrow[d, "\hom(A,f)"'] & \hom(B,X) \arrow[d, "\hom(B,f)"] \\ \hom(A,Y) \arrow[r, "\hom(h,Y)"'] & \hom(B,Y)\end{tikzcd}$$

Consider a morphism $g \in \hom(A,X)$. We will follow how it is mapped under this square along the two different paths, in a technique called *diagram chasing.*

Along the upper path, we have $g \mapsto \hom(h,X)(g) = g \circ h \mapsto \hom(B,f)(g \circ h) = f \circ (g \circ h)$. Along the lower path, we have $g \mapsto \hom(A,f)(g) = f \circ g \mapsto \hom(h,Y)(f \circ g) = (f \circ g) \circ h$. But, by associativity of morphism composition, these paths are equal, and we see that this diagram commutes for any choice of f, g, and h, implying that hom(−,−) is a functor $\mathscr{C}^{\text{op}} \times \mathscr{C} \to \mathbf{Set}$.

A functor $F : \mathscr{C} \to \mathbf{Set}$ is *representable* if $F \cong h_X$ (or h^X) for at least one choice of $X \in \text{ob}(\mathscr{C})$. The object X along with the natural transformation $F \Rightarrow h_X$ are then a *representation* of F. As it turns out, the object X is determined uniquely up to isomorphism in $\mathscr{C}$. We often call representable functors just *representables.*

As an example of a representable, the identity functor $\text{id}_{\mathbf{Set}} : \mathbf{Set} \to \mathbf{Set}$ is represented by the singleton set $\mathbf{1}$. Any function $\mathbf{1} \to X$ just picks elements from the set X, so there are exactly as many functions $\mathbf{1} \to X$ as there are elements of X, giving $\hom_{\mathbf{Set}}(\mathbf{1},X) \cong X = \text{id}_{\mathbf{Set}}(X)$, as required. Naturality also follows trivially as half of the functions to be considered are identities.

For a more interesting example, the forgetful functor $U : \mathbf{Grp} \to \mathbf{Set}$ is represented by the group $\mathbb{Z}$. Let G be a group. Because group homomorphisms send identities to identities, $0 \in \mathbb{Z}$ is always sent to the identity in G, so any homomorphism $\phi : \mathbb{Z} \to G$ is determined entirely by the image of 1 with the rest of the map following from the cyclic nature of $\mathbb{Z}$. This suggests that we send each homomorphism $\phi : \mathbb{Z} \to G$ to its determining value $\phi(1)$, giving us the components of a map $\alpha : \hom_{\mathbf{Grp}}(\mathbb{Z},-) \Rightarrow U$. The inverse map is then given by sending each $g \in U(G)$ to the homomorphism $z \mapsto g^z$. But, we still need naturality of this isomorphism.

Let $f : G \to H$ be a group homomorphism.

$$\begin{tikzcd}
G \arrow[d, "f"] & \hom_{\mathbf{Grp}}(\mathbb{Z},G) \arrow[r, "\alpha_G"] \arrow[d, "h_{\mathbb{Z}}(f)"'] & U(G) \arrow[d, "U(f)"] \\
H & \hom_{\mathbf{Grp}}(\mathbb{Z},H) \arrow[r, "\alpha_H"'] & U(H)
\end{tikzcd}$$

We will chase a homomorphism $\phi : \mathbb{Z} \to G$ through the diagram. Along the upper path, we have $\big(U(f) \circ \alpha_G\big)(\phi) = (f \circ \alpha_G)(\phi) = f(\alpha_G(\phi)) = f(\phi(1)) = (f \circ \phi)(1)$, and along the lower, we have, $\big(\alpha_H \circ h_{\mathbb{Z}}(f)\big)(\phi) = \alpha_H\big(h_{\mathbb{Z}}(f)(\phi)\big) = \big(h_{\mathbb{Z}}(f)(\phi)\big)(1) = (f \circ \phi)(1)$, so the diagram commutes, and $\hom_{\mathbf{Grp}}(\mathbb{Z},G) \cong U(G)$ naturally in G, as required. Through similar arguments, the forgetful functor $U : \mathbf{Ring} \to \mathbf{Set}$ is represented by the polynomial ring $\mathbb{Z}[x]$, and $U : \mathbf{Mon} \to \mathbf{Set}$ by the monoid $\mathbb{N}_0$ (you might notice that these are all free algebras on single generators – this is not a coincidence, §51.6.5).

As another example, the contravariant power set functor $\mathcal{P} : \mathbf{Set}^{\mathrm{op}} \to \mathbf{Set}$ sending sets to their power sets and functions to their inverse image is represented by the two element set $\mathbf{2}$, often depicted as $\{\top,\bot\}$ or $\{0,1\}$ with morphisms interpreted as an indicator functions of elements [Rie17].

For an example of a non-representable functor [Dot23], consider the functor $F : \mathbf{Set} \to \mathbf{Set}$ defined on objects by $X \mapsto X \coprod X$. Suppose there exists a set Y such that $\hom_{\mathbf{Set}}(Y,X) \cong X \coprod X$. If $X = \mathbf{1}$, then $\hom_{\mathbf{Set}}(Y,\mathbf{1}) \cong \{\mathbf{1}\} \cong \mathbf{1}$ is a singleton set, while $\mathbf{1} \coprod \mathbf{1} = \{\{0,1\},\{1,1\}\} \cong \mathbf{2}$ is a set with two elements, so they are not isomorphic and hence no such Y exists.

51.5 The Yoneda Lemma

It is an almost universal meta-problem in all of mathematics to describe and classify collections of mathematical objects [Hal20]. While a mathematical axiomatic definition of an object certainly distinguishes that object away from any others, this doesn't tell us much about the collection of all those objects as a whole. For example, while we can define a group in four short axioms, classifying all groups is a much harder problem. For a simpler example, imagine we are tasked with classifying the real numbers. The real number line is a classification of all real numbers by embedding them in some space that has more properties than the real numbers had alone. For instance, the number line is a metric space, a topological space, etc.

While we can define real number with Dedekind cuts, or with completeness axioms, this kind of embedding gives a lot of additional useful information that isn't visible from the axioms alone. Importantly, there is a bijection between the points on the number line and real numbers, but we also have the new information in that real numbers near each other on the number line are similar in magnitude. We can try extend this idea of a classifying space to other kinds of objects, where "nearby" objects have more similar properties than "distant" objects, and more generally, these spaces are called *moduli spaces* [Hal20]. Unfortunately, the moduli space for any kind of useful object is often completely unrecognisable, and has very few properties we can leverage to our advantage.

However, we can attempt to examine these spaces by looking at the maps from other spaces to them. Let $\mathbf{1}$ be the set with one element. Any map from $\mathbf{1}$ to $\mathbb{R}$ effectively amounts to picking an element from $\mathbb{R}$, so there is a bijection between the functions $\mathbf{1} \to \mathbb{R}$ and the points in $\mathbb{R}$. In fact, there's nothing specific about $\mathbb{R}$ here. More generally, the maps from the one-point space $\mathbf{1}$ to any space X amount to picking points from X. If X is, for example, a metric or topological space, then it is a set equipped with some extra structure in the form of a metric or a topology. By examining the maps from $\mathbf{1}$ to X, we can recover half of that information: just by looking at X from the simplest possible (non-empty) space, we recover all the points of X.

What if we look at the maps from a more complicated space? A map from the interval $[0,1]$ to X is just some parametrisation of a curve in X, so the maps $[0,1] \to X$ recover the paths in X, while a map from the circle S^1 to X is just a topological loop, so the maps $S^1 \to X$ recover the homotopy classes of loops on X. The point is, we get more and more information about X by examining how it appears from different choices of domains.

But exactly how much information can we recover? Is it always possible to obtain as much data from looking at maps as we would from just analysing the space itself? After all, we have no reason to expect that the entire structure of the space is always captured by these maps.

Except, it always is – and that, is the Yoneda lemma.*

The remarkable thing is that the Yoneda lemma is a proof at the level of categories, so it holds for any category of spaces.

We begin the lemma by asking what information representables recover. More precisely, let $\mathscr{C}$ be a locally small category, and fix an object $A \in \text{ob}(\mathscr{C})$, which induces the representable covariant functor h_A. For each covariant functor F, what are the natural transformations $h_A \Rightarrow F$ in the functor category $[\mathscr{C},\mathbf{Set}]$?

Lemma 51.5.1 (Yoneda)**.** *Let $\mathscr{C}$ be a locally small category. Then,*

$$\hom_{[\mathscr{C},\mathbf{Set}]}(h_A,F) \cong F(A)$$

naturally in $F \in \text{ob}([\mathscr{C},\mathbf{Set}])$ and $A \in \text{ob}(\mathscr{C})$.

Before we proceed with the proof, we should unwrap what this is saying, in exact terms. Firstly, there is an isomorphism of sets, so there is a bijective function between $\hom_{[\mathscr{C},\mathbf{Set}]}(h_A,F)$ and $F(A)$ – there are as many natural transformations from h_A to F as there are elements of $F(A)$. Moreover, the collection of natural transformations between two functors isn't guaranteed to be a set, even if the two associated categories are (locally) small, so the lemma also shows that hom-sets of this form can be put into bijection with proper sets.

Next, the isomorphism is said to be natural in F and A, suggesting that both sides are functorial in *both F and A* – any morphisms $F \Rightarrow G$ and $A \to B$ must induce maps

$$\hom_{[\mathscr{C},\mathbf{Set}]}(h_A,F) \to \hom_{[\mathscr{C},\mathbf{Set}]}(h_B,G) \quad \text{and} \quad F(A) \to G(B)$$

and not only does the isomorphism hold for every F and A, there exist isomorphisms $\hom_{[\mathscr{C},\mathbf{Set}]}(h_A,F) \to F(A)$ and $\hom_{[\mathscr{C},\mathbf{Set}]}(h_B,G) \to G(B)$ such that the induced square commutes for any choice of F and A.

More precisely, we can regard the left and right sides of the expression as bifunctors $[\mathscr{C},\mathbf{Set}] \times \mathscr{C} \to \mathbf{Set}$, mapping (F,A) to $\hom_{[\mathscr{C},\mathbf{Set}]}(h_A,F)$ and $F(A)$, respectively (in particular, this latter functor is known as the *evaluation functor*), and the Yoneda lemma states that these functors are naturally isomorphic.

Proof. Let $\eta : h_A \Rightarrow F$ be a natural transformation. Consider the following diagram:

$$\begin{tikzcd}
A \arrow[d, "f"] & h_A(A) \arrow[r, "\eta_A"] \arrow[d, "h_A(f)"'] & F(A) \arrow[d, "F(f)"] \\
B & h_A(B) \arrow[r, "\eta_B"'] & F(B)
\end{tikzcd}$$

[*] Or at least, part of it – it says a lot of things. The Yoneda lemma is very powerful in more advanced category theory, but this is one elementary application of it.

We chase the identity $\mathrm{id}_A \in \hom(A,A) = h_A(A)$ through the diagram. Along the upper path, we have $\mathrm{id}_A \mapsto \eta_A(\mathrm{id}_A) \mapsto F(f)(\eta_A(\mathrm{id}_A))$. Along the lower path, we have $\mathrm{id}_A \mapsto h_A(f)(\mathrm{id}_A) = f \circ \mathrm{id}_A = f$, followed by $f \mapsto \eta_B(f)$. From naturality of η, this diagram is commutative, so these two paths must be equal, giving $\eta_B(f) = F(f)(\eta_A(\mathrm{id}_A))$.

Remarkably, the input to the function on the right side is always $\eta_A(\mathrm{id}_A)$. This implies that any natural transformation $h_A \Rightarrow F$ is completely determined by its value at id_A. This naturally induces a function $\hom_{[\mathscr{C},\mathbf{Set}]}(h_A,F) \to F(A)$ defined by $\eta \mapsto \eta_A(\mathrm{id}_A)$, and moreover, this function is a bijection, as every value in $F(A)$ conversely extends to a unique natural transformation.

This establishes the required isomorphism, but we still need to show naturality.

First, we write both sides as functors $\vartheta, \mathrm{ev} : [\mathscr{C},\mathbf{Set}] \times \mathscr{C} \to \mathbf{Set}$. As mentioned before, the action of the two functors on objects is given by,

$$\vartheta(F,A) = \hom_{[\mathscr{C},\mathbf{Set}]}(h_A,F) \quad \text{and} \quad \mathrm{ev}(F,A) = F(A)$$

respectively. Now, we need to define their action on morphisms.

Being a product category, every morphism $(F,A) \to (G,B)$ in $[\mathscr{C},\mathbf{Set}] \times \mathscr{C}$ is of the form (α,f), where $\alpha : F \Rightarrow G$ is a morphism in $[\mathscr{C},\mathbf{Set}]$, and $f : A \to B$ is a morphism in $\mathscr{C}$. Fix two such morphisms, $\alpha : F \Rightarrow G$ and $f : A \to B$.

The first functor, ϑ, sends (α,f) to a function $\vartheta(\alpha,f) : \hom_{[\mathscr{C},\mathbf{Set}]}(h_A,F) \to \hom_{[\mathscr{C},\mathbf{Set}]}(h_B,G)$ defined by mapping each $\varphi : h_A \Rightarrow F$ to the composition $h_B \xrightarrow{h_f} h_A \xrightarrow{\varphi} F \xrightarrow{\alpha} G$. That is, $[\vartheta(\alpha,f)](\varphi) = \alpha \circ \varphi \circ h_f$.

The second functor, ev, sends the morphism (α,f) to a function $\mathrm{ev}(\alpha,f) : F(A) \to G(B)$. At this point, we should recall that α is a natural transformation, so the following diagram commutes:

$$\begin{tikzcd}
A \arrow[d, "f"] & F(A) \arrow[r, "\alpha_A"] \arrow[d, "F(f)"] & G(A) \arrow[d, "G(f)"] \\
B & F(B) \arrow[r, "\alpha_B"] & G(B)
\end{tikzcd}$$

From this, we see that there are two paths from $F(A)$ to $G(B)$, namely, $F(A) \xrightarrow{\alpha_A} G(A) \xrightarrow{G(f)} G(B)$, and $F(A) \xrightarrow{F(f)} F(A) \xrightarrow{\alpha_B} G(B)$. But, from naturality, these compositions are equal, so either choice yields the desired map. Next, we verify the functor axioms for ϑ and ev. First, the identity law:

$$\begin{aligned}
\vartheta(\mathrm{id}_F\,,\mathrm{id}_A)(\varphi) &= \mathrm{id}_F \circ \varphi \circ h_{\mathrm{id}_A} & \qquad \mathrm{ev}(\mathrm{id}_F\,,\mathrm{id}_A) &= F(\mathrm{id}_A) \circ (\mathrm{id}_A)_A \\
&= \varphi & &= \mathrm{id}_{F(A)} \circ \mathrm{id}_{F(A)} \\
&= \mathrm{id}_{[\hom(H_A,F)]}(\varphi) & &= \mathrm{id}_{F(A)}
\end{aligned}$$

where the first term on the right follows from the functoriality of F. So, ϑ and ev preserve identities.

Now, let $(F,A) \xrightarrow{(\alpha,f)} (G,B) \xrightarrow{(\beta,g)} (H,C)$ be morphisms.

$$\begin{aligned}
\vartheta\big((\beta,g) \circ (\alpha,f)\big)(\varphi) &= \vartheta(\beta \circ \alpha, g \circ f)(\varphi) & \qquad \mathrm{ev}\big((f,g) \circ (\alpha,f)\big) &= \mathrm{ev}(\beta \circ \alpha, g \circ f) \\
&= (\beta \circ \alpha) \circ \varphi \circ h_{g \circ f} & &= H(g \circ f) \circ (\beta \circ \alpha)_A \\
&= (\beta \circ \alpha) \circ \varphi \circ (h_f \circ h_g) & &= H(g) \circ H(f) \circ \beta_A \circ \alpha_A \\
&= \beta \circ (\alpha \circ \varphi \circ h_f) \circ h_g & &= H(g) \circ \beta_B \circ G(f) \circ \alpha_A \\
&= \beta \circ \big(\vartheta(\alpha,f)(\varphi)\big) \circ h_g & &= \big(H(g) \circ \beta_B\big) \circ \big(G(f) \circ \alpha_A\big) \\
&= \big[\vartheta(\beta,g) \circ \vartheta(\alpha,f)\big](\varphi) & &= \mathrm{ev}(\beta,g) \circ \mathrm{ev}(\alpha,f)
\end{aligned}$$

On the left, the expansion of $h_{g \circ f}$ follows from functoriality, with the reversal of the components resulting from contravariance. On the right, the expansion of $(\beta \circ \alpha)_A$ follows from the definition of vertical composition, and the replacement of $H(f) \circ \beta_A$ with $\beta_B \circ G(f)$ follows from the naturality of β. This last point is perhaps clearer as a diagram chase:

$$\begin{tikzcd}
F(A) \arrow[r, "(\beta\circ\alpha)_A"] \arrow[d, "F(g\circ f)"] \arrow[dr, "\mathrm{ev}((\beta\circ\alpha),(g\circ f))"] & H(A) \arrow[d, "H(g\circ f)"] \\
F(C) \arrow[r, "(\beta\circ\alpha)_A"] & H(C)
\end{tikzcd}
\qquad
\begin{tikzcd}
F(A) \arrow[r, "\alpha_A"] \arrow[d, "F(f)"] \arrow[dr, "\mathrm{ev}(\alpha,f)"] & G(A) \arrow[r, "\beta_A"] \arrow[d, "G(f)"] & H(A) \arrow[d, "H(f)"] \\
F(B) \arrow[r, "\alpha_B"] \arrow[d, "F(g)"] & G(B) \arrow[r, "\beta_B"] \arrow[d, "G(g)"] \arrow[dr, "\mathrm{ev}(\beta,g)"] & H(B) \arrow[d, "H(g)"] \\
F(C) \arrow[r, "\alpha_C"] & G(C) \arrow[r, "\beta_C"] & H(C)
\end{tikzcd}$$

The first two lines of the equation correspond to taking the upper path along the left diagram. The expansion in the third corresponds the uppermost path through the right diagram that passes through $H(A)$. But the upper right square commutes by the naturality of β, so we may route through $G(B)$ instead of $H(A)$. But then, this is just the route created by taking $\mathrm{ev}(\alpha,f)$ followed by $\mathrm{ev}(\beta,g)$, as required. If we take the other definition of the evaluation functor, we similarly use the functoriality of α along the lower path.

So, ϑ and ev preserve composition, finally verifying functoriality.

Now, we define a natural transformation $\Phi : \vartheta \Rightarrow \mathrm{ev}$. As stated earlier, we will map a natural transformation $\eta \in \vartheta(F,A)$ to its determining value $\eta_A(\mathrm{id}_A) \in \mathrm{ev}(F,A)$, giving us our definition of the component $\Phi_{(F,A)}$. All that remains is to show naturality:

$$\begin{tikzcd}
(F,A) \arrow[d, "(\alpha,f)"] \\
(G,B)
\end{tikzcd}
\qquad
\begin{tikzcd}
\vartheta(F,A) \arrow[r, "\Phi_{(F,A)}"] \arrow[d, "\vartheta(\alpha,f)"] & \mathrm{ev}(F,A) \arrow[d, "\mathrm{ev}(\alpha,f)"] \\
\vartheta(G,B) \arrow[r, "\Phi_{(G,B)}"] & \mathrm{ev}(G,B)
\end{tikzcd}$$

$$\begin{aligned}
\big(\mathrm{ev}(\alpha,f) \circ \Phi_{(F,A)}\big)(\eta) &= \mathrm{ev}(\alpha,f)\big(\eta_A(\mathrm{id}_A)\big) \\
&= \big(\alpha_B \circ F(f)\big)\big(\eta_A(\mathrm{id}_A)\big) \\
&= (\alpha_B \circ \eta_B \circ h_A(f))(\mathrm{id}_A) \\
&= (\alpha_B \circ \eta_B)\big(h_A(f)(\mathrm{id}_A)\big) \\
&= (\alpha_B \circ \eta_B)(f \circ \mathrm{id}_A) \\
&= (\alpha_B \circ \eta_B)(f) \\
&= (\alpha_B \circ \eta_B)(\mathrm{id}_B \circ f) \\
&= (\alpha_B \circ \eta_B)\big((h_f)_B(\mathrm{id}_B)\big) \\
&= (\alpha_B \circ \eta_B \circ (h_f)_B)(\mathrm{id}_B) \\
&= (\alpha \circ \eta \circ h_f)_B(\mathrm{id}_B) \\
&= \Phi_{(G,B)}(\alpha \circ \eta \circ h_f) \\
&= \big(\Phi_{(G,B)} \circ \vartheta(\alpha,f)\big)(\eta)
\end{aligned}$$

so the diagram commutes, as required. ■

51.5.1 The Yoneda Embedding

An important case of the Yoneda lemma is when the functor F is another hom-functor, h_B:

$$\hom_{[\mathscr{C},\mathbf{Set}]}(h_A,h_B) \cong \hom(B,A)$$

That is, the natural transformations between the two covariant hom-functors induced by A and B are in bijection with the morphisms between A and B in reverse direction: this is a contravariant(!) functor $\mathscr{C}^{\mathrm{op}} \to [\mathscr{C},\mathbf{Set}]$. This functor is denoted $h_\bullet$, defined on objects A by $h_\bullet(A) = h_A$, and on morphisms f by $h_\bullet(f) = h_f$. Similarly, applying the contravariant version of the Yoneda lemma to a contravariant hom-functor naturally gives rise to the covariant functor $h^\bullet : \mathscr{C} \to [\mathscr{C}^{\mathrm{op}},\mathbf{Set}]$.

In this context, the Yoneda lemma simply says that the functor $h_\bullet$ gives an embedding of $\mathscr{C}^{\mathrm{op}}$ into $[\mathscr{C},\mathbf{Set}]$. These functors are called the *Yoneda embeddings*, and are often denoted よ (hiragana *yo*) for Yoneda (from this point onwards, we will use よ wherever a proof applies to either functor).

Theorem 51.5.2 (Yoneda Embedding)**.** *Let $\mathscr{C}$ be a locally small category. Then, the Yoneda embeddings $\text{よ} : \mathscr{C} \hookrightarrow [\mathscr{C}^{\mathrm{op}},\mathbf{Set}]$ and $\text{よ} : \mathscr{C}^{\mathrm{op}} \hookrightarrow [\mathscr{C},\mathbf{Set}]$ are embeddings – that is, よ is fully faithful, and injective on objects up to isomorphism.*

Proof. よ is fully faithful if the induced mapping $\text{よ}_{A,B} : \hom(A,B) \to \hom(\text{よ}(A),\text{よ}(B))$ is a bijection for all objects $A,B \in \mathrm{ob}(\mathscr{C})$. But this is just the statement of the Yoneda lemma applied to hom-functors. Injectivity on objects up to isomorphism is proved in the corollary. ■

Corollary 51.5.2.1. *If $\hom(X,-) \cong \hom(Y,-)$ or $\hom(-,X) \cong \hom(-,Y)$, then $X \cong Y$.*

Proof. By the Yoneda lemma, any natural transformation $\eta : h_\bullet(X) \to h_\bullet(Y)$ (dually, $h^\bullet$) is induced by a morphism $Y \to X$ (resp. $X \to Y$). If η is an isomorphism, it follows that η and η^{-1} are both induced by inverse morphisms between X and Y, so $X \cong Y$. ■

At the beginning of this section, we asked how much information we get when we examine how an object looks from all other possible viewpoints. This corollary states that we recover the object, up to isomorphism – that is, the maps into or maps out of an object contain exactly as much information as that object itself.

Now, for one quick application of the Yoneda embedding, let $(G,\cdot\,)$ be a group of order n, interpreted as a category $\mathscr{G}$ with unique object $*$. We will write G to represent the set underlying the group $(G,\cdot\,)$.

Consider the action of the hom functor $h_* : \mathscr{G} \to \mathbf{Set}$ on the unique object of $\mathscr{G}$:

$$h_*(*) = \hom_{\mathscr{G}}(*,*)$$

and by construction, this hom-set is just the set G. Now, by the Yoneda lemma, we also have,

$$\hom_{[\mathscr{G},\mathbf{Set}]}(h_*,h_*) \cong \hom_{\mathscr{G}}(*,*) = G$$

First, note that this is exactly the statement that h_* is bijective on hom-sets, so h_* is fully faithful. Then, the left side is the set of natural transformations $h_* \Rightarrow h_*$, but since there is only one object, each of these natural transformations consists of a single component $G \to G$, and so,

$$\hom_{[\mathscr{G},\mathbf{Set}]}(h_*,h_*) \subseteq \hom_{\mathbf{Set}}(G,G) = S_n$$

Together, these equations give,

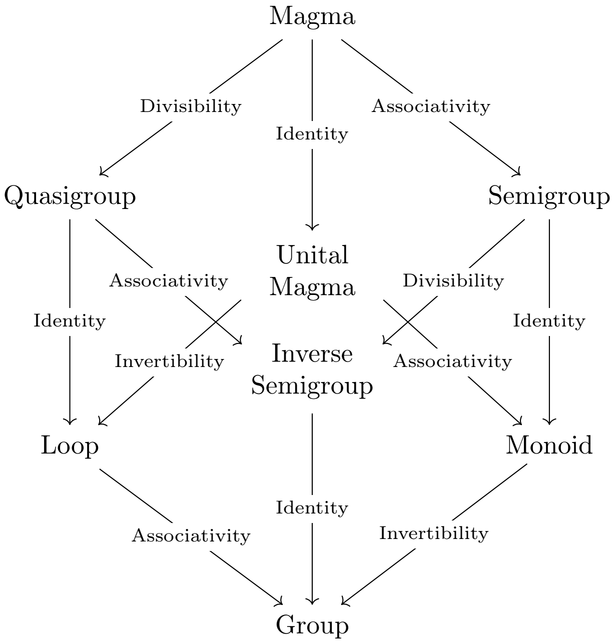
$$G \subseteq S_n$$

and moreover, because h_* is faithful, it provides an injection of $\hom_{\mathscr{G}}(*,*) \cong (G,\cdot\,)$ into $\hom_{\mathbf{Set}}(G,G) \cong S_n$, which is exactly the statement of Cayley's theorem.

51.6 Addendum

51.6.1 Group-Like Algebraic Structures

In category theory and abstract algebra, we often speak of sets with additional structure, usually in the form of a binary operation on that set. The simplest structure we begin with is a *magma*: a set that is closed under a binary operation. Adding in additional requirements produces a variety of group-like structures.



Of particular interest is the *group* and the *monoid*. The former is important to mathematics for obvious reason, and the latter is important in computer science – both theoretical and applied – as, for instance, the set of strings that can be constructed from a given set of characters forms a free monoid under the operation of concatenation.

51.6.2 Universal Set

We quickly recall three of the axioms of ZF(C).

- The axiom of extensionality states that two sets are contained by the same sets if they contain the same elements. (Informally, sets are determined entirely by their elements – two sets are equal if and only if they contain exactly the same elements.)
- The axiom of pairing states that if x and y are sets, then there exists a set that contains x and y as elements.
- The axiom of regularity states that every non-empty set x contains a member y such that x and y are disjoint.

Theorem 51.6.1. *Under the ZF(C) axiomatisation of set theory, there does not exist a universal set:*

$$\neg\exists S \forall x : x \in S$$

or, a set containing all sets.

Proof. Suppose S is a universal set. We can construct the set $\{S\}$ by applying the axiom of pairing to S with itself and removing the extra copy of S with the axiom of extensionality. Then, as $\{S\}$ contains only one element, regularity implies that S is disjoint from $\{S\}$, and hence S does not contain itself, contradicting the construction of S. It follows that S is not a set. ■

51.6.3 Set-Theoretic Problems

As mentioned before, the collections of objects and morphisms in a category do not generally form a set. The four main solutions [Bor21] to this are as follows:

- Ignore the problem;
- Use classes;
- Bounding the size of objects by some cardinal number, κ;
- Use Grothendieck universes (or other axiomatic solutions).

In this document, we mainly use a combination of the first two options: while we have recognised that these collections do not necessarily form sets, we also do not address the problem any further.

In our usage, this is acceptable as the categories we encounter are generally (locally) small and the classes we use are, for all intents and purposes, always sets wherever the distinction could matter. It is only in more advanced categorical constructions that the difference between sets and classes is of importance, but it is notable that many theorems in category theory are deeply intertwined with set-theoretic questions of size [Shu08] unlike in many other areas of mathematics. For instance, the Yoneda lemma demands that the categories used are locally small, while Freyd's celebrated *adjoint functor theorem* explicitly depends on a set of morphisms actually being a set and not a class.

One problem with our formulation of classes is that classes cannot contain other classes, or else we encounter problems when attempting to form the class of all classes. This causes some issues with constructing certain large categories which require collections of classes of objects or morphisms. The solution to this to use *conglomerates*, as in [Mac13] and [AHS90], which are to classes what classes are to sets. Since we mainly work with categories that have at most classes of objects and morphisms, conglomerates are generally a satisfactory solution to this problem, but we still run into issues when forming things like the category of all categories with this approach.

The third option, which we have opted not to use, turns the object and hom-classes of a category into sets by bounding the sizes of objects available. The hom-class is then bounded by the size of the power set of the object class, which is a set by the axiom of the power set. For instance, instead of considering the class of all sets to be the object class of **Set**, we pick a cardinal number κ, and only consider the set of sets of cardinality at most κ.

However, this is somewhat clumsy and artificial, as we need to keep track of extra data for every category we work with, and moreover, it involves making an arbitrary choice, with runs counter to the working principles of naturality.

The fourth option is to use a *Grothendieck universe* (or to resolve these problems in other axiomatic ways). Before we discuss Grothendieck universes, we must discuss model theory.

In order to mathematically encapsulate some concept, we begin with a list of *axioms*, which we take to be true by definition, and a list of *inference rules* that let us derive new statements from existing statements. Together, axioms and inference rules generate a *theory* consisting of all the statements that can be constructed from the axioms by applying inference rules to them. All the statements within a theory that are not axioms are called *theorems*.

For instance, we could have,

- All men are mortal (axiom);
- Socrates is a man (axiom);

- If "all A are B" and "X is A", then "X is B" (inference rule);
- Therefore, Socrates is mortal (theorem).

We can't do anything further with these axioms using our inference rule, so these three statements form our entire theory about Socrates, men, and mortality.

A *model* is any collection of objects that is consistent with a given theory. For instance, while our theory requires for us to have a mortal Socrates, it does not preclude the possibility of our model containing an immortal Cerberus, because the theory does not say anything about Cerberus, or about things that are not men.

For a more practical mathematical example, suppose we are trying to axiomatise the natural numbers. We begin by asserting that 0 is a number, then by saying that every number x has a successor, $S(x)$. The natural numbers are clearly a model of these two axioms, but they aren't the only model. For instance, a model consisting of a single number, 0, such that $S(0) = 0$, is consistent with our theory. The real numbers, or complex numbers are also consistent with our theory. So, the goal is to add just enough axioms to sufficiently constrain the possible models for our theory to be useful.

In much the same way, the axioms of ZFC are not assertions about "the real" universe of sets, because they are satisfied by many possible "universes of sets" [Shu08]. In fact, the Löwenheim-Skolem theorem states that any countable theory of first-order logic that admits an infinite model cannot have a unique model (up to isomorphism).

A Grothendieck universe U is a set that is *transitive*, closed under pairing, power sets, and indexed unions. That is,

- (transitive) $x \in U \land y \in x \to y \in U$;
- (pairing) $x \in U \land y \in U \to \{x,y\} \in U$;
- (power set) $x \in U \to (x) \in U$;
- (indexed unions) $I \in U \land \{x_i\}_{i \in I} \subseteq U \to \left(\bigcup_{i \in I} x_i\right) \in U$.

You may notice that several of these properties closely mirror axioms of ZFC, and as such, U will behave much like a "universal set" with respect to any element it contains. That is, for any element $x \in U$, U will contain all subsets of x, $\mathcal{P}(x)$, $\mathcal{P}(\mathcal{P}(x))$, etc., and it turns out that any uncountable Grothendieck universe is a model of ZFC itself.

Furthermore, the existence of non-trivial Grothendieck universes is not provable from within ZFC, as it would imply the existence of certain infinite cardinal numbers called *strongly inaccessible cardinals* that are not provable from ZFC, and in fact, it is possible to formulate Grothendieck universes as a type of inaccessible cardinal, as is done in [Shu08]. We can then add an axiom stating the existence of a Grothendieck universe.

Another popular extension of ZFC is *Tarski-Grothendieck set theory*, which is ZFC with an additional axiom that roughly says "for every set x, there exists a Grothendieck universe it belongs to", which states the existence of not just one Grothendieck universe, but an entire infinite hierarchy of Grothendieck universes.

In any case, once a Grothendieck universe is established, we may speak of *small* and *large* sets, which are sets that are and are not elements of the Grothendieck universe, respectively, instead of sets and (proper) classes.

Yet another approach is to abandon classical axiomatisations of set theory altogether, and formulate the foundations of mathematics in terms of category theory. There are several such systems with very different approaches, the most popular of which include *Elementary Theory of the Category of Sets* (ETCS), *First-Order Logic with Dependent Sorts* (FOLDS), and, most famously, *homotopy type theory.*

These topics are far beyond the scope of this essay, but they make for very compelling motivations for the study of category theory. For the interested reader, the following may make for useful further reading:

[Awo11] General background reading;

[LR03] Undergraduate textbook based in categorical foundations;

[Mak95] FOLDS;

[Uni13] Homotopy type theory.

51.6.4 Horizontal Composition

Given the name of vertical composition, it is unsurprising that we have a notion of *horizontal composition*, denoted by $\diamond$.

Fix categories $\mathscr{C}$, $\mathscr{D}$ and $\mathscr{E}$, and let F,$G : \mathscr{C} \to \mathscr{D}$ and F',$G' : \mathscr{D} \to \mathscr{E}$ be functors. Consider the natural transformations $\alpha : F \Rightarrow G$ and $\beta : F' \Rightarrow G'$.

$$\begin{tikzcd}
\mathscr{C} \arrow[r, bend left=50, "F"{name=F}] \arrow[r, bend right=50, "G"'{name=G}] \arrow[from=F, to=G, Rightarrow, "\alpha"] & \mathscr{D} \arrow[r, bend left=50, "F'"{name=Fp}] \arrow[r, bend right=50, "G'"'{name=Gp}] \arrow[from=Fp, to=Gp, Rightarrow, "\beta"] & \mathscr{E}
\end{tikzcd}$$

Because functors compose, we also have functors $F' \circ F : \mathscr{C} \to \mathscr{E}$ and $G' \circ G : \mathscr{C} \to \mathscr{D}$. The horizontal composition $\beta \diamond \alpha$ then maps $F' \circ F$ to $G' \circ G$.

