Notes on the Mathematical Tripos

Sky Wilshaw

Part IA

University of Cambridge
2020–2024

Contents

I. Numbers and Sets
Lectured in Michaelmas 2020 by Prof. I. B. Leader **5**

II. Differential Equations
Lectured in Michaelmas 2020 by Dr. J. R. Taylor **51**

III. Groups
Lectured in Michaelmas 2020 by Dr. A. Khukhro **125**

IV. Vectors and Matrices
Lectured in Michaelmas 2020 by Dr. J. M. Evans **193**

V. Dynamics and Relativity
Lectured in Lent 2021 by Prof. P. H. Haynes **269**

VI. Probability
Lectured in Lent 2021 by Dr. P. Sousi **343**

VII. Vector Calculus
Lectured in Lent 2021 by Dr. A. Ashton **437**

VIII. Analysis I
Lectured in Lent 2021 by Prof. G. Paternain **519**

I. Numbers and Sets

Lectured in Michaelmas 2020 by PROF. I. B. LEADER

This course gives an introduction to university level maths. We begin by presenting some basic rules, called axioms. Then, we carefully prove logical statements that follow from these axioms. We build upon our previous results iteratively until we have proven some important theorems. Almost every other course in an undergraduate maths course will follow this pattern, and this course acts as a prototypical example.

In the first part of the course, we rigorously define the natural numbers, integers, rationals, and reals. Using the axioms, we can prove facts about things like prime numbers, modular arithmetic, and limits. In the second half of the course, we establish the notion of a set, and define what concepts like functions are. At the end, we use the rules of sets to prove that there are different sizes of infinity.

Contents

1. Proofs **8**
1.1. Motivation for proof 8
1.2. Proofs and non-proofs 8
2. Elementary number theory **11**
2.1. The natural numbers 11
2.2. Strong induction 11
2.3. The integers and rationals 12
2.4. Primes 13
2.5. Highest common factors 13
2.6. The division algorithm 13
2.7. Euclid's algorithm 14
2.8. Linear Diophantine equations 15
2.9. The fundamental theorem of arithmetic 16
3. Modular arithmetic **18**
3.1. Introduction 18
3.2. Inverses 18
3.3. Invertibility 18
3.4. Euler's totient function 19
3.5. Fermat's little theorem and Fermat–Euler theorem 19
3.6. Square roots of one 19
3.7. Square roots of negative one 20
3.8. Solving congruence equations 21
3.9. Chinese remainder theorem 21
3.10. RSA encryption 22
4. The reals **23**
4.1. Motivation for the reals 23
4.2. Axioms of the reals 23
4.3. Examples of sets and least upper bounds 24
4.4. Sequences and limits 26
4.5. Series 27
4.6. Testing convergence of a sequence 27
4.7. Decimal expansions 29
4.8. The number e 30
4.9. Algebraic and transcendental numbers 30
4.10. Complex numbers 32
5. Sets **33**
5.1. Sets and subsets 33

5.2. Composing sets . . . 33
5.3. Russell's paradox . . . 34
5.4. Finite sets . . . 34
5.5. Binomial coefficients . . . 35
5.6. Computing binomial coefficients . . . 35
5.7. Binomial theorem . . . 36
5.8. Inclusion-exclusion theorem . . . 36
6. Functions . . . 38
6.1. Definition . . . 38
6.2. Injection, surjection and bijection . . . 39
6.3. Composition of functions . . . 40
6.4. Invertibility . . . 40
6.5. Relations . . . 41
6.6. Equivalence classes as partitions . . . 41
6.7. Quotients . . . 42
7. Countability . . . 43
7.1. Basic properties . . . 43
7.2. Products of countable sets . . . 44
7.3. Countable unions of countable sets . . . 44
7.4. Uncountable sets . . . 45
7.5. Comparing sizes of sets . . . 47
7.6. Schröder–Bernstein theorem . . . 48
7.7. Arbitrarily large sets . . . 49
7.8. What happens next? . . . 49

1. Proofs

1.1. Motivation for proof

Definition (Proof)**.** A proof is a logical argument that establishes a conclusion.

Clearly there are some things missing from this definition; we have not yet defined a 'logical argument' or a 'conclusion'; however we have to start somewhere, and assuming understanding of logic is a good place to start. There is a 3rd year course called 'Logic and Set Theory' that rigorously defines this.

There are two main reasons to want to prove things.

(i) To be sure that they are true; and

(ii) to understand why they are true.

For the first point, it is easy to make a contrived example that shows why we need to prove statements even though they appear to be true for small n, for example: 'all positive integers n are not equal to 100 trillion'. Understanding the reasoning behind why a statement is true is also very important; an example of this is at the end of this lecture.

1.2. Proofs and non-proofs

Claim. For any positive integer n, $n^3 - n$ is a multiple of 3.

Proof. Given some positive integer n, we have

$$n^3 - n = (n-1)n(n+1)$$

One of $n-1$, n, $n+1$ must be a multiple of 3 as they are 3 consecutive integers.

Therefore, $(n-1)n(n+1)$ must be a multiple of 3. □

There are a couple of things to note about this proof.

- The phrase 'given a positive integer' is important; we need to know where this variable n came from.
- We used the fact that three consecutive numbers contain a multiple of 3 here, but this was not proven. We must prove this fact elsewhere, or we cannot use it in this course!
- It is important to write proofs legibly and linearly down the page; don't just write a long line of symbols.

Claim. For any positive integer n, if n^2 is even then n is even.

Proof. Given a positive integer n that is even, we have $n = 2k$ for some integer k.

Thus $n^2 = (2k)^2 = 4k^2 = 2(2k^2)$,

so n^2 is even. □

Note. This is a false proof. We proved that $B \implies A$, but we want $A \implies B$. Our result wasn't false, but it didn't show what we set out to prove. The words 'for some integer k' are important: we must specify which set k belongs to. Our proof would be incorrect if we did not state this, as it would be unclear that $2(2k^2)$ is an even number.

Claim. For any positive integer n, if n^2 is a multiple of 9 then n is a multiple of 9.

Proof. Given a positive integer n that is a multiple of 9, we have $n = 9k$ for some integer k.

Therefore, $n^2 = (9k)^2 = 81k^2 = 9(9k^2)$,

so n^2 is a multiple of 9. □

Note. Not only does this fall for the same trap as the previous proof, but the original claim is false (e.g. $n = 6$)! It's entirely irrelevant that the claim is true for some positive integers, because even one counterexample disproves the claim.

Let's return now to the previous incorrect example: 'if n^2 even then n even for all positive integers n'.

Proof. Suppose that n is odd.

We have $n = 2k + 1$ for some integer k.

Therefore, $n^2 = (2k + 1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1$

n^2 is odd #

Therefore n is even. □

- We prove things to show *why* something is true. We can see why this claim was true here—it's really a statement about the properties of odd numbers, not the properties of even numbers.
- We started by saying that we need something tangible to work with: just stating that 'n^2 is even' is really hard to work with because square roots just get messy and don't yield any result. So we had to choose a clever first step.
- The symbol # shows that we have a contradiction.

This was a kind of proof by contradiction. Essentially, $A \implies B$ is the same as saying $\neg B \implies \neg A$. This is because:

- $A \implies B$ means that there is no case such that A is false and B is true.

- $\neg B \implies \neg A$ means that there is no case such that $\neg B$ is false and $\neg A$ is true. In other words, there is no case such that B is true and A is false. This is equivalent to the case with $A \implies B$.

Claim. The solution to the real equation $x^2 - 5x + 6 = 0$ is $x = 2$ or $x = 3$.

Note. This is really two assertions:

(i) $x = 2 \vee x = 3 \implies x^2 - 5x + 6 = 0$, and

(ii) $x^2 - 5x + 6 = 0 \implies x = 2 \vee x = 3$

We can denote this using a two-way implication symbol $\iff$:

$$x = 2 \vee x = 3 \iff x^2 - 5x + 6 = 0$$

Proof. We prove case i by expressing the left hand side as a product of factors: $(x-3)(x-2) = 0$. The other case may be proven using factorisation. □

We can do another kind of proof using $\iff$ symbols a lot. However, we need to be absolutely sure that each step really is a bi-implication.

Alternative Proof. For any real x:

$$\begin{aligned} x^2 - 5x + 6 = 0 &\iff (x-2)(x-3) = 0 \\ &\iff x - 2 = 0 \vee x - 3 = 0 \\ &\iff x = 2 \vee x = 3 \end{aligned}$$

□

Claim. Every positive real is at least 1.

Proof. Let x be the smallest positive real. We want to prove $x = 1$, so we prove this by contradiction.

Case 1: if $x < 1$ then $x^2 < x$ #

Case 2: if $x > 1$ then $\sqrt{x} < x$ #

Therefore $x = 1$ □

Note. The assertion that there exists a smallest positive real is not justified. This means that the proof is invalid in its entirety. It is important that every line in a proof must be justified.

2. Elementary number theory

2.1. The natural numbers

Each line in a proof must be justified. So, in number theory, what are you allowed to assume? We must begin with a set of axioms. We define that the natural numbers are a set denoted $\mathbb{N}$, that contains an element denoted 1, with an operation $+1$ satisfying:

(i) $\forall n \in \mathbb{N}, n+1 \neq 1$

(ii) $\forall m, n \in \mathbb{N}, m \neq n \implies m+1 \neq n+1$ (together with the previous rule, this captures the idea that all numbers in $\mathbb{N}$ are distinct)

(iii) For any property $p(n)$, if $p(1)$ is true and $p(n) \implies p(n+1)\, \forall n \in \mathbb{N}$, then $p(n)\, \forall n \in \mathbb{N}$ (induction axiom).

This list of rules is known as the Peano axioms. Note that we did not include 0 in this set. You can show that the list of natural numbers is complete and has no extras (like the rational number 3.5) by specifying $p(n) =$ 'n is on the list of natural numbers'.

Note that while numbers are defined as, for example, $1+1+1+1$, we are free to use whatever names we like, e.g. 4 or 3735928559.

We may also define our own operations, such as $+2$, which is defined to be $+1+1$. In fact, we can define the operation $+k$ for any $k \in \mathbb{N}$ by stating:

$$(n+k)+1 = n+(k+1) \quad (\forall n, k \in \mathbb{N})$$

and using induction to construct the $+k$ operator for all k. We can similarly construct multiplication and exponentiation operators for all natural numbers, although this is omitted here. We can also prove properties on these operators such as associativity, commutativity and distributivity.

We can also define the $<$ operator as follows: $a < b \iff \exists k \in \mathbb{N}$ s.t. $a+k=b$. Of course, we can also prove several properties using this rule, such as transitivity, and the fact that $a \not< a$, which are omitted here.

2.2. Strong induction

The induction axiom states that if we know

- $p(1)$ is true, and
- $p(n) \implies p(n+1)$ for any $n \in \mathbb{N}$

then we can conclude that $p(n)$ is true for all $n \in \mathbb{N}$. We can in fact prove a stronger statement using this axiom, known as 'strong induction'.

Claim. If we know that

- $p(1)$ is true, and
- the fact that $p(k)$ is true for all $k < n$ implies that $p(n)$ is true

then $p(n)$ is true for all $n \in \mathbb{N}$.