We again consider an object X in $\mathscr{C}$. F and G map X to a pair of objects in $\mathscr{D}$, and α gives the morphism between them. F', G' and β then map these objects and morphism to a square in $\mathscr{E}$.

$$\begin{tikzcd}
 & & \mathscr{D} & & & \\
 & & F(X) \arrow[r, dashed, "F'"] \arrow[dd, "\alpha_X"] \arrow[rr, bend left=40, dashed, "F'"] \arrow[r, Rightarrow, bend left=30, "\beta"] & F'(F(X)) \arrow[r, "\beta_{F(X)}"] \arrow[dd, "F'(\alpha_X)"] \arrow[rdd, dashed, "(\beta\diamond\alpha)_X"] & G'(F(X)) \arrow[dd, "G'(\alpha_X)"] & \\
\mathscr{C} & X \arrow[ru, dashed, "F"] \arrow[rd, dashed, "G"'] \arrow[r, Rightarrow, "\alpha"] & {} & & & \mathscr{E} \\
 & & G(X) \arrow[r, dashed, "G'"] \arrow[rr, bend right=40, dashed, "G'"'] \arrow[r, Rightarrow, bend right=30, "\beta"'] & F'(G(X)) \arrow[r, "\beta_{G(X)}"] & G'(G(X)) &
\end{tikzcd}$$

The square of morphisms in $\mathscr{E}$ commutes as β is a natural transformation, so we can define $(\beta \diamond \alpha)_X = \beta_{G(X)} \circ F'(\alpha_X) = G'(\alpha_X) \circ \beta_{F(X)}$.

Next, we show naturality of this assignment.

First, consider the naturality diagram of α.

$$\begin{tikzcd}
A \arrow[d, "f"] & & F(A) \arrow[r, "\alpha_A"] \arrow[d, "F(f)"] & G(A) \arrow[d, "G(f)"] \\
B & & F(B) \arrow[r, "\alpha_B"] & G(B)
\end{tikzcd}$$

Then,

$$\begin{tikzcd}
A \arrow[d, "f"] & F'(F(A)) \arrow[r, "\alpha_A"] \arrow[d, "F'(F(f))"] & F'(G(A)) \arrow[d, "F'(G(f))"] \\
B & F'(F(B)) \arrow[r, "\alpha_B"] & F'(G(B))
\end{tikzcd} \tag{1}$$

also commutes for any choice of $A \xrightarrow{f} B$ in $\mathscr{C}$ as F' is a functor (Theorem 51.2.2).

Next, we observe the naturality diagram of β.

$$\begin{tikzcd}
X \arrow[d, "g"] & F'(X) \arrow[r, "\beta_X"] \arrow[d, "F'(g)"] & G'(X) \arrow[d, "G'(g)"] \\
Y & F'(Y) \arrow[r, "\beta_Y"] & G'(Y)
\end{tikzcd}$$

This diagram commutes for choice of objects and morphism $X \xrightarrow{g} Y$ in $\mathscr{D}$, so, picking $X = G(A)$, $Y = G(B)$, and $g = G(f)$, we have that

$$\begin{tikzcd}
G(A) \arrow[d, "G(f)"] & F'(G(A)) \arrow[r, "\beta_{G(A)}"] \arrow[d, "F'(G(f))"] & G'(G(A)) \arrow[d, "G'(G(f))"] \\
G(B) & F'(G(B)) \arrow[r, "\beta_{G(B)}"] & G'(G(B))
\end{tikzcd} \tag{2}$$

commutes (again, for any choice of $A \xrightarrow{f} B$ in $\mathscr{C}$).

Pasting diagrams (1) and (2) together, we have,

$$\begin{tikzcd}
A \arrow[d, "f"] & F'(F(A)) \arrow[r, "F'(\alpha_A)"] \arrow[d, "F'(F(f))"] & F'(G(A)) \arrow[r, "\beta_{G(A)}"] \arrow[d, "F'(G(f))"] & G'(G(A)) \arrow[d, "G'(G(f))"] \\
B & F'(F(B)) \arrow[r, "F'(\alpha_B)"] & F'(G(B)) \arrow[r, "\beta_{G(B)}"] & G'(G(B))
\end{tikzcd}$$

We have just shown that the left and right squares commute, and hence the outer square also commutes.

This gives,

$$\begin{aligned}
(G' \circ G)(f) \circ (\beta \circ \alpha)_A &= G'(G(f)) \circ \beta_{G(A)} \circ F'(\alpha_A) \\
&= B_{G(B)} \circ F'(\alpha_B) \circ F'(F(f)) \\
&= (\beta \circ \alpha)_A \circ (F' \circ F)(f)
\end{aligned}$$

which is exactly the naturality condition.

Vertical and horizontal composition are related by the *interchange law*: given categories, functors, and natural transformations,

$$\begin{tikzcd}[column sep=huge]
\mathscr{C} \arrow[r, bend left=60, "F"] \arrow[r, "G" description] \arrow[r, bend right=60, "H"'] \arrow[r, phantom, shift left=4ex, "\Downarrow\alpha"] \arrow[r, phantom, shift right=4ex, "\Downarrow\beta"] & \mathscr{D} \arrow[r, bend left=60, "F'"] \arrow[r, "G'" description] \arrow[r, bend right=60, "H'"'] \arrow[r, phantom, shift left=4ex, "\Downarrow\alpha"] \arrow[r, phantom, shift right=4ex, "\Downarrow\beta"] & \mathscr{E}
\end{tikzcd}$$

we have,

$$(\beta' \circ \alpha') \diamond (\beta \circ \alpha) = (\beta' \diamond \beta) \circ (\alpha' \diamond \alpha)$$

In these situations, not only do we have objects and morphisms in the form of categories and functors, but we also have morphisms between morphisms in the form of natural transformations between those functors.

What we have really been examining is an example of a *2-category*, which is a generalisation of a category to include morphisms between morphisms. But of course, there are 3-categories, and now we've started counting. This line of inquiry quickly leads to ∞-categories, which are some of the objects of study in a generalisation of category theory called *higher category theory.*

51.6.5 Adjoint Functors

Fix categories $\mathscr{C}$ and $\mathscr{D}$, and let $\begin{tikzcd} \mathscr{C} \arrow[r, shift left, "F"] & \mathscr{D} \arrow[l, shift left, "G"] \end{tikzcd}$ be functors. F is *left adjoint* to G, and G is *right adjoint* to F, if,

$$\hom_{\mathscr{C}}(F(A),B) \cong \hom_{\mathscr{D}}(A,G(B))$$

naturally in $A \in \text{ob}(\mathscr{C})$ and $B \in \text{ob}(\mathscr{D})$, and we write $F \dashv G$ to denote this relationship.

Often, forgetful functors from a category $\mathscr{C}$ of algebraic objects to **Set** admit a left adjoint which is often given by the free functor that constructs the associated free algebraic object on any set.

Recall that a free group F_S on a set S consists of all words whose letters are either elements $s \in S$, or their formal inverses s^{-1}, modulo the equivalence relation that identifies xx^{-1} and $x^{-1}x$ with the empty string, ε. The group operation is then given by concatenation of words, and the identity element is given by ε. Note that the free group on a single generator is isomorphic to $(\mathbb{Z},+)$, with the isomorphism $\phi : F(\mathbf{1}) \to \mathbb{Z}$ given by mapping each word to its length.

As you'd might expect, this assignment is a functorial: there is a functor $\mathcal{F} : \mathbf{Set} \to \mathbf{Grp}$ called the *free group functor* that sends every set S to the free group F_S.

Let X be a set, Y a group, and $U : \mathbf{Grp} \to \mathbf{Set}$ the forgetful functor. Every group homomorphism $\phi : \mathcal{F}(X) \to Y$ is determined uniquely by the image of the generators of $\mathcal{F}(X)$, which are exactly the elements of the set underlying Y, or, $U(Y)$. That is, every group homomorphism $\phi \in \hom_{\mathbf{Grp}}(\mathcal{F}(X),Y)$ corresponds uniquely to a function $X \to U(Y)$, which is exactly the statement,

$$\hom_{\mathbf{Grp}}(\mathcal{F}(X),Y) \cong \hom_{\mathbf{Set}}(X,U(Y))$$

Through some tedious algebra, naturality can also be verified, and the free group functor $\mathcal{F} : \mathbf{Set} \to \mathbf{Grp}$ is left adjoint to the forgetful functor $U : \mathbf{Grp} \to \mathbf{Set}$. Again, similar arguments show that the forgetful and free functors for other algebraic structures like rings and monoids are all adjoint pairs.

Now, suppose a functor $F : \mathbf{Set} \to \mathscr{C}$ is left adjoint to a functor $G : \mathscr{C} \to \mathbf{Set}$, so we have,

$$\hom_{\mathscr{C}}(F(A),X) \cong \hom_{\mathbf{Set}}(A,G(X))$$

Note that the hom-set on the right is in **Set**, and we know that the set functions $\mathbf{1} \to X$ are in bijection with elements of X for any set X, so we have,

$$\hom_{\mathbf{Set}}(\mathbf{1},G(X)) \cong G(X)$$

so G is representable! Moreover, we have,

$$\hom_{\mathscr{C}}(F(\mathbf{1}),X) \cong G(X)$$

so G is specifically represented by $F(\mathbf{1})$. Because F and G were arbitrary, this shows that any such right adjoint is representable. In the case where F and G are a free and forgetful adjoint functor pair, this also shows that forgetful functors are always represented by free objects on single generators.

Bibliography

[Lei14] Leinster, T. *Basic Category Theory*. Cambridge University Press, 2014. ISBN: 9781107044241.

[Mac13] MacLane, S. *Categories for the Working Mathematician*. Springer New York, 2013. ISBN: 9781475747218.

[Bor+94] Borceux, F. et al. *Handbook of Categorical Algebra: Volume 1, Basic Category Theory*. Cambridge University Press, 1994. ISBN: 9780521441780.

[nLa23] nLab authors. *Yoneda embedding*. Revision 39. 2023. URL: https://ncatlab.org/nlab/show/Yoneda+embedding.

[Rie17] Riehl, E. *Category Theory in Context*. Dover Publications, 2017. ISBN: 9780486820804.

[Dot23] Dotto, E. *Personal communications*. 2023.

[Hal20] Halpern-Leistner, Daniel. *Moduli theory*. 2020. URL: http://pi.math.cornell.edu/~danielhl/modern_moduli_theory.pdf.

[Bor21] Borcherds, R.E. *Categories for the Idle Mathematician*. 2021. URL: https://www.youtube.com/watch?v=JOp7mH72Jlg.

[Shu08] Shulman, Michael A. *Set Theory for Category Theory*. 2008. arXiv: 0810.1279 [math.CT].

[AHS90] Adamek, J., Herrlich, H., and Strecker, G.E. *Abstract and Concrete Categories: The Joy of Cats*. Wiley, 1990. ISBN: 9780471609223.

[Awo11] Awodey, Steve. *From Sets to Types to Categories to Sets*. 2011. ISBN: 9789400704305.

[LR03] Lawvere, F.W. and Rosebrugh, R. *Sets for Mathematics*. Cambridge University Press, 2003. ISBN: 9780521010603.

[Mak95] Makkai, Michael. *First Order Logic with Dependent Sorts, with Applications to Category Theory*. 1995. URL: https://www.math.mcgill.ca/makkai/folds/foldsinpdf/FOLDS.pdf.

[Uni13] Univalent Foundations Program, The. *Homotopy Type Theory: Univalent Foundations of Mathematics*. Institute for Advanced Study: https://homotopytypetheory.org/book, 2013.

All diagrams were written in LaTeX with the `tikz` package.

Chapter 52

Category Theory II

52.1 Introduction

Suppose you were asked, "is $3 \in \mathbb{N}$?" Being a natural number, 3 is indeed a member of $\mathbb{N}$, so the answer is "yes". A little tricker is the question, "is $0 \in \mathbb{N}$?", but there is at least a meaningful and unambiguous answer as long as we are clear about the meaning of the symbol $\mathbb{N}$.[†] On the other hand, the question, "is $\pi \in \mathbb{Q}$?", would quickly receive an answer of "no".

Now, suppose you were then asked, "is $\pi \in \log$?"

You'd might pause for a moment, before again answering in the negative, but for a different reason than before. After all, π is a number, and log is a function, so π being a member of log – whatever that means – would be ridiculous! A better answer might be to declare the question as meaningless.

This illustrates the intuitive notion of *type*, which may be particularly familiar to programmers. Many programming languages (called *strongly typed* languages) require you to declare the type of a variable before using it, with the idea being that strictly enforcing the type of every variable stops the programmer from performing nonsensical operations like adding an `int` to a `bool`, or trying to divide by a `string`.

However, in the standard foundational framework of ZFC, Zermelo–Fraenkel set theory with Choice – the "assembly language" of mathematics, if you will – *everything* is a set, so the question "is $\pi \in \log$?" should have a yes-or-no answer.

For instance, in ZFC, membership is a global relation, so it is always a valid question to ask whether any two arbitrary objects are members of each other, even if the answer is entirely meaningless. Because of this, the way ZFC uses the word "set" is very different from what mathematicians usually mean when they say "set". In ZFC, π is a set, as is log – but ask any mathematician to list the elements of π or log, and you will likely have difficulties in receiving an answer.

The benefit of this style of axiomatisation is simplicity – everything is a set, so we don't have to have extra rules to deal with every possible different kind of object. On the other hand, we lose this basic notion of type, because everything is of type `set`. (We say that ZFC is a *single-sorted* theory.)

As we saw above, this isn't always sensible. The axioms of ZFC allow even more nonsensical questions beyond asking whether any two random objects contain each other or are equal, and many of the axioms of ZFC themselves are similarly incomprehensible in ordinary mathematical usage. For instance, the axiom of regularity states that every non-empty set X contains a member $x \in X$ such that $x \cap X = \varnothing$. But, pick any ordinary set, say, $\mathbb{R}$, and the resulting statement is difficult to interpret. What does an expression like $3 \cap \mathbb{R}$ even mean?

[†] We will take the answer to be "yes" in this paper.

One response to this problem might be to say that set theory offers not only a set of axioms, but also a collection of standard encodings of different mathematical objects. We can again compare this situation with computers: down in a hard drive, every file – text, image, audio, video, etc. – is ultimately encoded as bits: as pure combinations of 0s and 1s. So, one could argue that it doesn't matter that these questions don't have meaningful answers, because nobody is claiming that they should; just as how opening a text file with the wrong encoding results in garbled nonsense doesn't stop it from being useful when opened correctly.

But, even if we accept that the encodings are arbitrary, this problem of being able to create meaningless statements goes deeper than just posing questions about set memberships. One enlightening exercise, as posed by Benacerraf [Ben65], is to consider the question, "is $3 \in 17$?"

52.1.1 Is $3 \in 17$?

Benacerraf describes two children, Johnny and Ernie,[*] who have learnt mathematics from axiomatic set-theoretic foundations (as opposed to the more commonly preferred method of starting from "counting", which he calls the "vulgar way"), say for instance, ZFC. To introduce the notion of "counting", and other common uses of natural numbers to these children is simple, as their teachers merely need to point out the common "vulgar" names of set-theoretic constructions they already know. However, there is some choice in the matter here.

Johnny is taught that there is a set, N, which ordinary people call the "(natural) numbers", that is equipped with a well-ordering called the *less-than* relation. Furthermore, this set contains an element that ordinary people refer to as the natural number 0; the empty set, and the *successor* $s(n)$ for any set n is the set $s(n) = n \cup \{n\}$ – so every number n is simply the collection of numbers less than it.[†]

The normal properties of natural numbers assumed by ordinary people can then be exhibited as concrete theorems for Johnny. While the common "vulgar" explanation of addition, multiplication, exponentiation, etc., are informal recursive definitions, Johnny can concretely define these procedures in terms of the successor operation, so these operations are derivable from this theory. Restricting our focus to finite sets, Johnny can also encode the common notion of counting with *cardinality* – a set has n elements if it can be put in bijective correspondence with the set of natural numbers less than n, and this definition is well-defined as Johnny's first order theory is sufficiently powerful to construct such a correspondence for any finite n.

At this point, Johnny can now communicate with the vulgar, with all the common constructions and usages of numbers fully encoded within his first order theory of sets. Note that all we have done is specify the set N and explain the notions of 0 and successor to Johnny. The laws of arithmetic can be derived from there, as can any other "extramathematical" uses of numbers. For instance, the notion of counting can be similarly encoded with the additional provision of a definition of cardinality. It can be reasonably agreed that this information is both necessary and sufficient to completely characterise the natural numbers for common usage.

52.1.2 The Isomorphism Problem

Now, Ernie is also provided with a set to be labelled N, a designated element $0 \in N$, and a definition of a successor function, so all the previous statements apply similarly to Ernie. The two are thus equally knowledgeable about the natural numbers and can both prove numerous theorems about them; and in discussion with ordinary people, they are in agreement.

The problems first arise when they consider the statement, "is $3 \in 17$?"

[*] Named in reference to *John von Neumann* and *Ernst Zermelo*.
[†] This is the standard von Neumann construction of the naturals.

Johnny argues that the statement is true, while Ernie disagrees. Attempts to resolve this by consulting ordinary people are met with nothing but confusion – after all, to ordinary people, numbers are just that – *numbers* – and not sets.

Examining their given information reveals the origin of this discrepancy: by Johnny's definition of a successor function $s(n) = n \cup \{n\}$, every number is the set of numbers less than it, so $17 = \{0,1,2,3,\dots,15,16\}$; clearly, $3 \in 17$. However, Ernie's successor function is instead defined by $s(n) = \{n\}$,* so $17 = \{16\}$ and 3 is nowhere to be found; clearly, $3 \notin 17$. This isn't the only disagreement between the two systems either.

Johnny claims that a set has n elements if and only if it can be placed in bijection with the set of numbers less than n – and here, Ernie agrees; then, Johnny claims further that a set has n elements if and only if it can be placed in bijection with the number n itself – but for Ernie, every number contains only a single element (apart from zero, which is empty), so their notions of cardinality also disagree.

The source of the disagreements between Johnny and Ernie is obvious – the difference between their successor functions, and by extension, the set N. But what is *not* obvious, is how these disagreements should be reconciled.

Each account of the naturals is equally valid and correct when considered in isolation, with neither one to be preferred over the other. In more modern language, both constructions yield valid models of the Peano axioms – that is, the resulting semirings are isomorphic. So, if we accept Johnny's construction, there is no good reason why we shouldn't also accept Ernie's. Moreover, Johnny's and Ernie's accounts really are arbitrary, and there are infinitely many ways to assign sets to numbers – infinitely many choices of N, $0 \in N$, and $s : N \to N$ – that satisfy the Peano axioms.

Of course, we could choose to accept both accounts, and agree that $\big\{\varnothing,\{\varnothing\},\big\{\{\varnothing,\{\varnothing\}\}\big\}\big\} = 3 = \big\{\{\{\varnothing\}\}\big\}$, but this is clearly absurd, so we explore the alternative: at least one of the two accounts is false. We can actually make a stronger statement – at most one of the accounts (out of the infinite possibilities) can be "correct".

The belief that there *is* a true account is called set-theoretic Platonism – this is the idea that there is a particular set of sets somewhere in the universe which is the "real" set of natural numbers, regardless of whether there exists an argument to prove this or not, or even if we can ever find it.

Benacerraf rejects the possibility that there is no such argument, saying, "...if the number 3 is really one set rather than another, it must be possible to give some cogent reason for thinking so; for the position that this is an unknowable truth is hardly tenable. But there seems to be little to choose among the accounts. Relative to our purposes in giving an account of these matters, one will do as well as another, stylistic preferences aside."

This last sentence is at the heart of *structuralism.*

52.1.3 Structuralism

Mathematics, as mathematicians actually use it, does not demand of the natural numbers that they exist as some specific object, but only that they have the structure we require – when we work with the natural numbers, *we don't care* about the specific construction used; only that they have semiring structure; that they support recursive definitions and induction; that have a canonical embedding into the integers, etc.

In fact, this is how we usually describe and use objects in mathematics. Some basic examples of this are vectors and groups: a vector space is anything that satisfies the vector space axioms; and similarly, a group is anything that satisfies the group axioms. In neither definitions we do we prescribe what the vector space or group itself actually consists of, only requiring that whatever object or objects it is *behaves* in a certain way.

[*] This is the historical Zermelo construction of the naturals.

To the structuralist, mathematics is the study of structures independent of the things they are composed of. As seen above, this is the approach taken in many other mathematical contexts, so it is strange that the foundations of mathematics itself are commonly formulated in a way that is distinctly *not* structural in nature. But this does not have to be the case.

In a *structural set theory*, sets are objects that are characterised by their connections to other sets as prescribed by functions or relations – and this is essentially how sets are used in common practice of mathematics. Elements of sets themselves have no identity or internal structure beyond that which is given by functions and relations. In particular, this means that elements are not sets,* and cannot be members of other sets (not in the sense that it is false that they are, but in the sense that it is meaningless to ask whether they are [nLa23a]), so elements of different arbitrary sets are not comparable.

It is meaningless to ask whether $3 = \{\{\{\varnothing\}\}\}$ or not, because it is not asked in the context of the rest of the natural numbers – for instance, we know $2 \neq 3$, because, for example, 2 is strictly less than 3, which is a property of natural numbers; but it seems wrong to argue that $3 \neq \{\{\{\varnothing\}\}\}$ because, say, we know that 3 has three elements (or none, or seventeen, or infinitely many), while $\{\{\{\varnothing\}\}\}$ only has one, because *we don't know this.* The number of elements of 3 isn't a part of the structure of the natural numbers, so we cannot meaningfully deny that $3 = \{\{\{\varnothing\}\}\}$ on the grounds that 3 contains a different number of elements than $\{\{\{\varnothing\}\}\}$.

What makes the number 3 the number 3 is exactly its relations to other natural numbers, so structuralism tells us that a more sensible question than "what is 3?" would be "what are *all* the natural numbers?" – or more precisely, "what *structure* is the natural numbers?"

We just saw that the question "is $3 \in 17$?" has a different answer depending on which construction is chosen – Johnny says "yes"; and Ernie, "no" – despite the resulting collections being isomorphic. We say that ZFC is not *isomorphism invariant.*

In contrast, structural definitions are always isomorphism invariant (with respect to the relevant structure). For instance, we structurally characterise a *natural numbers object* as a triple $(N,0,s)$ consisting of a set N, a distinguished element $0 \in N$, and a *successor function* $s : N \to N$. The natural numbers object then expresses natural arithmetic in terms of these components. Importantly, in this characterisation, it doesn't actually matter what the elements of N actually are, only that they carry this structure – the elements themselves are meaningless in isolation.

The structuralist then says that the number "3" is "the third place in a natural numbers object", rather than any particular set like in ZFC. We don't have to argue what set 3 is, because 3 *isn't a set* – it is a relation or structure that some particular objects may exhibit. The statement "is $3 \in 17$?" is then not a well-formed statement, because the $\in$ relation isn't compatible with members of this structure in this way.

In this way, structural set theories are not only isomorphism invariant, but are also free from much of the arbitrary constructions and additional baggage of ZFC that are never actually used in common mathematics.

Mathematicians generally do not appeal to any kind of axioms when doing mathematics – even when working with sets – without any loss of accuracy to their work. In practice, we never actually think of functions or numbers as sets, and that doesn't ever seem to pose a problem. So, it appears that we all subconsciously have a set of operating principles we use for manipulating mathematical constructions that are *good enough* for almost all purposes. The idea is that the axioms of structural foundations are much closer to these intuitions because we formulate our axioms in terms of how we want objects to behave, rather than as a list of rules about which objects exist.

* Such an element is called a *urelement.* In classical material set theories, we usually have no urelements, as it is possible to embed material set theories with urelements into a version without urelements, simplifying the theory. In structural set theories, however, *every* element is a urelement.

52.1.4 Primitive Notions

In ZFC, and many other traditional *material* axiomatisations of set theory, the basic primitive notions are of *sets*, *elements*, and *membership*, and everything else is derived from there.

For instance, consider the notion of a *function*. Informally, a function is a special kind of correspondence between pairs of objects, where every given object is assigned exactly one corresponding object by the function. It is also helpful to view a function as an *operation* or as some kind of input-output process that is applied to an object to obtain its associated object (its *image*).

We can represent a function $f : A \to B$ as a *relation* – a subset of $A \times B$ – given by $\hat{f} = \big\{(x,y) : y \text{ is the } f\text{-image of } x\big\}$, where (a,b) is an ordered pair. Conversely, to distinguish which relations $R \subseteq A \times B$ represent functions, we use the property that functions assign exactly one image to each input: if a relation $\hat{f}$ satisfies the property that $(x,y) \in \hat{f}$ and $(x,z) \in \hat{f}$ imply that $y = z$, then $\hat{f}$ is a representation of some function.

This construction encodes our informal notion of a function as a set of ordered pairs satisfying a certain property. The next step is a trick commonly used in mathematics; we drop the distinction between the notion of a function and its abstraction as a set, and we say that this formal representation is itself the *definition* of a function [Gol84].

This definition works well on a technical level, and much theory can be developed with it, but there are several conceptual hurdles that come alongside it. One point of difficulty is with the *codomain* of the function: we can easily define the sets $\text{dom}(f) = \big\{x : \exists y : (x,y) \in f\big\}$ and $\text{im}(f) = \big\{y : \exists x : (x,y) \in f\big\}$, but there is no way to recover the codomain of a function from this definition.

This is not a problem in some branches of mathematics, such as analysis, or even much of set theory. However, in more algebraic or topological areas, this poses some difficulties.

Let $A \subset B$, and consider the functions $\text{id}_A : A \to A$ and $\iota : A \hookrightarrow B$ both defined by $x \mapsto x$. The former is the identity function on A, while the latter is the inclusion of A into B, with the usage of the term "inclusion" indicating that we should view the function as including the elements of A into B. These functions are conceptually very distinct, but they are both the set $\big\{(x,x) : x \in A\big\}$.

This is not only a conceptual problem, but a practical one in some cases: if, for instance, we take $A = S^1$ and $B = \mathbb{C}$, then the identity and inclusion maps yield very different induced homomorphisms in first homology.

Even if we patch this definition to specify the information of the codomain separately, this definition still fails to faithfully capture the *dynamic* quality of a function; we often speak of a function *acting* on or being *applied* to an input, and the symbol between the domain and codomain of a function is even an *arrow*! Even further, more specialised functions like some transformations in linear algebra, geometry, group theory, etc. are explicitly described as motions of space. In contrast, the characterisation of a function as a set is inherently static.

Because functions are exactly how sets relate to one another, they are very important in a structural context. In fact, in our structural axiomatisation of set theory, we will instead take sets and *functions* to be our primitive notions, with elements and the membership relation now being derived. This choice of primitive notions lends itself well to be described with the language of category theory.

52.2 Categories

We briefly state some standard categorical definitions, generally adapting those from [Lei14] (though with some notable differences). For a more introductory and motivated treatment of these definitions, see [Kit22].

Loosely speaking, a *category* consists of a collection of *objects*, with *morphisms* or *arrows* pointing between objects, subject to a couple of axioms pertaining to how morphisms compose. These axioms derive from the properties of function composition, so in many ways, a category is a vast generalisation of sets and set functions.

One of the basic precepts of category theory is that objects have no internal identity; in an arbitrary category, objects are not (necessarily) sets, so it makes no sense to try "look inside" an object. Even if they do happen to be sets, it turns out that looking at the morphisms connecting to that object is sufficient to determine it (almost) uniquely – in this way, category theory is inherently structural, making it well-suited for discussing structural foundations.

Formally, a *category* $\mathscr{C}$ consists of:

- A class $\mathrm{ob}(\mathscr{C})$ of *objects* in $\mathscr{C}$. We often write $A \in \mathscr{C}$ to abbreviate $A \in \mathrm{ob}(\mathscr{C})$.
- For all (ordered) pairs of objects $A,B \in \mathrm{ob}(\mathscr{C})$, a class $\hom_{\mathscr{C}}(A,B)$ of *maps* or *arrows* called *morphisms* from A to B, called the *hom-set* or *hom-class* of morphisms from A to B, also sometimes written $\mathscr{C}(A,B)$ or $\hom(A,B)$ if the ambient category is clear. If $f \in \hom(A,B)$, we write $f : A \to B$ or $A \xrightarrow{f} B$. The union of all of these classes is the hom-set of $\mathscr{C}$, and is written $\hom(\mathscr{C})$.
- For any three objects $A,B,C \in \mathrm{ob}(\mathscr{C})$, a binary operation, $\circ : \hom(A,B) \times \hom(B,C) \to \hom(A,C)$, $(f,g) \mapsto g \circ f$, called *composition*, such that,
  - (*associativity*) if $f : A \to B$, $g : B \to C$, and $h : C \to D$, then $h \circ (g \circ f) = (h \circ g) \circ f$;
  - (*identity*) for every object $X \in \mathrm{ob}(\mathscr{C})$, there exists a morphism $\mathrm{id}_X : X \to X$ called the *identity morphism* on X, such that every morphism $f : A \to X$ satisfies $\mathrm{id}_X \circ f = f$, and every morphism $g : X \to B$ satisfies $g \circ \mathrm{id}_X = g$.

Note that, despite the name, a hom-set is not necessarily a set (under ZFC), and may in fact be a proper class.* If the class of morphisms between any pair of objects does happen to be a set, then $\mathscr{C}$ is *locally small*; and if $\mathrm{ob}(\mathscr{C})$ is also a set, then $\mathscr{C}$ is additionally *small*.

We list a few illustrative examples of categories:

(i) The prototypical example of a category is the category of sets and set functions, **Set**. Identity morphisms are identity functions, and associativity follows from basic properties of set functions.

(ii) The category of groups and group homomorphisms, **Grp**. Every group is a set with extra structure, and every group homomorphism is a set function that happens to preserve this structure, so associativity and identity are inherited from **Set**.

(iii) Similarly, collections of sets with extra structure and maps that preserve that structure generally form categories called *concrete*† categories. For example, the category of:

- Monoids and monoid homomorphisms, **Mon**;
- Rings and ring homomorphisms, **Ring**;
- Metric spaces and non-expansive maps, **Met**;
- Topological spaces and continuous maps, **Top**;
- C^p-manifolds and p-times differentiable maps $\mathbf{Man}^p$;
- Measurable spaces and measurable functions, **Mea**;

* Given that we are attempting to axiomatise sets themselves, this statement may be somewhat confusing, but for now we will take the word "set" as a declaration of *intent* to form a collection that does not invoke paradoxes or contradictions – no universal or Russell sets. We can get surprisingly far with the intuitive idea of a set as a "bag of featureless dots".

† More properly, a concrete category is a category equipped with a faithful functor to **Set**, but informally, they are categories that "look like **Set** with extra structure".

- Vector spaces and linear maps over a fixed field K, $\mathbf{Vect}_K$; etc.

(iv) Let $(M, *)$ be a monoid, $\mathrm{ob}(\mathscr{C}) = \{\bullet\}$, and $\hom(\bullet,\bullet) = G$. For any two morphisms f and g, define the composition $f \circ g$ to be $f * g$. Identity and associativity follow from the monoid axioms, so $\mathscr{C}$ is a category. In this way, any monoid can be regarded as a category on a single object.

In particular, we see that the structure of this category is captured entirely within the morphisms, and the object itself is unimportant, so much so that we don't even assign it any characteristics beyond being a featureless point, $\bullet$. Note that this also means this category is *not* concrete, as the object $\bullet$ is not a set.

(v) Let X be any set equipped with a preorder (a reflexive and transitive relation), $\leq$. Let $\mathrm{ob}(\mathscr{C}) = X$, and for any two elements $x,y \in X$, define a unique morphism $f : x \to y$ if and only if $x \leq y$. Reflexivity gives identity morphisms, and transitivity guarantees that compositions exist, so any preordered set can be regarded as a category. More generally, a *thin* or *posetal* category is a category with at most one morphism in every hom-set, so up to isomorphism, every preordered set is a thin category.

(vi) For any topological space X, its fundamental groupoid $\Pi_1(X)$ is a category. Its objects are points in X, and morphisms are homotopy classes of paths, with composition given by path concatenation.

(vii) Any ordinal constructed in the von Neumann style, $n = \{m : m < n\}$, defines a category $\mathbb{n}$ on n objects with morphisms given by set inclusions. (This defines a poset, so this is a specific example of the above preordering categories.)

For instance, $\mathbb{0}$ is the category with zero objects and morphisms (the *empty category*); $\mathbb{1}$ is the category with one object (the *trivial* or *terminal category*); $\mathbb{2}$ is the category with two objects and a single non-identity morphism between them (the *arrow category*), often depicted as $0 \to 1$; etc.

More generally, $\mathbb{n}$ is the category freely generated by the graph

$$0 \longrightarrow 1 \longrightarrow 2 \longrightarrow 3 \longrightarrow \cdots \longrightarrow n-1 \longrightarrow n$$

in the sense that every non-identity morphism can be uniquely factored as a composite of morphisms in the displayed graph [Rie17]. For instance, $0 \subset 3$, so there should be a morphism $0 \to 3$ in this category, but this is just the composition of $0 \to 1$, $1 \to 2$, and $2 \to 3$, which are all displayed in the graph.

(viii) For any cardinal n, we can define a category with n objects and no morphisms aside from the required identities. Such a category is called a *discrete* category.

Note that an *indiscrete* or *codiscrete* category is *not* simply a category that is not discrete, but is instead a category where every hom-set is a singleton (i.e. the category forms a complete digraph).

A category can be interpreted as a *context* or a *universe* in which we perform some kind of mathematics. For instance, group theory is performed within **Grp**, topology within **Top**, differential geometry within $\mathbf{Man}^\infty$, etc. Category theory thus provides a unified language for dealing with all of these different contexts in a uniform manner, and for translating statements and methods between these different disciplines.

Category theory itself was developed from within algebraic topology, where it was used to apply the tools of abstract algebra to topological contexts [Gol84], but has since become a branch of pure mathematics in its own right. In particular, category theory has imparted the lesson of conceptually reframing existing theories in structural, arrow-theoretic terms, which has often proven to be a valuable way to obtain new insights and underlying connections.

The most general context for performing mathematics is the category of sets, **Set**: all mathematical objects can all be translated down into structures on sets (or in the case of material foundations, into sets themselves), so axiomatising the category of sets is one method of formalising alternative foundations of mathematics.

52.2.1 Diagrams

The structure of a collection of objects and morphisms in a category is often visually represented as a directed graph, called a *diagram.* We have already used $A \to B$ to denote a morphism from A to B, but we can also draw larger diagrams to represent more objects and morphisms. For instance, this diagram depicts 3 objects with morphisms between them:

$$\begin{tikzcd} X \arrow[r, "f"] \arrow[dr, "h"'] & Y \arrow[d, "g"] \\ & Z \end{tikzcd}$$

Note that it is standard to omit identity morphisms from these diagrams to reduce clutter.

Because categories require compositions to exist, following a path through a diagram always gives a valid morphism between the endpoint objects. For instance, there is a path from X to Z that passes through morphisms f and g, so there is a morphism $g \circ f : X \to Z$ in this category. Furthermore, a diagram is *commutative* if for every pair of objects in the diagram, all routes between them are equal. For instance, the diagram above is commutative if and only if $h = g \circ f$. This also justifies the omission of identity morphisms in general diagrams; they don't meaningfully add any additional paths to the diagram.

52.2.2 Constructing Categories

Given two categories, $\mathscr{C}$ and $\mathscr{D}$, the *product category* $\mathscr{C} \times \mathscr{D}$ is the category with objects $\mathrm{ob}(\mathscr{C} \times \mathscr{D}) = \mathrm{ob}(\mathscr{C}) \times \mathrm{ob}(\mathscr{D})$ and morphisms $\hom_{\mathscr{C}\times\mathscr{D}}((A,B),(A',B')) = \hom_{\mathscr{C}}(A,A') \times \hom_{\mathscr{D}}(B,B')$, with compositions defined componentwise [Mac13]. That is, if $A \xrightarrow{f} A'$ and $B \xrightarrow{g} B'$ are objects and morphisms in categories $\mathscr{C}$ and $\mathscr{D}$ respectively, then we have the objects and morphism $(A,B) \xrightarrow{(f,g)} (A',B')$ in the product category $C \times D$.

Example. Take the arrow category, $\mathbb{2}$, with objects and single non-identity morphism $0 \xrightarrow{f} 1$. The product category $\mathbb{2} \times \mathbb{2}$ can then be represented as the diagram,

$$\begin{tikzcd} (0,0) \arrow[r, "(\mathrm{id}_0, f)"] \arrow[d, "(f,\, \mathrm{id}_0)"'] \arrow[dr, "(f, f)" description] & (0,1) \arrow[d, "(f,\, \mathrm{id}_1)"] \\ (1,0) \arrow[r, "(\mathrm{id}_1, f)"'] & (1,1) \end{tikzcd}$$

The diagonal morphism is often omitted from diagrams of this category, replaced by the requirement that the square commutes. △

Another way to construct new categories from an existing category is to *dualise* the category. The *dual* or *opposite* category $\mathscr{C}^{\mathrm{op}}$ of a category $\mathscr{C}$ is the category with the same class of objects, but with the domains and codomains of all morphisms interchanged. That is, $\mathrm{ob}(\mathscr{C}) = \mathrm{ob}(\mathscr{C}^{\mathrm{op}})$, and $\hom_{\mathscr{C}}(A,B) = \hom_{\mathscr{C}^{\mathrm{op}}}(B,A)$ for all objects A and B.

More generally, the *principle of duality* states that every categorical definition and theorem has a *dual* definition and theorem, obtained by reversing the direction of all morphisms in the categories involved. Dual notions are often prefixed with *co*-, as in domains and *co*domains.

Theorem 52.2.1 (Conceptual Duality)**.** *Let Σ be a statement that holds in all categories. Then the dual statement Σ^* holds for all categories.*

Proof. [Bor+94, adapted][Kit22] If Σ holds in a category $\mathscr{C}$, then Σ^* holds in $\mathscr{C}^{\mathrm{op}}$. Every category is the dual of its dual, so Σ^* holds in all categories. ■

There is also a notion of a *subcategory*; given a category $\mathscr{C}$, a category $\mathscr{D}$ is a subcategory of $\mathscr{C}$ if $\mathrm{ob}(\mathscr{D})$ is a subcollection of $\mathrm{ob}(\mathscr{C})$ and $\hom_{\mathscr{D}}(A,B)$ is a subcollection of $\hom_{\mathscr{C}}(A,B)$ for any objects A and B in $\mathscr{D}$. A subcategory is furthermore said to be *full* if for every pair of objects A and B, every morphism $A \to B$ in $\mathscr{C}$ is also in $\mathscr{D}$. That is, a full subcategory $\mathscr{D}$ is a subcollection of objects of $\mathscr{C}$ with all possible morphisms included.

52.2.3 Morphisms

52.2.3.1 Isomorphisms

Suppose we have objects A and B and morphisms $f : A \to B$ and $g : B \to A$ such that the following diagram is commutative:

$$\begin{tikzcd}
A \arrow[loop left, "\mathrm{id}_A"] \arrow[r, shift left, "f"] & B \arrow[l, shift left, "g"] \arrow[loop right, "\mathrm{id}_B"]
\end{tikzcd}$$

That is, $f \circ g = \mathrm{id}_B$ and $g \circ f = \mathrm{id}_A$, so f and g are *mutually inverse*. Then, we say that f and g are *isomorphisms*, and we alternatively label g by f^{-1}. If an isomorphism between a pair of objects A and B exists, we say that A and B are *isomorphic* and we write $A \cong B$.

Isomorphic objects are, as far as the ambient category is concerned, effectively identical – anything you can say about one object will apply just as well to any other isomorphic object.

In algebra, we often see phrases such as "*the* group S_3", or "*the* (semiring of) naturals", despite the fact that there exists many ostensibly distinct objects to which these names could refer – for instance, the set of isometries that preserve an equilateral triangle and the set of automorphisms on a set of cardinality 3 could both reasonably be labelled "S_3".

This reflects the idea that we often do not know or care about whether two objects are literally equal, but only that they are isomorphic with respect to whatever property we care about. In contrast, in set theory, we sometimes do care about whether two elements of a set are exactly equal or not, since, for example, sets are entirely determined by their elements.

This indicates that for different contexts and types of data, we care about different degrees of likeness. For elements of sets, this notion of likeness is equality. For objects in a category (such as, say, groups in **Grp**), this notion is isomorphism. In general, equality is too strong of a requirement in category theory in the sense that effectively all categorical results still hold if we weaken any requirements of equality to isomorphism – and further still, depending on the foundations used, arbitrary categories may not even admit a notion of equality between objects. Conversely, any purely categorical notion should also not refer to strict equality at all.

52.2.3.2 Monics and Epics

Consider a morphism $f : A \to B$ in some category $\mathscr{C}$. Suppose that for every pair of parallel morphisms into A

$$\begin{tikzcd}
X \arrow[r, shift left, "g_2"] \arrow[r, shift right, "g_1"'] & A \arrow[r, "f"] & B
\end{tikzcd}$$

we have that $f \circ g_1 = f \circ g_2$ implies $g_1 = g_2$ (f is *left-cancellative*). Then, we say that f is a *monomorphism* (or is *monic*), and we write $f : A \rightarrowtail B$.

Monomorphisms generalise injective set functions, and in many categories where objects are structured sets and morphisms are structure preserving set functions, the two notions coincide.

Theorem 52.2.2. *The monomorphisms in* **Set** *are precisely the injections.*

Proof. Suppose $f : X \to A$ is injective, and let $g_1,g_2 : A \to B$ be functions such that $f \circ g_1 = f \circ g_2$. Then, for any $x \in X$, $f\big(g_1(x)\big) = f\big(g_2(x)\big)$, and by injectivity, $g_1(x) = g_2(x)$, so $g_1 = g_2$ and f is monic.

Now, suppose instead that $f : X \to Y$ is monic. Consider two elements $a,b \in X$ such that $f(a) = f(b)$ and define $g_1,g_2 : \{\bullet\} \to X$ by $g_1(\bullet) = a$ and $g_2(\bullet) = b$. Then, $f \circ g_1 = f \circ g_2$, and since f is monic, $g_1 = g_2$, giving $a = g_1(\bullet) = g_2(\bullet) = b$, and hence f is injective. ■

Note that we have now characterised injectivity in terms of functions into and out of sets, making no mention of the elements within the set; thus describing injectivity in **Set** structurally.

(Non-empty) injective functions in **Set** are also always left-invertible, hinting at another connection between monics and invertibility.

If $f : A \to B$ is a morphism such that there exists a morphism $s : B \to A$ such that the composite $s \circ f$ is the identity on A, then f is a *split monomorphism*, and we say that s is the *section* of f, and that f is the *retraction* of s.

Note that being a split monomorphism is distinct from being a monomorphism; the former requires having a left-inverse, while the latter requires being left-cancellative, which, in general, are not the same thing. However, we do have:

Theorem 52.2.3. *Split monomorphisms are monomorphisms.*

Proof. Let $f : A \to B$ be a split monomorphism with left inverse $\ell : B \to A$, so $\ell \circ f = \mathrm{id}_A$, and let $g_1,g_2 : X \to A$ be morphisms such that $f \circ g_1 = f \circ g_2$. Then,

$$\begin{aligned} f \circ g_1 &= f \circ g_2 \\ \ell \circ f \circ g_1 &= \ell \circ f \circ g_2 \\ g_1 &= g_2 \end{aligned}$$

so f is monic. ■

Note, however, that the converse does not hold in general; not all monomorphisms split, as demonstrated by the empty function in **Set**. For another example, let H be a subgroup of a group G in **Grp**, and consider the inclusion map $\iota : H \hookrightarrow G$. The inclusion map is injective as a set function, so it is monic in **Set**, which is inherited into **Grp**. But, ι has a left inverse if and only if $G \setminus H$ is normal in G, so this monomorphism does not split in general.

Dually, a morphism $f : A \to B$ is then an *epimorphism* (or is *epic*) if it is monic in $\mathscr{C}^{\mathrm{op}}$. That is, if for every pair of parallel morphisms from B

$$\begin{tikzcd} A \arrow[r, "f"] & B \arrow[r, shift left, "g_2"] \arrow[r, shift right, "g_1"'] & X \end{tikzcd}$$

we have $g_1 \circ f = g_2 \circ f$ implies $g_1 = g_2$ (f is *right-cancellative*), and we write $f : A \twoheadrightarrow B$.

Epimorphisms, like monomorphisms and injections, generalise surjective set functions, and in **Set**, the two notions coincide.

Theorem 52.2.4. *The epimorphisms in* **Set** *are precisely the surjections.*

Proof. Suppose $f : X \to Y$ is surjective, and let $g_1,g_2 : Y \to A$ be functions such that $g_1 \circ f = g_2 \circ f$. By surjectivity, for every $y \in Y$ there is an $x \in X$ such that $y = f(x)$, so $g_1(y) = g_1\big(f(x)\big) = g_2\big(f(x)\big) = g_2(y)$, so $g_1 = g_2$ and hence f is epic.

Suppose otherwise that $f : X \to Y$ is epic, and define $g_1,g_2 : Y \to 2 = \{\top,\bot\}$ by $g_1(y) = \top$ and

$$g_2(y) = \begin{cases} \top & \exists x \in X : f(x) = y \\ \bot & \text{otherwise} \end{cases}$$

That is, g_2 maps the elements in the image of f to $\top$ and those outside to $\bot$. Then, $g_1 \circ f = g_2 \circ f$ by construction, and as f is epic, we have $g_1 = g_2$ so g_2 is the constant map at $\top$. So, the image of f is equal to Y, and f is surjective. ■

However, in categories of structured sets, epimorphisms are generally *not* surjective (unlike monomorphisms, which generally *are* injective), so the analogy with surjectivity should be taken less literally with epimorphisms. However, we again have a notion of splitting for epimorphisms:

If $f : A \to B$ is a morphism such that there exists a morphism $s : B \to A$ such that the composite $f \circ s$ is the identity on B, then f is a *split epimorphism*, and we say that s is the *section* of f, and that f is the *retraction* of s.

Corollary 52.2.4.1. *Split epimorphisms are epimorphisms.*

Proof. Dual of previous theorem. ■

Again, the converse does not hold in general; not all epimorphisms will split in an arbitrary category.

The previous two results together imply the following:

Theorem 52.2.5. *Isomorphisms are monic and epic.*

In certain categories, such as **Set**, every morphism that is simultaneously monic and epic (a *bimorphism*) is an isomorphism, and such a category is called *balanced*, but in general, the converse of this result does not hold. For instance, consider the inclusion $\iota : \mathbb{Z} \hookrightarrow \mathbb{Q}$ in **Ring**.

The inclusion map is injective, so it is monic in **Set**, which is inherited into **Ring**. Now, consider two maps f and g from $\mathbb{Q}$ to some ring R:

$$\begin{tikzcd} \mathbb{Z} \arrow[r, hook, "\iota"] & \mathbb{Q} \arrow[r, shift left, "f"] \arrow[r, shift right, "g"'] & R \end{tikzcd}$$

Because every rational $\frac{a}{b} \in \mathbb{Q}$ can be written as the product $a \cdot b^{-1}$ of an integer and a multiplicative inverse of an integer, the ring homomorphism f must map the rational $\frac{a}{b}$ to,

$$f\left(\frac{a}{b}\right) = f(a) \cdot f(b)^{-1}$$

and similarly for g, so if f and g agree over the integers – that is, if $f \circ \iota = g \circ \iota$ – then they are equal everywhere, and hence ι is epic. So, ι is a bimorphism, but is clearly not an isomorphism, so **Ring** is not a balanced category.

52.2.4 Functors

A *functor*, $F : \mathscr{C} \to \mathscr{D}$ between categories $\mathscr{C}$ and $\mathscr{D}$, consists of a mapping on objects and a mapping on morphisms, such that,

- $F(\mathrm{id}_X) = \mathrm{id}_{F(X)}$ for every object X in $\mathscr{C}$;
- $F(g \circ f) = F(g) \circ F(f)$ for all morphisms $f : X \to Y$ and $g : Y \to Z$ in $\hom(\mathscr{C})$.

That is, the functor preserves identity morphisms and composition of morphisms. Equivalently, for every pair of objects $A,B \in \mathrm{ob}(\mathscr{C})$, the functor F induces a mapping $F_{A,B} : \hom_{\mathscr{C}}(A,B) \to \hom_{\mathscr{D}}(F(A),F(B))$ that respects the structure of the categories. If this induced function is surjective, then F is *full*; if it is injective, then F is *faithful*; and if it is bijective, then F is *fully faithful*. If a fully faithful functor $F : \mathscr{C} \to \mathscr{D}$ is additionally injective on isomorphism classes, it is called an *embedding*, and F is said to *embed* $\mathscr{C}$ into $\mathscr{D}$.

Functors encapsulate the idea that categorical constructions should also tell you what to do with mappings.

Example (Hom-Functor). Let $\mathscr{C}$ be locally small, and fix any object $A \in \text{ob}(\mathscr{C})$. For any object $X \in \text{ob}(\mathscr{C})$, we can form the set of morphisms $X \to A$, which is exactly $\hom_{\mathscr{C}}(X,A)$. Note that because $\mathscr{C}$ is locally small, this hom-set is a set and not a proper class, so it can be viewed as some element of **Set**.

This assignment of hom-sets to objects is functorial; there is a functor $\hom(A,-)$, also denoted h_A, that sends every object X to the hom-set $h_A(X) = \hom_{\mathscr{C}}(A,X)$. This functor sends every morphism $f : X \to Y$ to the function $\hom(A,f) : \hom_{\mathscr{C}}(A,X) \to \hom_{\mathscr{C}}(A,Y)$, also denoted $h_A(f)$, defined by mapping every $g \in \hom_{\mathscr{C}}(A,X)$ to its postcomposition by f to obtain $f \circ g \in \hom_{\mathscr{C}}(A,Y)$, thus defining the *covariant hom-functor.* △

A *covariant* functor from $\mathscr{C}$ to $\mathscr{D}$ is simply a functor $\mathscr{C} \to \mathscr{D}$. In contrast, a *contravariant* functor from $\mathscr{C}$ to $\mathscr{D}$ is a functor $\mathscr{C} \to \mathscr{D}^{\text{op}}$ (or equivalently, $\mathscr{C}^{\text{op}} \to \mathscr{D}$). That is, a contravariant functor reverses the direction of morphisms. This happens naturally in some constructions (particularly in many topological constructions involving preimages), and in those cases, it is easier to say that a functor is contravariant than to start appending "$^{\text{op}}$" to half the categories involved.

Contravariant functors with codomain **Set** are common enough that they have their own name: a *presheaf* on a category $\mathscr{C}$ is a functor $\mathscr{C}^{\text{op}} \to \textbf{Set}$, the name deriving from the notion of presheaves on topological spaces.

Example (Hom-Functor). Again, let $\mathscr{C}$ be locally small, and fix any object $A \in \text{ob}(\mathscr{C})$. For any other object X, we can similarly form the set of morphisms $A \to X$, which is exactly $\hom_{\mathscr{C}}(A,X) \in \text{ob}(\textbf{Set})$.

This assignment of hom-sets to objects is again functorial; the functor $\hom(-,A)$, also denoted h^A, sends every object X to the hom-set $h^A(X) = \hom_{\mathscr{C}}(X,A)$. Let $f : X \to Y$ be a morphism in $\mathscr{C}$. Unlike in the covariant case, there is no natural way to construct an induced function $h^A(f) = \hom(f,A) : \hom_{\mathscr{C}}(X,A) \to \hom_{\mathscr{C}}(Y,A)$, but we can easily construct one in the opposite direction by mapping every morphism $g \in \hom_{\mathscr{C}}(Y,A)$ to its precomposition by f to obtain $g \circ f \in \hom_{\mathscr{C}}(X,A)$. Thus, this construction sends objects in $\mathscr{C}$ to **Set** while reversing all morphisms, defining the *contravariant* hom-functor $h^A : \mathscr{C}^{\text{op}} \to \textbf{Set}$.

△

Example (Hom-Bifunctor). The notation $\hom(A,-)$ and $\hom(-,B)$ in the previous examples suggests that there might be a functor $\hom(-,-)$ that sends (ordered) pairs of objects to the hom-set between them. In order for this construction to be functorial, we also need to map the pairs of morphisms between these pairs of objects to some function between the hom-sets. That is, if $f : X \to Y$ and $h : B \to A$ are morphisms, then there should be an induced function $\hom(h,f) : \hom(A,X) \to \hom(B,Y)$, noting that the first argument is reversed due to contravariance. By alternatively fixing each component of the functor, we can construct the following square:

$$\begin{tikzcd}
\hom(A,X) \arrow[r, "\hom(h{,}X)"] \arrow[d, "\hom(A{,}f)"'] & \hom(B,X) \arrow[d, "\hom(B{,}f)"] \\
\hom(A,Y) \arrow[r, "\hom(h{,}Y)"'] & \hom(B,Y)
\end{tikzcd}$$

We will follow how a morphism $g \in \hom(A,X)$ is mapped under this square along the two different paths, in a technique called *diagram chasing.* The vertical arrows are the covariant hom-functors that precompose their inputs by f, and the horizontal arrows are the contravariant hom-functors that postcompose their inputs by h.

So, along the upper path, we have $g \mapsto g \circ h \mapsto f \circ (g \circ h)$, and along the lower path, we have $g \mapsto f \circ g \mapsto (f \circ g) \circ h$. But by the associativity of composition, these paths are equal, so the diagram commutes for all f, g, and h; there is a unique morphism from $\hom(A,X)$ to $\hom(B,Y)$ induced by h and f:

$$\begin{tikzcd}
\hom(A,X) \arrow[r, "\hom(h,X)"] \arrow[d, "\hom(A,f)"'] \arrow[dr, dashed, "\hom(h,f)" description] & \hom(B,X) \arrow[d, "\hom(B,f)"] \\
\hom(A,Y) \arrow[r, "\hom(h,Y)"'] & \hom(B,Y)
\end{tikzcd}$$

which is exactly the statement that the function $\hom(h,f)$ is well-defined. So, $\hom(-,-)$ is indeed a valid functor $\mathscr{C}^{\mathrm{op}} \times \mathscr{C} \to \mathbf{Set}$. Because this functor takes objects and morphisms from a product category, it is also called a *bifunctor*. The ordinary covariant and contravariant hom-functors are then just the partial applications of this bifunctor in the first and second arguments, respectively. △

Theorem 52.2.6. *Functors preserve commutative diagrams.*

Proof. [Kit22] Because functors preserve compositions, for any paths $a_1 \circ a_2 \circ \cdots \circ a_n$ and $b_1 \circ b_2 \circ \cdots \circ b_m$ connecting a pair of objects in a commutative diagram, we have,

$$\begin{aligned} F(a_1) \circ F(a_2) \circ \cdots \circ F(a_n) &= F(a_1 \circ a_2 \circ \cdots \circ a_n) \\ &= F(b_1 \circ b_2 \circ \cdots \circ b_m) \\ &= F(b_1) \circ F(b_2) \circ \cdots \circ F(b_m) \end{aligned}$$

so the image of the two paths are also equal, so the image of the diagram remains commutative. ■

In particular, this implies that isomorphism diagrams are also preserved, so

- If f is an isomorphism, then $F(f)$ is also an isomorphism;
- If $A \cong B$ are isomorphic objects, then $F(A) \cong F(B)$.

A functor that satisfies the converse of the first statement is said to *reflect* isomorphisms, and a functor that satisfies the converse of the second is said to *create* isomorphisms [Rie17].

Theorem 52.2.7. *Fully faithful functors (i) reflect and (ii) create isomorphisms.*

That is, for a fully faithful functor $F : \mathscr{C} \to \mathscr{D}$,

(i) If a morphism f in $\mathscr{C}$ is such that $F(f)$ is an isomorphism in $\mathscr{D}$, then f is an isomorphism;

(ii) If a pair of objects X and Y in $\mathscr{C}$ are such that $F(X) \cong F(Y)$, then $X \cong Y$.

Proof. Suppose $f : A \to B$ is a morphism such that $F(f) : F(A) \to F(B)$ is an isomorphism with inverse $\tilde{g} : F(B) \to F(A)$. As F is full, there exists a morphism $g : B \to A$ such that $F(g) = \tilde{g}$, so $F(g \circ f) = F(g) \circ F(f) = \tilde{g} \circ F(f) = \mathrm{id}_{F(A)} = F(\mathrm{id}_A)$, so by faithfulness, $g \circ f = \mathrm{id}_A$. Exchanging f and g in the previous yields $f \circ g = \mathrm{id}_B$, so f is an isomorphism.

Suppose $F(A) \cong F(B)$, so there is an isomorphism $\tilde{f} : F(A) \to F(B)$ with inverse $\tilde{g}$. As F is full, there exist morphisms $f : A \to B$ and $g : B \to A$ such that $F(f) = \tilde{f}$ and $F(g) = \tilde{g}$. Then, $F(g \circ f) = F(g) \circ F(f) = \tilde{g} \circ \tilde{f} = \mathrm{id}_{F(A)} = F(\mathrm{id}_A)$, so by faithfulness, $g \circ f = \mathrm{id}_A$. Again, exchanging f and g in the previous yields $f \circ g = \mathrm{id}_B$, so f and g are isomorphisms and hence $A \cong B$. ■

Note that the converse of this theorem does not hold in that functors that create or reflect isomorphisms are not necessarily full or faithful.

These two conditions may seem similar, but they do not imply each other in general. For instance, let $\mathscr{C}$ be a category in which every object is isomorphic, but there exist non-isomorphism morphisms, e.g. the category $\mathbf{Set}_n$ of sets of cardinality n. Because every object is isomorphic, any functor $F : \mathscr{C} \to \mathscr{D}$ to any category $\mathscr{D}$ trivially creates isomorphisms, but will not, in general, reflect isomorphisms.

52.2.5 Natural Transformations

Given categories and functors $\mathscr{C} \underset{G}{\overset{F}{\rightrightarrows}} \mathscr{D}$, a *natural transformation* is a mapping $\begin{tikzcd} \mathscr{C} \arrow[r, bend left, "F"{name=U}] \arrow[r, bend right, "G"'{name=D}] & \mathscr{D} \arrow[Rightarrow, from=U, to=D, "\eta"] \end{tikzcd}$ or $\eta : F \Rightarrow G$ between functors.

The functors F and G map objects and morphisms in $\mathscr{C}$ to objects and morphisms in $\mathscr{D}$, so to define a mapping $F \Rightarrow G$, we want to associate the images of objects and morphisms under F to their images under G. For objects, this just means that if X is in $\mathscr{C}$, then $F(X)$ should be associated with $G(X)$ – this is just a morphism in $\hom_{\mathscr{D}}(F(X),G(X))$. So, the natural transformation η associates each object $X \in \mathrm{ob}(\mathscr{C})$ to a morphism $h_X : F(X) \to G(X)$ called the *component* of η at X.

$$\begin{tikzcd}
 & & F(X) \arrow[dd, "\eta_X"] & \\
\mathscr{C} & X \arrow[ur, dashed, "F"] \arrow[dr, dashed, "G"'] \arrow[r, Rightarrow, "\eta"] & {} & \mathscr{D} \\
 & & G(X) &
\end{tikzcd}$$

[Kit22]

However, there could be many morphisms $F(X) \to G(X)$ we could choose. We need a way of selecting these components that is consistent throughout the whole category.

Consider a morphism $f : A \to B$ in $\mathscr{C}$. Under F and G, we have the images $F(f) : F(A) \to F(B)$ and $G(f) : G(A) \to G(B)$. Along with the components $\eta_A : F(A) \to G(A)$ and $\eta_B : F(B) \to G(B)$, this completes the square

$$\begin{tikzcd}
 & A \arrow[r, dashed, "F"] \arrow[dd, "f"] \arrow[rr, dashed, bend left=50, "G"] \arrow[rr, Rightarrow, bend left=20, "\eta"] & F(A) \arrow[r, "\eta_A"] \arrow[dd, "F(f)"] \arrow[ddr, dashed, "(\exists!)?"] & G(A) \arrow[dd, "G(f)"] & \\
\mathscr{C} & & & & \mathscr{D} \\
 & B \arrow[r, dashed, "F"] \arrow[rr, dashed, bend right=50, "G"'] \arrow[rr, Rightarrow, bend right=20, "\eta"'] & F(B) \arrow[r, "\eta_B"] & G(B) &
\end{tikzcd}$$

[Kit22]

In this diagram, there are two paths from $F(A)$ to $G(B)$, namely, $\eta_B \circ F(f)$, and $G(f) \circ \eta_A$, and without any further conditions on the components of η, these paths may be distinct. However, if we require that these paths are equal – that the diagram commutes – then this forces our selection of components to be consistent throughout the whole category. This coherency condition is called the *naturality* requirement.

So, overall, a natural transformation a natural transformation $\eta : F \Rightarrow G$ between functors F,$G : \mathscr{C} \to \mathscr{D}$ is a collection of morphisms $\left(F(X) \xrightarrow{\eta_X} G(X)\right)_{X \in \mathrm{ob}(\mathscr{C})}$ indexed by the objects of $\mathscr{C}$ such that the following diagram commutes:

$$\begin{tikzcd}
A \arrow[d, "f"] & F(A) \arrow[r, "\eta_A"] \arrow[d, "F(f)"] & G(A) \arrow[d, "G(f)"] \\
B & F(B) \arrow[r, "\eta_B"] & G(B)
\end{tikzcd}$$

That is, $\eta_B \circ F(f) = G(f) \circ \eta_A$ for all $f : A \to B$ in $\hom(\mathscr{C})$.

Natural transformations are collections of morphisms between the images of two functors that are canonical or consistent in some way in that they preserve some of the functoriality. Informally, a construction involving a collection of mappings between objects is said to be “natural” if those mappings can be extended to some natural transformation over the whole category, and “unnatural” otherwise. Often, unnatural constructions depend on some arbitrary choice e.g. of basis, generator, relations, etc. while natural constructions are independent of these choices.

Consider the following diagram of categories, functors, and natural transformations:

$$\begin{tikzcd}
\mathscr{C} \arrow[r, bend left=60, "F"{name=F}] \arrow[r, "G"{description, name=G}] \arrow[r, bend right=60, "H"'{name=H}] & \mathscr{D}
\arrow[Rightarrow, from=F, to=G, "\alpha"] \arrow[Rightarrow, from=G, to=H, "\beta"]
\end{tikzcd}$$

From the diagram, it would seem that we should be able to compose α and β to obtain a natural transformation $\beta \circ \alpha : F \Rightarrow H$. Such a composition is called a *vertical composition.*[*]

Consider an object X in $\mathscr{C}$. The components of α and β at X are the morphisms $\alpha_X : F(X) \to G(X)$ and $\beta_X : G(X) \to H(X)$ – these morphisms are compatible in that we can compose them, so we can define the component $(\beta \circ \alpha)_X$ to be $\beta_X \circ \alpha_X : F(X) \to H(X)$. For naturality, consider the following diagram:

$$\begin{tikzcd}
A \arrow[d, "f"] & F(A) \arrow[r, "\alpha_A"] \arrow[d, "F(f)"] & G(A) \arrow[r, "\beta_A"] \arrow[d, "G(f)"] & H(A) \arrow[d, "H(f)"] \\
B & F(B) \arrow[r, "\alpha_B"] & G(B) \arrow[r, "\beta_B"] & H(B)
\end{tikzcd}$$

α and β are natural transformations, so each square individually commutes, and hence the outer square also commutes, so $\beta \circ \alpha$ is natural, as required.

The collection of natural transformations between functors between two fixed categories $\mathscr{C}$ and $\mathscr{D}$, under vertical composition, forms the *functor category* $[\mathscr{C},\mathscr{D}]$ (also written as $\mathscr{D}^{\mathscr{C}}$) that has functors from $\mathscr{C}$ to $\mathscr{D}$ as objects, natural transformations as morphisms, and vertical composition as composition. Identities in this category are given by identity natural transformations that associate every object with the identity morphism on their image.

An isomorphism in a functor category is then called a *natural isomorphism.* That is, if $\eta : F \Rightarrow G$ and $\vartheta : G \Rightarrow F$ are natural transformations such that $\vartheta \circ \eta = \mathrm{id}_F$ and $\eta \circ \vartheta = \mathrm{id}_G$, then η and ϑ are natural isomorphisms, and we write η^{-1} for ϑ. If two functors F and G are naturally isomorphic, then we write $F \cong G$, or, we say that $F(X) \cong G(X)$ *naturally in* X whenever we need to bind a variable.

Note that the statement “$F(X) \cong G(X)$” is a statement about the objects $F(X)$ and $G(X)$ in $\mathscr{D}$, while “$F(X) \cong G(X)$ naturally in X” is a much stronger statement about the functors F and G in $[\mathscr{C},\mathscr{D}]$. In particular, $F(X) \cong G(X)$ naturally in X not only requires that there are isomorphisms $F(X) \cong G(X)$ for every X, but also that these individual isomorphisms can be selected in some consistent way such that all naturality diagrams commute; it is entirely possible for $F(X) \cong G(X)$ to hold for all X, but for no such selection of isomorphisms to exist and for $F \ncong G$.

Example (Dual Vector Spaces). In linear algebra, the *dual* V^* of a vector space V over a field K is the vector space of linear functionals $V \to K$ equipped with pointwise addition and scalar multiplication.

It is well known that any finite-dimensional vector space V is isomorphic to its dual V^*, and also to the dual of its dual, or the *double dual*, $V^{**} = (V^*)^*$, the space of linear functionals $V^* \to K$. These

[*] There is also a related notion of *horizontal composition* that combines natural transformations

$$\begin{tikzcd}
\mathscr{C} \arrow[r, bend left=50, "F"{name=F}] \arrow[r, bend right=50, "G"'{name=G}] & \mathscr{D} \arrow[r, bend left=50, "F'"{name=F2}] \arrow[r, bend right=50, "G'"'{name=G2}] & \mathscr{E}
\arrow[Rightarrow, from=F, to=G, "\alpha"] \arrow[Rightarrow, from=F2, to=G2, "\beta"]
\end{tikzcd}$$

into a natural transformation $\alpha \diamond \beta : F' \circ F \Rightarrow G' \circ G$ that we will not use.

isomorphisms follow from a standard construction that, given a basis of V, yields a *dual basis* of V^* of the same dimension. However, in many ways, V^{**} has a lot more in common with V than the dual V^* does, and we can make this idea precise by showing that the collection of isomorphisms $V \cong V^{**}$ is natural in the sense that it can be extended to a natural transformation, while the isomorphisms $V \cong V^*$ cannot.

There is a contravariant functor $(-)^* : \mathbf{Vect}^{\mathrm{op}} \to \mathbf{Vect}$ that sends vector spaces V to their dual V^*, and linear maps $f : U \to V$ to their transpose $f^* : V^* \to U^*$, defined by precomposing linear functionals $\omega \in V^*$ by f to obtain $\omega \circ f \in U^*$. Applying this functor again yields the covariant double dual functor $(-)^{**} : \mathbf{Vect}_K \to \mathbf{Vect}_K$ that maps vector spaces V to their double dual V^{**}. Note that the elements of V^* are themselves functions $V \to K$, so the elements of V^{**} are functionals that send functions $V \to K$ to elements in K.

One obvious way to map these functionals to elements is just to supply the functions in V^* with an input $v \in V$ – that is, for any $\omega \in V^*$, we have $\omega(v) \in K$ by definition, so the evaluation mapping $\omega \mapsto \omega(v)$ is an element of V^{**}. So, for each vector space V, we have a linear map $\eta_V : V \to V^{**}$ that sends vectors to their associated evaluation mappings [Per21]. We show that these maps are natural in the formal sense.

Let $f : V \to W$ be a linear transformation, and consider the following diagram:

$$\begin{tikzcd}
V \arrow[r, "\eta_V"] \arrow[d, "f"] & V^{**} \arrow[d, "f^{**}"] \\
W \arrow[r, "\eta_W"] & W^{**}
\end{tikzcd}$$

Let $v \in V$ and $\omega \in W^*$. Along the lower path, we have,

$$\big[\eta_W\big(f(v)\big)\big](\omega) = \omega\big(f(v)\big)$$

and along the upper path, we have,

$$\begin{aligned}
\big[f^{**}\big(\eta_V(v)\big)\big](\omega) &= \big[\eta_V(v) \circ f^*\big](\omega) \\
&= \big[\eta_V(v)\big]\big(f^*(\omega)\big) \\
&= \big[\eta_V(v)\big](\omega \circ f) \\
&= (\omega \circ f)(v) \\
&= \omega\big(f(v)\big)
\end{aligned}$$

so the diagram commutes. If we view the objects on the left side of the commutative square as the images of objects under the identity functor, then we see that every map η_V is a morphism $\mathrm{id}_{\mathbf{Vect}_K}(V) \to (V)^{**}$, so the entire collection $(\eta_V)_{V \in \mathrm{ob}(\mathbf{Vect}_K)}$ defines a natural transformation $\mathrm{id}_{\mathbf{Vect}_K} \Rightarrow (-)^{**}$, with the above diagram verifying naturality.