Proof. Consider the predicate $q(n)$ defined as: '$p(k)$ is true for all $k < n$'. Given that $p(1)$ is true, $q(1)$ is trivially true since there are no k below 1. Since $q(n) \implies q(n+1)$, we can use the induction axiom, showing that $q(n)$ is true for all n, so $p(n)$ is true for all n. □

This provides a very useful alternative way of looking at induction. Instead of just considering a process from n to $n+1$, we can inject an inductive viewpoint into any proof. When proving something on the natural numbers, we can always assume that the hypothesis is true for smaller n than what we are currently using. This allows us to write very powerful proofs because in the general case we are allowed to refer back to other smaller cases—but not just $n-1$, any k less than n.

We may rewrite the principle of strong induction in the following ways:

(i) If $p(n)$ is false for some n, there must be some m where $p(m)$ is false and $p(k)$ is true for all $k < m$. In other words, if a counterexample exists, there must exist a minimal counterexample.

(ii) If $p(n)$ is true for some n, then there is a smallest n where $p(n)$. In other words, if an example exists, there must exist a minimal example. This is known as the 'well-ordering principle'.

2.3. The integers and rationals

The integers $\mathbb{Z}$ consist of the set of natural numbers $\mathbb{N}$, their additive inverses, and an identity element denoted 0. In other words, $(\mathbb{Z}, +)$ is the group generated by $\mathbb{N}$ and the addition operator: $\mathbb{Z} = \langle \mathbb{N} \rangle$. We define operations in a familiar way, for example $a < b \iff \exists c \in \mathbb{N}$ s.t. $a + c = b$.

The rational numbers $\mathbb{Q}$ consist of all expressions denoted $\frac{a}{b}$ where $a, b \in \mathbb{Z}$ with $b \neq 0$; with $\frac{a}{b}$ regarded as the same as $\frac{c}{d}$ if and only if $ad = bc$. We define, for example,

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}$$

Note that is important to verify with each operation that it does not matter how you write a given rational number. For example, $\frac{1}{2} + \frac{1}{2} = \frac{2}{4} + \frac{3}{6}$. This means that operations such as $\frac{a}{b} \mapsto \frac{a^3}{b^2}$ cannot exist because then it would depend on how you write the rational number.

2.4. Primes

Proposition. Every $n \geq 2$ is expressible as a product of primes.

Proof. We use induction on an integer n, starting at 2, a trivial case. Given $n > 2$, we have two cases:

- n is prime. Therefore, n is a product of primes as required.
- n is composite. We know that n can be split into two factors, denoted here as a, b. Using (strong) induction, we know that because both a and b are smaller than n, they are expressible as a product of primes. We simply multiply these products together to express n as a product of primes.

□

Proposition. There are infinitely many primes.

Proof. Assume there exists a largest prime. Then, the list of primes is $p_1, p_2 \cdots p_k$. Let $n = p_1 p_2 \cdots p_k + 1$. Then n has no prime factor. This is a contradiction immediately because we know that every number greater than two has a factorisation, but this doesn't. □

We want to prove that prime factorisation is unique (up to the ordering). We need that $p \mid ab \implies p \mid a \vee p \mid b$. However, this is hard to answer—p is defined in terms of what divides it, not what it divides. This is the reverse of its definition, so we need to prove it in a more round-about way.

2.5. Highest common factors

For $a, b \in \mathbb{N}$, a number $c \in \mathbb{N}$ is defined to be the highest common factor if:

- $c \mid a$ and $c \mid b$, and
- For all other factors d ($d \mid a$ and $d \mid b$), we have that $d \mid c$.

The second point implies that it is the *highest* common factor, but it is actually slightly stronger. Note that, for example, if a pair's common factors were 1, 2, 3, 4, 6 then the numbers would not have a highest common factor, because 4 does not divide 6.

2.6. The division algorithm

The division algorithm allows us to write any number $n \in \mathbb{N}$ as a multiple $q \in \mathbb{N}$ of $k \in \mathbb{N}$ with some remainder $r \in \mathbb{N}$ such that $0 \leq r < k$; this can be shortened to $n = qk + r$. We begin by writing 1 in this form: $1 = 0k + 1$. Inductively, n can be written as:

$$n = (n-1) + 1 = q_0 k + r_0 + 1$$

where q_0 and r_0 are the results of q and r for $n - 1$. Note that we have two cases:

- If $r_0 + 1 < k$: the result is simply $n = q_0 k + (r_0 + 1)$
- Else ($r_0 + 1 = k$): the result is $n = (q_0 + 1)k + 0$

2.7. Euclid's algorithm

We can find the highest common factor of two natural numbers a and b (without loss of generality, we assume that $a \leq b$). This process is known as Euclid's algorithm.

- Write a as some multiple q_1 of b, with remainder r_1.
- Write b as some multiple q_2 of r_1, with remainder r_2.
- Write r_1 as some multiple q_3 of r_2, with remainder r_3.
- Continue until $r_{n+1} = 0$. Then, r_n is the highest common factor of a and b. We know that the algorithm terminates because $r_k < r_{k-1}$ so it will terminate in at most b steps.

We now prove that the algorithm works.

Proof. We need to prove that it is a common factor and then that it divides all other common factors.

- On the last line of the algorithm, we have $r_{n-1} = q_{n+1} r_n + 0$, so we know that $r_n \mid r_{n-1}$. On the second last line, we have $r_{n-2} = q_n r_{n-1} + r_n$, but r_n divides r_{n-1}, so r_n must divide r_{n-2}. We can continue this logic up to the start of the algorithm, where we can see that $r_n \mid a$ and $r_n \mid b$. So r_n is a common factor of a and b.
- Given some other common factor $d \neq r_n$, we can look at the first line of the algorithm to see that $d \mid r_1$. Using this, we can use the next line to see that $d \mid r_2$. Continuing to the last line, we have $d \mid r_n$.

So r_n is the highest common factor of a and b. Therefore, the highest common factor exists and is unique for any natural numbers a and b. □

Consider running Euclid's algorithm on the numbers 87 and 52.

$$
\begin{aligned}
87 &= 1 \cdot 52 + 35 \\
52 &= 1 \cdot 35 + 17 \\
35 &= 2 \cdot 17 + 1 \\
17 &= 17 \cdot 1 + 0
\end{aligned}
$$

1 is the highest common factor of 87 and 52. Now, we can write 1 as a linear combination of 87 and 52 by looking at each line of this algorithm in the reverse direction (ignoring the

bottom line).

$$\begin{aligned}1 &= 35 - 2 \cdot 17 \\ &= 35 - 2 \cdot (52 - 1 \cdot 35) \\ &= -2 \cdot 52 + 3 \cdot 35 \\ &= -2 \cdot 52 + 3 \cdot (87 - 1 \cdot 52) \\ &= 3 \cdot 87 - 5 \cdot 52\end{aligned}$$

Each two lines of this equation represents one line on Euclid's algorithm. We end up with a linear combination of the two input numbers. We can prove that this linear combination exists in the general case.

Theorem. Let $a, b \in \mathbb{N}$. Then there exist some $x, y \in \mathbb{Z}$ such that $xa + yb = \text{HCF}(a, b)$.

Proof. Run Euclid's algorithm on a and b, and let the output be r_n. Then we have $r_n = xr_{n-1} + yr_{n-2}$ for some $x, y \in \mathbb{Z}$. So, r_n can be written as a linear combination of r_{n-1} and r_{n-2}. Also, from the previous line we know that $r_{n-1} = xr_{n-2} + yr_{n-3}$ for some other x and y. So we can rewrite r_n as a linear combination of r_{n-2} and r_{n-3}. Inductively, we can rewrite r_n as a linear combination of a and b by moving up the lines of the algorithm. □

We can also make an alternate proof without using Euclid's algorithm. Note that this algorithm does not show how to generate this linear combination, it just shows that one exists.

Alternate Proof. Let h be the least positive linear combination of a and b. We want to prove that $h = \text{HCF}(a, b)$.

- Assume that there exists some common factor d of a and b, so that $d \mid a$ and $d \mid b$. Then for some x and y, $d \mid (xa + yb)$. So $d \mid h$.
- Suppose h does not divide a. Then $a = qh + r$ where q is the quotient and r is the remainder ($r \neq 0$). Then $r = a - qh = a - q(xa + yb)$ for some integers x and y. So r is a linear combination of a and b. But this is a contradiction because we said that h was the smallest one. So h divides a.

Therefore h is the highest common factor. □

2.8. Linear Diophantine equations

Suppose a, b and c are natural numbers. When can we solve $ax + by = c$ for $x, y \in \mathbb{Z}$? Well, by looking at the previous theorem, we might guess that c must be some multiple of the highest common factor of a and b. This can be proven in the general case.

Corollary (Bézout's Theorem)**.** Let $a, b, c \in \mathbb{N}$. Then $ax + by = c$ where $x, y \in \mathbb{Z}$ has a solution if and only if $\text{HCF}(a, b) \mid c$.

Proof. Let $h = \text{HCF}(a, b)$. We must prove this bi-implication in both directions.

- First, let us assume that $ax + by = c$ has a solution for some integers x and y. Since $h \mid a$ and $h \mid b$ then $h \mid (ax + by)$ so $h \mid c$.
- Conversely, we know that $h = ax + by$ for some x and y by the above theorem. We can multiply both sides by the integer c/h (this is an integer because $h \mid c$). Then we have an expression for c as a linear combination of a and b as required.

□

2.9. The fundamental theorem of arithmetic

Lemma. Let p be a prime, let $a, b \in \mathbb{N}$. Then $p \mid ab$ implies $p \mid a$ or $p \mid b$.

Proof. Let $p \mid ab$. Then we have two cases, either p divides a or it does not divide a. If it does, our statement is trivially true. Otherwise, we want to prove that p divides b.

Now $\text{HCF}(p, a) = 1$ as p is a prime, and it does not divide a. So 1 can be written as some linear combination of p and a: $px + ay = 1$ for some $x, y \in \mathbb{Z}$.

Now we can multiply both sides by b, giving $pbx + aby = b$. Since p divides ab, p must divide the left hand side. So p divides b. □

Note that we started with a kind of 'negative' statement: 'p does not divide a'; this told us that we cannot do something (namely, factorise it). We turned it into a 'positive' statement: '$px + ay = 1$'; this allows us to rearrange to find out information about these variables. Converting 'negative' statements to 'positive' statements is a useful tool in making proofs.

Theorem (the fundamental theorem of arithmetic)**.** Every $n \in \mathbb{N}$ is uniquely expressible as a product of primes.

Proof. Note that we have already proven that a prime factorisation is possible in Section 3.4; we just need to prove uniqueness of a factorisation (at least, down to its order). We will use induction on some integer n that we wish to factorise. Clearly the theorem is true for $n = 1$ (assuming empty products are valid) and $n = 2$.

So given that $n > 2$ we suppose that there exist two possible factorisations:

$$n = p_1 p_2 \cdots p_k = q_1 q_2 \cdots q_l$$

We want to prove that $k = l$ and that (after reordering) $p_i = q_i$ for all valid i.

We know that $p_1 \mid n$, so $p_1 \mid (q_1 \cdots q_l)$. So there must exist some i where $p_1 \mid q_i$. But since q_i is prime, $p_1 = q_i$. Let us reorder the list such that q_i is moved to the front, so that $p_1 = q_1$.

$$n = p_1 p_2 \cdots p_k = p_1 q_2 \cdots q_l$$

Now, we divide the entire equation by p_1 to give

$$\frac{n}{p_1} = p_2 \cdots p_k = q_2 \cdots q_l$$

The integer $\frac{n}{p_1}$ is smaller than n, so we can use induction to assume that its factorisation is unique. Therefore

$$[p_2, p_3 \cdots p_k] = [q_2, q_3 \cdots q_l]$$

So the prime factorisation of n is unique. □

The common factors of two numbers $m = p_1^{a_1} \cdots p_k^{a_k}$ and $n = p_1^{b_1} \cdots p_k^{b_k}$ where a and b are zero or above is given by $p_1^{c_1} \cdots p_k^{c_k}$ where $c_i \leq \min(a_i, b_i)$ So the highest common factor is given by $c_i = \min(a_i, b_i)$.

The common multiples of those two numbers is given by $d_i \geq \max(a_i, b_i)$. So analogously the lowest common multiple is given by $d_i = \max(a_i, b_i)$.

We have the interesting property that $\mathrm{HCF}(m, n)\,\mathrm{LCM}(m, n) = mn$. This is true because any term p_i is given by $p_i^{\min(a_i,b_i)} p_i^{\max(a_i,b_i)} = p_i^{a_i+b_i}$.

3. Modular arithmetic

3.1. Introduction

In modular arithmetic, we need to prove that things like addition and multiplication are valid. In order to do this, we need to show that if $a \equiv a' \mod n$ and $b \equiv b' \mod n$ then, for example, $ab \equiv a'b'$. We can prove these statements trivially by writing $a' = a + kn$ where k is some integer, then evaluating the left and right hand sides in $\mathbb{Z}$.

Many rules of arithmetic are inherited from $\mathbb{Z}$; for example, addition is commutative. This is easy to realise: to prove that $a + b = b + a$ in $\mathbb{Z}_n$ it is sufficient to prove the statement is true in the whole of $\mathbb{Z}$.

As another example, we can transform the unique prime factorisation lemma into $\mathbb{Z}_p$. In $\mathbb{Z}_p$ where p is prime,

$$ab = 0 \implies (a = 0) \vee (b = 0)$$

In general, $\mathbb{Z}_p$ where p is prime is a very well behaved and convenient-to-use subset of $\mathbb{Z}$.

3.2. Inverses

For any $a, b \in \mathbb{Z}_n$, b is an inverse of a if $ab = 1$. Note that unlike in group theory, it is not necessarily the case that all elements will have inverses. For example, in $\mathbb{Z}_{10}$, the elements 3 and 7 are inverses, but 4 has no inverse. Note that:

- Invertible integers are cancellable. For example, $ab = ac \implies b = c$ if a is invertible (by left-multiplying by its inverse).
- In general, you cannot simply cancel an integer multiple in the realm of modular arithmetic. For example $4 \cdot 5 = 2 \cdot 5$ does not imply $4 = 2$.
- Invertible numbers are also called 'units'.

3.3. Invertibility

Proposition. Let $n \geq 2$. Then every $a \not\equiv 0\,(n)$ is invertible modulo n if and only if $(a, n) = 1$. Note that the parenthesis notation means the highest common factor of the parameters. In particular, if n is prime, then all $1 \leq a < n$ are invertible.

Proof. This first proof uses Euclid's algorithm. If a and n satisfy $(a, n) = 1$ then $ax + ny = 1$ for some $x, y \in \mathbb{Z}$. So $ax = 1 - ny$, so $ax \equiv 1\,(n)$. So x is the inverse of a. □

Proof. This alternate proof only works for $n = p$ where p is a prime; our whole proof lies entirely within $\mathbb{Z}_p$. Consider $0a, 1a, 2a, \cdots, (p-1)a$. Take two numbers i, j between 0 and $p-1$, then consider the condition $ia = ja$. This implies that $(i-j)a = 0$, but $a \neq 0$, so $i = j$. So this list $0a, 1a, \cdots$ contains all distinct elements, all of which must be between 0 and $p-1$.

Therefore, by the pigeonhole principle, one of these elements must be equal to 1. Therefore there exists an inverse for a. □

3.4. Euler's totient function

Definition. Let $\varphi(n)$ be the amount of natural numbers less than or equal to n that are coprime to n.

Here are some examples.

- If p is prime, then $\varphi(p) = p - 1$ since all naturals less than p are coprime to it.
- $\varphi(p^2) = p^2 - p$ because there are p numbers in this range who shares the common factor p with p^2, specifically the numbers $p, 2p, 3p, \cdots, (p-1)p, p^2$.
- If a, b are coprime, $\varphi(ab) = ab - a - b + 1$. There are ab numbers in total to pick from. There are a multiples of b and b multiples of a, and since we discounted ab itself twice we need to count it again. Note that $\varphi(ab) = \varphi(a)\varphi(b)$.

3.5. Fermat's little theorem and Fermat–Euler theorem

Theorem. Let p be a prime. Then in $\mathbb{Z}_p$, $a \neq 0 \implies a^{p-1} = 1$.

This is actually a special case of the following theorem:

Theorem (Fermat–Euler Theorem)**.** Let $n \geq 2$. Then in $\mathbb{Z}_n$, any unit a satisfies $a^{\varphi(n)} = 1$.

Proof. Let the set of units $\mathbb{Z}_n \supset X = \{x_1, x_2, \cdots, x_{\varphi(n)}\}$. Consider multiplying each unit by a. We have $Y = \{ax_1, ax_2, \cdots, ax_{\varphi(n)}\}$. Since a is invertible, this set is comprised of distinct elements. Further, since they are all products of units, they are all units. So Y is a list of $\varphi(n)$ distinct units, so this list must be equal to X. Now, since the lists are the same, the product of all their elements must be the same. So $\prod X = \prod Y = a^{\varphi(n)} \prod X$. We can cancel the factor of $\prod X$ because it is a product of invertibles, leaving $1 = a^{\varphi(n)}$ as required. □

If alternatively we wanted to prove this just for p prime, then we could replace $\varphi(n)$ with $p - 1$, and $\prod X$ with $(p-1)!$.

3.6. Square roots of one

Lemma. Let p be prime. Then in $\mathbb{Z}_p$, $x^2 = 1$ has solutions 1 and -1 only.

Note. In $\mathbb{Z}_8$, for example, we have $1^2 = 3^2 = 5^2 = 7^2 = 1$, so obviously this does not hold in the general case.

Proof. $x^2 = 1$ implies that $(x-1)(x+1) = 0$. Because of the $p \mid ab \implies (p \mid a) \vee (p \mid b)$ lemma, we know that $(x-1) = 0$ or $(x+1) = 0$, so -1 and 1 are the only solutions. □

3.7. Square roots of negative one

Theorem (Wilson's Theorem)**.** Let p be prime. Then $(p-1)! \equiv -1\ (p)$.

Proof. Since this is obviously true for $p = 2$, we will suppose that $p > 2$. In $\mathbb{Z}_p$, let us consider the list $1, 2, 3 \cdots (p-1)$. We can pair each a with its inverse a^{-1} for all $a \neq a^{-1}$. Note that $a = a^{-1} \iff a^2 = 1$ so in this case $a = 1$ or $a = -1$. So let us now multiply each element together, to get

$$(p-1)! = (aa^{-1})(bb^{-1}) \cdots 1 \cdot -1 = (1) \cdot (1) \cdots 1 \cdot -1 = -1$$

□

Proposition. Let $p > 2$ be prime. Then -1 is a square number modulo p if and only if $p \equiv 1\ (4)$.

Proof. If $p > 2$ then p is odd. There are therefore two cases, either $p \equiv 1$ or $p \equiv 3$ modulo 4. Each case is proven individually.

- $(p = 4k + 3)$ Suppose that $x^2 = -1$ in $\mathbb{Z}_p$. The only thing we know about powers in modular arithmetic is Fermat's Little Theorem, so we will have to use this. So, $x^{p-1} = x^{4k+2} = 1$. Therefore, $(x^2)^{2k+1} = 1$. But we know that $x^2 = -1$, and we raise this -1 to an odd power, which is -1. So this is a contradiction.
- $(p = 4k + 1)$ By Wilson's Theorem, we know that $(4k)! = -1$. We intend to show that this is a square number in the world of $\mathbb{Z}_p$. We will compare the termwise expansion of $(4k)!$ and $[(2k)!]^2$ on consecutive lines.

$$\begin{aligned} (4k)! &= 1 \cdot 2 \cdot 3 \cdots (2k) \cdot (2k+1) \cdot (2k+2) \quad \cdots (4k-1) \quad \cdot (4k) \\ [(2k)!]^2 &= 1 \cdot 2 \cdot 3 \cdots (2k) \cdot 1 \qquad\quad \cdot 2 \qquad\quad \cdots (2k-1) \quad \cdot (2k) \end{aligned}$$

By writing each term as an equivalent negative:

$$= 1 \cdot 2 \cdot 3 \cdots (2k) \cdot (-4k) \quad \cdot (-4k+1) \cdots (-2k-2) \cdot (-2k-1)$$

Extracting out the negatives:

$$= 1 \cdot 2 \cdot 3 \cdots (2k) \cdot (4k) \qquad \cdot (4k-1) \quad \cdots (2k+2) \quad \cdot (2k+1) \cdot (-1)^{2k}$$

which is equal to the first line by rearranging. So $[(2k)!]^2 = (4k)! = -1$. So -1 is a square number modulo p.

□

3.8. Solving congruence equations

Let us try to solve the equation $7x \equiv 4\ (30)$. We take a two-phase approach: first, we will find a single solution, and then we will find all of the other solutions.

Since 7 and 30 are coprime, we can use Euclid's algorithm to find a way of expressing 1 in terms of 7 and 30, in particular $13 \cdot 7 - 3 \cdot 30 = 1$. This allows us to solve $7y \equiv 1\ (30)$, by setting $y = 13$. Then, of course, we can multiply both sides by 4: $7y \cdot 4 \equiv 4\ (30)$, so $x = y \cdot 4 = 13 \cdot 4 = 22$.

We can now find other solutions (apart from trivially adding $30k$). Suppose that there exists some other solution x', i.e. $7x' \equiv 4\ (30)$. Then $7x \equiv 7x'\ (30)$. As 7 is invertible modulo 30, we can simply multiply by the inverse of 7 to give $x \equiv x'\ (30)$. So x is unique modulo 30. Alternatively, we could solve the equation without any of this working out by noticing that 7 is invertible! However, this is not very likely to happen in the general case, since it requires that the coefficient of x is coprime to the modulus.

Now, let's try a different equation, $10x = 12\ (34)$. Since 10 is not invertible, we can't do quite the same thing as above. We can't also just divide the whole thing by 2, there isn't a rule for that in general. We can, however, move into $\mathbb{Z}$ and manipulate the expression there. $10x = 12 + 34y$ for some $y \in \mathbb{Z}$, so we can divide the equation by 2 to get $5x = 6 + 17y$, so $5x = 6\ (17)$ and we can solve from there.

3.9. Chinese remainder theorem

Is there a solution for the simultaneous congruences

$$x \equiv 6\ (17); \quad x \equiv 2\ (19)$$

17 and 19 are coprime, so congruence mod 17 and congruence mod 19 are independent of each other. How about

$$x \equiv 6\ (34); \quad x \equiv 11\ (36)$$

In this instance, there is obviously no solution; should x be even or odd? We can see that, the smallest amount we can adjust x by in one equation while retaining congruence in the other equation is HCF(34, 36), which is 2.