For finite-dimensional spaces, η is furthermore a isomorphism; if we restrict our attention to the finite-dimensional case in the subcategory $\mathbf{FinVect}_K$, then η defines a natural isomorphism $\mathrm{id}_{\mathbf{Vect}_K} \cong (-)^{**}$. In contrast, the dual V^* is *not* naturally isomorphic to V, even in finite dimensions, simply because the single dual functor $(-)^*$ is contravariant and lives in a different functor category than the identity. Even if we extend the notion of naturality to cover contravariance, there is a deeper reason why the single dual is not natural or “canonical”, unlike η:

Consider the case where f is an endomorphism, interpreted as a *change of basis*:

$$\begin{tikzcd}
V \arrow[r, "\eta_V"] \arrow[d, "f"] & V^{**} \arrow[d, "f^{**}"] \\
V \arrow[r, "\eta_V"] & V^{**}
\end{tikzcd}$$

What this diagram is saying is that if we change the basis of V, and also the basis of V^{**} with the induced function f^{**}, then the map η is completely unaffected: that is, η does not depend on the choice of basis [Per21]. In contrast, every isomorphism $V \to V^*$ that can be constructed (even in finite dimensions) will depend on some choice of basis of V, and moreover, changing the basis of V does not change the basis of V^* in a way that is compatible with these isomorphisms. △

Natural transformations can also be composed with functors, in a sense. Let F,G : $\mathscr{C} \to \mathscr{D}$ and H : $\mathscr{D} \to \mathscr{E}$ be functors, and $\eta : F \Rightarrow G$ be a natural transformation. The *whiskering* $H \cdot \eta$ of H by η is the natural transformation $H \circ F \Rightarrow H \circ G$ defined by $(H \cdot \eta)_X = H(\eta_X)$.

$$\begin{tikzcd}
\mathscr{C} \arrow[r, bend left, "F"{name=U}] \arrow[r, bend right, "G"'{name=D}] & \mathscr{D} \arrow[r, "H"] & \mathscr{X}
\arrow[Rightarrow, from=U, to=D, "\eta"]
\end{tikzcd}
\qquad
\begin{tikzcd}
\mathscr{C} \arrow[r, bend left, "H\circ F"{name=U}] \arrow[r, bend right, "H\circ G"'{name=D}] & \mathscr{X}
\arrow[Rightarrow, from=U, to=D, "H\cdot\eta"]
\end{tikzcd}$$

(This is a special case of the horizontal composition where one of the natural transformations is the identity natural transformation, so $H \cdot \eta = \eta \diamond \mathrm{id}_H$.)

52.2.6 Equivalence of Categories

We have seen numerous examples of structures and structure-preserving maps forming categories, and the same holds true for categories and functors themselves: the collection of small categories and functors forms a large category, **Cat**.[*]

If there is an isomorphism between two categories $\mathscr{C}$ and $\mathscr{D}$ in **Cat**, then $\mathscr{C}$ and $\mathscr{D}$ are *isomorphic categories* – categories which differ only in the labelling of their objects and morphisms.

$$\begin{tikzcd}
\mathscr{C} \arrow[r, shift left, "F"] & \mathscr{D} \arrow[l, shift left, "G"]
\end{tikzcd}$$

$$G \circ F = \mathrm{id}_{\mathscr{C}} \qquad \text{and} \qquad F \circ G = \mathrm{id}_{\mathscr{D}}$$

Again, we write $\mathscr{C} \cong \mathscr{D}$ if there exists an isomorphism between $\mathscr{C}$ and $\mathscr{D}$. Like with other isomorphic objects, results about one immediately gives identical results about the other, but unfortunately, isomorphism of categories tends to be too strong of a requirement in that very few useful categories are isomorphic to each other.

However, as mentioned earlier, we care about different degrees of likeness for different types of data. For elements of a set, this notion is strict equality, while for objects in a category, this notion is isomorphism. Applying this idea to a functor category, we see that natural isomorphism is the appropriate degree of likeness for functors.

Now, looking back at the definition of isomorphic categories above, we notice that the compositions of F and G are required to be *equal* to the identity functors, but as we have just seen, this degree of likeness is unreasonably strict. For functors, we really should be using natural isomorphisms, and indeed, if we

[*] There is not a simple category of all categories for the same reason that there is no set of all sets, but given a choice of Grothendieck universe, a similar category can be constructed, and is denoted **CAT**.

only require the compositions to be naturally isomorphic to the identity functor, we obtain a weaker but much more useful notion of likeness called *equivalence of categories.*

$$\begin{tikzcd}\mathscr{C} \arrow[r, shift left, "F"] & \mathscr{D} \arrow[l, shift left, "G"]\end{tikzcd}$$

$$G \circ F \cong \mathrm{id}_{\mathscr{C}} \qquad \text{and} \qquad F \circ G \cong \mathrm{id}_{\mathscr{D}}$$

In this case, we say that $\mathscr{C}$ and $\mathscr{D}$ are *equivalent*, and we write $\mathscr{C} \simeq \mathscr{D}$. We also call the functors F and G *equivalences.*

Example. Consider the category $\mathbf{FinVect}_k$ of finite-dimensional vector spaces over a field k, and the category $\mathbf{Mat}(k)$, where morphisms are matrices with entries in k, and objects are the dimensions of those matrices (so, if n and m are objects, then $\hom(n,m)$ is the collection of $m \times n$ matrices). That $\mathbf{FinVect}_k$ and $\mathbf{Mat}(k)$ are not isomorphic categories can be deduced by observing that there are no isomorphic objects in $\mathbf{Mat}(k)$ – but, these categories are clearly related in some way as matrices are well known to represent linear transformations, and indeed, these categories are not isomorphic, but equivalent, with the equivalence $\mathbf{FinVect}_k \to \mathbf{Mat}_k$ sending each vector space to its dimension, and each linear transformation to its corresponding matrix. Each choice of basis for each vector space provides a different equivalence, also demonstrating that an equivalence is not unique. △

Example. Up to isomorphism, every thin category is a preordered set, but up to equivalence, every thin category is a partially ordered set. △

52.2.7 The Yoneda Lemma

Let $1 = \{\bullet\}$ be the set with one element. For any set X, a function $1 \to X$ amounts to selecting an element of X since the only data required to uniquely characterise such a function is just the image of $\bullet$ in X. Similarly, in any space X, the functions $1 \to X$ (where 1 is the one-point space) are essentially just points in X. In fact, in arbitrary categories with terminal objects 1, we will *define* a map $1 \to X$ to be an *element* of X.

We can extend this idea of "objects as functions" by picking different choices of domain spaces. For instance, the functions $[0,1] \to X$ are just paths in X; the functions $\mathbb{N} \to X$ are the sequences in X; and the functions $S^1 \to X$ are the topological loops in X. In some of these cases, these are even the *definitions* of these constructions.

More generally, given an object A in a category $\mathscr{C}$, a *generalised element* of A is any morphism with codomain A, and the domain of such a morphism is called the *domain of variation* of the element [Kos12]. We also alternatively call a morphism $S \to A$ a generalised element of A with *shape* S [Lei14]. Note that there really isn't any mathematical difference between a "generalised element" and a morphism, but the change in naming represents a change in perspective, as above.

Note, however, that this is more than just a semantic trick; this concept of treating objects as special cases of functions – or more generally, of morphisms – is a fundamental idea in category and structural set theory, and arguably the most important result in category theory [Rie17], which we display below, expands on this idea.

Now, we can exhibit basic objects like points or paths as certain types of maps, but how can we apply this idea to arbitrary objects X? That is, we have captured certain basic features within X with the maps $1 \to X$, $[0,1] \to X$, $S^1 \to X$, etc., but is it possible to characterise the entirety of X itself using these maps? A priori, there is no reason we should expect that the entire structure of X is contained within these maps. This is the content of the next result.

Lemma 52.2.8 (Yoneda)**.** *Let $\mathscr{C}$ be a locally small category. Then,*

$$\hom_{[\mathscr{C},\mathbf{Set}]}(h_A,F) \cong F(A)$$

naturally in $F \in \mathrm{ob}([\mathscr{C},\mathbf{Set}])$ and $A \in \mathrm{ob}(\mathscr{C})$.

Proof. [Kit22, abridged] We give a proof of the isomorphism only.

Let $f : A \to B$ be a morphism, $\eta : h_A \Rightarrow F$ be a natural transformation, and consider the naturality diagram of η:

$$\begin{tikzcd}
h_A(A) \arrow[r, "\eta_A"] \arrow[d, "h_A(f)"'] & F(A) \arrow[d, "F(f)"] \\
h_A(B) \arrow[r, "\eta_B"'] & F(B)
\end{tikzcd}$$

We chase the identity $\mathrm{id}_A \in \hom(A,A) = h_A(A)$ around the diagram:

$$\begin{tikzcd}
\mathrm{id}_A \arrow[r, mapsto, "\eta_A"] \arrow[d, mapsto, "h_A(f)"'] & \eta_A(\mathrm{id}_A) \arrow[d, mapsto, "F(f)"] \\
f \arrow[r, mapsto, "\eta_B"'] & \eta_B(f) = F(f)\big(\eta_A(\mathrm{id}_A)\big)
\end{tikzcd}$$

The input to the function on the right side is always $\eta_A(\mathrm{id}_A)$, implying that any natural transformation $h_A \Rightarrow F$ is completely determined by its value at id_A. This naturally induces a function $\phi : \hom_{[\mathscr{C},\mathbf{Set}]}(h_A,F) \to F(A)$ defined by $\eta \mapsto \eta_A(\mathrm{id}_A)$. Conversely, given an element $u \in F(A)$, we can define the components of a unique natural transformation $\eta : h_A \Rightarrow F$ by $\eta_B(f) = F(f)(u)$, where $f : A \to B$, defining a mapping $\psi : F(A) \to \hom_{[\mathscr{C},\mathbf{Set}]}(h_A,F)$. Then, $\psi\big(\phi(\eta)\big) = \psi\big(\eta_A(\mathrm{id}_A)\big) = \eta$, and $\phi\big(\psi(u)\big) = \phi(\eta) = \eta_A(\mathrm{id}_A) = u$, so the functions are mutually inverse, hence defining an isomorphism $F(A) \cong \hom_{[\mathscr{C},\mathbf{Set}]}(h_A,F)$.

For a proof of the naturality of this isomorphism, see [Kit22]. ■

If we take F to be another hom-functor in the Yoneda lemma, we obtain:

$$\hom_{[\mathscr{C},\mathbf{Set}]}(h_A,h_B) \cong h_B(A) = \hom_{\mathscr{C}}(B,A)$$

so the natural transformations $h_A \Rightarrow h_B$ are in bijection with the morphisms $B \to A$. This assignment of natural transformations from morphisms is the action of the contravariant functor $h_\bullet : \mathscr{C}^{\mathrm{op}} \to [\mathscr{C},\mathbf{Set}]$ defined on objects by $h_\bullet(A) = h_A$. In fact, we have already seen this functor – $h_\bullet$ is exactly the partial application of the hom-bifunctor $\hom(-,-)$ in the first argument (e.g., $h_\bullet(A) = \hom(-,-)(A,-) = \hom(A,-) = h_A$). Dually, we can also partially apply the hom-bifunctor in the second argument to obtain the covariant functor $h^\bullet : \mathscr{C} \to [\mathscr{C}^{\mathrm{op}},\mathbf{Set}]$.

The Yoneda lemma then says that the functors $h_\bullet$ and $h^\bullet$ give embeddings of $\mathscr{C}^{\mathrm{op}}$ and $\mathscr{C}$ into $[\mathscr{C},\mathbf{Set}]$ and $[\mathscr{C}^{\mathrm{op}},\mathbf{Set}]$, respectively – there is a copy of every (locally small) category contained within the collection of functors between its dual and **Set**.

These functors are called the covariant and contravariant *Yoneda embeddings*, and are often collectively denoted よ, with context disambiguating between the two.

Theorem 52.2.9 (Yoneda Embedding). *Let $\mathscr{C}$ be a locally small category. Then, the Yoneda embeddings $\text{よ} : \mathscr{C} \hookrightarrow [\mathscr{C}^{\mathrm{op}},\mathbf{Set}]$ and $\text{よ} : \mathscr{C}^{\mathrm{op}} \hookrightarrow [\mathscr{C},\mathbf{Set}]$ are embeddings – that is,* よ *is fully faithful, and injective on objects up to isomorphism.*

Proof. See [Kit22]. ■

From functoriality, the Yoneda embeddings imply that if $X \cong Y$, then $\hom(X,-) \cong \hom(Y,-)$ and $\hom(-,X) \cong \hom(-,Y)$. More notably, full faithfulness also implies the converse – that is, the Yoneda embeddings create isomorphisms:

Corollary 52.2.9.1. *If* $\hom(X,-) \cong \hom(Y,-)$ *or* $\hom(-,X) \cong \hom(-,Y)$*, then* $X \cong Y$*.*

That is, the maps in to and maps out from an object contain exactly as much information as the object itself; the collections of these maps are isomorphic if and only if the associated objects are, so objects are completely characterised by their generalised elements.

52.3 Universal Properties

52.3.1 Terminal and Initial Objects

An object $T \in \text{ob}(\mathscr{C})$ is *terminal* if for every object $X \in \text{ob}(\mathscr{C})$ there exists a unique morphism $X \to T$. Dually, an object $I \in \text{ob}(\mathscr{C})$ is *initial* if for every object $X \in \text{ob}(\mathscr{C})$, there exists a unique morphism $I \to X$ (or equivalently, if it is terminal in $\mathscr{C}^{\text{op}}$). Terminal and initial objects are also sometimes collectively called *universal* objects, with context disambiguating between the two cases.

In **Set**, any singleton set $1 = \{\bullet\}$ is terminal as for any set X, a function $X \to 1$ exists, defined by mapping every element of X to $\bullet$, and is unique as there are is only one possible target for every input. Conversely, the empty set is initial as for any set X, the empty function $\varnothing \to X$ exists and is unique.

In many categories of structured sets such as **Set**, **Top**, **Ring**, and **Grp**, terminal objects are often singleton sets, so terminal objects are often denoted by 1. Initial objects are slightly less well-behaved, but are often the empty set, as is in **Set** or **Top**,* and are often denoted by 0.

Terminal and initial objects may not necessarily exist; for instance, in the preorder category $(\mathbb{N}, \leq)$, 0 is initial, but no terminal object exists; and in $(\mathbb{Z}, \leq)$, terminal and initial objects both fail to exist. However, if initial or terminal objects *do* exist, then they are unique up to unique isomorphism – that is, if C and C' are distinct and terminal (initial), then there is a unique isomorphism $C \to C'$ – and we say that they are *essentially unique.*

Theorem 52.3.1. *Terminal (initial) objects are essentially unique.*

Proof. Suppose C and C' are distinct and terminal. Because C is terminal, the morphisms $f : C' \to C$ and $\text{id}_C : C \to C$ are unique, and because C' is terminal, the morphisms $g : C \to C'$ and $\text{id}_{C'} : C' \to C'$ are unique. Then, the compositions $g \circ f$ and $f \circ g$ must be identities, so they form a unique isomorphism. The essential uniqueness of initial objects follows from duality. ■

Terminal and initial objects can, however, coincide. For instance, in **Grp**, the trivial group is both terminal and initial. In these cases, the object is called a *null* or *zero* object.

52.3.2 Representability

By the Yoneda lemma, objects are determined entirely by the maps into or out from them; informally, a *universal property* is a description of these maps.

Terminal and initial objects are examples of objects characterised by universal properties – in this case, the universal property that the maps to or from them exist uniquely – and in fact, we can reverse this somewhat by saying that an object has a universal property, or is *universal*, if we can find some category in which it is initial or terminal.

Universal properties are a way of describing the “best”, “largest”, “smallest”, “most _____” etc. object in a category – or in the case of terminal and initial objects, the “final” and “first” objects – and the exact notion of “best” will define different types of objects.

* Rings and groups require identities, so the empty set is not in **Ring** or **Grp**. Instead, the initial object in **Ring** is $\mathbb{Z}$; and in **Grp**, the trivial group.

Through almost the exact same reasoning as for terminal and initial objects, objects characterised by universal properties are essentially unique (though we will prove this more formally soon). This also provides another way of proving that a collection of objects are isomorphic; just find a universal property they satisfy in common.

Once a universal property of some object has been identified, we often then ignore the specifics of how it was constructed in the first place, as the universal property alone is sufficient information to recover the object essentially uniquely. This allows us to more easily work with constructions that have unwieldy definitions but simple universal properties.

Now, the maps into or out from an object A are captured by the hom-functors h_A and h^A (or more concisely, by the Yoneda functors $よ(A)$), so a unversal property is a description of these functors. We formally give this description in terms of an isomorphism.

A covariant or contravariant functor F from a locally small category $\mathscr{C}$ to **Set** is *representable* if $F \cong よ(A)$ for some $A \in \text{ob}(\mathscr{C})$ (where the Yoneda embedding necessarily matches the variance of F). The object A, along with the natural transformation $F \Rightarrow よ(A)$ is then called a *representation* of F. Representable functors then carry information about the hom-functors they are isomorphic to, hence encoding a universal property about their representing object.

We rephrase the universal property of initial objects using representations as an example. Let $\Delta 1 : \mathscr{C} \to \mathbf{Set}$ be the *constant functor* that sends every object in $\mathscr{C}$ to a fixed singleton set 1, and every morphism to the identity function id_1. Then, an object A is initial if and only if $h_A \cong \Delta 1$. The natural isomorphism requires that $\text{hom}(A,B)$ contains exactly a single morphism for every object B, which is precisely our previous definition of an object being initial. Equivalently, we could say that a category $\mathscr{C}$ has an initial object if and only if the constant functor $\Delta 1 : \mathscr{C} \to \mathbf{Set}$ is representable.

Theorem 52.3.2. [Rie17] *Let X and Y be objects of a locally small category $\mathscr{C}$. If either the covariant or contravariant functors represented by X and Y are naturally isomorphic, then X and Y are isomorphic.*

Proof. The Yoneda embeddings are fully faithful, and hence create isomorphisms. It follows that an isomorphism between represented functors must be induced by a unique isomorphism between the representing objects. ■

In particular, if X and Y represent the same functor, then X and Y are isomorphic; and moreover, this isomorphism is unique. That is to say, the object representing a functor is essentially unique, thus extending the claim from terminal and inital objects to all objects characterised by universal properties.

We will now explore some of the many important constructions that can be characterised in this way.

52.3.3 Products

Let X and Y be objects in a category $\mathscr{C}$. We should intuitively expect that any notion of a *product* should consist of another object $P \in \text{ob}(\mathscr{C})$ that is related to X and Y; that is, an object equipped with a pair of morphisms, called *projections*, to X and Y:

$$\begin{tikzcd}
& P \arrow[dl, "\pi_1"'] \arrow[dr, "\pi_2"] & \\
X & & Y
\end{tikzcd}$$

This resulting diagram shape of one object pointing to two others is called a *span*.

But not just any object with maps to X and Y can be the product – we need the product to be universal; or, the "best" one possible. The notion of "best" here is that for all other spans

$$\begin{tikzcd}
 & A \arrow[dl, "g"'] \arrow[dr, "h"] & \\
X & & Y
\end{tikzcd}$$

we require for there to exist a unique *factorisation* through P. That is, there exists a unique morphism $f : A \to P$ such that the following diagram commutes:

$$\begin{tikzcd}
 & A \arrow[d, dashed, "\exists! f"] \arrow[ddl, "g"'] \arrow[ddr, "h"] & \\
 & P \arrow[dl, "\pi_1"] \arrow[dr, "\pi_2"'] & \\
X & & Y
\end{tikzcd}$$

and we say that A *factors through* P. We then write $X \times Y$ for the product object P, and $\langle g,h \rangle$ for the unique morphism f, and we call g and h the *components* of the *pairing* $\langle g,h \rangle$. This latter notation is justified as f is uniquely determined by g and h, as prescribed by the universal property. Conversely, given any map into a product, we can reconstruct its components by postcomposing it by each projection map.

There are notable similarities to terminal objects with this definition, in that we require a unique morphism to exist; P appears to be "terminal" with respect to other objects that have maps to X and Y. (We formalise this observation later §52.4.3, with the notion of a category of cones.)

Just like with terminal objects, products do not always exist in any given category. For instance, for any pair of distinct objects in a discrete category, there is no way to form a span connecting the two objects, as every morphism in a discrete category is an identity, so these products do not exist. However, if the product *does* exist, then it is essentially unique due to the universal property.

Example (Products in **Set**). The Cartesian product $X \times Y$ is a categorical product.

The Cartesian product $X \times Y$ comes naturally equipped with two projection mappings to the component sets defined by $(x,y) \mapsto x$ and $(x,y) \mapsto y$. Now, suppose there is another set A with maps $X \xleftarrow{g} A \xrightarrow{h} Y$. Let $a \in A$, so $g(a) \in X$ and $h(a) \in Y$. To make the diagram commute, the obvious – and only – choice for $f : A \to X \times Y$ is to have $f(a) = \big(g(a),h(a)\big)$.

For uniqueness, suppose there exists another another mapping $f' : A \to X \times Y$ such that the product diagram commutes, and let $f'(a) = (x,y)$. Then, by commutativity, $g(a) = \big(\pi_1 \circ f'\big)(a) = \pi_1(x,y) = x$, and similarly, $h(a) = \big(\pi_2 \circ f'\big)(a) = \pi_2(x,y) = y$, so $f'(a) = \big(g(a),h(a)\big) = f(a)$, and the factorisation is unique. △

Note that this characterisation of the Cartesian product as a categorical product makes no mention of the elements of $A \times B$ at all (or any other set), depending entirely on mappings between sets, and specifically on how the product *interacts* with other sets; thus formulating the Cartesian product structurally.

Example (Products in **Top**). If we take A to be the one-point space $\mathbf{1}$, then the maps g, h, and f are just elements of X, $P = X \times Y$, and Y, respectively. The product then says that there is a bijection

$$\hom_{\mathbf{Top}}(\mathbf{1},X \times Y) \cong \hom_{\mathbf{Top}}(\mathbf{1},X) \times \hom_{\mathbf{Top}}(\mathbf{1},Y)$$

so the points of $X \times Y$ correspond to the points in the Cartesian product of the underlying sets of X and Y. Then, if A is the set $X \times Y$ equipped with different topologies, the existence of f then requires that sets in A are open whenever they are open in P, so the topological product P must have the coarsest topology on $X \times Y$ possible such that the projection maps are still continuous, thus defining the product topology. This similarly extends to infinite products of topological spaces $(X_i)_{i\in\mathcal{I}}$, where the topology on the product space $\prod_{i\in\mathcal{I}} X_i$ is defined to be the coarsest topology such that each projection $\pi_j : \prod_{i\in\mathcal{I}} X_i \to X_j$ is continuous.

$\triangle$

Example (Products in **Cat**). A product category $\mathscr{C} \times \mathscr{D}$ of two small categories $\mathscr{C}$ and $\mathscr{D}$ is a categorical product in the category **Cat** of small categories. $\triangle$

Consider a pair of morphisms $g : A \to X$ and $h : B \to Y$. We can construct the product $A \times B$ and $X \times Y$ to obtain:

$$\begin{tikzcd}
& A \times B \arrow[dl, "\pi_1"'] \arrow[dr, "\pi_2"] \arrow[dd, dashed, "g\times h"] & \\
A \arrow[dd, "g"'] & & B \arrow[dd, "h"] \\
& X \times Y \arrow[dl, "p_1"] \arrow[dr, "p_2"'] & \\
X & & Y
\end{tikzcd}$$

Now, notice that the compositions $g \circ \pi_1$ and $h \circ \pi_2$ forms a span into X and Y, so by the universal property of the product, there exists a unique map $A \times B \to X \times Y$. Again, this map is entirely determined by g and h, so we write $g \times h$ for this *product morphism.* Note that this is distinct from the pairing $\langle g,h\rangle$, which is a map from a single object into a product, while a morphism of the form $g \times h$ is a map from a product into another product.

It may be helpful to consider these morphisms in **Set**: as before, a pairing function $\langle f,g\rangle : A \to X \times Y$ acts on elements $a \in A$ by $\langle f,g\rangle(a) = (f(a),g(a))$. On the other hand, a product function $s \times t : A \times B \to X \times Y$ acts on pairs $(a,b) \in A \times B$ pointwise, $(s \times t)(a,b) = \big(s(a),t(b)\big)$.

Another special case of the product is given by taking the product of an object X with itself, forming the span with two other copies of X using the projection maps:

$$\begin{tikzcd}
& X \times X \arrow[dl, "\pi_1"'] \arrow[dr, "\pi_2"] & \\
X & & X
\end{tikzcd}$$

Another span can be given by three copies of X equipped with a pair of identity maps:

$$\begin{tikzcd}
& X \arrow[dl, "\mathrm{id}_X"'] \arrow[dr, "\mathrm{id}_X"] & \\
X & & X
\end{tikzcd}$$

By the universal property of the product, there must exist a unique morphism $X \to X \times X$ such that the following diagram commutes:

$$\begin{tikzcd}
& X \arrow[d, dashed, "\Delta_X"] \arrow[ddl, "\mathrm{id}_X"'] \arrow[ddr, "\mathrm{id}_X"] & \\
& X \times X \arrow[dl, "\pi_1"] \arrow[dr, "\pi_2"'] & \\
X & & X
\end{tikzcd}$$

This morphism is called the *diagonal morphism* of X, denoted by $\langle \mathrm{id}_X, \mathrm{id}_X \rangle$, Δ, or Δ_X.

For example, in **Set**, the diagonal function is given by the function $x \mapsto (x,x)$ that sends every element to the corresponding diagonal subset of the Cartesian square.

The diagonal morphism also provides a link between the product morphism and the pairing morphism: suppose we have a product $A \times B$, and a pair of maps $f : X \to A$ and $g : X \to B$. Then, the following diagram commutes:

$$\begin{tikzcd}
X \arrow[r, "\Delta_X"] \arrow[dr, "\langle f{,}g \rangle"'] & X \times X \arrow[d, "f \times g"] \\
& A \times B
\end{tikzcd}$$

52.3.3.1 Coproducts

We can dualise the notion of a product into a *coproduct* (or *categorical sum*) by reversing the direction of all morphisms in the previous definition. Given two objects X and Y, the coproduct is an object P equipped with *insertion* maps,

$$\begin{tikzcd}
& P & \\
X \arrow[ur, "\iota_1"] & & Y \arrow[ul, "\iota_2"']
\end{tikzcd}$$

with the universal property that for all objects and maps

$$\begin{tikzcd}
& A & \\
X \arrow[ur, "g"] & & Y \arrow[ul, "h"']
\end{tikzcd}$$

of the same shape (a *cospan*), there exists a unique factorisation of P through A. That is, there exists a unique map $f : P \to A$ such that the following diagram commutes:

$$\begin{tikzcd}
& A & \\
& P \arrow[u, dashed, "\exists! f"] & \\
X \arrow[uur, "g"] \arrow[ur, "\iota_1"'] & & Y \arrow[uul, "h"'] \arrow[ul, "\iota_2"]
\end{tikzcd}$$

Then, we write $X \amalg Y$, or less commonly $X + Y$, for P, and $[g,h]$ for the *copairing* map f.

Example (Coproducts in **Set**)*.* The disjoint union $X \sqcup Y$ is a categorical coproduct.

The obvious choice for the insertion maps is just to send every element of X and Y to their embedded copy in $X \sqcup Y$. For the map f, consider an element $x \in X$, and its image $g(x) \in A$. The insertion map ι_1 sends x to its copy in $X \sqcup Y$, so for the left triangle to commute, this copy needs to be mapped to $g(x)$ by f, and similarly for every copy of $y \in Y$ in $X \sqcup Y$. So, we define $f : X \sqcup Y \to A$ by,

$$f(u) = \begin{cases} g(u) & u \in X \\ h(u) & u \in Y \end{cases}$$

Again, this is the only possible map we could define that makes the diagram commute, so the disjoint union of sets is a categorical coproduct. △

52.3.4 Pullbacks

Let $\mathscr{C}$ be a category, and consider the following span:

$$\begin{tikzcd}
 & Y \arrow[d, "t"] \\
X \arrow[r, "s"'] & Z
\end{tikzcd}$$

The *pullback* or *fibred product* of this diagram is an object $P \in \mathrm{ob}(\mathscr{C})$ equipped with a pair of projection maps $\pi_1 : P \to X$ and $\pi_2 : P \to Y$ such that the *pullback square* below commutes:

$$\begin{tikzcd}
P \arrow[r, "\pi_2"] \arrow[d, "\pi_1"'] \arrow[rd, phantom, very near start, "\lrcorner"] & Y \arrow[d, "t"] \\
X \arrow[r, "s"'] & Z
\end{tikzcd}$$

But again, not just any object P with maps to X and Y will suffice; we want the "best" such square amongst all similar diagrams. The notion of "best" here is the same as for products; every other similar square should uniquely factorise through P. That is, for any commutative square

$$\begin{tikzcd}
A \arrow[r, "g"] \arrow[d, "h"'] & Y \arrow[d, "t"] \\
X \arrow[r, "s"'] & Z
\end{tikzcd}$$

in $\mathscr{C}$, there must exist a unique map $f : A \to P$ such that

$$\begin{tikzcd}
A \arrow[rrd, bend left, "h"] \arrow[rdd, bend right, "g"'] \arrow[rd, dashed, "\exists! f"] & & \\
 & P \arrow[r, "\pi_2"] \arrow[d, "\pi_1"'] \arrow[rd, phantom, very near start, "\lrcorner"] & Y \arrow[d, "t"] \\
 & X \arrow[r, "s"'] & Z
\end{tikzcd}$$

commutes. The symbol $\lrcorner$ marks the square as a pullback and not as a simple commutative square – that is, that P, π_1, and π_2 satisfy a universal property. We then say that π_1 is the pullback of t along s, and similarly, that π_2 is the pullback of s along t, and the object P is then also written as $X \times_Z Y$. The pullback of a morphism along itself is also called the *kernel pair* of that morphism.

Note that if Z is terminal, then the entire diagram commutes trivially as every map from any given object into the terminal object must be equal, so the pullback in this case is exactly the ordinary product. Or, put another way, the product $X \times Y$ is just the pullback of $X \to 1$ and $Y \to 1$.

Example (Pullbacks in **Set**). We need to pick a set for P and a pair of projection maps such that the lower right square of the diagram commutes in the best possible way.

If we ignore the maps to Z for a moment, we could simply pick P to be the product $X \times Y$. The maps to Z then add the extra constraints in that when mapping elements from X to Z and Y to Z, they must be equal. So, intuitively, we'd might think that the pullback of a diagram $X \xrightarrow{s} Z \xleftarrow{t} Y$ in **Set** should be:

$$P = \big\{(x,y) \in X \times Y : s(x) = t(y)\big\}$$

with π_1 and π_2 given by the standard projection maps, as they were for products.

To verify universality of this construction, suppose there is another set A with maps $X \xleftarrow{g} A \xrightarrow{h} Y$ such that the outer square commutes. Let $a \in A$, so $g(a) \in X$ and $h(a) \in Y$. To make the two triangles commute, this forces f to be defined by $f(a) = \big(g(a),h(a)\big)$. We then need to check that the image of f is within P; that is, that $s\big(g(a)\big) = t\big(h(b)\big)$. But this is just the commutativity requirement of the outer square, which holds by assumption. $\triangle$

Example (Preimages). The preimage of a set under a function can be exhibited as a type of pullback:

Given a function $f : X \to Y$, and a subset $B \subseteq Y$, we can use the inclusion map $\iota : B \hookrightarrow Y$ to form the span

$$\begin{tikzcd}
 & B \arrow[d, hook, "\iota"] \\
X \arrow[r, "f"'] & Y
\end{tikzcd}$$

The pullback of this diagram is then given by the preimage of B, with the projection π_1 given by the inclusion mapping $j : f^{-1}[B] \hookrightarrow X$, and the projection π_2 given by the restriction of f to $f^{-1}[B]$:

$$\begin{tikzcd}
f^{-1}[B] \arrow[r, "f\big|_{f^{-1}[B]}"] \arrow[d, hook, "j"'] \arrow[dr, phantom, "\lrcorner", very near start] & B \arrow[d, hook, "\iota"] \\
X \arrow[r, "f"'] & Y
\end{tikzcd}$$

$\triangle$

Theorem 52.3.3. *Monomorphisms are stable under pullback. That is, if s is a monomorphism, then π_2 is also a monomorphism. Similarly, if t is a monomorphism, then so is π_1.*

Proof. Let $s : X \rightarrowtail Z$ be a monomorphism, and suppose there is an object A with two maps $p,q : A \to P$ such that $\pi_2 \circ p = \pi_2 \circ q$. Then, $t \circ \pi_2 \circ p = t \circ \pi_2 \circ q$, and the commutativity of the pullback square yields $s \circ \pi_1 \circ p = s \circ \pi_1 \circ q$. Since s is monic, we then have $\pi_1 \circ p = \pi_1 \circ q$.

So, A has a pair of morphisms to X and Y, given by $(\pi_1 \circ p = \pi_1 \circ q) : A \to X$ and $(\pi_1 \circ p = \pi_1 \circ q) : A \to Y$, and hence forms a commutative square with Z. But, by the universal property of the pullback, this square must factor uniquely through P, so the map $A \to P$ is unique, and hence $p = q$, so π_2 is monic. The proof for t and π_1 is similar. ■

52.3.4.1 Pushouts

Dually, the *pushout* of a cospan diagram

$$\begin{tikzcd}
 & Y \\
X & Z \arrow[l, "s"] \arrow[u, "t"']
\end{tikzcd}$$

is an object $P \in \mathrm{ob}(\mathscr{C})$ with morphisms $\iota_1 : X \to P$ and $\iota_2 : Y \to P$ such that the pushout square below commutes and is universal:

$$\begin{tikzcd}
P \arrow[dr, phantom, "\lrcorner", very near start] & Y \arrow[l, "\iota_2"'] \\
X \arrow[u, "\iota_1"] & Z \arrow[l, "s"] \arrow[u, "t"']
\end{tikzcd}$$

That is, for any other commutative square

$$\begin{tikzcd}
A & Y \arrow[l, "h"'] \\
X \arrow[u, "g"] & Z \arrow[l, "s"] \arrow[u, "t"']
\end{tikzcd}$$

there exists a unique map $f : P \to A$ such that

$$\begin{tikzcd}
A & & \\
 & P \arrow[ul, dotted, "\exists! f"] \arrow[dr, phantom, "\lrcorner", very near start] & Y \arrow[l, "\iota_2"'] \arrow[ull, bend right, "h"'] \\
 & X \arrow[u, "\iota_1"] \arrow[uul, bend left, "g"] & Z \arrow[l, "s"] \arrow[u, "t"']
\end{tikzcd}$$

commutes. Again, the *cokernel pair* of a morphism is the pushout of the morphism against itself.

Corollary 52.3.3.1. *Epimorphisms are stable under pushout.*

Proof. Dual of previous theorem. ■

Example (Pushouts in **Set**). As for pullbacks, ignoring Z yields $P = X \sqcup Y$, with ι_1 and ι_2 the usual insertion maps. The commutativity of the pushout square then requires that for any $z \in Z$, $s(z)$ is equal to $t(z)$ in $X \sqcup Y$; However, by construction, these elements are never equal in a disjoint union.