Theorem. Let u, v be coprime. Then for any a, b, there exists a value x such that

$$x \equiv a\ (u); \quad x \equiv b\ (v)$$

and that this value is unique modulo uv.

Proof. We first prove existence of such an x. By Euclid's Algorithm, we have $su + tv = 1$ for some integers s, t. Note that therefore:

$$su \equiv 0\ (u); \quad tv \equiv 0\ (v); \quad su \equiv 1\ (v); \quad tv \equiv 1\ (u);$$

Therefore we can make a linear combination of su and tv that is the required size in each congruence, specifically

$$x = (su)b + (tv)a$$

Now we prove that this value x is unique modulo uv. Suppose there was some other solution x'. Also, $x' \equiv x\ (u)$ and $x' \equiv x\ (v)$. So we have $u \mid (x' - x)$ and $v \mid (x' - x)$ but as u and b are coprime we have $uv \mid (x' - x)$. So x is unique modulo uv. □

3.10. RSA encryption

A practical use of number theory is RSA encryption, which is an asymmetric encryption protocol that allows encryption by using a public and private key pair. We will begin by first choosing two large distinct primes p and q. By large, we mean primes that are hundreds of digits long; in practice, these primes are between around 512 bits and 2048 bits long when represented in binary. Let $n = pq$, and pick a 'coding exponent' e. Our message that we want to send must be an element of $\mathbb{Z}_n$, so if it is not representable in this form we must break it apart into several smaller messages, or perhaps use RSA to share some kind of small symmetric key for another encryption algorithm. Let this message be x, so $x < n$.

To encode x, we raise it to the power e in $\mathbb{Z}_n$. To efficiently compute large powers of x, we can use a repeated squaring technique. For example, we can find x, x^2, x^4, x^8, x^{16} through repeated squaring, and then for example we can calculate $x^{19} = x^{16}x^2x^1$.

To decode x^e, we ideally want some number d such that $(x^e)^d = x$. By the Fermat–Euler Theorem, we have $x^{\varphi(n)} = 1$, so clearly $x^{k\varphi(n)+1} = x$. In other words, we want $ed \equiv 1 \mod \varphi(n)$. By running Euclid's algorithm on e and $\varphi(n)$, we can find such a d. Note that this requires e and $\varphi(n)$ to be coprime; in practice we would choose e after we have chosen n such that this is the case.

Now, we can see that to encode a message, all you need is n and e. However, to decode, you need to also know d, which means you need to know $\varphi(n) = \varphi(pq) = pq - p - q + 1$ which requires that you know the original p and q. If we pick sufficiently large p and q, our n will be so big as to be almost impossible to factorise in any decent length of time. So we can publish n and e as our public key, and anyone may use these numbers to encrypt a message that then only we can decode.

4. The reals

4.1. Motivation for the reals

Why do we need the real numbers in the first place? Well, we introduce new sets of numbers when there are equations that we cannot solve using our current number system. For example, the equation $x+2=0$ is not solvable in $\mathbb{N}$, so we constructed $\mathbb{Z}$. Then we could not solve equations like $2x=3$, so we created the rationals, $\mathbb{Q}$. Now, we cannot solve equations such as $x^2=2$, so we must create a new set of numbers that contains this solution.

Proposition. There does not exist a $q \in \mathbb{Q}$ such that $q^2=2$. Note that in this proposition we make no assumption that $q^2=2$ is solvable, or that a solution if one exists does not lie within $\mathbb{Q}$; we simply state that confined to the realm of $\mathbb{Q}$ the equation is unsolvable.

Proof 1. Suppose that such a $q \in \mathbb{Q}$ exists, such that $q^2=2$. Without loss of generality, we will assume that $q>0$ because $(-q)^2=q^2$. So let q be written as a/b where $a,b \in \mathbb{N}$. Then $a^2/b^2=2$, so $a^2=2b^2$. If we factorise each side as a product of primes, the exponent of the prime 2 on the left hand side must be even, but on the right hand side it must be odd. This contradicts the unique factorisation of natural numbers. So such a q does not exist. □

Proof 2. Suppose that there exists some $q \in \mathbb{Q}$ written similarly to above as a/b. Note that for any $c,d \in \mathbb{Z}$, $cq+d$ is of the form e/b for some integer e. Therefore, if $cq+d>0$ then $cq+d \geq 1/b$.

Now, note that $0<(q-1)<1$, so for a suitably large n, we have $0<(q-1)^n<1/b$. However, $(q-1)^n$ is of the form $cq+d$ because $q^2=1$ so we can eliminate all exponents. This is a contradiction so such a q does not exist. □

We can see from the proofs above that $\mathbb{Q}$ has a 'gap' at $\sqrt{2}$. How can we express this fact without mentioning $\mathbb{R}$? We can't just say plainly that $\sqrt{2} \notin \mathbb{Q}$ because as far as we know from $\mathbb{Q}$, there is no reason to assume that such a number called $\sqrt{2}$ even exists! We need to find a way to express the concept of $\sqrt{2}$ in the language of $\mathbb{Q}$. One way to do this is by creating sone set $S=\{q \in \mathbb{Q} : q^2<2\}$. Then we can write down some upper bounds for this set. For example, 2 is a trivial upper bound, as is 1.5, and as is 1.42. In fact, we can continue making smaller and smaller upper bounds. We can see therefore that there exists no least upper bound in $\mathbb{Q}$.

4.2. Axioms of the reals

We define the reals as follows: the reals are a set written $\mathbb{R}$ with elements 0 and 1 with $0 \neq 1$; with operations $+$ and $\cdot$; and an ordering $<$; such that:

(i) $+$ is commutative, associative, has identity 0, and there are inverses for all elements;

(ii) $\cdot$ is commutative, associative, has identity 1, and there are inverses for all nonzero elements;

(iii) $\cdot$ is distributive over $+$;

(iv) for all a and b in $\mathbb{R}$, exactly one of $a < b$, $a = b$ and $a > b$ are true, and that $a < b$ and $b < c$ implies $a < c$;

(v) for all $a, b, c \in \mathbb{R}$, $a < b$ implies $a + c < b + c$, and $a < b$ implies $ac < bc$ when $c > 0$; and

(vi) for any set S of reals that is non-empty and bounded above, S has a least upper bound.

There are some notable immediate remarks about the definitions of the reals.

- We can contain the rationals inside the reals: $\mathbb{Q} \subset \mathbb{R}$
- The least upper bound axiom is false in $\mathbb{Q}$, which is why it's so important in $\mathbb{R}$.
- Why did we specify 'non-empty' and 'bounded above' in the least upper bound axiom? Of course, if a set is not bounded above, then it has no upper bound, so clearly it can have no least upper bound. If a set is empty, then every real is an upper bound for this set, and as there is no least real number, there is no least upper bound.
- It is possible to construct $\mathbb{R}$ out of $\mathbb{Q}$, and check that the above axioms hold. However, this is a rare example where the construction of $\mathbb{R}$ is complicated and irrelevant, so it is not covered here.

The reals do not contain infinitely big or infinitesimally small elements.

Proposition (the axiom of Archimedes)**.** $\mathbb{N}$ is not bounded above in $\mathbb{R}$.

Proof. If there were some upper bound $c = \sup \mathbb{N}$, then $c-1$ is clearly not an upper bound for $\mathbb{N}$. So there exists some natural number n such that $n > c-1$. But then clearly $n+1 \in \mathbb{N} > c$ contradicting the existence of this upper bound. □

Corollary. For each $t \in \mathbb{R} > 0$, $\exists n \in \mathbb{N}$ such that $\frac{1}{n} < t$.

Proof. We have some $n \in \mathbb{N}$ with $n > \frac{1}{t}$ by the above proposition. So $\frac{1}{n} < t$. □

4.3. Examples of sets and least upper bounds

Note that a common way to write 'least upper bound' is the word supremum, denoted $\sup S$.

(i) Let $S = \{x \in \mathbb{R} : 0 \leq x \leq 1\} = [0, 1]$. The least upper bound of S is 1, because:

- 1 is an upper bound for S; $\forall x \in S, x \leq 1$; and
- Every upper bound y must have $y \geq 1$ because $1 \in S$.

(ii) Let $S = \{x \in \mathbb{R} : 0 < x < 1\} = (0, 1)$. $\sup S = 1$ because:

- 1 is an upper bound for S; $\forall x \in S, x \leq 1$; and

- No upper bound c has $c < 1$. Indeed, certainly $c > 0$ ($c > \frac{1}{2}$ since $\frac{1}{2} \in S$). So if $c < 1$, then $0 < c < 1$, so the number $\frac{1+c}{2} \in S$ and is larger than c, so it is not an upper bound.

(iii) Let $S = \left\{1 - \frac{1}{n} : n \in \mathbb{N}\right\}$. $\sup S = 1$ because:

- 1 is clearly an upper bound.
- Let us suppose $c < 1$ is an upper bound. Then $\forall n \in \mathbb{N}, 1 - \frac{1}{n} < c$ so $1 - c < \frac{1}{n}$. From the corollary of the Axiom of Archimedes above, this is a contradiction.

Remark. If S has a greatest element, then this element is the supremum of the set: $\sup S \in S$. But if S does not have a greatest element, then $\sup S \notin S$. Also, we do not need any kind of 'greatest lower bound' axiom—if S is a non-empty, bounded below set of reals, then the set $\{-x : x \in S\}$ is non-empty and bounded above, and so has a least upper bound, so S has a greatest lower bound equivalent to its additive inverse. This is commonly called the 'infimum', or $\inf S$.

Theorem. $\exists x \in \mathbb{R}$ with $x^2 = 2$.

Proof. Let S be the set of all real numbers such that $x^2 < 2$. Of course, it is non-empty (try $x = 0$) and bounded above (try $x = 2$). So let $c = \sup S$; we want to show that $c^2 = 2$. We prove this by eliminating all alternatives; clearly either $c^2 < 2$, $c^2 = 2$ or $c^2 > 2$.

- ($c^2 < 2$) We want to prove that $(c + t)^2 < 2$ for some small t. For $0 < t < 1$, we have $(c + t)^2 = c^2 + 2ct + t^2 \leq c^2 + 5t$, since c is at most 2, and t^2 is at most t. So this value is less than 2 for some suitably small t, contradicting the least upper bound—we have just shown that $(c + t) \in S$.
- ($c^2 > 2$) We want to prove that $(c - t)^2 > 2$ for some small t. For $0 < t < 1$, we have $(c-t)^2 = c^2-2ct+t^2 \geq c^2-4t$, since c is at most 2, and t^2 is at least zero. So this value is greater than 2 for some suitably small t, contradicting the least upper bound—we have just created a lower upper bound.

So $c^2 = 2$. □

This same kind of proof works for a lot of real values, for example $\sqrt[n]{x}$ for $n \in \mathbb{N}, x \in \mathbb{R}, x < 0$. Reals that are not rational are called irrational. This is a negative statement however, so it is better in proofs to suppose that something is rational, and then show a contradiction.

Also, the rationals are 'dense'; for any $a, b \in \mathbb{R}$, there is another rational between them. We may assume without loss of generality that they are both non-negative and that $a < b$. Then pick some $n \in \mathbb{N}$ with $\frac{1}{n} < b - a$. Among the list $\frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \ldots$, there is a final one that is less than or equal to a, which we will denote $\frac{q}{n}$ (otherwise a is an upper bound to this list, contradicting the axiom of Archimedes). So $a < \frac{q+1}{n} < b$ as required.

The irrationals are also dense; for any reals a and b with the same conditions above, these exists some irrational c with $a < c < b$. We know that there exists a rational c with $a\sqrt{2} < c < b\sqrt{2}$, so $a < \frac{c}{\sqrt{2}} < b$.

4.4. Sequences and limits

How can we ascribe meaning to expressions like this?

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots$$

Certainly, we have a concept of addition, and we can keep adding as many terms as we like, but there is no implicit definition of an infinite sum from the aforementioned axioms.

A definition that makes sense would involve partial sums x_n of this infinite series. However, we could not just say that the partial sums get progressively closer to a value, because then trivially something like $\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \ldots$ tends to 107, even though they're clearly getting closer.

A more accurate definition would be to state that we can get arbitrarily close (within some given ε) to a 'limit value' c by taking some amount of terms n of this series: $c - \varepsilon < x_n < c + \varepsilon$. But this is still wrong: the sequence $\frac{1}{2}, 10, \frac{2}{3}, 10, \frac{3}{4}, 10, \frac{4}{5}, 10, \ldots$ could then tend to 1 even though every other term is 10.

The best definition would state that the sequence of partial sums would *stay* within ε of c for all x_k where $k \geq n$ for some $n \in \mathbb{N}$. In less formal words, for any $\varepsilon > 0$, x_n will eventually stay within ε of c. Equivalently, $\forall \varepsilon > 0, \exists N \in \mathbb{N}$ such that $\forall n > N$ we have $|x_n - c| < \varepsilon$.

(i) Consider the sequence $\frac{1}{2}, \frac{1}{2} + \frac{1}{4}, \frac{1}{2} + \frac{1}{4} + \frac{1}{8}, \ldots$. This is $x_1, x_2, x_3, \ldots$ where $x_n = 1 - \frac{1}{2^n}$ (inductively on n). We want to show that x_n tends to 1. Given some $\varepsilon > 0$, we choose some $N \in \mathbb{N}$ with $N > \frac{1}{\varepsilon}$. Then, for every $n \geq N$, $|x_n - 1| = \frac{1}{2^n} \leq \frac{1}{n} \leq \frac{1}{N} < \varepsilon$.

(ii) Consider the constant sequence $c, c, c, c, \ldots$. We want to show that $x_n \to c$. Given some $\varepsilon > 0$, we have $|x_n - c| < \varepsilon$ for all n; $N = 1$ is the time after which the sequence stays within ε of c.

(iii) Consider now $x_n = (-1)^n$, i.e. $-1, 1, -1, 1, \ldots$. We want to show that this does not tend to a limit. Suppose $x_n \to c$ as $n \to \infty$. We may choose some ε that acts as a counterexample—for example, $\varepsilon = 1$. So $\exists N \in \mathbb{N}$ such that $\forall n \geq n$ we have $|x_n - c| < 1$. In particular, $|1 - c| < 1$ and $|-1 - c| < 1$ so $|1 - (-1)| < 2$, by the triangle inequality. This is a contradiction.

(iv) The sequence x_n given by

$$x_n = \begin{cases} \frac{1}{n} & n \text{ odd} \\ 0 & n \text{ even} \end{cases}$$

should tend to zero. Given some $\varepsilon > 0$, we will choose $N \in \mathbb{N}$ with $\frac{1}{N} < \varepsilon$. Then for all $n \geq N$, either $x_n = \frac{1}{n}$ or 0. In either case, $|x_n - 0| \leq \frac{1}{n} \leq \frac{1}{N} < \varepsilon$.

We can denote the entirety of a sequence $x_1, x_2, \ldots$ as

$$(x_n) \quad \text{or} \quad (x_n)_{n=1}^{\infty}$$

For example, $((-1)^n)_{n=1}^{\infty}$ is divergent. This isn't saying that it goes to infinity, just that it doesn't converge. Note also that if $x_n \to c$ and $x_n \to d$, then $c = d$. Suppose that $c \neq d$. Then pick $\varepsilon = \frac{|c-d|}{2}$. Then $\exists N \in \mathbb{N}$ with $|x_n - c| < \varepsilon$, and $\exists M \in \mathbb{N}$ with $|x_n - d| < \varepsilon$. After the point $\max(N, M)$, the points must be within ε of both c and d, but as c and d are 2ε apart this is a contradiction (by the triangle inequality).

4.5. Series

A sequence given in the form $x_1,\ x_1 + x_2,\ x_1 + x_2 + x_3, \ldots$ is called a series. They are often written $\sum_{n=1}^{\infty} x_n$. The kth term of the sequence, given by $\sum_{n=1}^{k} x_n$, is called the kth partial sum. If the series converges to some value c, then we can write $\sum_{n=1}^{\infty} x_n = c$. Note that we cannot use this notation to denote the limit until we know that the limit actually exists. This is just the same as with sequences, where we cannot write $\lim_{n\to\infty} x_n$ until we know that the limit exists.

Limits behave as we would expect. For example, if $x_n \leq d$ for all n, and $x_n \to c$, then $c \leq d$. Suppose $c > d$. Then we will choose $\varepsilon = \frac{|c-d|}{2}$. Then there are no points x_n within this bound of c #.

Proposition. If $x_n \to c$ and $y_n \to d$, then $x_n + y_n \to c + d$.

Proof. Given some $\varepsilon > 0$, let $\zeta = \frac{1}{2}\varepsilon$. Then, after some term x_N, $|x_n - c| < \zeta$, and after some term y_M, $|y_m - d| < \zeta$. So for every $n \geq \max(M, N)$, by the triangle inequality, $|(x_n + y_n) - (c + d)| < 2\zeta = \varepsilon$ as required. □

This is commonly known as an $\varepsilon/2$ argument. Also, if we had instead not taken any ζ value and just stuck with ε, it would still be a good proof because we could just have divided ε at the beginning—it's not expected that you completely rewrite the proof to add in this division.

4.6. Testing convergence of a sequence

A sequence $x_1, x_2, \ldots$ is called 'increasing' if $x_{n+1} \geq x_n$ for all n.

Theorem. If $x_1, x_2, \ldots$ is increasing and bounded above, it converges to a limit.

This is a very important theorem that we will refer back to time and time again.

Note. If we were in $\mathbb{Q}$, this would not necessarily hold. For example, consider the decimal expansion of $\sqrt{2}$.

$$1, 1.4, 1.41, 1.414, 1.4142, \ldots$$

They don't converge to a limit in $\mathbb{Q}$. So our proof will have to be more rigorous than just 'they have to tend to somewhere below the upper bound'; we must use a property that $\mathbb{R}$ has that $\mathbb{Q}$ does not have, i.e. the least upper bound axiom.

Proof. Let $c = \sup\{x_1, x_2, \ldots\}$. We want to prove that $x_n \to c$. Given some $\varepsilon > 0$, there exists some n such that $x_n > c - \varepsilon$ (else, $c - \varepsilon$ would be a smaller upper bound #). As the sequence is increasing, all x_k where $k > n$ are at least x_n. So $|x_k - c| < \varepsilon$ as required. □

Of course, a decreasing sequence works in an identical way; if it is bounded below then it converges. More compactly, a bounded monotone sequence is convergent (where monotone means either increasing or decreasing).

Proposition. The harmonic series

$$\sum_{n=1}^{\infty} \frac{1}{n}$$

diverges; the solution to the Basel problem

$$\sum_{n=1}^{\infty} \frac{1}{n^2}$$

converges.

There is no closed form for the nth term of either of these sequences, which is one reason that series are often more challenging to work with than regular sequences.

Proof. Since the harmonic series is difficult to deal with, we will compare it to a sequence that we understand easier. Therefore, we show that the first sequence diverges using a comparison test with powers of 2, one of the simplest series.

$$\begin{aligned} &1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \frac{1}{9} + \cdots \\ \geq\ &1 + \frac{1}{2} + \underbrace{\frac{1}{4} + \frac{1}{4}}_{\frac{1}{2}} + \underbrace{\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}}_{\frac{1}{2}} + \frac{1}{16} + \cdots \end{aligned}$$

By inspection, we can see that the harmonic series is larger than the sum of an infinite amount of $\frac{1}{2}$, so surely it must diverge. More rigorously:

$$\begin{aligned} \frac{1}{3} + \frac{1}{4} &\geq \frac{1}{2} \\ \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} &\geq \frac{1}{2} \\ \frac{1}{2^n + 1} + \frac{1}{2^n + 2} + \cdots + \frac{1}{2^{n+1}} &\geq \frac{2^n}{2^{n+1}} = \frac{1}{2} \end{aligned}$$

So the partial sums of the series are unbounded, so the series diverges. For the sum of reciprocals of squares, we want to do a similar thing because again the only simple sequence we have to work with is the powers of 2.

$$\begin{aligned} &1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{6^2} + \frac{1}{7^2} + \frac{1}{8^2} + \frac{1}{9^2} + \cdots \\ \leq\ &1 + \underbrace{\frac{1}{2^2} + \frac{1}{2^2}}_{\frac{2}{2^2}} + \underbrace{\frac{1}{4^2} + \frac{1}{4^2} + \frac{1}{4^2} + \frac{1}{4^2}}_{\frac{4}{4^2}} + \frac{1}{8^2} + \frac{1}{8^2} + \cdots \end{aligned}$$

The bottom sequence simplifies to just the sequence $1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots \to 2$, and the upper sequence is bounded above by the lower sequence. More rigorously:

$$\begin{aligned} \frac{1}{2^2} + \frac{1}{3^2} &\leq \frac{2}{2^2} = \frac{1}{2} \\ \frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{6^2} + \frac{1}{7^2} &\leq \frac{4}{4^2} = \frac{1}{4} \\ \frac{1}{(2^n)^2} + \frac{1}{(2^n+1)^2} + \cdots + \frac{1}{(2^{n+1}-1)^2} &\leq \frac{2^n}{(2^n)^2} = \frac{1}{2^n} \end{aligned}$$

So the partial sums are bounded, and hence the series converges by the above theorem. □

In fact, $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$. This is proved in the Linear Analysis course in Part II.

4.7. Decimal expansions

What should $0.a_1a_2a_3\ldots$ mean (where each a is a digit from 0 to 9)? It should be the limit of $0.a_1$, $0.a_1a_2$, $0.a_1a_2a_3$ and so on. We will define it by

$$0.a_1a_2a_3\cdots := \sum_{n=1}^{\infty} \frac{a_n}{10}$$

This clearly converges as the partial sums are increasing and bounded above by 1, so infinite decimal expansions are valid. Conversely, given some $x \in \mathbb{R}$ with $0 < x < 1$, we can certainly write it as a (potentially infinite) decimal. We will start by choosing the greatest a_1 from 0 to 9 such that $\frac{a_1}{10} \leq x$. Thus $0 < x - \frac{a_1}{10} < \frac{1}{10}$. Now, we can pick the greatest a_2 in the set such that $\frac{a_1}{10} + \frac{a_2}{100} \leq x$. Therefore, $0 \leq x - \frac{a_1}{10} - \frac{a_2}{100} < \frac{1}{100}$. Continue inductively, and then we obtain a decimal expansion $0.a_1a_2a_3\ldots$ such that $0 \leq x - \sum_{n=1}^{k} \frac{a_n}{10^n} < \frac{1}{10^k}$ for any given k. By the definition of convergence, the sequence given for a tends to x as required.