To force this equality, we quotient out by the equivalence relation $\sim$ generated by $s(z) \sim t(z)$ for all $z \in Z$, so the pushout is given by $A \sqcup B\big/\sim$. △

Example (Unions and Intersections). Consider two sets X and Y, and their intersection $X \cap Y$ and (ordinary) union $X \cup Y$. The intersection is naturally equipped with inclusions into X and Y, and similarly, X and Y have inclusions into $X \cup Y$. The induced square is then simultaneously a pullback and a pushout:

$$\begin{tikzcd}
X \cap Y \arrow[r, hook] \arrow[d, hook] \arrow[dr, phantom, "\lrcorner", very near start] & Y \arrow[d, hook] \\
X \arrow[r, hook] & X \cup Y \arrow[ul, phantom, "\ulcorner", very near start]
\end{tikzcd}$$

△

52.3.5 Equalisers

A *fork* is a collection of objects and morphisms such that the following diagram commutes:

$$\begin{tikzcd}
A \arrow[r, "g"] & X \arrow[r, shift left, "s"] \arrow[r, shift right, "t"'] & Y
\end{tikzcd}$$

That is, this diagram is a fork if and only if $s \circ g = t \circ g$, and we say that g *equalises* s and t.

Let X and Y be objects in a category $\mathscr{C}$ with parallel morphisms $s,t : X \to Y$:

$$\begin{tikzcd}
X \arrow[r, shift left, "s"] \arrow[r, shift right, "t"'] & Y
\end{tikzcd}$$

The *equaliser* of s and t is an object E equipped with a map $\iota : E \to X$ such that the following diagram is a fork, and is universal:

$$\begin{tikzcd}
E \arrow[r, "\iota"] & X \arrow[r, shift left, "s"] \arrow[r, shift right, "t"'] & Y
\end{tikzcd}$$

That is, every other fork through X and Y beginning at an object A factors uniquely through E:

$$\begin{tikzcd}
A \arrow[d, dashed, "\exists! f"'] \arrow[dr, "g"] & & \\
E \arrow[r, "\iota"'] & X \arrow[r, shift left, "s"] \arrow[r, shift right, "t"'] & Y
\end{tikzcd}$$

and we write $\mathrm{eq}(s,t)$ for E.

Example (Equalisers in **Set**). In set theory, the notion of an equaliser is often defined to be the set of values upon which two functions agree. That is, given two functions $f,g : X \to Y$, the equaliser is the subset $\big\{x \in X : f(x) = g(x)\big\}$ of X. This function equaliser is a categorical equaliser, with ι given by the obvious inclusion map.

△

Example (Kernels). Let G and H be groups, and let $\varepsilon : G \to H$ be the trivial homomorphism defined by $g \mapsto \mathrm{id}_H$ for all $g \in G$. The equaliser of ε and any group homomorphism $\phi : G \to H$ is exactly the kernel of ϕ equipped with the inclusion mapping $\ker(\phi) \hookrightarrow G$:

$$\begin{tikzcd}
\ker(\phi) \arrow[r, hook, "\iota"] & G \arrow[r, shift left, "\varepsilon"] \arrow[r, shift right, "\phi"'] & H
\end{tikzcd}$$

△

Theorem 52.3.4. *The equaliser of two morphisms is monic.*

Proof. [Gol84, adapted] Let $\iota : E \to X$ equalise s,$t : X \to Y$, and let g,$h : A \to E$ be morphisms such that $\iota \circ g = \iota \circ h$, so the following diagram commutes:

$$\begin{tikzcd} A \arrow[r, shift left, "g"] \arrow[r, shift right, "h"'] & E \arrow[r, "\iota"] & X \arrow[r, shift left, "s"] \arrow[r, shift right, "t"'] & Y \end{tikzcd}$$

Then,

$$\begin{aligned} s \circ (\iota \circ g) &= (s \circ \iota) \circ g \\ &= (t \circ \iota) \circ g \\ &= t \circ (\iota \circ g) \end{aligned}$$

so $\iota \circ g$ equalises s and t and defines a fork $A \to X \rightrightarrows Y$. By the universal property of the equaliser, this fork must then factorise uniquely through E, so the map $A \to E$ that makes the diagram

$$\begin{tikzcd} A \arrow[d, dashed, "g=h"'] \arrow[dr, "\iota \circ g"] & & \\ E \arrow[r, "\iota"] & X \arrow[r, shift left, "s"] \arrow[r, shift right, "t"'] & Y \end{tikzcd}$$

commute is unique and hence ι is monic. ■

A morphism that is the equaliser of some pair of parallel morphisms is called a *regular monomorphism.* As shown above, regular monomorphisms are monomorphisms, but the converse does not always hold. For instance, in **Top**, the monomorphisms are injective continuous functions, while the regular monomorphisms are topological embeddings (injective continuous functions that are homeomorphic on their image equipped with the subspace topology).

52.3.6 Coequalisers

Dualising, a *cofork* is a collection of objects and morphisms such that the following diagram commutes:

$$\begin{tikzcd} X \arrow[r, shift left, "s"] \arrow[r, shift right, "t"'] & Y \arrow[r, "g"] & A \end{tikzcd}$$

and we say that g *coequalises* s and t.

The *coequaliser* of a pair of parallel morphisms s,$t : X \rightrightarrows Y$ is an object C equipped with a map $\pi : Y \to C$ such that the following diagram is a universal cofork:

$$\begin{tikzcd} X \arrow[r, shift left, "s"] \arrow[r, shift right, "t"'] & Y \arrow[r, "\pi"] & C \end{tikzcd}$$

and we write coeq(s,t) for C.

Dual to regular monomorphisms, a morphism that is the coequaliser of some pair of parallel morphisms is called a *regular epimorphism.* By the dual of the previous theorem, every regular epimorphism is an epimorphism, but again, the converse does not necessarily hold.

52.4 Limits

In the previous constructions, we have started with some initial data – for products; a discrete pair of objects; for pullouts, a span; for equalisers, a parallel pair of morphisms – and we construct a new object equipped with maps to the given data in the most "general" way possible. That is, any other similar object will factor through our universal construction.

The notion of a *limit* unifies these three constructions. But first, we more precisely formulate how this initial data is specified.

52.4.1 Diagrams

So far, we have been frequently representing collections of objects and morphisms with the use of directed graphs, which we called diagrams. We formalise this notion in more detail now.

Recall that elements of a set X is can be viewed as functions $1 \to X$. Analogously, objects in a category $\mathscr{C}$ can be viewed as a functor $\mathbb{1} \to \mathscr{C}$ from the trivial category $\mathbb{1}$ – the functor just sends the unique object of $\mathbb{1}$ to some object of $\mathscr{C}$. A morphism in $\mathscr{C}$ can then similarly be viewed as a functor $\mathbb{2} \to \mathscr{C}$, where $\mathbb{2}$ is the arrow category $[\,\bullet \to \bullet\,]$ – such a functor then picks out two objects, and a morphism between them. More generally, this suggests that we can view any collection of objects and morphisms to be a functor from some indexing category to the target category:

A *diagram* (*of shape* $\mathscr{I}$) in a category $\mathscr{C}$ is a functor $D : \mathscr{I} \to \mathscr{C}$, where $\mathscr{I}$ is called the *indexing category.* If the indexing category $\mathscr{I}$ is small, then the diagram D is said to be small.

The directed graph representations used previously is then obtained by drawing the images of these functors, and functoriality requires that any compositions that exist in $\mathscr{I}$ also exist in this image graph, so the resulting graph is always commutative.

Theorem 52.4.1. *Functors preserve commutative diagrams.*

Proof. A commutative diagram in $\mathscr{C}$ is given by a functor $D : \mathscr{I} \to \mathscr{C}$. Given any functor $F : \mathscr{C} \to \mathscr{D}$, the composition $F \circ D : \mathscr{I} \to \mathscr{D}$ is by definition a commutative diagram in $\mathscr{D}$. ■

We proved this result earlier by applying the functor to pairs of paths between objects and moving composition operations about with functoriality, but with this characterisation of diagrams, the result is easily apparent.

One important type of diagram is as follows: let $\mathscr{C}$ and $\mathscr{I}$ be categories, and let $\mathscr{I}$ be small; then, for any object $C \in \mathrm{ob}(\mathscr{C})$, the *constant functor* (or *constant diagram*) $\Delta C : \mathscr{I} \to \mathscr{C}$ maps every object in $\mathscr{I}$ to C, and every morphism in $\mathscr{I}$ to id_C, similarly to an ordinary constant set function.

The assignment of objects to their constant functors is itself functorial; the *diagonal functor* $\Delta : \mathscr{C} \to [\mathscr{I},\mathscr{C}]$ sends every object $C \in \mathrm{ob}(C)$ to the constant functor $\Delta C : \mathscr{I} \to \mathscr{C}$, and every morphism $f : A \to B$ to a natural transformation $\Delta f : \Delta A \to \Delta B$ that takes the same value f at every object in $\mathscr{I}$.

For an explanation to the diagonal functor's name, consider the case where $\mathscr{I} = \mathbf{2}$ is the discrete category on two objects. First, note that any functor $\mathbf{2} \to \mathscr{C}$ simply picks out pairs of objects of $\mathscr{C}$, so $[\mathbf{2},\mathscr{C}] \cong \mathscr{C} \times \mathscr{C}$. This gives the *binary* diagonal functor, $\Delta : \mathscr{C} \to \mathscr{C} \times \mathscr{C}$, defined by $\Delta(X) = (X,X)$ and $\Delta(f) = \langle f,f \rangle$ for objects X and morphisms f.

For objects X, the pair (X,X) is really just the image of a constant functor from $\mathbf{2}$ to $\mathscr{C}$, consistent with the above, and for morphisms f, the pairing $\langle f,f \rangle$ is similarly just the components of a natural transformation $\Delta X \Rightarrow \Delta X$. This may seem similar to diagonal morphisms as defined earlier (§52.3.3), and in fact, diagonal functors are exactly the diagonal morphisms in **Cat**.

52.4.2 Cones

A *cone* over a diagram $F : \mathscr{I} \to \mathscr{C}$ with *summit* $X \in \mathrm{ob}(\mathscr{C})$ is a natural transformation $\phi : \Delta X \to F$, and the components of this natural transformation are called the *legs* of the cone.

It may be helpful here to think of a natural transformation purely in terms of its components. That is, a cone over a diagram $F : \mathscr{I} \to \mathscr{C}$ is an assignment of a morphism $\phi_J : X \to F(J)$ for each object $J \in \mathrm{ob}(\mathscr{I})$ in the diagram. Because one of the functors sends everything to a single object, the naturality square is contracted into a triangle; so, naturality is just the requirement that for every morphism $f : J \to K$ in $\mathscr{I}$, the following triangle in $\mathscr{C}$ commutes:

$$\begin{tikzcd}
 & X \arrow[dl, "\phi_J"'] \arrow[dr, "\phi_K"] & \\
F(J) \arrow[rr, "F(f)"'] & & F(K)
\end{tikzcd}$$

As an example, suppose $\mathscr{I} = \mathbb{2} \times \mathbb{2}$, so the diagram in $\mathscr{C}$ is some commutative square:

$$\begin{tikzcd}
 & 0 \arrow[r] \arrow[d] & 1 \arrow[d] & & A \arrow[r] \arrow[d] & B \arrow[d] & \\
\mathscr{I} & 2 \arrow[r] & 3 \arrow[rr, "F"] & & C \arrow[r] & D & \mathscr{C}
\end{tikzcd}$$

(Here, $F(0) = A$, $F(1) = B$, etc.) A cone over this diagram with summit X is a collection of morphisms from X to each object of the diagram:

$$\begin{tikzcd}
 & & & & X \arrow[dd] \arrow[ddrr] \arrow[dddl] \arrow[dddr] & & & \\
 & 0 \arrow[r] \arrow[dd] & 1 \arrow[dd] \arrow[urr, bend left, "\Delta X"] & & & & & \\
\mathscr{I} & & & & A \arrow[rr] \arrow[dl] & & B \arrow[dl] & \mathscr{C} \\
 & 2 \arrow[r] & 3 \arrow[r, "F"] & C \arrow[rr] & & D & &
\end{tikzcd}$$

subject to the constraint that every triangle in the diagram involving two legs of the cone commutes. (The visual appearance of the resulting diagram also lends this construction its name.)

As with every other categorical construction, we can dualise the notion of a cone. A *cocone* under a diagram $F : \mathscr{I} \to \mathscr{C}$ with *nadir* $X \in \mathrm{ob}(\mathscr{C})$ is a natural transformation $\psi : F \to \Delta X$. The standard visualisation here is to place the object X below the diagram, drawing morphisms down to it from above.

$$\begin{tikzcd}
 & A \arrow[rr] \arrow[dl] \arrow[ddr] & & B \arrow[dl] \arrow[ddl] \\
C \arrow[rr] \arrow[dr] & & D \arrow[d] & \\
 & & X &
\end{tikzcd}$$

For a slight technicality, we've been saying that a cone with summit X over a diagram $F : \mathscr{I} \to \mathscr{C}$ is a natural transformation $\Delta X \Rightarrow F$, which is perfectly fine for almost all use cases, because X is completely determined by ϕ whenever $\mathscr{I}$ is non-empty. However, if this is the case, then ΔX and F are both the empty functor, which has nothing to do with X, so the natural transformation doesn't actually contain enough information to recover X.

For this reason, a cone is more properly a *pair* (X,ϕ), so that the summit is specified separately. This also necessitates a small modification to ΔX – to account for the case of the empty diagram, we use not the constant functor, but instead the unique functor from $\mathscr{I}$ to the terminal category $\mathbb{1} = \{\bullet\}$ composed with the inclusion $\iota_X : \mathbb{1} \to \mathscr{C}$ defined by $\bullet \mapsto X$ – but for non-empty diagrams, they function exactly as we have described previously.

We will interchangably refer to a (co)cone both as a natural transformation ϕ, and as a pair (X,ϕ) that includes the summit/nadir separately, depending on context.

52.4.3 Universal Cones

Let $F : \mathscr{I} \to \mathscr{C}$ be a diagram. The *limit* of F is the universal cone $\phi : \Delta X \Rightarrow F$ over the diagram F, and we denote the summit X of this cone by $\lim F$. Dually, the *colimit* is the universal cocone under F, with nadir denoted $\operatorname{colim} F$. By a slight abuse of language, sometimes, we call the summit X of this universal cone "the limit" by itself, but the limit is properly the entire data of the cone. To distinguish the two notions, we call this universal cone the *limit cone*, and the summit the *limit object*, or just *limit*.

The notion of universality for cones is again of unique factorisation; a cone over a diagram D with summit X is universal if every other cone over D with summit X' factors through it uniquely. That is, there exists a unique morphism $f : X' \to X$

$$\begin{tikzcd}
 & & X' \arrow[d, dashed, "\exists! f"] \arrow[dl, bend right] \arrow[dr, bend left] \arrow[ddll, bend right] \arrow[dd, bend left] & \\
 & & X \arrow[dl] \arrow[dr] \arrow[ddll] \arrow[dd] & \\
 & A \arrow[rr] \arrow[dl] & & B \arrow[dl] \\
C \arrow[rr] & & D &
\end{tikzcd}$$

such that every leg of the non-universal cone factors through the corresponding leg of the universal cone, so the triangle

$$\begin{tikzcd}
X' \arrow[d, dashed, "f"'] \arrow[ddr, bend left, "\phi'_J"] & \\
X \arrow[dr, "\phi_J"'] & \\
 & F(J)
\end{tikzcd}$$

commutes for all $J \in \operatorname{ob}(\mathscr{I})$. Comparing X' and X with the constant functors $\Delta X'$ and ΔX, we see that f must be given by the single component of the constant natural transformation Δf.

In this case, we also call f a *cone morphism*, as it defines a way of mapping between cones. In fact, this allows us to define a *category of cones* – if $F : \mathscr{I} \to \mathscr{C}$ is a diagram, then the collection of cones and cone morphisms over that diagram forms a category. Associativity is inherited from $\mathscr{C}$, and identities are given by identity morphisms on the summits.

This allows us to characterise limits as terminal objects in the category of cones over a diagram. This justifies the use of the phrasing "*the* limit", rather than "*a* limit", as any two ostensibly distinct limits over the same diagram will be isomorphic up to unique isomorphism:

Theorem 52.4.2. *(Co)limits are essentially unique.*

That is, given any two universal cones $\phi : \Delta X \Rightarrow F$ and $\psi : \Delta Y \Rightarrow F$ over the same diagram $F : \mathscr{I} \to \mathscr{C}$, there is a unique cone isomorphism $\Delta X \cong \Delta Y$ (a cone morphism in the sense described above that additionally has an inverse).

Proof. A limit of F is a terminal object in the category of cones over F. But by Theorem 52.3.1, terminal objects are essentially unique. Essential uniqueness of colimits follows from duality. ■

Conversely, if we begin with a map $f : X' \to X$, where X is the summit of a limit cone $\phi : \Delta X \Rightarrow F$, then the universal property says that the induced natural transformation $\psi : \Delta X' \Rightarrow F$ with components defined by precomposing every component of ϕ with f is also a cone, and furthermore, the uniqueness of the factorisation through a limit cone implies that this association between maps into X and cones over F is a bijection.[*] More concisely,

Theorem 52.4.3. *The maps into a limit object* $\lim F$ *are precisely the cones over the diagram* F.

52.4.4 Examples

Consider the most trivial possible case of $\mathscr{I}$ being the empty category, so all diagrams $F : \mathscr{I} \to \mathscr{C}$ are empty. A cone over the empty diagram with summit X is just the object X with no other data, so the limit of this diagram is then just an object such that any cone – which is just another object – factors through it uniquely. That is, there is a unique morphism to the limit object from every other object. This is exactly the terminal object of $\mathscr{C}$, so the limit of the empty diagram is the terminal object. Dually, a cocone under the empty diagram with nadir X is again just the object X with no other data, so the colimit of the empty diagram is the initial object.

If we take $\mathscr{I}$ to be the trivial category, then a diagram $F : \mathscr{I} \to \mathscr{C}$ consists of a single object, say, X_1. A cone over F with summit X is then just the object X equipped with a morphism to X_1 that every other object uniquely factors through it. But, we can just take $X = X_1$, as every object factors through the identity on X_1, so the limit is just the object itself. The colimit is also just the object X_1, for the same reason.

If $\mathscr{I}$ is the discrete category on two objects, then a cone over $F : \mathscr{I} \to \mathscr{C}$ with summit X is the object X equipped with a pair of morphisms to the diagram:

$$\begin{tikzcd} & X \arrow[dl] \arrow[dr] & \\ X_1 & & X_2 \end{tikzcd}$$

The limit is then the object through which every other similar span factors – which is exactly the product of the two objects in the diagram, so a product is a special case of a limit. Similarly, the colimit of the two object diagram is a universal cospan, or, the coproduct.

We can generalise the binary (co)product to take n input objects by taking $\mathscr{I}$ to be the discrete category on n objects. The limit of the resulting diagram with objects $(X_i)_{i \in \mathscr{I}}$ is then the n-ary product, written

[*] The precomposition here may be reminiscent of the Yoneda embeddings; in fact, this assignment is functorial, and limits can be expressed as a type of representation, though we will not be doing this here.

as $\prod_{i \in \mathscr{I}} X_i$, and similarly, the colimit is the n-ary coproduct, written as $\coprod_{i \in \mathscr{I}} X_i$.

$$\begin{tikzcd}
 & & & \prod\limits_{i \in \mathscr{I}} X_i \arrow[dlll] \arrow[dll, "\pi_{-2}"] \arrow[dl, "\pi_{-1}"] \arrow[d, "\pi_0"] \arrow[dr, "\pi_1"] \arrow[drr, "\pi_2"] \arrow[drrr] & & & \\
\cdots \arrow[drrr] & X_{-2} \arrow[drr, "\iota_{-2}"] & X_{-1} \arrow[dr, "\iota_{-1}"] & X_0 \arrow[d, "\iota_0"] & X_1 \arrow[dl, "\iota_1"] & X_2 \arrow[dll, "\iota_2"] & \cdots \arrow[dlll] \\
 & & & \coprod\limits_{i \in \mathscr{I}} X_i & & &
\end{tikzcd}$$

This generalisation encompasses the previous two examples as well; the empty and trivial categories can be viewed as discrete categories on zero and one objects, respectively, so the *nullary* product of the empty family is exactly the terminal object (which is one of the reasons why the terminal object is denoted 1), and the *unary* product is just the original input object.

If we instead take the limit of a span $\bullet \to \bullet \leftarrow \bullet$, we obtain the diagram:

$$\begin{tikzcd}
\lim F \arrow[r, "\pi_2"] \arrow[d, "\pi_1"'] \arrow[dr, "\pi_0"] & \bullet \arrow[d] \\
\bullet \arrow[r] & \bullet
\end{tikzcd}$$

The projection π_0 can be omitted as it is implied by commutativity, so the limit cone is just a commutative square. The universal property then says that every similar square factors through $\lim F$, so this diagram is exactly the pullback of the two morphisms in the original span. Similarly, the colimit of a span $\bullet \leftarrow \bullet \to \bullet$ is then the pushout of the two morphisms.

We can also take the limit over a parallel pair of morphisms $\bullet \rightrightarrows \bullet$ to obtain the equaliser:

$$\begin{tikzcd}
\lim F \arrow[d, "\iota"'] \arrow[dr, "f"] & \\
\bullet \arrow[r, shift left, "s"] \arrow[r, shift right, "t"'] & \bullet
\end{tikzcd}$$

Commutativity implies that ι forms a fork with s and t, and that f can be omitted to obtain the usual equaliser diagram. Colimits over parallel pairs then similarly yield coequalisers.

52.4.5 Completeness

Diagrams can be of any shape, but limits over arbitrary diagrams do not always exist in arbitrary categories. For example, we have already seen that products of distinct objects do not exist in discrete categories due to a lack of morphisms to use as projections.

A category $\mathscr{C}$ *has limits of shape* $\mathscr{I}$ if every diagram F of shape $\mathscr{I}$ in $\mathscr{C}$ admits a limit in $\mathscr{C}$. Similar variations on this wording can be applied to special classes of limits like products ("$\mathscr{C}$ has products") [Lei14]. One important case is of *small limits*, where the indexing category of the diagram is small.

A category $\mathscr{C}$ is *complete* if it has all small limits. That is, if it has limits of shape $\mathscr{I}$ for every small category $\mathscr{I}$. Dually, $\mathscr{C}$ is *cocomplete* if it has all small colimits. If a category is both complete and cocomplete, it is *bicomplete.* The condition of having *all* (large) limits is too strong of a condition for any useful categories to satisfy, so we often do not consider it.

A weaker form of completeness is of *finite completeness* – a category is *finite* if it has finitely many *morphisms* (which also implies that there are finitely many objects), and a *finite limit* is a limit of shape $\mathscr{I}$ for a finite indexing category $\mathscr{I}$. A category $\mathscr{C}$ is then *finitely complete* if it has all finite limits, *finitely cocomplete* if it has all finite colimits, and *finitely bicomplete* if it is finitely complete and cocomplete.

Theorem (Existence Theorem for Limits)**.**

- *If a category $\mathscr{C}$ has all products and binary equalisers, then $\mathscr{C}$ is complete.*
- *If $\mathscr{C}$ has binary products, a terminal object and binary equalisers, then $\mathscr{C}$ is finitely complete.*

Proof. Let $\mathscr{C}$ be a category with all products and equalisers, and let $F : \mathscr{I} \to \mathscr{C}$ be a diagram. If we ignore all morphisms in $\mathscr{I}$, then the limit of F would just be the product $\prod_{I \in \mathscr{I}} F(I)$. The morphisms then add the additional constraints that for each morphism $f : A \to B$ in $\mathscr{I}$, the triangle

$$\begin{tikzcd}
\prod\limits_{I \in \mathscr{I}} F(I) \arrow[d, "\pi_A"'] \arrow[dr, "\pi_B"] & \\
F(A) \arrow[r, "F(f)"'] & F(B)
\end{tikzcd}$$

must commute.

We construct another product $\prod_{(f:A\to B) \text{ in } \mathscr{I}} F(B)$ indexed over the morphisms in $\mathscr{I}$. Now, recall that a map from an object X into a product $\prod X_i$ is determined entirely by the component maps $X \to X_i$. The two routes $\prod_{I\in\mathscr{I}} F(I) \to F(B)$ through the triangle above, given by π_B and $F(f) \circ \pi_A$, can be reindexed by morphisms in $\mathscr{I}$ to obtain:

$$\begin{aligned} s_{(f:A\to B)} &= \pi_B \\ t_{(f:A\to B)} &= F(f) \circ \pi_A \end{aligned}$$

and hence these components define a pair of maps into the product indexed by morphisms:

$$\begin{tikzcd}
\prod\limits_{I \in \mathscr{I}} F(I) \arrow[r, shift left, "s"] \arrow[r, shift right, "t"'] & \prod\limits_{(f:A\to B) \text{ in } \mathscr{I}} F(B)
\end{tikzcd}$$

Because we want the triangles of the form given above to commute, these paths should be equal, so we construct the equaliser $L \xrightarrow{p} \prod_I F(I)$ of s and t. We claim that the components of p form a limit cone on F.

First, note that $s \circ p = t \circ p$ are morphisms into a product, so their components also agree by the universal property of the product. Writing π' for the projections of $\prod_f F(B)$, we have,

$$\begin{aligned} s \circ p &= t \circ p \\ \pi'_f \circ s \circ p &= \pi'_f \circ t \circ p \\ s_f \circ p &= t_f \circ p \\ \pi_B \circ p &= F(f) \circ \pi_A \circ p \\ p_B &= F(f) \circ p_A \end{aligned}$$

which is exactly the commutativity requirement of a cone, so the components of p form a cone $\phi : \Delta L \Rightarrow F$. Any other cone $\psi : \Delta L' \Rightarrow F$ must induce a map that equalises s and t, so L' factors through L by the universal property of the equaliser, and hence ϕ is a limit cone.

$$\begin{tikzcd}
L' \arrow[d, dashed, "\exists!"'] \arrow[dr, "p'"] & & \\
L \arrow[r, "p"'] & \prod\limits_{I\in\mathscr{I}} F(I) \arrow[r, shift left, "s"] \arrow[r, shift right, "t"'] & \prod\limits_{(f:A\to B)\text{ in }\mathscr{I}} F(B)
\end{tikzcd}$$

We have now expressed an arbitrary limit in terms of products and equalisers, so $\mathscr{C}$ is complete.

Now suppose that $\mathscr{C}$ only admits products that are binary. By induction, $\mathscr{C}$ also has n-ary products for finite $n \geq 1$, and since $\mathscr{C}$ has a terminal object (the nullary product), it has all finite products. The previous argument then applies, but only for finite indexing categories, and hence $\mathscr{C}$ is finitely complete. ■

The converse of the propositions are also clearly true, so the implications are biconditional; (finite) completeness is equivalent to having (finite) products and binary equalisers.

52.4.6 Limits and Functors

Let $D : \mathscr{I} \to \mathscr{C}$ be a diagram in $\mathscr{C}$ with limit object X and limit cone ϕ. A functor $F : \mathscr{C} \to \mathscr{D}$ *preserves* this limit if $F(X)$ is the limit object of the composite diagram $F \circ D : \mathscr{I} \to \mathscr{D}$, and the whiskering $F \cdot \phi$ (that is, the natural transformation with components defined by $(F \cdot \phi)_A = F(\phi_A)$) is a limit cone on $F \circ D$.

Theorem 52.4.4. *The hom-bifunctor* $\hom(-,-) : \mathscr{C}^{\text{op}} \times \mathscr{C} \to \mathbf{Set}$ *preserves limits in both arguments.*

This means that, for example, $\hom(X,A) \times \hom(X,B) \cong \hom(X,A \times B)$ and $\hom(A,X) \times \hom(B,X) \cong \hom(A \amalg B,X)$, with the product being transformed into a coproduct in the first argument by contravariance.

Proof. Recall that the morphisms into a limit object $\lim D$ correspond to the cones over the diagram D (Theorem 52.4.3). More specifically, the morphisms $X \to \lim D$ are the cones over D with apex X, so we have

$$\hom_{\mathscr{C}}(X, \lim D) \cong \hom_{[\mathscr{I},\mathscr{C}]}(\Delta X, D) \tag{52.1}$$

naturally in X.

If we have $\mathscr{C} = \mathbf{Set}$ and $X = 1$ in the previous, then

$$\begin{aligned}\hom_{\mathbf{Set}}(1, \lim D) &\cong \hom_{[\mathscr{I},\mathbf{Set}]}(\Delta 1, D)\\ \lim D &\cong \hom_{[\mathscr{I},\mathbf{Set}]}(\Delta 1, D)\end{aligned} \tag{2}$$

so the set of cones over a diagram with apex 1 is precisely the limit of that diagram.

Now, consider a cone ϕ over the composite diagram $\hom(X,-) \circ D = \hom\big(X,D(-)\big) : \mathscr{I} \to \mathbf{Set}$ with apex 1. Every morphism $f : I \to J$ between objects I,J in $\mathscr{I}$ induces a function $\hom(\mathrm{id}_X,f) = f \circ (-)$

from $\hom\bigl(X,D(I)\bigr)$ to $\hom\bigl(X,D(J)\bigr)$, so such a cone is a collection of components such that the triangle

$$\begin{tikzcd}
 & 1 \arrow[dl, "\phi_I"'] \arrow[dr, "\phi_J"] & \\
\hom\bigl(X,D(I)\bigr) \arrow[rr, "f\circ(-)"'] & & \hom\bigl(X,D(J)\bigr)
\end{tikzcd}$$

commutes for all $I,J \in \mathrm{ob}(\mathscr{I})$ and $f \in \hom_{\mathscr{I}}(I,J)$.

Because 1 is a singleton set, the component ϕ_I effectively just selects a single function from the hom-set $\hom\bigl(X,D(I)\bigr)$. We can alteratively interpret this as ϕ assigning a function $\psi_I : X \to D(I)$ to each object I such that $f \circ \psi_I = \psi_J$ for all objects I,J in $\mathscr{I}$ i.e., such that

$$\begin{tikzcd}
 & X \arrow[dl, "\psi_I"'] \arrow[dr, "\psi_J"] & \\
D(I) \arrow[rr, "f"'] & & D(J)
\end{tikzcd}$$

commutes, which is exactly a cone over D. Thus, the set of cones over F with apex X is isomorphic to the set of cones over the hom-sets $\hom\bigl(X,D(-)\bigr)$ with apex 1:

$$\hom_{[\mathscr{I},\mathscr{C}]}(\Delta X,D) \cong \hom_{[\mathscr{I},\mathbf{Set}]}\bigl(\Delta 1, \hom_{\mathscr{C}}\bigl(X,D(-)\bigr)\bigr) \tag{3}$$

Combining the above, we deduce that $\hom_{\mathscr{C}}(X, \lim D)$ is the limit of

So combining the previous isomorphisms, we have,

$$\begin{aligned}
\hom_{\mathscr{C}}(X, \lim D) &\cong \hom_{[\mathscr{I},\mathscr{C}]}(\Delta X,D) && \text{[by (1)]}\\
&\cong \hom_{[\mathscr{I},\mathbf{Set}]}\bigl(\Delta 1, \hom_{\mathscr{C}}\bigl(X,D(-)\bigr)\bigr) && \text{[by (3)]}\\
&\cong \lim\bigl(\hom\bigl(X,D(-)\bigr)\bigr) && \text{[by (2) in reverse]}
\end{aligned}$$

Preservation of limits in the first argument follows from duality. ■

52.5 Adjunctions

Let $\mathscr{C} \underset{R}{\overset{L}{\rightleftarrows}} \mathscr{D}$ be categories and opposing functors. We say that L is *left adjoint* to R, and that L is *right adjoint* to R, if,

$$\hom_{\mathscr{C}}(L(A),B) \cong \hom_{\mathscr{D}}(A,R(B))$$

naturally in $A \in \mathrm{ob}(\mathscr{C})$ and $B \in \mathrm{ob}(\mathscr{D})$, and we write $L \dashv R$ to denote this relationship.

The image of a morphism under this isomorphism is called its *adjoint transposition.* For a morphism $f : A \to R(B)$ in $\mathscr{C}$, its adjoint transposition $L(A) \to B$ in $\mathscr{D}$ is called its *left adjunct*, and is denoted $f^\sharp$, and similarly, the adjoint transposition of a morphism $g : L(A) \to B$ in $\mathscr{D}$ is called its *right adjunct*, and is denoted $g^\flat$.

Note that the functors L and R are *adjoint*, while the morphisms f and $f^\sharp$ or $g^\flat$ and g are *adjunct* [Mac13].

Adjoint functors correspond to a weak form of equivalence between categories, in that every equivalence functor and its inverse are adjoint, and conversely, if L and R are both fully faithful, then they form an equivalence of categories.

Example.

- The functor $\mathscr{C} \to \mathbb{1}$ to the trivial category has a right adjoint if and only if $\mathscr{C}$ has a terminal object, and a left adjoint if and only if $\mathscr{C}$ has an initial object.
- If $\mathscr{C}$ has binary products, then the *cartesian product bifunctor* $- \times - : \mathscr{C} \times \mathscr{C} \to \mathscr{C}$ is right adjoint to the diagonal functor $\Delta : \mathscr{C} \to \mathscr{C} \times \mathscr{C}$.
- A so-called *forgetful functor* is a functor that does nothing to objects and morphisms apart from "forgetting" some of the structure of the original category. For instance, the forgetful functor $U : \mathbf{Grp} \to \mathbf{Set}$ sends groups to their underlying sets – "forgetting" the group structure – and group homomorphisms to themselves, but considered as functions. The left adjoint to a forgetful functor from a category of algebraic objects to **Set** is given by the functor that constructs the free object of the appropriate type on any given set.

△

Theorem 52.5.1. *A left or right adjoint, if it exists, is unique up to natural isomorphism.*

Proof. We give the proof for the uniqueness of a right adjoint. The uniqueness of left adjoints follows similarly.

Suppose $R_1,R_2 : \mathscr{C} \to \mathscr{D}$ are right adjoint to $L : \mathscr{D} \to \mathscr{C}$, so for each object D in $\mathscr{D}$, we have the natural isomorphisms of functors

$$\hom\big(-,R_1(D)\big) \cong \hom\big(L(-),D\big) \cong \hom\big(-,R_2(D)\big)$$

also natural in D. By the Yoneda lemma, this isomorphism is induced by an isomorphism $R_1(D) \cong R_2(D)$. As the Yoneda embedding is fully faithful, $R_1(D) \cong R_2(D)$ is also natural in D, and we have $R_1 \cong R_2$. ■

Theorem 52.5.2. *Left adjoints preserve colimits and right adjoints preserve limits.*

Proof. Let $(L \dashv R) : \mathscr{C} \to \mathscr{D}$ be a pair of adjoint functors, and let $F : \mathscr{I} \to \mathscr{D}$ be a diagram in $\mathscr{D}$ whose limit $\lim F$ exists. Then, for every C in $\mathscr{C}$,

$$\begin{aligned} \hom_{\mathscr{C}}\big(C,R(\lim F)\big) &\cong \hom_{\mathscr{D}}\big(L(C), \lim F\big) \\ &\cong \lim \hom_{\mathscr{D}}\big(L(C),F(-)\big) \\ &\cong \lim \hom_{\mathscr{C}}\big(C,(R \circ F)(-)\big) \\ &\cong \hom_{\mathscr{C}}\big(C, \lim(R \circ F)\big) \end{aligned}$$

with all isomorphisms natural in C, so by the Yoneda lemma, $R(\lim F) \cong \lim(R \circ F)$. That left adjoints preserve colimits follows from duality. ■

52.6 Subobjects

Recall that an element $x \in X$ is simply a map $x : 1 \to X$, where 1 is terminal. We have also seen the notion of a generalised element of shape S in which we identify certain features of a space with maps $S \to X$ into the space, but can also abstract the notion of a *subset* into categories other than **Set** in a similar way.