Note, if $0.a_1a_2\ldots$ and $0.b_1b_2\ldots$ are different decimal expansions of the same number, then there exists some $N \in \mathbb{N}$ such that $a_n = b_n$ for all $n < N$ and $a_N = b_N - 1$ and $a_n = 9, b_n = 0$ for all $n > N$ (or vice versa). For example, $0.99999\ldots$ is equivalent to $1.00000\ldots$

4.8. The number e

We define

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \ldots$$

The partial sums are increasing and bounded above by the powers of two after the first term, so it converges.

4.9. Algebraic and transcendental numbers

A real x is called algebraic if it is a root of a nonzero polynomial with integer coefficients. Otherwise, it is called transcendental. For example, any rational $\frac{p}{q}$ is algebraic as it is the root of $qx - p = 0$. As another example, $\sqrt{2} + 1$ is algebraic as it is a root of the equation $x^2 - 2x - 1 = 0$. The logical next question to ask is whether all reals are algebraic.

Proposition. e is not rational.

Proof. Suppose that e is rational, let it be written $\frac{p}{q}$, where $q > 1$ (if $q = 1$, rewrite it as $\frac{2p}{2q}$). Multiplying up by $q!$ (easier than just q because then we can compare factorials) gives

$$\sum_{n=0}^{\infty} \frac{q!}{n!} \in \mathbb{Z}$$

We know that $\sum_{n=0}^{q} \frac{q!}{n!} \in \mathbb{Z}$. The next terms are:

$$\begin{aligned}
\frac{q!}{(q+1)!} &= \frac{1}{q+1} \\
\frac{q!}{(q+2)!} &= \frac{1}{(q+1)(q+2)} \leq \frac{1}{(q+1)^2} \\
\frac{q!}{(q+3)!} &= \frac{1}{(q+1)(q+2)(q+3)} \leq \frac{1}{(q+1)^3} \\
\frac{q!}{(q+n)!} &\leq \frac{1}{(q+1)^n}
\end{aligned}$$

So the next partial sums are bounded above by the geometric series.

$$\sum_{n=q+1}^{\infty} \frac{q!}{n!} \leq \frac{1}{q} < 1$$

So the whole series multiplied by $q!$ is a whole number plus a fractional part, which is not an integer #. □

Ideally now we'd have a proof that e is transcendental. However, even though the terms of e tend to zero quickly, they don't tend to zero quite quickly enough for us to be able to prove it using what we know now. We instead prove that there exists some transcendental number using a different example, one whose terms tend to zero very quickly indeed.

Theorem. Liouville's constant $c = \sum_{n=1}^{\infty} \frac{1}{10^{n!}}$ is transcendental. As a decimal expansion:

$$c = 0.11000100000000000000000010\ldots$$

This is a long proof, the hardest in this course. We will cherry-pick some important results about polynomials in order to make this proof, without a proper introduction to features of polynomials.

- For any polynomial P, $\exists k \in \mathbb{R}$ such that $|P(x) - P(y)| \leq k|x - y|$ for all $0 \leq x, y \leq 1$. Indeed, say $P(x) = a_d x^d + \cdots + a_0$, then

$$\begin{aligned} P(x) - P(y) &= a_d(x^d - y^d) + a_{d-1}(x^{d-1} - y^{d-1}) + \cdots + a_1(x - y) \\ &= (x - y)[a_d(x^{d-1} + x^{d-2}y + \cdots + y^{d-1}) + \cdots + a_1] \\ |P(x) - P(y)| &\leq |x - y|[(|a_d| + |a_{d-1}| + \cdots + |a_1|)d] \end{aligned}$$

  because x and y are between 0 and 1.

- A nonzero polynomial of degree d has at most d roots. Given some polynomial P of degree d:
  - If P has no roots, we are trivially done.
  - If P has some root a, then P can be written as $(x - a)Q(x)$. Inductively, $Q(x)$ has at most $d - 1$ roots, so P has at most d roots.

Now we can prove the above theorem.

Proof. We will write $c_n = \sum_{k=0}^{n} \frac{1}{10^{k!}}$, such that $c_n \to c$. Suppose that some polynomial P has c as a root. Then $\exists k$ such that $|P(x) - P(y)| \leq k|x - y|$ when $0 \leq x, y \leq 1$. Let P have degree d, such that

$$P(x) = a_d x^d + \cdots + a_0$$

Now, $|c - c_n| = \sum_{k=n+1}^{\infty} \frac{1}{10^{k!}} \leq \frac{2}{10^{(n+1)!}}$. This is a trivial upper bound, of course better upper bounds exist.

Also, $c_n = \frac{a}{10^{n!}}$ for some $a \in \mathbb{Z}$. So $P(c_n) = \frac{b}{10^{dn!}}$ for some $b \in \mathbb{Z}$ (since $P(\frac{s}{t}) = \frac{q}{t^d}$ for some integer q, where $\frac{s}{t} \in \mathbb{Q}$).

For n large enough, c_n is not a root, because P only has finitely many roots. So

$$|P(c) - P(c_n)| = |P(c_n)| \leq \frac{1}{10^{dn!}}$$

Therefore

$$\frac{1}{10^{dn!}} \leq k\frac{2}{10^{(n+1)!}}$$

which is a contradiction if n is large enough. □

Here are some remarks about this proof.

- This same proof shows that any real x such that $\forall n \exists \frac{p}{q} \in \mathbb{Q}$ with $0 < \left|x - \frac{p}{q}\right| < \frac{1}{q^n}$ is transcendental. Informally, x has very good rational approximations.
- Such x are often called Liouville numbers; the proof works for all Liouville numbers.
- This proof does not show that e is transcendental (even though it is), because the terms do not go to zero fast enough.
- We now know that there exist some transcendental numbers. Another proof of existence of transcendental numbers will be seen in a later lecture.

4.10. Complex numbers

Some polynomials have no real roots, for example $x^2 + 1$. We'll try to 'force' an x with the property $x^2 = -1$. Note that for example we could not force an x into existence with the property $x^2 = 2, x^3 = 3$; how do we know introducing i will not lead to a contradiction? We will define $\mathbb{C}$ to consist of the plane $\mathbb{R}^2$, i.e. pairs of real numbers, with operations $+$ and $\cdot$ which satisfy:

- $(a, b) + (c, d) = (a + c, b + d)$
- $(a, b) \cdot (c, d) = (ac - bd, ad + bc)$

We can view $\mathbb{R}$ as being contained within $\mathbb{C}$ by identifying the real number a with $(a, 0)$. Note that the rules of arithmetic of the reals are inherited inside the first element of the complex plane, so there is no contradiction here. Then let $i = (0, 1)$. Trivially then, any point (a, b) in the complex numbers may be written as $a + bi$ where $a, b \in \mathbb{R}$. And, of course, $i^2 = -1$.

All of the basic rules like associativity and distributivity work in the complex plane. There are multiplicative inverses: given $a + bi$, we know that $(a + bi)(a - bi) = a^2 + b^2$ so $\frac{a-bi}{a^2+b^2}$ is the inverse (provided the point is nonzero). This kind of structure with familiar properties is known as a field, for example $\mathbb{C}$, $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{Z}_p$ where p is prime. The fundamental theorem of algebra states that any nonzero polynomial with complex coefficients has a complex root; this is proven in the IB course Complex Analysis.

5. Sets

5.1. Sets and subsets

A set is any* collection of mathematical objects. $(\forall x, x \in A \iff x \in B) \iff (A = B)$. In words, two sets which have the same members are considered to be the same; order of members is not important in a set. There is no 'multiple membership' of a set, $\{a, a\} = \{a\}$.

Given a set A and a property $p(x)$, we can form $\{x \in A : p(x)\}$; the subset of all members of A with property p. This is sometimes called the 'subset selection' rule or axiom. We can say that B is a subset of A if $\forall x, x \in B \implies x \in A$, written $B \subseteq A$. Further, $A = B \iff A \subseteq B, B \subseteq A$.

5.2. Composing sets

Given sets A and B, we can form their union $A \cup B = \{x : x \in A \vee x \in B\}$. We can also form their intersection $A \cap B = \{x : x \in A \wedge x \in B\}$. If $A \cap B = \varnothing$, we say A and B are disjoint. Note that we could consider $A \cap B$ as a special case of subset selection; the subset of A with the property that the element is in B. Therefore, $A \cap B \subseteq A$, and $A \cap B \subseteq B$. We define the set difference $A \setminus B = \{x \in A : x \notin B\}$.

Note that $\cap$ and $\cup$ are commutative and associative. Also, $\cup$ is distributive over $\cap$, and $\cap$ is distributive over $\cup$. For example, let us prove that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

- (LHS $\subseteq$ RHS) Given $x \in A \cap (B \cup C)$, we have $x \in A$ and also either $x \in B$ or $x \in C$. If $x \in B$ then $x \in A \cap B$ so $x \in (A \cap B) \cup (A \cap C)$; and vice versa for C.
- (RHS $\subseteq$ LHS) Given $x \in (A \cap B) \cup (A \cap C)$, either $x \in A \cap B$ or $x \in A \cap C$. If $x \in A \cap B$ then $x \in A$ and $x \in B \cup C$ as required; and vice versa for the other case.

As the union is associative, we can have bigger unions of more sets. For example, if we let $A_n = \{n^2, n^3\}$ for each $n \in \mathbb{N}$, the infinite union

$$A_1 \cup A_2 \cup A_3 \cup \cdots = \bigcup_{n=1}^{\infty} A_n = \bigcup_{n \in \mathbb{N}} A_n = \{x \in N : x \text{ is a square or a cube}\}$$

When we use the $n \in \mathbb{N}$ on the large union symbol, we call $\mathbb{N}$ the 'index set'. Note that the infinite union is not defined as a limit of finite unions; it is simply defined using set comprehension. In general, given a set I, and sets A_i, $i \in I$, we can form

$$\bigcup_{i \in I} A_i = \{x : \exists i \in I, x \in A_i\}$$

and

$$\bigcap_{i \in I} A_i = \{x : \forall i \in I, x \in A_i\}$$

Note that we cannot form an intersection when $I = \varnothing$, as will be explained later.

For any a, b, we can form the ordered pair (a, b), where equality is checked component-wise. For sets A, B, we can form their product $A \times B = \{(a, b) : a \in A, b \in B\}$. For example, $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ can be viewed as a plane. We can form other sizes of tuples similarly.

For any set X, we can form the power set $\mathcal{P}(X)$ consisting of all subsets of X.

$$\mathcal{P}(X) = \{Y : Y \subseteq X\}$$

For example:

$$\mathcal{P}(\{1, 2\}) = \{\varnothing, \{1\}, \{2\}, \{1, 2\}\}$$

5.3. Russell's paradox

For a set A, we can always form the set $\{x \in A : p(x)\}$ for any property p. We cannot, however, form the set $\{x : p(x)\}$. Suppose we could form such a set, then we could form the set $X = \{x : x \notin x\}$. Now, is $X \in X$? If this is true, then it fails the defining property $x \notin x$. If this is false, then the defining property is true, so it must be in the set. This is a contradiction in both cases.

Similarly, there is no 'universal' set $\mathcal{E}$, meaning $\forall x, x \in \mathcal{E}$. Otherwise we could form the X above by $\{x \in \mathcal{E} : p(x)\}$. To guarantee that a given set exists, we need to obtain it in some way from known sets.

5.4. Finite sets

We will write $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For $n \in \mathbb{N}_0$, we can say that a set A has size n if we can write $A = \{a_1, a_2, \cdots, a_n\}$ where the a_i are distinct. A set is called finite if it has a size $n \in \mathbb{N}_0$.

Note that a set cannot have size n and size m for $n \neq m$. Suppose that A has size n and size m where $n, m > 0$. Then, removing an element, we obtain a set that has size $n - 1$ and $m - 1$. By induction on the larger of n and m, we will eventually reach a size of both zero and nonzero which is a contradiction.

Proposition. A set of size n has exactly 2^n subsets.

Proof 1. We may assume that our set is simply $\{1, 2, \cdots, n\}$ by relabelling. When constructing a subset S from this set, there are n independent binary choices for whether a given element should be within this subset, since for example either $1 \in S$ or $1 \notin S$ must be true. So there are 2^n distinct choices of subset you can make. □

Proof 2. We will prove this inductively on n, noting that $n = 0$ is trivial. For any subset $T \subseteq \{1, 2, \cdots n - 1\}$, how many $S \subseteq \{1, \cdots, n\}$ have $S \cap \{1, 2, \cdots n - 1\} = T$? Exactly two: T and $T \cup \{n\}$. So there are two choices for how to extend this subset to the new element n. So the number of subsets is $2 \cdot 2^{n-1} = 2^n$. □

In some sense Proof 2 is a more 'formal' version of Proof 1, using induction rather than intuition. We sometimes say that if A has size n, then $|A| = n$, and that A is an n-set.

5.5. Binomial coefficients

For $n \in \mathbb{N}_0$ and $0 \leq k \leq n$, we can write $\binom{n}{k}$ for the number of subsets of an n-set that are of size k.

$$\binom{n}{k} = |\{S \subseteq \{1, 2, \dots, n\} : |S| = k\}|$$

For example, there are six 2-sets in a 4-set. There is a formula for this, but generally this definition is a lot easier to use. Note that $\binom{n}{0} = 1$, $\binom{n}{n} = 1$, and $\binom{n}{1} = n$ where $n > 0$.

Note that $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} = 2^n$ as each side counts the number of subsets in an n-set. Also:

(i) $\binom{n}{k} = \binom{n}{n-k}$ $(\forall n \in N_0, 0 \leq k \leq n)$. Indeed, specifying which k members to pick for a subset is equivalent to specifying which $n - k$ members not to pick.

(ii) $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$ $(\forall n \in \mathbb{N}, 0 < k < n)$. Indeed, the number of k-subsets of $\{1, 2, \dots, n\}$ without n is $\binom{n-1}{k}$. The number of k-subsets of $\{1, 2, \dots, n\}$ that do contain n is $\binom{n-1}{k-1}$ as we must pick the remaining $k-1$ elements of this new subset. So in total, $\binom{n-1}{k-1} + \binom{n-1}{k}$ encapsulates both possibilities.

This last point illustrates that Pascal's Triangle will give all the binomial coefficients since it perfectly encapsulates the relationship between a given element of the triangle with two elements from the previous row. The exact proof follows from the other known properties of the binomial coefficients.

5.6. Computing binomial coefficients

Proposition.

$$\binom{n}{k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k(k-1)(k-2)\cdots(1)}$$

Proof. The number of ways to name a k-set is $n(n-1)(n-2)\cdots(n-k+1)$ because there are n ways to choose a first element, $n-1$ ways to choose a second element, and so on. We have overcounted the k-sets, though—there are $k(k-1)(k-2)\cdots(1)$ ways to name a given k-set because you have k choices for the first element, $k-1$ choices for the second element, and so on. Hence the number of k-sets in $\{1, 2, \dots, n\}$ is the required result. □

Note that we can also write

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

but this is a very unwieldy formula to use especially by hand, so will be rarely used. Further, we can make asymptotic approximations using this formula, for example $\binom{n}{3} \sim \frac{n^3}{6}$ for large n.

5.7. Binomial theorem

Theorem. For all $a, b \in \mathbb{R}, n \in \mathbb{N}$, we have

$$(a+b)^n = \binom{n}{0}a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \cdots + \binom{n}{n}b^n$$

Proof. When we expand $(a+b)^n = (a+b)(a+b)\dots(a+b)$, we obtain terms of the form $a^k b^{n-k}$. To get a single term of this form, we must choose k brackets for which to take the a value in the expansion, and the other $n-k$ brackets will take the b value. The number of terms of the form $a^k b^{n-k}$ for a fixed k is therefore the amount of ways of choosing k brackets out of a total of n, which is $\binom{n}{k}$. So

$$(a+b)^n = \sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k} = \sum_{k=0}^{n}\binom{n}{n-k}a^k b^{n-k}$$

□

For example, we can tell that $(1+x)^n$ reduces to

$$1 + nx + \frac{1}{2}n(n-1)x^2 + \frac{1}{3!}n(n-1)(n-2)x^3 + \cdots + nx^{n-1} + x^n$$

So when x is small, a good approximation to $(1+x)^n$ is $1+nx$.

5.8. Inclusion-exclusion theorem

Given two finite sets A, B, we have

$$|A \cup B| = |A| + |B| - |A \cap B|$$

For three sets, we have

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |B \cap C| - |C \cap A| + |A \cap B \cap C|$$

Theorem. Let $S_1, \dots, S_n$ be finite sets. Then,

$$\left|\bigcup_{S \in S_n} S\right| = \sum_{|A|=1}|S_A| - \sum_{|A|=2}|S_A| + \sum_{|A|=3}|S_A| - \cdots$$

where

$$S_A = \bigcap_{i \in A} S_i$$

and

$$\sum_{|A|=k}$$

is a sum taken over all $A \subseteq \{1, 2, \dots, n\}$ of size k.

Proof. Let x be an element of the left hand side. We wish to prove that x is counted exactly once on the right hand side. Without loss of generality, let us rename the sets that x belongs to as $S_1, S_2, \dots, S_k$.

Then the number of sets A with $|A| = 1$ such that $x \in S_A$ is k. The number of sets A with $|A| = 2$ such that $x \in S_a$ is $\binom{k}{2}$, since we must choose two of the sets $S_1, \dots, S_k$, so there are $\binom{k}{2}$ ways to do this. So in general, the amount of A with $|A| = r$ with $x \in S_A$ is just $\binom{k}{r}$.

So the number of times x is counted on the right hand side is

$$k - \binom{k}{2} + \binom{k}{3} - \cdots + (-1)^{k+1}\binom{k}{k}$$

But $(1 + (-1))^k$ by the binomial expansion is

$$1 - \binom{k}{1} + \binom{k}{2} - \binom{k}{3} + \cdots + (-1)^k\binom{k}{k}$$

So the number of times x is counted on the right hand side is $1-(1+(-1))^k = 1-0 = 1$. □

6. Functions

6.1. Definition

For sets A and B, a function f from A to B is a rule that assigns to each $x \in A$ a unique value $f(x) \in B$. More precisely, a function from A to B is a set $f \subseteq A \times B$ such that for every $x \in A$, there is a unique $y \in B$ with $(x, y) \in f$. Of course therefore, if $(x, y) \in f$ then we can write $f(x) = y$. Here are some examples.

(i) $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$, or using an alternative notation, $x \mapsto x^2$ is a function.

(ii) A non-example is $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = \frac{1}{x}$ since it is undefined at $x = 0$.

(iii) Another non-example is $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = \pm\sqrt{|x|}$ since it does not define a unique value in the output space for a given input, such as $x = 2$.

(iv) $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \begin{cases} 1 & x \in \mathbb{Q} \\ 0 & \text{otherwise} \end{cases}$$

is a function since it clearly satisfies the second definition. Note that even though we don't know if $e + \pi$ is rational or not, the function is still well defined since it produces a unique solution for $f(e + \pi)$, we just don't know which output value it gives.

(v) $A = \{1, 2, 3, 4, 5\}$, $B = \{1, 2, 3, 4\}$, and $f : A \to B$ is given by

$$\begin{aligned} f(1) &= 1 \\ f(2) &= 4 \\ f(3) &= 3 \\ f(4) &= 3 \\ f(5) &= 4 \end{aligned}$$

(vi) $A = \{1, 2, 3\}$, $f : A \to A$ is given by

$$\begin{aligned} f(1) &= 1 \\ f(2) &= 3 \\ f(3) &= 2 \end{aligned}$$

(vii) $A = \{1, 2, 3, 4\}$, $f : A \to A$ is given by

$$\begin{aligned} f(1) &= 1 \\ f(2) &= 3 \\ f(3) &= 3 \\ f(4) &= 4 \end{aligned}$$

(viii) $A = \{1, 2, 3, 4\}$, $B = \{1, 2, 3\}$, $f : A \to B$ is given by

$$f(1) = 3$$
$$f(2) = 3$$
$$f(3) = 2$$
$$f(4) = 1$$

6.2. Injection, surjection and bijection

Definition. A function $f : A \to B$ is

- injective, if $\forall a, a' \in A$, we have $a \neq a' \implies f(a) \neq f(a')$, or equivalently, $f(a) = f(a') \implies a = a'$, or in words, 'different points stay different' (e.g. example 6 above).
- surjective, if $\forall b \in B$, $\exists a \in A$ such that $f(a) = b$, or in words, 'everything in B is hit' (e.g. examples 6 and 8).
- bijective, if it is injective and surjective, or in words, 'everything in B is hit exactly once', or 'f pairs up elements of A and elements of B' (e.g. example 6, or $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^3$).

Definition. For a function $f : A \to B$, A is the domain, B is the range, and $\{b \in B : \exists a \in A \text{ s.t. } f(a) = b\}$ is the image.

We must always provide the domain and range of a function; a function's properties depend on this. For example, is the function f defined by $f(x) = f^2$ injective? If $f : \mathbb{N} \to \mathbb{N}$, then it is injective, but if $f : \mathbb{Z} \to \mathbb{Z}$, then it is not.

There are a number of properties that hold specifically for finite sets A, B:

(i) There is no surjection $A \to B$ if $|B| > |A|$.

(ii) There is no injection $A \to B$ if $|A| > |B|$.

(iii) For a function $f : A \to A$, f injective $\iff$ f surjective. Hence, if f is either injective or surjective, it is bijective.

(iv) There is no bijection from A to any proper subset of A.

As counterexamples for infinite sets:

(i) We define $f_0 : \mathbb{N} \to \mathbb{N}$ by $f_0(x) = x + 1$. Then, f_0 is injective but not surjective.

(ii) We define $f_1 : \mathbb{N} \to \mathbb{N}$ by $f_0(x) = x - 1$, or 1 if $x = 1$. Then, f_0 is surjective but not injective.

(iii) We define $g : \mathbb{N} \to \mathbb{N} \setminus \{1\}$ by $g(x) = x + 1$. Then, g is bijective between $\mathbb{N}$ and a proper subset of $\mathbb{N}$.