Note that the notion of a subset in material set theories is not isomorphism invariant: for instance, consider the sets $\{1\}$, $\{2\}$, $\{\text{cat}\}$, and $X = \{1,2,3\}$. We have that $\{1\} \subseteq X$ and $\{2\} \subseteq X$ are distinct subsets of X, and $\{\text{cat}\} \not\subseteq X$, but these singleton sets are all isomorphic! Categories don't care about how we label elements (and in arbitrary categories, it currently doesn't make sense to ask about elements of objects anyway).

What we *do* care about is how these sets *embed* into X or not – just as elements of sets are not themselves sets, subsets of sets are also not sets.

Consider the class of monomorphisms into an object, X. We can define a preorder on this class as follows. Let $f : A \rightarrowtail X$ and $g : B \rightarrowtail X$ be monomorphisms into X. Then, let $f \leq g$ if f factors through g – that is, there exists a morphism $h : A \to B$ such that $f = g \circ h$ (h is necessarily unique if it exists, since g is monic, and must also be a monomorphism, as f is monic). If both $f \leq g$ and $g \leq f$, or equivalently, if h is an isomorphism, then f and g are *isomorphic* morphisms, and we write $f \cong g$.

A *subobject* of an object X is then an isomorphism class of monomorphisms into X. If $S : A \rightarrowtail X$ is a monomorphism into X, then we write $[S] \subseteq X$ for the isomorphism class/subset represented by S. As a small abuse of notation, we sometimes pick a representative monomorphism S and just write $S \subseteq X$.

If we interpret this definition in **Set**, a monomorphism $f : A \rightarrowtail X$ describes (or rather, represents) the subset $f(A)$, which, as a set, is isomorphic to A via f. Note that the object A is entirely irrelevant here; a different function $g : A \rightarrowtail X$ with image distinct from f represents a different subset from f; and conversely, a function $h : B \rightarrowtail X$ that agrees with f will witness the same subset as f. (In fact, they will be isomorphic, and will reside in the same isomorphism class, as we'd might hope.)

This is what distinguishes, say, $\{1\}$ from $\{2\}$ in the context of being a subset of $X = \{1,2,3\}$; these subsets are distinct because the ways they embed into X do not admit a factorisation in both directions – being a subobject is not really a relation on objects, but rather on morphisms.

This definition also helps to resolve some problems in material set theories. For instance, one question that inevitably arises in material set theory, is to ask whether $\mathbb{Z}$ is a *really* subset of $\mathbb{R}$ or not. The former is a set of equivalence classes of von Neumann ordinals (or Zermelo ordinals, etc.), while the latter is a set of Dedekind cuts (or equivalence classes of Cauchy sequences, or elements of a complete ordered field, etc.), so it seems that $\mathbb{Z}$ cannot be a subset of $\mathbb{R}$ – but when integers and real numbers are interpreted in the conventional non-set-theoretic way, every integer is clearly a real number. (This is also very similar in flavour to our earlier question of whether $3 \in 17$ or not.)

The structural viewpoint says that it doesn't make sense to ask whether $\mathbb{Z} \subseteq \mathbb{R}$ or not because their *elements* are the same or not, but instead to ask if there is some *map* $\mathbb{Z} \to \mathbb{R}$ that *witnesses* that $\mathbb{Z} \subseteq \mathbb{R}$. Note that the maps $\mathbb{Z} \to \mathbb{R}$ also depend on the structure being considered, or rather, the ambient category – if $\mathbb{Z}$ and $\mathbb{R}$ are being considered as, say, groups or rings (i.e. objects in **Grp** or **Ring**, respectively), then we can only ask if there are any group or ring homomorphisms (i.e. morphisms in the respective category) that provide a reason for us to write $\mathbb{Z} \subseteq \mathbb{R}$, in the context of being a subgroup or subring.

Furthermore, this definition includes information about *how* an object embeds into another, rather than just *that it does* in some unspecified way. This definition of a subobject in more general categories of structured sets is also often more natural than the set-theoretic notion of having the underlying set of a structure be a subset of some larger structure (as well as generalising to non-concrete categories).

This does, however, have the somewhat odd side effect that, for example, $\{0,1\} \subseteq \mathbb{Z}$ is not the same as the subset $\{0,1\} \subseteq \mathbb{R}$ because one is an isomorphism class of monomorphisms into $\mathbb{Z}$, and the other of monomorphisms into $\mathbb{R}$. However, they are related in that we can lift the former isomorphism class to the latter with a suitable function $\mathbb{Z} \to \mathbb{R}$. This further reflects the structural idea that objects should be characterised by their connections to other objects – the ambient containing set is different, so these subsets hold different structures and are hence distinct entities.

Now, the $\leq$ relation on monomorphisms is only a preorder as it is not antisymmetric: $f \leq g$ and $g \leq f$

only imply that f and g are isomorphic and not strictly equal. However, since subobjects are isomorphism classes of monomorphisms, isomorphic morphisms do induce strictly equal subobjects, so the preorder $\leq$ induces a partial order $\subseteq_X$ on the subobjects of X. Again, we often pick representative monomorphisms A and B and abbreviate $[A] \subseteq_X [B]$ to just $A \subseteq_X B$. This notion of containment of subsets is inherently local to subsets of an ambient set X, unlike in a material set theory, where the subset relation, being defined in terms of the global membership relation, is compatible with any two arbitrary sets.

Importantly, the subset relation $\subseteq_X$ between *two subobjects* of X is distinct from the symbol $\subseteq$ indicating that a subobject belongs to a true *object* X. The former is a relation, local to the collection of subobjects of some given object, while the latter is just notation for a class of morphisms.

For the notion of membership on subsets, we say that *x is a member of a* and write $x \in a$ if $x \in X$, $a : S \rightarrowtail X$ is a subset of X, and there exists an element $\bar{x} \in S$ such that $a(\bar{x}) = x$:

$$\begin{tikzcd}
 & S \arrow[d, "a"] \\
1 \arrow[ur, "\bar{x}"] \arrow[r, "x"'] & X
\end{tikzcd}$$

That is, x is a member of a if x lifts through a. Again, membership for subsets is only defined locally within a containing set S, unlike in a material set theory, so it doesn't make sense to ask whether $x \in y$ or not for arbitrary sets x and y.

We can also dualise the notion of a subobject. The collection of epimorphisms from an object X are similarly preordered by factorisation – that is, we write $f \leq g$ for epimorphisms $f : X \twoheadrightarrow A$ and $g : X \twoheadrightarrow B$ if there exists a (necessarily unique and epic) morphism $h : A \to B$ such that $f = h \circ g$, and two epimorphisms are isomorphic if they factor through each other, or equivalently, if h is an isomorphism. A *quotient object* of X is then an isomorphism class of epimorphisms from X.

One important kind of subobject is given by the notion of an *image*. In **Set**, we can identify the image of a function $f : A \to B$ with a particular subset of B, namely, the subset consisting of all the elements of B of the form $f(a)$ for some $a \in A$. We can describe this situation more generally, without reference to elements, as follows.

The *image* of a morphism $f : A \to B$ is the minimal subobject of B through which f factorises universally into a composition $A \xrightarrow{e} \operatorname{im}(f) \overset{m}{\rightarrowtail} B$. That is, $f = m \circ e$, and for every other factorisation $A \xrightarrow{e'} S \overset{m'}{\rightarrowtail} B$, we have $\operatorname{im}(f) \subseteq_B S$. Then, the morphism $e : A \to \operatorname{im}(f)$ is called the *corestriction* of f.

Dually, the *coimage* of a morphism is the image of the corresponding morphism in the opposite category, or equivalently, the maximal quotient object of A through which f factors through universally.

52.6.1 The Subobject Classifier

For **Set** in particular, another way to characterise a subset A of a given set X is as a function $\chi_A : X \to 2$, where $2 = \{\top,\bot\}$, by taking χ_A to be the indicator function of A defined by

$$\chi_A(x) = \begin{cases} \top & x \in A \\ \bot & x \notin A \end{cases}$$

So, there is a bijection between the subobjects $A \rightarrowtail X$ and the functions $\chi_A : X \to 2$ given by $\chi_A \mapsto A = \chi_A^{-1}\left[\{\top\}\right]$. Now, recall that preimages are a special case of pullbacks, so this bijection says that for every subset $A \subseteq X$, there is a unique function $\chi : X \to 2$ such that

$$\begin{tikzcd}
A = f^{-1}\left[\{\top\}\right] \arrow[r, "!"] \arrow[d, hook] \arrow[dr, phantom, "\lrcorner", very near start] & 1 \arrow[d, hook, "\top"] \\
X \arrow[r, "\chi"'] & 2
\end{tikzcd}$$

is a pullback [Lei11]. Nothing here is really specific to **Set**, so we can abstract this diagram into any arbitrary category that admits pullbacks and has a terminal object to obtain the following definition:

A *subobject classifier* in a category $\mathscr{C}$ is an object Ω and map $\top : 1 \to \Omega$ such that every monomorphism $m : A \rightarrowtail X$ is the pullback of $\top$ along a unique morphism $\chi_m : X \to \Omega$ called the *characteristic morphism.*

That is, for every monomorphism $m : A \rightarrowtail X$, there exists a unique morphism $\chi_m : X \to \Omega$ such that

$$\begin{tikzcd}
A \arrow[r, "!"] \arrow[d, tail, "m"'] \arrow[dr, phantom, "\lrcorner", very near start] & 1 \arrow[d, "\top"] \\
X \arrow[r, "\chi_m"'] & \Omega
\end{tikzcd}$$

is a pullback square.

The object Ω is then called the *object of truth values*, a morphism $X \to \Omega$ a *truth value*, and the morphism $\top : 1 \to \Omega$ the truth value *true.*

In a concrete category, commutativity of the square, $\chi_m \circ m = \top \circ !$, intuitively means that χ_m is "true" everywhere over the image of m. The diagram being a pullback then means that A is the "largest" subobject of X with this property, so χ_m is true exactly in the image of A, and if we have any other object with a map into X that makes χ_m similarly true, then it will factor uniquely through m.

We give another characterisation of a subobject classifier [Lei11].

For any object $X \in \mathrm{ob}(\mathscr{C})$, we write $\mathrm{Sub}_{\mathscr{C}}(X)$, or just $\mathrm{Sub}(X)$, to denote the collection of subobjects of X. If this collection is a set (as opposed to a proper class) for every object in a category, then the category is called *well-powered.*

Suppose $\mathscr{C}$ has finite limits and is locally small. Then, every map $f : X \to Y$ in $\mathscr{C}$ induces a map $f^* : \mathrm{Sub}(Y) \to \mathrm{Sub}(X)$ between the subobject posets in the reverse direction by pullback. That is, if $g : B \rightarrowtail Y$ is a subobject in $\mathrm{Sub}(Y)$, then it is a monomorphism, and because pullbacks preserve monomorphisms, the pullback

$$\begin{tikzcd}
X \times_Y B \arrow[r, "\pi_2"] \arrow[d, tail, "\pi_1"'] \arrow[dr, phantom, "\lrcorner", very near start] & B \arrow[d, tail, "g"] \\
X \arrow[r, "f"'] & Y
\end{tikzcd}$$

of g along f is another monomorphism into X, which is just a subobject in $\mathrm{Sub}(X)$. This defines a functor $\mathrm{Sub} : \mathscr{C}^{\mathrm{op}} \to \mathbf{Set}$.

A subobject classifier is then exactly a representation of this functor.

Recall that for Sub to be representable, there must exist an object Ω such that $\mathrm{Sub}(X) \cong \hom_{\mathscr{C}}(X,\Omega)$ naturally in X. This intuitively corresponds to the previous idea that the subsets of X correspond to maps $X \to 2$ (this also implies that $\mathscr{C}$ is well-powered), and furthermore, that this correspondence is canonical. This holds similarly in arbitrary categories, in that the subobjects of an object X naturally correspond to morphisms $X \to \Omega$; hence the name subobject *classifier.*

Subobject posets also allow us to abstract various other familiar operations on subsets. For instance, the *intersection* of two subobjects with representing monomorphisms f and g in $\mathrm{Sub}(X)$ should be the maximal subobject that factors through both f and g, which is exactly their meet $f \wedge g$ in the order-theoretic sense – but meets are precisely the products in a thin category. Similarly, the *union* of the representatives f and g should intuitively be the minimal subobject that f and g both factor through, which is exactly their order-theoretic join, $f \vee g$.

Given (two representative monomorphisms of) two subobjects $A \rightarrowtail X$ and $B \rightarrowtail X$, we write $A \cap_X B$ or $A \wedge_X B$ for their intersection and $A \cup_X B$ or $A \vee_X B$ for their union. (When the object X is clear, the subscripts are often suppressed.)

These definitions are defined in terms of operations internal to the subobject poset, but it turns out that these intersections and unions may also be expressed externally in terms of limits in the ambient category $\mathcal{C}$.

Theorem 52.6.1. *The product of two subobjects $A \rightarrowtail X$ and $B \rightarrowtail X$ in $\mathrm{Sub}_{\mathscr{C}}(X)$ is given by their pullback in $\mathscr{C}$. Conversely, the coproduct in $\mathrm{Sub}_{\mathscr{C}}(X)$ is given by the image of the induced map $A \amalg B \to X$.*

Proof. Monomorphisms are stable under pullback, so the pullback of two objects is also a subobject in $\mathrm{Sub}(X)$. The universal property of the pullback then says that it factors through every other pair of monomorphisms into X, which is exactly a product in $\mathrm{Sub}(X)$.

Unfortunately, monomorphisms are not stable under pushouts, so the naïve pushout of two subobjects is generally not a subobject, and we instead have to pass through an image.

The detailed proof that this construction is valid is somewhat involved, and is omitted. However, we later show how to construct indexed unions, from which we may recover binary unions if desired. ■

The above correspondence between pullbacks in $\mathscr{C}$ and products in $\mathrm{Sub}(X)$ allow us to transport some properties of $\mathscr{C}$ into this thin subobject poset category: if $\mathscr{C}$ is finitely (co-,bi-)complete, then the collection of subobjects is not just a poset, but is furthermore a meet-semilattice (join-semilattice, lattice, respectively). Given some favourable conditions* which we will assume, these meets and joins also distribute over each other, so the collection of subobjects is additionally a distributive lattice.

A *Boolean category* is a category in which every subobject $A \rightarrowtail X$ has a *complement* subobject $B \rightarrowtail X$ such that $A \wedge B \cong \varnothing$ is initial in the subobject lattice, and $A \vee B \cong X$ – that is, the subobject lattice $\mathrm{Sub}(X)$ of any object X is a *Boolean* lattice.

52.6.2 Power Objects

We can also abstract the notion of a power set into categories other than **Set**. In set theory, the power set $\mathcal{P}(A)$, also written perhaps more suggestively as 2^A, of a set A is the set of all subsets of A. This is a definition reliant on the set-theoretic subset relation, which is not a categorical notion, so we want to find a way to characterise 2^A with the maps to or from it.

Also note that the notion of a power object is far more specific to **Set** than subobjects are. For instance, there isn't really a notion of a "power group", in that the collection of all subgroups of a group does not have group structure itself – collections of subobjects in this sense generally do not inherit the structure required to be an object in their own right.

* The category must be at least a *coherent category*, which we have not discussed, but all the categories we will see later will be coherent. In particular, topoi are always coherent.

The idea here is that the functions $B \to 2^A$ are naturally isomorphic to subobjects of $A \times B$:

$$\hom(B{,}2^A) \cong \mathrm{Sub}(A \times B)$$

Or equivalently, that the power set of a set A is exactly a representation of the functor $\mathrm{Sub}(A \times -)$. This is already a complete description of power sets that generalises to arbitrary categories, but we can again give a more concrete definition in terms of pullbacks.

Consider a set B, along with a function $f : B \to 2^A$ that maps elements of B to subsets of A. This induces a relation $R \subseteq A \times B$ that identifies which elements of A belong to the images of elements in B:

$$R = \big\{(a{,}b) \in A \times B : a \in f(b)\big\}$$

That is, aRb if and only if $a \in f(b)$. There is also the canonical relation $\in_A \subseteq A \times 2^A$ that identifies which elements of A belong to which subsets S of A:

$$\in_A = \big\{(a{,}S) \in A \times 2^A : a \in S\big\}$$

So $a \in_A S$ if and only if... $a \in S$.

Now, we can define a function $R \to \in_A$ by applying f to the second component of elements in R, so $(a{,}b) \in R$ if and only if $\big(a{,}f(b)\big) \in (\in_A)$. Then, we see that R is the preimage of $\in_A$ by $\mathrm{id}_A \times f$ (with the appropriate restrictions), so we have the following pullback:

$$\begin{tikzcd}
R = f^{-1}[\in_A] \arrow[r, "\mathrm{id}_A \times f\big|_R"] \arrow[d, hook] \arrow[dr, phantom, "\lrcorner", very near start] & \in_A \arrow[d, hook] \\
A \times B \arrow[r, "\mathrm{id}_A \times f"'] & A \times 2^A
\end{tikzcd}$$

Again, we can abstract this diagram into other categories.

A *power object* of an object A consists of an object Ω^A and a monomorphism $\in_A \rightarrowtail A \times \Omega^A$ such that for every other object B and monomorphism $m : R \rightarrowtail A \times B$, there exists a unique morphism $\chi_m : B \to \Omega^A$ such that

$$\begin{tikzcd}
R \arrow[r] \arrow[d, tail, "m"'] \arrow[dr, phantom, "\lrcorner", very near start] & \in_A \arrow[d, tail] \\
A \times B \arrow[r, "\mathrm{id}_A \times \chi_m"'] & A \times \Omega^A
\end{tikzcd}$$

is a pullback square.

Now, the notation 2^A suggests a connection between how we classified subsets of A with maps $A \to 2$ before, and in fact, this definition is compatible with the notion of a subobject classifier in that if $A \cong 1$ is terminal, then $1 \times \Omega^1 \cong \Omega$ and $\mathrm{id}_A \times \chi_m \cong \chi_m$, so the pullback reduces to the subobject classifier pullback square from before, and the power object of a terminal object is exactly the subobject classifier.

52.7 Monoidal Categories

A monoidal category is a category that has properties similar to an algebraic monoid; it is equipped with a binary endofunctor that satisfies the monoid axioms in a certain sense.

A *monoidal category* $(\mathscr{C}, \otimes, I, \alpha, \lambda, \rho)$ consists of:

- A category $\mathscr{C}$;
- A bifunctor $\otimes : \mathscr{C} \times \mathscr{C} \to \mathscr{C}$ called the *tensor product*, written in infix notation;
- A designated object I in $\mathscr{C}$ called the *unit*;
- A natural isomorphism $\alpha : \big((-) \otimes (-)\big) \otimes (-) \Rightarrow (-) \otimes \big((-) \otimes (-)\big)$ with components of the form $\alpha_{A,B,C} : (A \otimes B) \otimes C \to A \otimes (B \otimes C)$ called the *associator*;
- A natural isomorphism $\lambda : I \otimes (-) \Rightarrow (-)$ with components of the form $\lambda_A : (I \otimes A) \to A$ called the *left unitor*;
- A natural isomorphism $\rho : (-) \otimes I \Rightarrow (-)$ with components of the form $\rho_A : (A \otimes I) \to A$ called the *right unitor*;

subject to the *coherence conditions* that the following diagrams commute:

- the *triangle identity*:

$$\begin{tikzcd}
(A \otimes I) \otimes B \arrow[rr, "\alpha_{A,I,B}"] \arrow[dr, "\rho_A \otimes \mathrm{id}_B"'] & & A \otimes (I \otimes B) \arrow[dl, "\mathrm{id}_A \otimes \lambda_A"] \\
& A \otimes B &
\end{tikzcd}$$

- the *pentagon identity*:

$$\begin{tikzcd}
& (A \otimes B) \otimes (C \otimes D) \arrow[dr, "\alpha_{A,B,C \otimes D}"] & \\
((A \otimes B) \otimes C) \otimes D \arrow[ur, "\alpha_{A \otimes B,C,D}"] \arrow[d, "\alpha_{A,B,C} \otimes \mathrm{id}_D"'] & & A \otimes (B \otimes (C \otimes D)) \arrow[d, "\mathrm{id}_A \otimes \alpha_{B,C,D}"] \\
(A \otimes (B \otimes C)) \otimes D \arrow[rr, "\alpha_{A,B \otimes C,D}"'] & & A \otimes ((B \otimes C) \otimes D)
\end{tikzcd}$$

The tensor product being a bifunctor means that the category is closed with respect to the tensor product, while the left and right unitors say that $I \otimes A \cong A$ and $A \otimes I \cong A$ for any object A, so I acts as the identity of the tensor product. The associator then says that $(A \otimes B) \otimes C \cong A \otimes (B \otimes C)$ for all objects A, B, and C, so the tensor product is also associative. This is all the structure that a monoid demands, so, why do we have the additional coherence conditions?

The analogue of this expression in a monoidal category is an object of the form $(A \otimes I) \otimes B$; the right unitor guarantees that the unit I acts like the identity, giving $A \otimes I \cong A$, so $(A \otimes I) \otimes B \cong A \otimes B$, but again, we could use the associator to first rebracket $(A \otimes I) \otimes B \cong A \otimes (I \otimes B)$ before reducing with the left unitor. However, there's no reason why we should expect that these two orderings will produce exactly equal objects. The triangle identity is exactly the condition that this equality *does* hold, and similarly, the pentagon identity guarantees that every way we rebracket an expression yields isomorphic objects.

For similar reasons, monoidal categories also admit several notions of commutativity. A *braided monoidal category* is a monoidal category equipped with an additional natural isomorphism with components of the form $\beta_{A,B} : A \otimes B \to B \otimes A$ called the *braiding*, subject to two additional coherence conditions called the *hexagon identities* that ensure compatibility with the associator.

The tensor product in a braided monoidal category is then commutative in the sense that reversing the order of a tensor product yields isomorphic objects, as given by the braiding. However, applying the braiding twice may yield objects that are not strictly equal, but only isomorphic. If these objects *are* strictly equal – that is, $\beta_{A,B} \circ \beta_{B,A} = \mathrm{id}_{A\otimes B}$ – then the category is furthermore a *symmetric monoidal category*. The tensor product in a symmetric monoidal category is then "as commutative as possible".

One special case of a monoidal category is if the monoidal structure is given by the categorical product; such a category is called a *cartesian monoidal category*. Because categorical products are essentially unique, every cartesian monoidal category is necessarily symmetric monoidal.

Example. **Set** is monoidal, with the tensor product given by the categorical product, so **Set** is cartesian monoidal. Note that $(A \times B) \times C \neq A \times (B \times C)$, but there is an obvious isomorphism between them that we can use as the associator. Similarly, the unitors are given by the isomorphisms $1 \times A \cong A$ and $A \times 1 \cong A$. △

Categories can often be monoidal in multiple ways; for instance, **Set** is also monoidal if we take the tensor product to be the categorical coproduct (we say that **Set** is *cocartesian monoidal*). Again, we don't have strict equality here, with $(A \sqcup B) \sqcup C \neq A \sqcup (B \sqcup C)$, but there is again a canonical isomorphism between the two objects. The unitors are then given by the isomorphisms $\varnothing \sqcup A \cong A$ and $A \sqcup \varnothing \cong A$.

For another example, the category $\mathbf{Vect}_K$ of vector spaces over a field K is also monoidal in multiple ways, with the tensor product given by either the traditional tensor product, or the direct sum of vector spaces. In this case, these two structures are actually compatible in that the tensor product distributes over the direct sum, giving the category an additional semiring structure.

52.8 Internalisation

Recall the standard definition of a group:

> A *group* $(G,*)$ is a set G equipped with a binary operation $* : G \times G \to G$ that is associative, admits an identity element $e \in G$ (is *unitary*), and every element $g \in G$ has an inverse $g^{-1} \in G$ under $*$.

At this point, we should be used to viewing various mathematical constructions as morphisms, and we might be tempted to do the same here. The binary operation is already a function, and the identity element can, as usual, also be viewed as a function $e : 1 \to G$. Similarly, we have the function $(-)^{-1} : G \to G$ that sends an element to its inverse.

Now, because G is a set and e, $*$, and $(-)^{-1}$ are set functions, the associativity, identity, and inverse axioms can be entirely encoded by the requirement that certain diagrams in **Set** relating the three functions commute [nLa23b]:

- Unitality:

$$\begin{tikzcd}
G \times 1 \arrow[r, "\mathrm{id} \times e"] \arrow[dr, "\cong"'] & G \times G \arrow[d, "*"] & 1 \times G \arrow[l, "e \times \mathrm{id}"'] \arrow[dl, "\cong"] \\
 & G &
\end{tikzcd}$$

- Associativity:

$$\begin{tikzcd}
G \times G \times G \arrow[r, "\mathrm{id}\times *"] \arrow[d, "*\times\mathrm{id}"'] & G \times G \arrow[d, "*"] \\
G \times G \arrow[r, "*"'] & G
\end{tikzcd}$$

- Inverses:

$$\begin{tikzcd}
& & G \arrow[dll, "\exists!"'] \arrow[drr, "\Delta"] & & \\
1 \arrow[dr, "e"'] & & & & G \times G \arrow[dl, "(-)^{-1}\times\mathrm{id}"] \\
& G & & G \times G \arrow[ll, "*"] &
\end{tikzcd}
\qquad
\begin{tikzcd}
& & G \arrow[dll, "\exists!"'] \arrow[drr, "\Delta"] & & \\
1 \arrow[dr, "e"'] & & & & G \times G \arrow[dl, "\mathrm{id}\times(-)^{-1}"] \\
& G & & G \times G \arrow[ll, "*"] &
\end{tikzcd}$$

For instance, in **Set**, we can interpret the two paths in the left inverse diagram as the chains of functions $g \mapsto (g,g) \mapsto (g^{-1},g) \mapsto g^{-1} * g$ and $g \mapsto 1 \mapsto e$, so commutativity says that $g^{-1} * g = e$.

However, we should notice that this characterisation of groups does not explicitly refer to the elements within the group – all the requirements are now to do with how the group interacts with these three functions. Furthermore, the only structure of **Set** that is used in the above characterisation is the existence of finite products. This is not specific to **Set**, and indeed, there is no reason why this definition needs to be tied to **Set** at all; all the previous diagrams make sense in any arbitrary category that admits these limits, even if we cannot necessarily interpret G to be a set in that category.

An *internal group* in a category $\mathscr{C}$ that admits finite products is an object G equipped with morphisms $e : 1 \to G$ (where 1 is terminal in $\mathscr{C}$), $* : G \times G \to G$, and $(-)^{-1} : G \to G$ such that the diagrams above commute.

If $\mathscr{C}$ is **Set**, then we just have the definition of an ordinary group; if $\mathscr{C} = \mathbf{Top}$, we obtain topological groups; if $\mathscr{C} = \mathbf{Man}^{\infty}$, we obtain Lie groups; if $\mathscr{C} = \mathbf{Grp}$, we obtain abelian groups, etc.

This process of abstracting a structure like a group into an object or objects within a general category is called *internalisation* – and we can do this with many other constructions, creating internal monoids,[*] rings, lattices, etc. For instance, the subobject classifier in a Boolean category is exactly an internal Boolean algebra.

These diagrams can also be dualised to obtain so-called cogroups, comonoids, corings, etc. but these dual objects often do not correspond to any standard algebraic structures [For02], though cogroups do arise naturally in algebraic topology. For example, the n-sphere S^n is precisely a cogroup object in the homotopy category of pointed topological spaces, and is related to why the higher homotopy groups are in fact groups. Categories themselves can also be internalised within categories with sufficient pullbacks. For instance, small categories are precisely the categories internal to **Set**. In general, the more structure the ambient category has, the more that is able to be internalised.

We can abstract further and replace the products with tensor products to produce internal objects in general monoidal categories. For instance, an internal monoid in **Ab** with monoidal structure given by the tensor product $\otimes_{\mathbb{Z}}$, is precisely a ring(!); and an internal monoid in $\mathbf{Vect}_K$, with monoidal structure given by the tensor product $\otimes_K$ of vector spaces, is exactly a K-algebra, etc.

[*] Dropping the commutative diagram for inverses in the above definition of an internal group yields this particular construction.

52.8.1 Internal Homs

Clearly, internalisation is very useful, as it unifies many seemingly distinct constructions. But for now, we are interested in the internalisation of a categorical hom-set. We begin by abstracting the similar notion of a *function set.*

Given two sets A and B, the collection of functions from A to B form a set $[A,B]$, called the function set. We consider the functions into $[A,B]$ from another set X.

Such a function takes an argument from X, and returns a function $A \to B$. We can alternatively interpret this as a function that takes an argument from *both* X and A, and returns an element in B, so there is a bijection between functions $X \to [A,B]$ and $X \times A \to B$.*

This allows us to easily abstract a function set into any arbitrary category that admits products as follows: given a pair of objects A and B, the *internal hom-object*, or just *internal hom*, is an object $[A,B]$ such that

$$\hom(X,[A,B]) \cong \hom(X \times A,B)$$

naturally in X. This assignment of objects to internal hom objects is also functorial, defining the *internal hom-functor* $[-,-] : \mathscr{C}^{\mathrm{op}} \times \mathscr{C} \to \mathscr{C}$ that sends pairs of objects to their internal homs, much like the ordinary hom-bifunctor.

Replacing the product in the above with a tensor product, internal hom-functors further generalise to categories that may not admit products.

If a monoidal category admits all internal hom objects, it is called a *closed monoidal category.* More precisely, a monoidal category is closed monoidal if for every object A, the functor $(-) \otimes A : \mathscr{C} \to \mathscr{C}$ that sends objects to their right tensor product by A has a right adjoint, $[A,-] : \mathscr{C} \to \mathscr{C}$, that sends objects to the internal hom out from A. That is,

$$\hom(X,[A,B]) \cong \hom(X \otimes A,B)$$

naturally in all three variables. These categories are “closed” in the sense that taking hom-sets leaves you within the category.

If the category is further cartesian monoidal, then it is called a *cartesian closed category.*

Example. Any locally small category has at most a set of morphisms between any pair of objects – which is an object of **Set**. **Set** itself is locally small, so the hom-set between every pair of objects is just another object of **Set** (specifically, the function set between them), so **Set** has all internal homs, and is hence closed monoidal. **Set** is also cartesian monoidal, and so is furthermore cartesian closed. △

In cartesian closed categories, the internal hom $[A,B]$ is also written as B^A and is called an *exponential object.* This notation is compatible with the notation Ω^A for power objects (when they exist), as the power object of A is precisely the exponential object of A into the subobject classifier Ω, and this can be taken as an alternative definition of the power object. More generally, this notation is also compatible with the categorical product in that we have $A^1 \cong A$, $A^2 \cong A \times A$ (where $2 \coloneqq 1 \amalg 1$), etc. In more detail,

$$\begin{aligned}\hom(X,A^2) &\cong \hom(X \times (1 \amalg 1),A)\\ &\cong \hom((X \times 1) \amalg (X \times 1),A)\\ &\cong \hom(X \times 1,A) \times \hom(X \times 1,A)\\ &\cong \hom(X,A) \times \hom(X,A)\\ &\cong \hom(X,A \times A)\end{aligned}$$

so by the Yoneda lemma, $A^2 \cong A \times A$ (and so on, by induction).

* Particularly in computer science and formal logic, the reverse direction of this bijection is called *currying*, and is related to the notion of partial application.

This compatibility with categorical products allows for a more concise characterisation of cartesian closed categories: a category is cartesian closed if and only if it has finite products, and the right cartesian product functor $(-) \times A$ admits a right adjoint $(-)^A$ for every object A. That is, for every pair of objects A and B, there is an object B^A such that

$$\hom(X, B^A) \cong \hom(X \times A, B)$$

naturally in all three variables. The left adjunct $f^\flat : X \to B^A$ of a morphism $f : X \times A \to B$ is also called the *exponential transpose* of f, and similarly, the right adjunct is called the *exponential cotranspose*.

In the special case where $X \cong 1$ is terminal, this isomorphism becomes

$$\hom(1, B^A) \cong \hom(1 \times A, B) \cong \hom(A, B)$$

In other words, the elements of B^A (that is, the morphisms $1 \to B^A$) are naturally isomorphic to the morphisms $A \to B$, so the exponential object B^A can be thought of as the "object of morphisms $A \to B$", much like a function set in **Set**. Given a morphism $f : A \to B$, we write $[f]$ for its isomorphic copy or "label" in B^A.

Another interesting case is given by $X = B^A$, with the isomorphism then being:

$$\hom(B^A, B^A) \cong \hom(B^A \times A, B)$$

The image of the identity map on B^A is called the *evaluation map*, denoted by $\mathrm{ev} : B^A \times A \to B$. Concretely, in **Set**, or more generally on generalised elements, the evaluation map is given by evaluating a function $[f] \in B^A$ at a value $a \in A$; $([f], a) \mapsto f(a)$, hence the name.

The evaluation map also satisfies the universal property that given any object X and map $e : X \times A \to B$, there is a unique morphism $u : X \to B^A$ such that $\mathrm{ev} \circ (u \times \mathrm{id}_A) = e$:

$$\begin{tikzcd}
X \times A \arrow[d, dashed, "u \times \mathrm{id}_A"'] \arrow[dr, "e"] & \\
B^A \times A \arrow[r, "\mathrm{ev}"'] & B
\end{tikzcd}$$

52.9 ETCS

52.9.1 Topoi

The notion of a *topos* (plural *topoi*) was first introduced in algebraic geometry by Grothendieck in the early 1960s as a generalisation of sheaves of sets in topology. Every topological space induces a topos, and conversely, every topos, as defined by Grothendieck, behaves much like a generalised topological space. These topoi are called *Grothendieck topoi*, and are now prevalent in modern algebraic geometry.

A more general notion of a topos was soon developed by Lawvere and Tierny over the same decade, which we will now introduce.

An (*elementary* or *Lawvere–Tierny*) *topos* is a category that:

- is finitely complete;
- is cartesian closed;
- has a subobject classifier.

This definition seems rather short for a structure we claim to be so important – and it is, deceptively so; a topos carries a vast amount of additional rich structure that just happens to follow from these few axioms. We give a few basic properties of topoi.

Lemma 52.9.1. *Every monomorphism in a topos is regular. That is, every monomorphism occurs as the equaliser of some pair of parallel morphisms.*

Proof. Let $m : X \to Y$ be a monomorphism. Then, it is a subobject of Y, so it is classified by the unique map $\chi_m : X \to \Omega$ that makes the following diagram a pullback square:

$$\begin{tikzcd}
X \arrow[r, "!_A"] \arrow[d, tail, "m"'] \arrow[dr, phantom, "\lrcorner", very near start] & 1 \arrow[d, "\top"] \\
Y \arrow[r, "\chi_m"'] & \Omega
\end{tikzcd}$$

We claim that m is the equaliser of χ_m and $\top \circ !_Y$ (where $!_Y$ is the unique map $Y \to 1$): Let $f : A \to Y$ also equalise χ_m and $\top \circ !_Y$, i.e. $\chi_m \circ f = \top \circ !_Y \circ f$. Our maps are now:

$$\begin{tikzcd}
A \arrow[dr, dashed, "g"] \arrow[drr, bend left, "!_A"] \arrow[ddr, bend right, "f"'] & & \\
& X \arrow[r, "!_X"] \arrow[d, tail, "m"] \arrow[dr, phantom, "\lrcorner", very near start] & 1 \arrow[d, "\top"] \\
& Y \arrow[ur, "!_Y"] \arrow[r, "\chi_m"'] & \Omega
\end{tikzcd}$$

Because 1 is terminal, $!_Y \circ f =!_A$, so we have $\chi_m \circ f = \top \circ !_A$ and the outer square commutes, so the universal property of the pullback yields a unique map $g : A \to X$ making the whole diagram commute.

By commutativity of the left triangle, $f = m \circ g$, and hence

$$\begin{tikzcd}
A \arrow[d, dashed, "g"] \arrow[dr, "f"] & & \\
X \arrow[r, tail, "m"'] & Y \arrow[r, shift left, "\chi_m"] \arrow[r, shift right, "\top \circ !_Y"'] & \Omega
\end{tikzcd}$$

commutes, so m is an equaliser, as required. ■

Corollary 52.9.1.1. *Topoi are balanced. That is, every bimorphism is an isomorphism.*

Proof. From the previous lemma, every bimorphism in a topos is an epic regular monomorphism.

Let $f : E \to A$ be an epic regular monomorphism. As f is a regular monomorphism, it is the equaliser of a pair of parallel morphisms g,$h : A \to B$, so we have $g \circ f = h \circ f$. Since f is epic, we have $g = h$.

The equaliser of $g = h$ is the identity map, and by the universal property of the equaliser, E must factor through A essentially uniquely, so $f : E \to A$ must be this isomorphism. ■

Despite starting with only a finitely complete category, the subobject classifier and cartesian closed structure together also imply that a topos also has all finite colimits:

Theorem 52.9.2. *[Par74] Every topos is finitely cocomplete.*

We also have a result stating that morphisms in a topos satisfy a epi-mono factorisation structure:

Theorem 52.9.3 (Image Factorisation)**.** *In a topos, every arrow factors essentially uniquely through its image into the composition of an epimorphism with a monomorphism.*

Proof. [MM12] Let $f : A \to B$ be a morphism. We construct the following diagram in stages.

$$\begin{tikzcd}
A \arrow[r, dashed, "e"'] \arrow[rr, bend left, "f"] \arrow[d, equal] & M \arrow[r, dashed, tail, "m"'] & B \arrow[r, shift left, "x"] \arrow[r, shift right, "y"'] \arrow[d, equal] & X \arrow[d, dashed, "u"] \\
A \arrow[r, "g"'] & N \arrow[r, tail, "h"'] & B \arrow[r, shift left, "s"] \arrow[r, shift right, "t"'] & Y
\end{tikzcd}$$

First, construct the cokernel pair $x{,}y : B \to X$ of f, and let $m : M \to B$ be the equaliser of x and y. By the universal property of the equaliser, f factors uniquely through the equaliser m, so $f = m \circ e$ for some $e : A \to M$. Note that, as an equaliser, m is monic.

Now, suppose f also factorises as $f = h \circ g$ with h monic. As every monomorphism in a topos is regular, h is the equaliser of some pair of morphisms $s{,}t : B \to Y$, so we have $s \circ h = t \circ h$, and precomposing by g yields $s \circ f = t \circ f$. Then, because $x{,}y$ is the cokernel pair of f, X factors through Y via a unique map $u : X \to Y$, giving

$$\begin{aligned} s \circ m &= u \circ x \circ m \\ &= u \circ y \circ m \\ &= t \circ m \end{aligned}$$

so m also equalises s and t and therefore factors uniquely through h. As h is arbitrary, we have that m is the minimal subobject of B and hence $M = \mathrm{im}(f)$. It remains to show that e is epic.

Perform this construction again on e to obtain the chain

$$\begin{tikzcd}
A \arrow[r, "e'"] & \mathrm{im}(e) \arrow[r, tail, "m'"] & \mathrm{im}(f) \arrow[r, tail, "m"] & B
\end{tikzcd}$$

equal to f. In particular, f factors through the monomorphism (subobject) $m \circ m'$, so the image also factors through $m \circ m'$, as it is the minimal subobject, so $m = (m \circ m') \circ v$ for some $v : M \to \mathrm{im}(e)$. It follows that $m' \circ v = \mathrm{id}_M$, so m' is an isomorphism.

As before, m' is the equaliser of the cokernel pair x', y' of e. But, because m' is an isomorphism, we have $x' = y'$, so the cokernel pair of e is x', x', and hence e is epic, as required. ■

The prototypical example of a topos is **Set**, but **Set** also has a couple of special properties it does not share with most other topoi, which we will explore soon. On the other hand, the existence of terminal objects allow us to consider elements of objects in arbitrary topoi; the subobject classifier Ω allows us to consider subobjects; and exponentials allow us to consider objects of morphisms from one object to another, as well as power objects in the form of exponentials of the subobject classifier; so, along with finite completeness and the cartesian closed structure, arbitrary topoi behave in many ways like **Set**.

In particular, this means that almost all categorical constructions in **Set** can readily be internalised in an arbitrary topos, and many theorems about these constructions similarly apply to their internalised variants. In this sense, topoi are just a kind of well-behaved generalised space in which objects behave “like sets”.

52.9.2 Set

We give some characteristics of **Set** that distinguish it from other topoi [Lei11], appealing only to "obvious" properties that sets and functions should satisfy.

Firstly, **Set** is non-trivial: that is, **Set** $\not\simeq \mathbb{1}$. Another way to characterise this property is that the terminal and inital objects of **Set** both exist, and are not isomorphic:

(i) $0 \not\cong 1$.

In **Set**, the terminal object 1 also has another special property: if the parallel morphisms $f,g : X \to Y$ are such that every map $x : 1 \to X$ equalises f and g, then $f = g$.

More generally, an object S is called a *separator* or is said to *separate morphisms* if for every pair of parallel morphisms $f,g : X \to Y$, if $f \circ s = g \circ s$ for every $s : S \to X$, then $f = g$. So, if a category admits a separator, then just by looking at (compositions with) the generalised elements of shape S, we can distinguish all morphisms in that category.

The next property of **Set** is then:

(ii) The terminal object 1 is a separator.

In **Set**, however, this has an additional important interpretation: recall that maps $1 \to X$ are elements of X, so, given a function $f : X \to Y$, the composition $f \circ x : 1 \to Y$ is an element of Y, which we might choose to write as $f(x)$. Thus, not only are elements a special case of functions, but evaluation of functions is a special case of composition. The property above then says that if $f(x) = g(x)$ for all x, then f and g are the same function – this is saying that functions have no internal identity, and are completely defined by their effects on elements (and implicitly, the data of their (co)domains). This property is similar to the axiom of extensionality in ZFC, but for functions instead of sets.

A topos that satisfies properties (i) and (ii) is called a *well-pointed* topos.

The next property quite specific to **Set** is, roughly speaking, the existence of the natural numbers. To state this more formally, we need to characterise the natural numbers categorically. One feature of the natural numbers that we often use, particularly with induction, is that they support recursive definitions.

Given a set X, and an element $x \in X$, every function $r : X \to X$ generates a unique sequence of elements $(x_i)_{i=1}^{\infty} \subseteq X$ such that $x_0 = x$ and $x_{n+1} = r(x_n)$. Note that such a sequence is indexed by the natural numbers, so this yields a correspondence between applications of r to x, and applications of the successor function to 0 in the subscripts. Moreover, a sequence in X is just a generalised element of shape $\mathbb{N}$, or a morphism $f : \mathbb{N} \to X$, so this is really a statement about the natural numbers. If we write $s : \mathbb{N} \to \mathbb{N}$ for the successor function, then the previous just says that the following diagram commutes:

$$\begin{tikzcd}
 & \mathbb{N} \arrow[r, "s"] \arrow[d, "f"] & \mathbb{N} \arrow[d, "f"] \\
1 \arrow[ur, "0"] \arrow[dr, "x"'] & & \\
 & X \arrow[r, "r"'] & X
\end{tikzcd}$$

where 1 is terminal.

A *natural numbers object* in a category $\mathscr{C}$ is a triple $(\mathbb{N},0,s)$ consisting of an object $\mathbb{N} \in \mathrm{ob}(\mathscr{C})$, a morphism $0 : 1 \to \mathbb{N}$ from the terminal object 1, and a *successor morphism* $s : \mathbb{N} \to \mathbb{N}$ with the universal property that all other similar triples (X,x,r) factor through $(\mathbb{N},0,s)$ uniquely. That is, there exists a unique morphism $f : \mathbb{N} \to X$ such that the previous diagram commutes. This universal property also means that natural numbers objects are essentially unique, so we are safe to speak about *the* natural numbers.

The sequence f given by this axiom is said to be defined by *simple recursion* with *starting value* x and *transition rule* r.

Arithmetic operations $\mathbb{N} \times \mathbb{N} \to \mathbb{N}$, such as addition, multiplication, exponentiation, etc. can then be defined in terms of their exponential transpositions $\mathbb{N} \to \mathbb{N}^{\mathbb{N}}$ by simple recursion. For instance, the following diagram defines addition of natural numbers:

$$\begin{tikzcd}
& \mathbb{N} \arrow[r, "s"] \arrow[d, "+^\flat"] & \mathbb{N} \arrow[d, "+^\flat"] \\
1 \arrow[ur, "0"] \arrow[dr, "\mathrm{id}_{\mathbb{N}}^\flat"'] & & \\
& \mathbb{N}^{\mathbb{N}} \arrow[r, "s^{\mathbb{N}}"'] & \mathbb{N}^{\mathbb{N}}
\end{tikzcd}$$

[Kos12, adapted]

Commutativity of the left triangle says that adding zero is the identity function, so $0 + n = n$, and commutativity of the square on the right says $(s \circ n) + m = s \circ (n + m)$, which is precisely the Peano definition of addition.

The third distinguishing property of **Set** is then:

(iii) **Set** has a natural numbers object.

The last special property of **Set** that we will need is that every surjective function $f : A \twoheadrightarrow B$ has a right section – a function $s : B \to A$ such that $f \circ s = \mathrm{id}_B$. This can be stated in categorical terms as:

(iv) Epimorphisms split.

The function s is defined by assigning each element $b \in B$ an element from $f^{-1}[b]$, (which is non-empty as f is surjective). However, this implies the existence of a choice function for any arbitrary f and thus, the statement that every epimorphism splits is precisely the Axiom of Choice. More generally, a category is said to *satisfy the Axiom of Choice*, or to *have Choice*, if every epimorphism splits.

So, the content of this section can be stated concisely as:

Sets and set functions form a well-pointed topos
with natural numbers object and Choice.

The category of sets of course has more properties than this; for instance, power objects always exist; the category is balanced; the subobject classifier has two objects, $\Omega \cong 2 := 1 \amalg 1$; and the topos is Boolean, etc. but these properties all follow from the previous conditions.

The question is now, what conditions do we need to enforce on sets and set functions to ensure that they *do* form such a topos?

One answer is of course, ZFC – and indeed, any model of ZFC will satisfy these properties, so the category of ZFC sets will satisfy the above.

This is the answer many mathematicians recognise and know in the back of their mind, but often do not like to concern themselves with, because the axioms of ZFC are generally quite far removed from the study of mathematics – the specific axioms seem unimportant compared to the need for the end result to satisfy these requirements.

This answer is thus rather unsatisfying, or even irrelevant, because all of the above requirements were derived from "obvious" properties of sets that we often use – extensionality of functions, existence of natural numbers, etc. At no point did we have to consult with a list of axioms to decide these properties, because they all follow from our informal idea of what sets should be and how set functions should behave.

In particular, this means that anything that satisfies the above requirements will behave *like a set.* This is the idea behind the alternative answer given by Lawvere in his *Elementary Theory of the Category of Sets*, or ETCS: we take these properties *as our axioms.* That is, we do not require that sets satisfy the axioms of ZFC, but instead require that sets and set functions form a well-pointed topos with natural numbers object and Choice.

At this point, one may think that there is some circularity here: that ETCS depends on the notion of a category, which itself depends on the notion of "collections" of objects and morphisms, which seem quite similar to "sets".

The straightforward formalist response is that category theory (and specialisations thereof, like ETCS) and ZFC are *first-order theories*, so they are all just collections of sentences in the first-order language over some signature – at a fully formalised level, none of these theories mention or depend on any prior notion of sets, because they are just alphabets of symbols, together with lists of axioms.

However, outside of formal logic, we usually don't think of theories in this way – as manipulations of strings of abstract symbols – but instead as descriptions of some universe of interest. This answer may thus be somewhat unsatisfactory in that it doesn't answer the question intuitively, so an alternative explanation for ETCS in particular is that, although motivated by category-theoretic ideas, ETCS does not intrinsically depend on the notion of a category – category theory is just a convenient language with which we can express the axioms of ETCS concisely. It is certainly possible to state the axioms of ETCS without mentioning categories at all.

For reference, the axioms stated without mentioning categories are, informally,

1. Function composition is associative and has identities
2. There exists an empty set
3. There exists a set with one element
4. Functions are completely characterised by their actions on elements
5. Given sets X and Y, we may form the Cartesian product $X \times Y$
6. Given sets X and Y, we may form the set Y^X of functions from X to Y
7. Given a function $f : X \to Y$ and $y \in Y$, we may form the fibre $f^{-1}[y]$
8. The subsets of a set X correspond to the functions $X \to \{0,1\}$
9. The natural numbers form a set
10. Every surjection admits a section

Stated in this way, the comparison with ZFC is now more obvious: ZFC says "there are things called sets; there is a binary relation $\in$ defined on sets; and some axioms hold.", and ETCS says "there are things called sets; for every pair of sets there are things called functions; there is a (partial) binary operation $\circ$ on functions called composition; and some axioms hold." In neither case do we specify what these "things" are, nor do we suppose that these "things" form any structure like a set or category beyond what the axioms demand. The point is, circularity is no more of a problem for ETCS than it is for ZFC.

As noted in [Lei12], the axioms of ETCS as stated above also appear to be more *fundamental* in some way than ZFC. Suppose that one day, we find that ZFC had been proved to be inconsistent: that some logician had started with the axioms of ZFC, and had irrefutably derived a logical contradiction from them. It is likely that most mathematicians, being generally detached from ZFC in the first place, would not be deeply bothered by this fact, and could continue on, generally confident that their theorems and results still hold true in the sense that their negations do not.

In contrast, the axioms of ETCS are modelled on core properties of sets and functions that we often use – a proof that ETCS were inconsistent would be devastating. We would no longer be able to safely assume that function composition is associative, that products or function sets exist, etc.

As an aside, note that in this paper, we are describing ETCS as a *two-sorted* first-order theory, roughly meaning that we have two distinct "kinds" of things – namely, objects (sets) and morphisms (functions).

Many introductions to logic, however, only discuss single-sorted theories (like ZFC, or fragments thereof), so some may object to this usage of a two-sorted theory. Fortunately, it is in fact possible to express ETCS as a single-sorted theory, where objects in the sense of the two-sorted theory are just a special type of morphism in the single-sorted theory. Specifically, objects are in bijection with identity morphisms, so they can be treated as a special case of morphisms. We then just add a source, target, and composition predicate to our underlying logic to obtain the desired single-sorted theory.

While we will not be discussing this style of axiomatisation, it is interesting that the primitive notion of this theory is not of sets, as in ZFC, but of functions – this is yet another illustration of the structural idea that connections are more important than objects.

52.9.3 Constructing the Universe

We prove some standard set theory machinery used for constructing common mathematical objects and the rest of the set-theoretic universe. Most of the proofs in this section have been roughly adapted from [LC05], with most modifications due to differences in definitions and conventions. (In particular, the ETCS axioms are very different.)

We have already constructed (local) intersections and unions as meets and joins in the subobject lattice, but we would like to extend this to indexed families of sets. In fact, we can prove (a structural version of) ZFC's axiom of the union in ETCS.

First, recall that a subobject $m : A \rightarrowtail X$ is classified by a characteristic morphism $\chi_m : X \to \Omega$ such that

$$\begin{tikzcd}
A \arrow[r, "!"] \arrow[d, tail, "m"'] \arrow[dr, phantom, "\lrcorner", very near start] & 1 \arrow[d, "\top"] \\
X \arrow[r, "\chi_m"'] & \Omega
\end{tikzcd}$$

is a pullback square. Then, given a subset $\alpha : I \rightarrowtail \Omega^X$ of the power object of X, i.e., an indexed family of subsets of X, the union of the α_i is then a subset $a : \bigcup_\alpha \to X$ of X such that for any $x \in X$, there exists an index $i \in I$ such that the function labelled by $\alpha(i)$ sends x to $\top \in \Omega$. That is,

$$x \in a \leftrightarrow \exists i \in I : \mathrm{ev}_{\Omega^X}(\alpha(i), x) = \top$$

where ev_{Ω^X} is the evaluation map on Ω^X, and $\top : 1 \to \Omega$ is the truth value true.

Theorem 52.9.4 (Indexed Unions)**.** *Given $\alpha : I \rightarrowtail \Omega^X$, the union $\bigcup_\alpha$ as defined above exists.*

Proof. Taking the exponential transpose, $\alpha^\flat : I \times X \to \Omega$, the desired property of $\bigcup_\alpha$ simplifies to:

$$x \in a \leftrightarrow \exists i \in I : \alpha^\flat(i, x) = \top$$

First, construct the equaliser of $\alpha^\flat$ and $\top \circ !_{I \times X}$, and consider the projection $\pi_2 : I \times X \to X$:

$$\begin{tikzcd}
E \arrow[r, "k"] \arrow[d, dashed, two heads, "q"'] & I \times X \arrow[r, shift left, "\alpha^\flat"] \arrow[r, shift right, "\top\circ!_{I\times X}"'] \arrow[d, "\pi_2"] & \Omega \\
\bigcup_\alpha \arrow[r, dashed, tail, "a"'] & X &
\end{tikzcd}$$

By image factorisation (Theorem 52.9.3), the composition $\pi_2 \circ k : E \to X$ factors through its image (essentially uniquely) into the epimorphism q and monomorphism a.

Let $x \in a$, i.e., x is a function $1 \to X$ and there exists a lift $\bar{x} \in \bigcup_\alpha$ of x. Then, since q is an epimorphism, there exists $\hat{x} \in E$ such that $q(\hat{x}) = \bar{x}$, so we have $x = a \circ q \circ \hat{x} = \pi_2 \circ k \circ \hat{x}$.

$$\begin{tikzcd}
 & & I & & \\
 & E \arrow[r, "k"] \arrow[d, dashed, two heads, "q"'] & I \times X \arrow[u, "\pi_1"] \arrow[d, "\pi_2"] \arrow[r, "!_{I \times X}"'] \arrow[rr, bend left, "\alpha^\flat"] & 1 \arrow[r, "\top"'] & \Omega \\
 & \bigcup_\alpha \arrow[r, dashed, tail, "a"'] & X & & \\
1 \arrow[uur, bend left, "\hat{x}"] \arrow[ur, "\bar{x}"'] \arrow[urr, bend right, "x"'] & & & &
\end{tikzcd}$$

Now, define $i := \pi_1 \circ k \circ \hat{x}$. Then,

$$\begin{aligned}
\alpha^\flat(i,x) &= \alpha^\flat(\pi_1 \circ k \circ \hat{x}, \pi_2 \circ k \circ \hat{x}) \\
&= \alpha^\flat \circ k \circ \hat{x} \\
&= \top \circ !_{I \times X} \circ k \circ \hat{x} \\
&= \top
\end{aligned}$$

since $!_{I \times X} \circ k \circ \hat{x} : 1 \to 1$ is necesssarily the identity. So, if $x \in a$, then there exists $i := \pi_1 \circ k \circ \hat{x} \in I$ such that $\alpha^\flat(i,x) = \top$, as required.

For the reverse implication, suppose there exists an index $i \in I$ such that $\alpha^\flat(i,x) = \top$. Then, by the universal property of the equaliser, there is some $\hat{x} \in E$ such that $k(\hat{x}) = (i,x)$. Then, applying q to $\hat{x}$, we have $x \in a$, as required. ■

Next, we prove that primitive recursion can be performed in ETCS, allowing the construction of a vast class of important functions. For instance, multiplication, division, the factorial function, the exponential function, and the function that returns the nth prime are all primitive recursive; in fact, most of the computable functions encountered in mathematics are primitive recursive.

Theorem 52.9.5 (Primitive Recursion)**.** *Given a pair of morphisms $f_0 : A \to B$ and $u : \mathbb{N} \times A \times B \to B$, there is a unique morphism $f : \mathbb{N} \times A \to B$ such that for all $n \in \mathbb{N}$ and $a \in A$,*

- *$f(0,a) = f_0(a)$;*
- *$f\big(s(n),a\big) = u\big(n,a,f(n,a)\big)$.*

Proof. Primitive recursion is more complicated than the simple recursion given by the natural numbers object in two main ways.

Firstly, the values of f depend not only on $n \in \mathbb{N}$, but also on the members a of the parameter object A. To simplify, we instead find the exponential transpose of f. That is, the function $f^\flat : \mathbb{N} \to B^A$ such that

- $f^\flat(0) = [f_0]$;
- For all $n \in \mathbb{N}$ and $a \in A$, $\mathrm{ev}\big(y(s(n)),a\big) = u\big(n,a,\mathrm{ev}(f^\flat(n),a)\big)$, where ev is the evaluation map for B^A.

The next complication is that the transition rule now depends not only on the value of the previous step, but also on the number n of previous steps that have already been taken. For this, we instead construct a sequence $F : \mathbb{N} \to \mathbb{N} \times B^A$ of ordered pairs where the first coordinate is just used to track the number of previous steps.

In other words, we define by simple recursion the graph of $f^\flat$, and from there, we can then recover $f^\flat$ by composing with a projection map. Explicitly, we require

- $F(0) = \big\langle 0, f^\flat(0)\big\rangle = \big\langle 0, [f_0]\big\rangle$;
- For all $n \in \mathbb{N}$ and $a \in A$,
  - *(i)* $(\pi_1 \circ F)\big(s(n)\big) = s(n)$;
  - *(ii)* $\mathrm{ev}\big((\pi_2 \circ F)(s(n)), a\big) = u\big(n, a, \mathrm{ev}\big((\pi_2 \circ F)(n), a\big)\big)$.

By the universal property of the natural numbers object, the existence of F can be given by finding a map $r : \mathbb{N} \times B^A \to \mathbb{N} \times B^A$ such that for any $n \in \mathbb{N}$, $[h] \in B^A$, and $a \in A$,

(i) $(\pi_1 \circ r)(n, [h]) = s(n)$;

(ii) $\mathrm{ev}\big((\pi_2 \circ r)(n, [h]), a\big) = u\big(n, a, \mathrm{ev}([h], a)\big)$.

We claim that such a map is given by $r := \langle s \circ \pi_1, G^\flat\rangle$, where $G^\flat$ is the exponential transpose of the map $G : A \times \mathbb{N} \times B^A \to B$ defined by the chain:

$$A \times \mathbb{N} \times B^A \xrightarrow{\Delta_A \times \mathrm{id}_{\mathbb{N}} \times \mathrm{id}_{B^A}} (A \times A) \times \mathbb{N} \times B^A \overset{\substack{\text{associator } \alpha, \\ \text{braiding } \beta}}{\cong} \mathbb{N} \times A \times (B^A \times A) \xrightarrow{\mathrm{id}_{\mathbb{N}} \times \mathrm{id}_A \times \mathrm{ev}} \mathbb{N} \times A \times B \xrightarrow{u} B$$

(Multiple associator components have been omitted to significantly simplify the chain.)

We verify that r satisfies the two desired properties:

(i)

$$\begin{aligned} (\pi_1 \circ r)(n, [h]) &= (s \circ \pi_1)(n, h) \\ &= s\big(\pi_1(n, h)\big) \\ &= s(n) \end{aligned}$$

(ii)

$$\begin{aligned} \mathrm{ev}\big((\pi_2 \circ r)(n, [h]), a\big) &= \mathrm{ev}(G^\flat(n, [h]), a) \\ &= G(a, n, [h]) \\ &= \big(u \circ (\mathrm{id}_{\mathbb{N}} \times \mathrm{id}_A \times \mathrm{ev}) \circ (\overset{\alpha;\beta}{\ldots}) \circ (\Delta_A \times \mathrm{id}_{\mathbb{N}} \times \mathrm{id}_{B^A})\big)(a, n, [h]) \\ &= \big(u \circ (\mathrm{id}_{\mathbb{N}} \times \mathrm{id}_A \times \mathrm{ev}) \circ (\overset{\alpha;\beta}{\ldots})\big)\big((a, a), n, [h]\big) \\ &= \big(u \circ (\mathrm{id}_{\mathbb{N}} \times \mathrm{id}_A \times \mathrm{ev})\big)\big(n, a, ([h], a)\big) \\ &= u\big(n, a, \mathrm{ev}([h], a)\big) \end{aligned}$$

as required.

Finally, uniqueness is given by function extensionality. ■

Note that existence only depends on finite completeness, exponentials, and the existence of the natural numbers object, and uniqueness on well-pointedness, so primitive recursion can actually be performed in much more general categories than just those that are models of ETCS; for instance, the functor category $[\mathscr{C}, \mathbf{Set}]$ for any small category $\mathscr{C}$.

Next, we verify that the natural numbers object in fact behaves as we would like:

Theorem 52.9.6 (Peano Postulates)**.** *The natural numbers object* $(\mathbb{N}, 0, s)$ *satisfies the Peano postulates. That is,* (*i*) *The successor function* $s : \mathbb{N} \to \mathbb{N}$ *is monic;*

(*ii*) *If* $m = s(n)$ *for some* $n \in \mathbb{N}$*, then* $m \neq 0$*;*

(*iii*) *If* $S \subseteq \mathbb{N}$ *and for all* $n \in \mathbb{N} : n \in S \to s(n) \in S$*, then* $S = \mathrm{id}_{\mathbb{N}}$*.*

Proof.

(*i*) The predecessor function p can be defined by primitive recursion with $f_0(-) = 0$ and $u(n,-,-) = n$. The parameter object A isn't actually used here, so suppressing it from the arguments, primitive recursion gives

- $p(0) = 0$;
- $p(s(n)) = n$.

Since s has a left-inverse, it is injective and hence monic.

(*ii*) Suppose $0 = s(n)$ for some $n \in \mathbb{N}$. Then,

$$\begin{aligned} n &= p\big(s(n)\big) \\ &= p(0) \\ &= 0 \end{aligned}$$

so $s(0) = 0$.

Let X be some set, $x \in X$ an element, and $r : X \to X$ an endofunction on X, and let $f : \mathbb{N} \to X$ be the unique morphism given by the universal property of the natural numbers object. As $\mathbb{N}$ factors through X, we have $f(0) = x$ and $r \circ f = f \circ s$, so,

$$\begin{aligned} r(x) &= (r \circ f)(0) \\ &= (f \circ s)(0) \\ &= f(0) \\ &= x \end{aligned}$$

Since $x \in X$ is arbitrary, $r = \mathrm{id}_X$ (as $x : 1 \to X$ equalises r and id_X and 1 is a separator), and since X is arbitrary, this implies that every endofunction is an identity function, which is a contradiction.

(*iii*) Let $S : A \rightarrowtail \mathbb{N}$ be a (representing monomorphism of a) subset of $\mathbb{N}$ with $0 \in S$, i.e., there exists $\bar{0} \in A$ such that $S(\bar{0}) = 0$; and such that $\forall n \in \mathbb{N} : n \in S \to s(n) \in S$.

The latter requirement implies that S is contained in its preimage $s^{-1}[S]$, so the map sending the lift to n to the lift of $s(n)$ is total, and extends to a morphism $t : A \to A$, satisfying $S \circ t = s \circ S$.

By simple recursion, $\bar{0}$ and t define a unique sequence $f : \mathbb{N} \to A$ such that $f(0) = \bar{0}$ and $(f \circ s)(n) = (t \circ f)(n)$ for all $n \in \mathbb{N}$.

Then, we have $(S \circ f)(0) = S(\bar{0}) = 0$ and

$$\begin{aligned} (S \circ f) \circ s &= S \circ (f \circ s) \\ &= S \circ (t \circ f) \\ &= (S \circ t) \circ f \\ &= (s \circ S) \circ f \\ &= s \circ (S \circ f) \end{aligned}$$

The identity $\mathrm{id}_{\mathbb{N}}$ also satisfies these equations, so $S \circ f = \mathrm{id}_{\mathbb{N}}$ by uniqueness of the map given by recursion. Then, for any $n \in \mathbb{N}$, we have $n = (S \circ f)(n) = S\big(f(n)\big)$, so $f(n)$ is a lift of n and hence $n \in S$.

■

In particular, this third point allows us to perform induction in any model of ETCS.

At this point, we have now developed sufficient machinery to construct much of modern set theory – Cantor's theorem, the Cantor–Schröder–Bernstein theorem, Zorn's lemma, etc. – as well as embedding the rest of mathematics into sets. Most of the set-theoretic universe at this point is constructed similarly to ZFC, constructing new sets by taking products and quotients of existing sets.

We end with a metatheorem that quantifies how strongly the axioms of ETCS characterise its models:

Theorem 52.9.7. *[New14] If $\mathscr{C}$ and $\mathcal{S}$ are models of* ETCS*, then $\mathscr{C} \simeq \mathcal{S}$, with the equivalence given by the adjoint functors $T \dashv \hom_{\mathcal{S}}(1,-)$, where*

$$T(X) \coloneqq \coprod_{x \in X} 1$$

52.10 Discussion

52.10.1 Relative Strength

ETCS is slightly weaker than ZFC, in the sense that there are statements provable in ZFC that are not provable in ETCS, but only to a slight extent, as these statements are generally beyond the interest of even most researching mathematicians (outside of those studying set theory/model theory/foundations). Undergraduate mathematics in particular (again, outside of a course on ZFC) also assumes no properties of sets beyond ETCS, so it would seem that ETCS is more than sufficient for most practical applications – in exchange for a very slightly weaker ontology, we obtain a great deal of conceptual clarity.

However, if one still needs these extra statements, then ETCS can be extended to encompass them. This is not unusual for a set of axioms; for instance, the (generalised) continuum hypothesis has been famously proven to be independent from ZFC, and is often taken as an additional axiom on top of ZFC when working with large cardinals in set theory.

The relationship between ETCS and ZFC has been well-studied, and it is known that ETCS is equivalent to the fragment of ZFC called *Restricted Zermelo with Choice* [Lei11][MM12], and it is also known what extra conditions need to be added on top of ETCS to obtain an axiom system with strength equivalent to ZFC (in the formal sense that a proposition is provable in this extended ETCS if and only if it is provable in ZFC).

This condition missing from ETCS is some form of an axiom of *collection* – axiom schemata that permit the construction of certain new sets from existing sets. These axiom schemata hence contribute to the size of the universe of constructible sets substantially, but conversely, this expansiveness is often not of particular importance in the practice of "ordinary" mathematics, so the omission of collection in ETCS is not damaging for most applications [nLa23c].

In ZFC, a form of a collection axiom is given by the axiom schema of replacement – informally, given a first-order formula φ, and a set x, we are permitted to form the set $\{\varphi(y) : y \in x\}$.

ETCS can similarly be extended with a collection axiom, which is exhibited as a form of cocompleteness [Osi74]. Informally, the axiom of collection in ETCS+C states that the category of sets has all coproducts $\coprod_{i \in I} X_i$ of families $(X_i)_{i \in I}$ of sets specified by first-order formulae. With this additional axiom, ETCS+C is then equivalent to the entirety of ZFC.

For another example, we can also augment ETCS with the Continuum Hypothesis, just like with ZFC. An elementary topos with collection is said to *satisfy the Continuum Hypothesis* (CH) if for all objects X, if there exists monomorphisms $\mathbb{N} \rightarrowtail X \rightarrowtail \Omega^{\mathbb{N}}$, then there exists either a monomorphism $X \rightarrowtail N$, or a monomorphism $\Omega^N \rightarrowtail X$, where $\mathbb{N}$ is the natural numbers object and Ω is the subobject classifier.

If the topos is Boolean, as in the case of any topos that satisfies ETCS, then (a categorical version of) the Cantor–Schröder–Bernstein theorem holds, so the existence of these latter two monomorphisms imply that there exists either an isomorphism $X \cong \mathbb{N}$, or an isomorphism $X \cong \Omega^{\mathbb{N}}$ (not necessarily equal to either monomorphism in either case), thus recovering the ordinary set-theoretic Continuum Hypothesis.

52.10.2 Material and Structural Sets

ETCS and ZFC both deal with "sets", but these notions are so distinct that it seems unhelpful to call them both by the same name. We will call a set in the style of ZFC a *material-set* and a set in the style of ETCS a *structural-set*.

In ZFC, we have the axiom of extensionality, which says that two material-sets are equal if and only if they have exactly the same elements. That is, material-sets are determined entirely by their elements. However, in ETCS, a weak extensionality principle is given by the Yoneda lemma: structural-sets are characterised only up to isomorphism by their generalised elements. However, we often only ask if two sets contained within a larger ambient set are equal. In that case, the strong extensionality principle for functions given by well-pointedness characterises structural-(sub)sets up to true equality.

Because elements of structural-sets are functions, this means that they themselves are never structural-sets, unlike in ZFC, where elements of material-sets are always themselves material-sets. This is perhaps closer to how we often use sets in ordinary mathematics; we never actually treat, say, the (real, natural, etc.) number "3" as a set itself.

In the introduction, we saw an argument of Benacerraf's that numbers cannot be sets, since numbers have no properties beyond arithmetic relations amongst themselves, and sets *do* have properties other than that. In this view, the natural numbers are envisioned as elements of an *abstract structure*, where elements have no properties beyond what is endowed upon them by that structure.

In ZFC, we define $\mathbb{N}$ to be some particular material-set, say $\{\varnothing,\{\varnothing\},\{\varnothing,\{\varnothing\}\},\ldots\}$, with all the arithmetic relations constructed on top of the chosen encoding, but this yields unwanted additional properties, like $3 \in 17$ (or not, given a different encoding) that we have to ignore.

In contrast, the natural numbers object in ETCS is a structural-set $\mathbb{N}$, equipped with an element 0 and a successor function s. Natural arithmetic is expressed in terms of the zero element 0 and the successor function s, so natural numbers, or elements of this abstract structure, have arithmetic relations between each other, but no additional properties beyond that.

More generally, *any* structural-set is precisely an abstract structure in this sense. An element $x \in X$ – a function $x : 1 \to X$ – has no identity or internal content except that it is an element of X, and is distinct from other elements of X.

Another effect of this is that, elements of structural sets, being functions, must also always be inherently attached to a set, unlike elements of material-sets, which are themselves objects that may exist in isolation; in ETCS, we can never refer to any element as existing by itself as some kind of free-floating Platonic essence, void of any connecting structure.

Given a material-set X, then for any other set A, we can ask whether $A \in X$ or not, regardless of any prior relations between A and X. This statement, "$A \in X$", is then a proposition in the formal sense: that is, it has a truth value, can be proven, can be combined with logical connectives, quantified over, etc.

In contrast, if X is a structural-set, then there are some things which are *intrinsically* elements of X, namely, the functions $1 \to X$. If a thing is not *given* as an element of X, then it *is not* an element of X, and similarly, an element $1 \to X$ cannot also be an element $1 \to Y$ of a different structural-set Y.

Thus, the statement $A \in X$ is not something one would ever *prove* about two pre-existing objects A and X. Consequently, the statement $x \in X$ is *not a proposition* in a structural set theory.

As an illustration of this difference, consider the statement "for all $x \in \mathbb{R}$, $x^2 \geq 0$" [Shu13]. If $\mathbb{R}$ is a material-set, then this statement could be read as "for all things x, if $x \in \mathbb{R}$, then $x^2 \geq 0$". Formally, the corresponding sentence in first order logic is $\forall x : x \in \mathbb{R} \to x^2 \geq 0$.

However, if $\mathbb{R}$ is a structural-set, then $x \in \mathbb{R}$ is a logical atom and cannot be the premise of an implication. Thus, the statement should be read as "it is a property of every real number that its square is non-negative."

Arguably, this is closer to how quantification is used in practice: we generally don't mean, "it is a property of any and all things in mathematics that *if* it happens to be a real number, *then* its square is non-negative." For instance, under the material interpretation, one particular instance of "for all $x \in \mathbb{R}$, $x^2 \geq 0$" is, "if the ring $\mathbb{Q}[x]$ happens to be a real number, then its square is non-negative", which could be reasonably agreed is a statement that most mathematicians would not naturally regard part of the content of "for all $x \in \mathbb{R}$, $x^2 \geq 0$".

Conversely, sometimes, we *do* want to regard $A \in X$ as a proposition. For instance [Shu13], if L is the set of complex numbers with real part $\frac{1}{2}$, then we are very interested in proving that "for all $z \in \mathbb{C}$, if $\zeta(z) = 0$ and z is not an even negative integer, then $z \in L$". In this statement, the first $\in$ is read structurally – z is *given* to be a complex number – while the second $\in$ is read materially; being the consequent of an implication, $z \in L$ should certainly be read as a proposition.

The observation here is that L is a subset of $\mathbb{C}$: z is already given to be an element of the structural-set $\mathbb{C}$ (i.e. it is a function $1 \to \mathbb{C}$), so it is possible to ask whether it happens to belong to this subset L (does it lift through the function witnessing $L \subseteq \mathbb{C}$?). As seen earlier, function extensionality characterises subsets up to true equality, so ETCS supports this use of material-subsets and propositional membership.

All this previous discussion also applies similarly to the subset relation: like elements, subobjects are (classes of) morphisms, so the statement $A \subseteq X$ for structural-sets X is similarly not a proposition – it is just notation for a class of monomorphisms with codomain X. However, the relation $\subseteq_X$ between subsets of a fixed set X *can* be used propositionally since it is a proper relation, so ETCS also supports propositional containment.

There are, however, some constructions that are somewhat less natural as structural-sets – in particular, function sets and power sets. In ETCS, function sets are given by exponential objects. We have a natural isomorphism

$$\hom(1, B^A) \cong \hom(A, B)$$

characterising the elements of B^A as being isomorphic to functions $A \to B$; but, this is only an isomorphism, and not true equality – elements of B^A are not literally functions $A \to B$, instead only being "labels" for them. To access the functions they reference, we need to invoke the evaluation map $\mathrm{ev} : B^A \times A \to B$.

Power sets have a similar problem to function sets in ETCS: the elements of a power set are characterised by the isomorphism

$$\hom(1, \Omega^A) \cong \mathrm{Sub}(A)$$

but this is again only up to isomorphism and not true equality, so elements of Ω^A are also only "labels" for subsets of A. This is one place where material-sets really are more conceptually clear: the elements of a material-power set $\mathcal{P}(X)$ are genuine subsets of X.

On the other hand, function sets are also rather unnatural in ZFC: we first have to pick an arbitrary encoding of an ordered pair, then define a function to be a special type of set of ordered pairs. This chain of encodings also results in lots of undesirable side-effects. At least in ETCS, the set of labels is given by a universal property and is hence isomorphism invariant.

52.10.3 Types

The problem here is that ZFC concerns itself exclusively with material-sets, and ETCS with structural-sets, when in mathematics, we often need to work with both. The awkwardness in these constructions, comes predominantly from forcing us to interpret structural-sets within a framework that only supports material-sets, or the reverse.

There are various solutions to this. For one, we could just keep adding more and more primitive notions to our systems until we have everything we need, i.e., add atoms to ZFC, and define the naturals to be a set of atoms; add functions to avoid encoding ordered pairs, etc., or, add a membership predicate to ETCS. However, a more systematic approach is to be desired.

It turns out that such a foundational system already exists, namely, *type theory.* In particular, variants such as *Martin–Löf dependent type theory* (MLTT) or *Homotopy Type Theory* (HoTT).

Type theories generally work like structural set theories. For instance, we have *types* which behave very much like structural-sets, and "elements" of types are called *terms*, and we write the *type declaration* $x : X$ for a term x of type X.

Just like with the structural usage of the $\in$ relation, a type declaration is not a proposition, as terms are intrinsically of some given immutable type, just like elements are intrinsically attached to structural-sets. Instead, these kinds of statements are called *judgements* – meta-assertions that cannot be proven within the theory.

Terms may also have *some* internal structure, unlike elements of structural-sets, though they do not *have* to, also unlike elements of material-sets, and the kind of internal structure they may have is controlled by their type. Then, *type constructors* can be used similarly to "adding primitive notions" to a set theory, i.e., a type constructor for ordered pairs, etc.

It turns out that category theory is also the natural language for the semantics of type theory; and conversely, type theory is a natural language for the syntax of category theory. This is beyond the scope of this discussion, but informally, we may interpret the objects of a category as types, and a subobject $\phi \rightarrowtail A$ can be regarded as a proposition by interpreting ϕ as the collection of terms of type A for which ϕ is true. Logical operations are then given by various limits in the subobject lattice.

This type system associated to each category is called that category's *internal logic*, and different kinds of categories induce different kinds of internal logics. For instance, Boolean categories induce type theories equivalent to classical first-order logics, while elementary topoi generate constructive higher-order logics. In particular, the internal logic of an ∞-topos is a model of a variant of Homotopy Type Theory.

Conversely, any type theory can be converted into a category by constructing objects from types, subobjects from relations, morphisms from functions, etc. It turns out that set theories can also be embedded within type theories, and type theories can also be embedded within sets: sets, categories, and types, can all be implemented within each other; one explicit construction is given in [Awo11].

52.10.4 Final Remarks

Once we have membership, functions, unions, quotients, products, etc. in whichever choice of set-theoretic foundations, the following development of mathematics is mostly the same: at a certain point, once basic mathematical structures have been constructed and encoded, for all practical purposes, it matters not if one begins with ZFC or ETCS.

After all, asking "is $3 \in 17$?" is not really a problem of practical concern in ZFC. However, it is still pedagogically fruitful to ask such questions. One advantage of teaching ETCS as a foundation is that it introduces the notions of isomorphisms and universal properties to students early on. It can also clarify why some material constructions are constructed in the way they are (i.e. "they are arbitrary choices of models of a (co)limit", "they satisfy the relevant structural property", etc.), even if we do not choose to use ETCS in practice.

Beyond this, the significance of ETCS is not from its use (or non-use) as a foundation of mathematics, but moreso from the research into topos theory that followed. ETCS was one of the first attempts of a categorical analysis of logic, and though it did not see much use as a foundation itself, the more general theory of topoi that followed is now the main language of categorical logic.

Bibliography

[Ben65] Benacerraf, Paul. *What Numbers Could Not Be*. 1965.

[nLa23a] nLab authors. *structural set theory*. Revision 37. 2023. URL: https://ncatlab.org/nlab/show/structural+set+theory.

[Gol84] Goldblatt, Robert. *Topoi: The Categorial Analysis of Logic*. Elsevier, 1984. ISBN: 9780444867117.

[Lei14] Leinster, T. *Basic Category Theory*. Cambridge University Press, 2014. ISBN: 9781107044241.

[Kit22] Kit. *The Yoneda Lemma*. University of Warwick. 2022.

[Rie17] Riehl, E. *Category Theory in Context*. Dover Publications, 2017. ISBN: 9780486820804.

[Mac13] MacLane, S. *Categories for the Working Mathematician*. Springer New York, 2013. ISBN: 9781475747218.

[Bor+94] Borceux, F. et al. *Handbook of Categorical Algebra: Volume 1, Basic Category Theory*. Cambridge University Press, 1994. ISBN: 9780521441780.

[Per21] Perrone, Paolo. *Notes on Category Theory with examples from basic mathematics*. 2021. arXiv: 1912.10642 [math.CT].

[Kos12] Kostecki, Ryszard. *An Introduction to Topos Theory*. 2012.

[Lei11] Leinster, Tom. *An informal introduction to topos theory*. 2011. arXiv: 1012.5647 [math.CT].

[nLa23b] nLab authors. *internalization*. Revision 88. 2023. URL: https://ncatlab.org/nlab/show/internalization.

[For02] Forrester-Barker, Magnus. *Group Objects and Internal Categories*. 2002. arXiv: math/0212065 [math.CT].

[Par74] Paré, Robert. "Colimits in topoi". In: *Bulletin of the American Mathematical Society* 80.3 (1974), pp. 556–561.

[MM12] MacLane, S. and Moerdijk, I. *Sheaves in Geometry and Logic: A First Introduction to Topos Theory*. Springer New York, 2012. ISBN: 9781461209270.

[Lei12] Leinster, Tom. *Rethinking set theory*. 2012. arXiv: 1212.6543 [math.LO].

[LC05] Lawvere, F.W. and C., McLarty. *An elementary theory of the category of sets (long version) with commentary*. Reprints in Theory and Applications of Categories, 2005.

[New14] Newstead, C. *An Elementary Theory of the Category of Sets*. 2014. URL: https://golem.ph.utexas.edu/category/2014/01/an_elementary_theory_of_the_ca.html.

[nLa23c] nLab authors. *axiom of replacement*. Revision 30. 2023. URL: https://ncatlab.org/nlab/show/axiom+of+replacement.

[Osi74] Osius, Gerhard. "Categorical set theory: A characterization of the category of sets". In: (1974). ISSN: 0022-4049. DOI: https://doi.org/10.1016/0022-4049(74)90032-2.

[Shu13] Shulman, Michael A. *From Set Theory to Type Theory*. 2013. URL: https://golem.ph.utexas.edu/category/2013/01/from_set_theory_to_type_theory.html.

[Awo11] Awodey, Steve. *From Sets to Types to Categories to Sets*. 2011. ISBN: 9789400704305.

All diagrams were written in LaTeX with the `tikz` package.

Chapter 53

Internal Logics of Categories

work in progress from this point onwards

53.1 The Algebra of Logic

53.1.1 Boolean and Heyting Algebras

In order theory, a poset $(L, \leq)$ is a *meet-semilattice* if every pair of elements $a,b \in L$ has a least upper bound or *meet* $a \wedge b$; a *join-semilattice* if every pair of elements $a,b \in L$ has a greatest lower bound or *join* $a \vee b$; and a *lattice* if it is simultaneously a meet-semilattice and join-semilattice. A lattice is furthermore *bounded* if there exist elements $0,1 \in L$ such that $0 \leq x \leq 1$ for all $x \in L$.

Because order-theoretic meets and joins are just (co)products, we can characterise lattices in purely categorical terms as follows: a lattice L is a thin category which has binary (co)products, and is furthermore bounded if it also has terminal and initial objects (or equivalently, a bounded lattice is a thin category that has finite (co)products).

The order-theoretic and categorical lattice are the kind of lattice we have been using so far, but we can also characterise lattices as a type of algebraic structure:

A lattice is an algebraic structure $(L, \wedge, \vee)$, consisting of a set L, and two commutative and associative binary operations $\wedge,\vee : L \times L \to L$, such that for all $a,b \in L$,

- $a \vee (a \wedge b) = a$;
- $a \wedge (a \vee b) = a$.

and the lattice is furthermore *bounded* if there exist two distinguished elements $0,1 \in L$, such that

- $a \wedge 1 = a$;
- $a \vee 0 = a$.

The existence of all meets and joins in a order-theoretic lattice imply that $\wedge$ and $\vee$ are binary operations, and it can be verified that they satisfy the axioms of an algebraic lattice, and conversely, the binary operations of an algebraic lattice induce a relation given by $a \leq b$ if $a = a \wedge b$ or $b = a \vee b$; and again, it is easy to verify that this relation is a partial ordering whose meets and joins are compatible with the binary operations. The three definitions are hence equivalent, but the algebraic definition is more convenient for modification:

A lattice is *distributive* if the binary operations $\vee$ and $\wedge$ distribute over each other. That is, for all $a,b,c \in L$,

$$a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c) \qquad \text{and} \qquad a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$$

(Note that any lattice satisfying one of the two conditions above must necessarily also satisfy the other, so only one of the above needs to be verified when checking whether a lattice is distributive or not.)

In a bounded lattice L, the *complement* of an element $a \in L$ is an element x such that

- $a \wedge x = 0$;
- $a \vee x = 1$.

A lattice is *complemented* if every element a has a complement, denoted as $\neg a$.

Lemma 53.1.1. *A complement in a distributive lattice is unique if it exists.*

Proof. Suppose x and x' are complements of a. Then,

$$\begin{aligned} x' &= 1 \wedge x' \\ &= (a \vee x) \wedge x' \end{aligned}$$

$$\begin{aligned}
&= (a \wedge x') \vee (x \wedge x') \\
&= 0 \vee (x \wedge x') \\
&= (x \wedge a) \vee (x \wedge x') \\
&= x \wedge (a \vee x') \\
&= x \wedge 1 \\
&= x
\end{aligned}$$

■

The meet and join symbols are notably the same as the symbols for logical conjunction and disjunction, and in fact, classical propositional logic can be modelled as a kind of lattice:

A *Boolean algebra* is a complemented distributive bounded lattice. In the two-element Boolean algebra, interpreting 0 as false; 1 as true; $\vee$ as disjunction; $\wedge$ as conjunction; and $\neg$ as logical negation, algebraic expressions in these symbols correspond to logical statements, in that two algebraic expressions are equal if and only if their corresponding logical statements are logically equivalent.

In classical logic, we define the implication operation $a \to b$ as $\neg a \vee b$, but we also could have started with a lattice equipped with an implication operation, and derive the complement operation from there, rather than the reverse:

A *Heyting algebra* is a bounded lattice equipped with a binary operation $\to$ such that $c \leq a \to b$ if and only if $c \wedge a \leq b$ for all a,b,c.

This means that, by definition, a Heyting algebra is the weakest structure in which modus ponens (a,$a \to b \vdash b$) is a sound inference rule. We also have that $1 \leq 0 \to a$ for any a, or "any statement a is implied by a contradiction 0", corresponding to the logical principle of explosion.

Heyting algebras can also be characterised categorically: a Heyting algebra is a bounded lattice that is cartesian closed when considered as a category, and the exponential b^a is written as $a \to b$. From the definition of an exponential, we also have that implication $a \to (-)$ is precisely the right adjoint to the meet (or product) functor $(-) \wedge a$, so $\hom(c,a \to b) \cong \hom(c \wedge a, b)$ naturally in all 3 variables; or equivalently, $c \leq a \to b$ if and only if $c \wedge a \leq b$, recovering the usual algebraic characterisation.

Also note that,

Theorem 53.1.2. *Every Boolean algebra is a Heyting algebra.*

Proof. The meet functor is a left adjoint so it preserves colimits, and in particular, coproducts, so meets distribute over joins in a Heyting algebra, and by duality, joins distribute over meets, so a Heyting algebra is necessarily a distributive lattice. It remains to show that the implication in a Boolean algebra defined by $\neg a \vee b$ satisfies the requirements for a Heyting algebra.

For the forward direction, suppose $c \leq \neg a \vee b$. Then,

$$\begin{aligned}
c \wedge a &\leq (\neg a \vee b) \wedge a \\
&\leq (\neg a \wedge a) \vee (b \wedge a) \\
&\leq 0 \vee (b \wedge a) \\
&\leq b \wedge a \\
&\leq b
\end{aligned}$$

For the reverse direction, suppose $c \wedge a \leq b$. Then,

$$\begin{aligned}
c &= c \wedge 1 \\
&= c \wedge (\neg a \vee a) \\
&= (c \wedge \neg a) \vee (c \wedge a)
\end{aligned}$$

$$\begin{aligned}&\leq (c \wedge \neg a) \vee b\\&\leq \neg a \vee b\end{aligned}$$ ■

Negations can also be defined in Heyting algebras as

$$\neg a \coloneqq a \to 0$$

(and logically, this can be interpreted as "$\neg a$ is the proposition that assuming a would imply a contradiction"), but this time, negations are *not* exactly equivalent to their Boolean counterparts of complements. However, the reuse of the notation is justified by:

Theorem 53.1.3. *The complement of an element in a Heyting algebra is, if it exists, its negation.*

That is, a complement may not exist for an element a in a Heyting algebra, but if it does, then it is precisely the negation $\neg a$. Conversely, a negation $\neg a$ always exists but is not necessarily always the complement of a. For this reason, the negation is sometimes called the *pseudo-complement.*

Proof. Suppose x is a complement of a, so,

$$\begin{aligned}a \wedge x &= 0\\x &\leq a \to 0\\x &\leq \neg a\end{aligned}$$

and also,

$$\begin{aligned}\neg a &= \neg a \vee 1\\&= \neg a \wedge (a \vee x)\\&= (\neg a \wedge a) \vee (\neg a \wedge x)\\&= 0 \vee (\neg a \wedge x)\\&= \neg a \wedge x\\&\leq x\end{aligned}$$

so $x = \neg a$. ■

Like complements, we have $a \wedge \neg a = 0$ for negations, and $a \leq \neg\neg a$, but not $\neg\neg a \leq a$ in general, so in Heyting algebras, $\neg\neg a = a$ does not generally hold. That is, unlike in Boolean algebras, double negation elimination does not hold in Heyting algebras. Another statement of Boolean algebra that does not hold in Heyting algebras is $a \vee \neg a = 1$: the law of the excluded middle also fails in Heyting algebras.

Theorem 53.1.4. *A Heyting algebra is a Boolean algebra if and only if either of the following equivalent statements are satisfied:*

- $\forall a : \neg\neg a = a$;
- $\forall a : a \vee \neg a = 1$.

Proof. [MM12] Since complements are unique in a Boolean algebra, a is the complement of $\neg a$, so $\neg\neg a = a$. Conversely, if $\neg\neg a = a$ in a Heyting algebra, then,

$$\begin{aligned}a \vee \neg a &= \neg\neg(a \vee \neg a)\\&= \neg(\neg a \wedge \neg\neg a)\\&= \neg(\neg a \wedge a)\\&= \neg 0\end{aligned}$$

$$= 1$$

$a \wedge \neg a = 0$ holds dually, so $\neg a$ is a complement of a and the lattice is Boolean. The proof for $a \vee \neg a = 1$ is analogous. ■

Because of this, the Boolean and Heyting algebras model different axiomatic systems of logic. Boolean algebras, as discussed, model *classical logic*; Heyting algebras instead model *intuitionistic logic.*

53.1.2 Intuitionistic Logic

Intuitionistic logic can be viewed as a fragment of classical logic in which the double negation elimination and law of excluded middle axioms do not hold. Intuitionistic logic is also sometimes called *constructive logic*, because it more closely follows the approach behind a constructive proof.

Beyond merely being an axiomatic system of symbolic logic, intuitionistic logic also corresponds to a philosophy of mathematics called *intuitionism* or *constructivism.*

Intuitionism says that mathematical objects and structures do not exist unless explicitly constructed. (Compare to (set-theoretic) Platonism in the introduction.) Acccording to intuitionism, mathematicians do not work in an ideal Platonic universe, discovering hidden, preexisting truths, but instead create the ontology themselves. Hence, to *prove* something in intuitionistic logic is to *construct* it explicitly.

In classical logic, the inference rules and operations have been carefully designed to preserve *truth values* with respect to proof. In intuitionistic logic, the inference rules instead preserve *justification* with respect to evidence and construction.

Because of this change in objective, the meanings of the logical symbols $\wedge$, $\vee$, $\neg$, and $\rightarrow$ also change. The standard *Brouwer-Heyting-Kolmogorov* (BHK) *interpretation* of intuitionistic logic assigns the following meanings to the symbols:

- A proof of $p \wedge q$ is a pair of proofs of p and q.
- A proof of $p \vee q$ is either a proof of p or a proof of q.
- A proof of $p \rightarrow q$ is a function that transforms a proof of p into a proof of q.
- A proof of $\neg p$ is a proof of $p \rightarrow \bot$, or a function that transforms a proof of p into a proof of $\bot$.
- There is no proof of $\bot$.

In all of the above, “proof” may also be read as “construction”.

For example, the identity function is a proof of the formula $p \rightarrow p$ for any p. In contrast, $\neg\neg p$ expands to $(p \rightarrow \bot) \rightarrow \bot$, which in general has no proof. More colloquially, a proof of $\neg\neg p$ is a proof that there is no proof that there is no proof of p, which is not the same as a proof of p, so double negation elimination $\neg\neg p \vdash p$ fails to hold intuitionistically.

On the other hand, the law of non-contradiction $\neg(p \wedge \neg p)$ expands to $\big(p \wedge (p \rightarrow \bot)\big) \rightarrow \bot$. A proof of this statement is a function that transforms the pair $\langle a, b \rangle$ – where a is a proof of p and b is a proof of $p \rightarrow \bot$ (i.e., a function that transforms a proof of p into a proof of $\bot$) – into a proof of $\bot$. The function $\langle a,b \rangle \mapsto b(a)$ does this, hence proving the law of non-contradiction intuitionistically.

Theorem 53.1.5. *For any object A in a topos $\mathcal{E}$, the subobject poset* $\mathrm{Sub}(A)$ *is a Heyting algebra. Furthermore, this structure is natural in the sense that for every morphism $f : A \rightarrow B$, the induced map* $f^* : \mathrm{Sub}(A) \rightarrow \mathrm{Sub}(B)$ *is a homomorphism of Heyting algebras.*

Now, recall that there is a natural isomorphism $\mathrm{Sub}(A) \cong \hom(A,\Omega)$ (natural in A), so the set of morphisms from any object into the subobject classifier also has the structure of a Heyting algebra.

53.1.3 TEMP

Let us consider the operations of intersection and union of subobjects.

Recall that the intersection and union of subobjects make the subobject poset $\mathrm{Sub}(X)$ of any object X a lattice.

Because $\mathrm{Sub}(X) \cong \hom(X,\Omega)$, these operations $\cap,\cup : \mathrm{Sub}(X) \times \mathrm{Sub}(X) \to \mathrm{Sub}(X)$ induce functions $\hom(X,\Omega) \times \hom(X,\Omega) \to \hom(X,\Omega)$. Furthermore, $\cong \hom(X,\Omega \times \Omega)$

An *ordinary predicate* of *type* A is a generalised element of Ω of shape A – that is, a morphism $\varphi : A \to \Omega$ – or equivalently, an ordinary element $\varphi' : 1 \to \Omega^A$, or a subobject $[\varphi] : S \subseteq A$.

A *generalised predicate* of *type* A is a generalised element of Ω^A of shape X – a morphism $\varphi' : X \to \Omega^A$ – or equivalently, a generalised element $\varphi : X \times A \to \Omega$, or a subobject $[\varphi] : S \subseteq X \times A$.

53.1.4 Categorical Logic

53.1.5 Internal Lattices

53.1.6 Structures and Interpretations

Here, we will discuss the formalisation of the language of a *mathematical theory*. Mathematical theories discuss mathematical structures (or an "object of interest"), but also depend on a *metamathematical* language and interpretation which provides the context with which these structures can be analysed.

The formalisation of mathematical theories was originally motivated by the discovery of paradoxes in foundational set theory, but now also provides a way for us to express a mathematical theory in an abstract form that allows it to be interpreted (or *modelled*) in other contexts, e.g. a theory originally expressed in terms of sets can be formalised, abstracted, and interpreted in, say, a topos.

A *structure* $\mathfrak{U} = (U,P,F)$ is a universe consisting of [Kos12]

- a set of *individuals* or *constants* (e.g. elements of a group);
- a set of *properties* of or *relations* on individuals (e.g. equality of elements in a group);
- a set of *functions* on atoms (e.g. group composition).

A *formal language* is a list of symbols from a collection called an *alphabet* that concatenate into strings according to some *grammar* or *syntax* rules. Note, however, that a formal language by itself only specifies the syntax of these strings, and not their semantics.

Example. The following list of syntax rules describes a formal language $\mathcal{L}$ over the alphabet $\Sigma = \{0,1,2,3,4,5,6,7,8,9,+,=\}$:

- Every non-empty string that does not contain "+" or "=" and does not start with "0" is in $\mathcal{L}$.
- The string "0" is in $\mathcal{L}$.
- A string containing "=" is in $\mathcal{L}$ if and only if there is exactly one "=", and it separates two valid strings of $\mathcal{L}$.
- A string containing "+" but not "=" is in $\mathcal{L}$ if and only if every "+" in the string separates two valid strings of $\mathcal{L}$.
- No string is in $\mathcal{L}$ unless implied by the previous rules.

The strings "$2+2=4$" and "$12+34+5=678$" are in $\mathcal{L}$, but the string $+ = +12 =$ is not. This formal language expresses natural numbers, well-formed additions, and well-formed addition equalities, but it only expresses what they look like (syntax) and not what they mean (semantics). △

A theory of any given structure (the collection of all statements about that structure) forms a formal language. The symbols in the alphabet of a formal language $\mathcal{L}$ corresponding to a structure can be more finely classified into:

- propositional variables;
- n-ary predicative or relational symbols;
- n-ary function symbols;
- auxiliary signs (like commas and brackets).

Propositional variables vary over particular individuals of a structure. Predicative symbols are intended to represent properties of the structure, and are intended to return a truth value (such as "=" either being true or false in arithmetic), while function symbols are those intended to return other constants (such as "+"). In particular, nullary or *constant* function symbols are identified with constants.

So in the above example of "$2+2=4$", if we think of this as a sentence describing natural addition, then "2" and "4" are both constants (constant function symbols), "+" is a function symbol, "=" is a predicative symbol, and there are no propositional variables. Again, however, note that a formal language only specifies the syntax of sentences: there is nothing indicating that the symbol "2" represents the natural number two, the symbol "+" means addition, "$2 + 2 = 4$" is true, etc.

Strings in $\mathcal{L}$ consisting of variables, predicatives, and functions are called a *formulae*, and strings consisting of just variables and functions are called *terms.* A *sentence* is a formula with propositional variables replaced with specific constants.

An assignment of meanings to symbols in a formal language is called an *interpretation.* We can interpret the language $\mathcal{L}$ over a structure $\mathfrak{U}_{\mathbb{N}} = (\mathbb{N}, = ,+)$ as follows:

- "2" and "4" mean the natural numbers two and four, respectively;
- "+" means the addition of natural numbers;
- "=" means the equality relation over the natural numbers.

so for example, the sentence "$2 + 2 = 4$" would be true in this interpretation.

Another interpretation could also be given by:

- "2" and "4" mean the natural numbers two and four, respectively;
- "+" means the *subtraction* of natural numbers;
- "=" means the equality relation over the natural numbers.

or even:

- "2" and "4" mean "a person" and "lunch", respectively;
- "+" means "talks to";
- "=" means "while eating"

In the former interpretation, "$2 + 2 = 4$" is now false, and in the latter, the string doesn't even have a meaningful truth value.

Logical connectives like $\neg$, $\wedge$, or $\vee$ can also be expressed as a kind of predicative symbol.

The *first order formal propositional language* $\mathcal{L}_{\text{FOFP}}$ consists of an alphabet Σ given by,

- individual variable symbols $(x,y,z,\ldots)$;
- n-ary function symbols $(= , \leq , \ldots)$;
- n-ary predicative symbols $(\neg, \wedge , \vee , \rightarrow , \ldots)$;
- unary quantifier symbols $(\forall,\exists,\ldots)$;
- auxiliary (e.g. commas, brackets).

a set T of terms such that

- all individual variables and constants belong to T;
- all terms built from function symbols substituted with elements of T belong to T.

and a set F of formulae such that

- all predicative symbols with arguments substituted by elements of T belong to F;
- all connectives with arguments substituted by elements of F belong to F;
- all quantifiers with arguments substituted by individual variables and acting on formulae of F belong to F.

But again, this formal language needs to be equipped with some extra data in the form of *inference rules* in order to perform logical deduction. More precisely, the inference rules are a set of operations that, given a sequence of formulae in $\mathcal{L}$ called *premises*, return a formula called a (*logical*) *consequence*. The map sending every set X of formulae to the set $C(X)$ of its logical consequences is called the *consequence operator*. The pair $(\mathcal{L},C)$ is then called a *formal deductive system*. A set A of formulae assumed to be true (i.e., that constrains the possible interpretations) is called a set of *axioms*. The triple $(\mathcal{L},C,A)$ is then called a *formal theory*, and the formulae in $C(A)$ are called the *theorems* of that theory.

... WIP

53.1.7 Lindenbaum-Tarski Algebras

53.1.8 Sieves and Sheaves

Let us dispense with the axioms of ETCS for the following sections, and consider a general topos.

Let $\mathscr{E}_1$ and $\mathscr{E}_2$ be topoi. A *geometric morphism* $g : \mathscr{E}_1 \to \mathscr{E}_2$ is a pair of adjoint functors $g^* \dashv g_*$ between $\mathscr{E}_1$ and $\mathscr{E}_2$ such that g^* preserves finite limits (and g_* preserves finite colimits).

A *sieve* on an object X in $\mathscr{C}^{\text{op}}$ is a collection S of morphisms with codomain X that is closed under precomposition. That is, if $f \in S$ and $g : A \to X$, then $f \circ g \in S$. A *maximal* or *principal* sieve on X, denoted $\uparrow X$ is then the collection of all morphisms into X.

53.1.9 Internal Logic of a Topos

Recall that the subobject classifier can be characterised as a representation of the functor $\text{Sub} : \mathscr{C}^{\text{op}}\mathbf{Set}$ that sends objects X to their subobjects posets $\text{Sub}(X)$, for a well-powered category $\mathscr{C}$:

$$\hom_{\mathscr{C}}(-,\Omega) \cong \text{Sub}(-)$$

In other words, external operations on subobject posets correspond naturally to internal operations on Ω.

Now, by the Yoneda lemma, the natural isomorphism

$$\eta : \hom(-,\Omega) \Rightarrow \text{Sub}(-)$$

is determined entirely by its value at the identity id_Ω, namely $\eta_\Omega(\text{id}_\Omega)$ (this follows almost identically to (the dual of) the proof of Lemma Lemma 52.2.8).

In fact, we've already seen an example of this: the intersection operation.

The intersection operation is a binary operation on a subobject poset,

$$\cap_X : \text{Sub}(X) \times \text{Sub}(X) \to \text{Sub}(X)$$

It turns out that these binary operations assemble into a natural endotransformation. In more detail, by the isomorphism above, we may replace $\text{Sub}(X)$ by hom-functors

$$\cap_X : \hom(X,\Omega) \times \hom(X,\Omega) \to \hom(X,\Omega)$$

and as the hom-functor preserves limits in its second argument, we have

$$\cap_X : \hom(X,\Omega \times \Omega) \Rightarrow \hom(X,\Omega)$$

and we can see that $\cap_X$ is really a natural transformation between these hom-sets involving Ω.

Because the intersection operation takes two subobjects in $\mathrm{Sub}(X)$ to return another subobject in $\mathrm{Sub}(X)$, we can consider it as a binary operation $\cap_X : \mathrm{Sub}(X) \times \mathrm{Sub}(X) \to \mathrm{Sub}(X)$. Because $\mathrm{Sub}(-)$ is a functor, it turns out that these binary operations assemble into a natural endotransformation.

In more detail: recall that the subobject classifier can be characterised as a representation of the functor $\mathrm{Sub} : \mathscr{C}^{\mathrm{op}} \to \mathbf{Set}$ that sends objects to their subobject posets, where $\mathscr{C}$ is a well-powered category. That is,

$$\hom_{\mathscr{C}}(X,\Omega) \cong \mathrm{Sub}(X)$$

naturally in X.

So, the components

$$\cap_X : \mathrm{Sub}(X) \times \mathrm{Sub}(X) \to \mathrm{Sub}(X)$$

correspond to transformations

$$\hom(X,\Omega) \times \hom(X,\Omega) \to \hom(X,\Omega)$$

One important interpretation of this representation is that "external" operations on $\mathrm{Sub}(X)$ naturally correspond to "internal" operations on Ω. More precisely, the isomorphism

$$\phi : \hom_{\mathscr{C}}(-,\Omega) \to \mathrm{Sub}(-)$$

is completely determined by its value at id_Ω, namely $\phi_\Omega(\mathrm{id}_\Omega)$,* which is just a subset of Ω, or more specifically, the monomorphism $\top : 1 \to \Omega$. In other words, the Yoneda lemma implies that operations on subobject posets $\mathrm{Sub}(X)$ are entirely determined by their action on $\top$.

For instance, by the representation of $\mathrm{Sub}(-)$, the "external" intersection natural transformation $\cap : \mathrm{Sub}(-) \times \mathrm{Sub}(-) \to \mathrm{Sub}(-)$ corresponds to a natural transformation

$$\phi : \hom(-,\Omega) \times \hom(-,\Omega) = \hom(-,\Omega \times \Omega) \to \hom(-,\Omega) \tag{53.1}$$

and the Yoneda lemma then tells us that this natural transformation is furthermore determined entirely by $\phi_\Omega(\mathrm{id}_\Omega)$ a function $\Omega \times \Omega \to \Omega$. In fact, this function is exactly the meet operation, $\wedge$.

* This follows similarly to the proof of the (dual) Yoneda Lemma Lemma 52.2.8, noting that $\hom_{\mathscr{C}}(-,\Omega) = h^{\Omega}$.

Chapter 54

Homotopy Type Theory

Index