We provide some more examples of functions.

(i) For any set X we have $1_X : X \to X$ defined by $1_X(x) = x$. This is known as the identity function on X.

(ii) For any set X, and $A \subset X$, we have an indicator function (or characteristic function) $\chi_A : X \to \{0, 1\}$ defined by

$$\chi_A(x) = \begin{cases} 0 & x \notin A \\ 1 & x \in A \end{cases}$$

(iii) A sequence of reals $x_1, x_2, \ldots$ is a function $f : \mathbb{N} \to \mathbb{R}$ defined by $f(n) = x_n$.

(iv) The operation $+$ on $\mathbb{N}$ is a function $\mathbb{N}^2 \to \mathbb{N}$.

(v) A set X has size $n \iff$ there is a bijection between X and $\{1, 2, \ldots, n\}$.

6.3. Composition of functions

Given $f : A \to B$ and $g : B \to C$, we define the composition $g \circ f : A \to C$, given by $(g \circ f)(a) = g(f(a))$. For example, if $f : \mathbb{R} \to \mathbb{R}$, $f(x) = 2x$, $g : \mathbb{R} \to \mathbb{R}$, $g(x) = x + 1$, then $(f \circ g)(x) = 2(x + 1)$, and $(g \circ f)(x) = 2x + 1$.

In general, the operation $\circ$ is not commutative, as we can see from this example. However, $\circ$ is associative. Given $f : A \to B$, $g : B \to C$, $h : C \to D$, we have $h \circ (g \circ f) = (h \circ g) \circ f$. Indeed, for any input $x \in A$,

$$(h \circ (g \circ f))(x) = h((g \circ f)(x)) = h(g(f(x))) = (h \circ g)(f(x)) = ((h \circ g) \circ f)(x)$$

Thus $(h \circ (g \circ f))(x) = ((h \circ g) \circ f)(x)$ for every $x \in A$, so $h \circ (g \circ f) = (h \circ g) \circ f$.

6.4. Invertibility

We say that a function $f : A \to B$ is invertible if there exists some $g : B \to A$ such that $g \circ f = 1_A$ and $f \circ g = 1_B$. For example $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = 2x + 1$ has inverse $g : \mathbb{R} \to \mathbb{R}$ given by $g(x) = \frac{x-1}{2}$. We can prove that this is correct by showing for all real numbers that $(g \circ f)(x) = x$ and vice versa as required.

As an example, consider $f_0 : \mathbb{N} \to \mathbb{N}$ given by $f_0(x) = x + 1$, and $f_1 : \mathbb{N} \to \mathbb{N}$ given by $f_1(x) = x - 1$ if $x \neq 1$ and 1 if $x = 1$. $f_1 \circ f_0 = 1_{\mathbb{N}}$ but $f_0 \circ f_1 \neq 1_{\mathbb{N}}$ because they disagree at 1. So we must check inverses both ways.

In fact, $f : A \to B$ is invertible if and only if it is a bijection.

- (forward implication) Let g be the inverse of f. It is surjective because $\forall b \in B$, we have $b = f(g(b))$. It is injective because given two elements a, a' such that $f(a) = f(a')$, we have $g(f(a)) = g(f(a')) = a = a'$ as required. So it is bijective.
- (backward implication) Suppose f is bijective. Let $g(b)$ be the unique point $a \in A$ with $f(a) = b$ for all $b \in B$. Then this g is the inverse of f.

6.5. Relations

A relation on a set X is a subset of $R \subseteq X \times X$. We usually write aRb to denote $(a, b) \in R$. Here are some examples.

(i) On $\mathbb{N}$, aRb if $a \equiv b$ (5). For example, $2R12$ but not $2R11$.

(ii) On $\mathbb{N}$, aRb if $a \mid b$.

(iii) On $\mathbb{N}$, aRb if $a \neq b$.

(iv) On $\mathbb{N}$, aRb if $a = b \pm 1$.

(v) On $\mathbb{N}$, aRb if $|a - b| \leq 2$.

(vi) On $\mathbb{N}$, aRb if either $a, b \leq 6$ or $a, b > 6$.

A relation may have a number of important properties:

- (reflexive) If $\forall x \in X$, xRx, e.g. examples 1, 2, 5, 6.
- (symmetric) If $\forall x, y \in X$, $xRy \implies yRx$, e.g. examples 1, 3, 4, 5, 6.
- (transitive) If $\forall x, y, z \in X$, $xRy, yRz \implies xRz$, e.g. examples 1, 2, 6.

An equivalence relation is a relation that is reflexive, symmetric and transitive. Examples 1, 6 above are equivalence relations. Here are some more examples.

(i) On $\mathbb{N}$, xRy if $x = y$.

(ii) Considering a partition of set X into subsets $C_1, C_2, \ldots, i \in I$ where the C_i are non-empty and disjoint, and their union is X. Then consider the relation aRb if $\exists i$ such that $a \in C_i$ and $b \in C_i$. aRb is an equivalence relation. In fact, all equivalence relations can be considered to be in this form; we will prove this shortly.

For an equivalence relation R on a set X, and $x \in X$, we define the equivalence class $[x] = \{y \in X : yRx\}$. In the first example 1 above, $[2] = \{y \in \mathbb{N} : y \equiv 2\ (5)\}$.

6.6. Equivalence classes as partitions

Proposition. Let R be an equivalence relation on a set X. Then the equivalence classes of R partition X.

Proof. Each equivalence class $[x]$ is non-empty, since $x = x$. Further,

$$\bigcup_{x \in X} = X$$

since $x \in [x]$ for all $x \in X$. Now we must show that the classes are disjoint, or are equal. Given x, y with $[x] \cap [y] \neq \emptyset$, we need to show that $[x] = [y]$. Choose some z such that $z \in [x] \cap [y]$. Then, zRx and zRy, so xRy. Thus for any t, $tRx \implies tRy$ due to transitivity, and $tRy \implies tRx$ for the same reason. So $[x] = [y]$. □

As an example, does there exist an equivalence relation on $\mathbb{N}$ with three equivalence classes, two of which are infinite, and one of which is finite? Yes—we can break up $\mathbb{N}$ into three parts, for example positive numbers, negative numbers and zero. This defines an equivalence relation.

6.7. Quotients

Given an equivalence relation R on a set X, the quotient of X by R is

$$X/R = \{[x] : x \in X\}$$

The map $q : X \to X/R$ given by $x \mapsto [x]$ is called the 'quotient map' or 'projection map'. As an example, on $\mathbb{Z} \times \mathbb{N}$, let us define $(a, b)R(c, d)$ to be true if $ad = bc$. This is an equivalence relation that demonstrates equivalence of rational numbers, where a, c are the numerators and b, d are denominators. Here, $\mathbb{Z} \times \mathbb{N}/R$ is a copy of $\mathbb{Q}$, associating $[(a, b)]$ with a/b. Then, $q : \mathbb{Z} \times \mathbb{N} \to \mathbb{Q}$ would map (a, b) to a/b.

7. Countability

7.1. Basic properties

We have a notion of 'size' for finite sets. Is there such an analogous notion for infinite sets? We will say that a set X is countable if X is finite, or it bijects with $\mathbb{N}$. Equivalently, we can list out the elements of the set, and each element will appear in the list. Here are some examples.

(i) Clearly any finite set is countable.

(ii) $\mathbb{N}$ is countable.

(iii) $\mathbb{Z}$ is countable, let us construct the list of numbers

$$0, 1, -1, 2, -2, 3, -3, 4, -4, \ldots$$

It makes sense now to consider two sets to have the same size if they biject with each other.

Proposition. A set X is countable if and only if it injects into $\mathbb{N}$.

Proof. The forward implication is trivial: if X is finite, then there must be an injection in to $\mathbb{N}$, and if it bijects with $\mathbb{N}$ then that bijection is a valid injection. This encompasses both cases of countable sets.

Now let us consider the reverse implication. We may assume X is infinite, since if X is finite then by definition X is countable. We know that X injects onto $\mathbb{N}$ under some injective function f, so X bijects with $\operatorname{Im} f$. So it is enough to show that the image $\operatorname{Im} f$ is countable. We will now set a_1 to be the least element of $\operatorname{Im} f$, and a_2 to be the least element not equal to a_1, and so on. In general, $a_n = \min(\operatorname{Im} f \setminus \{a_i : 0 \leq i < n\})$. Then $\operatorname{Im} f$ is the set $\{a_1, a_2, \ldots\}$. There are no extra elements that we have not covered, since each $a \in X$ is a_n for some n, because $a = a_n, n \leq a$. So we have listed elements of $\operatorname{Im} f$, so $\operatorname{Im} f$ is countable, so X is countable. □

Thus, we can view countability as being 'at most as large as $\mathbb{N}$'. For instance, any subset of a countable set is also countable.

Remark. In $\mathbb{R}$, let

$$X = \left\{\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \ldots\right\} \cup \{1\}$$

Then X is countable, as we can list it as

$$1, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \ldots$$

But if we counted from 'least element' to 'most element', we would never reach the element 1 in countable time. Note further that if we find it difficult to construct a list for a set, it does not mean it is uncountable, it could just mean that we haven't found the right list yet.

7.2. Products of countable sets

Theorem. $\mathbb{N} \times \mathbb{N}$ is countable.

Proof 1. We will define $a_1 = (1, 1)$, and inductively define

$$a_n = \begin{cases} (p-1, q+1) & \text{if } p > 1 \\ (q+1, 1) & \text{if } p = 1 \end{cases}$$

where $a_{n-1} = (p, q)$. Therefore, we are essentially moving across antidiagonals of the plane. This does hit every point $(x, y) \in \mathbb{N} \times \mathbb{N}$, for example by induction on $x + y$, so we have listed all elements of $\mathbb{N} \times \mathbb{N}$. □

Proof 2. If we can define an injective function $\mathbb{N} \times \mathbb{N} \to \mathbb{N}$, then it is countable. For example, let $f = 2^x 3^y$. f is injective, so $\mathbb{N} \times \mathbb{N}$ is countable. □

7.3. Countable unions of countable sets

Proof 2 is also a way to show the following theorem:

Theorem. Let $A_1, A_2, A_3, \ldots$ be countable sets. Then $A_1 \cup A_2 \cup A_3 \cup \ldots$ is countable. Less formally, 'a countable union of countable sets is countable'.

Proof. For each i, A_i is countable, so we can list A_i as $a_{i1}, a_{i2}, a_{i3}, \ldots$ which may or may not terminate. We can then define

$$f : \bigcup_{n \in \mathbb{N}} A_n \to \mathbb{N}; \quad f(x) = 2^i 3^j$$

where $x = a_{ij}$. If x is in more than one set, just take the least i that is valid. Then f is an injection so the union is countable. □

Here are some examples of using this theorem by partitioning sets as a countable union of countable subsets.

(i) $\mathbb{Q}$ is countable, since it is a countable union of countable sets:

$$\mathbb{Q} = \mathbb{Z} \cup \frac{1}{2}\mathbb{Z} \cup \frac{1}{3}\mathbb{Z} \cup \ldots$$

Each $\frac{1}{n}\mathbb{Z}$ is countable, since they biject with $\mathbb{Z}$ which is a countable set. It doesn't matter if we've counted an element in $\mathbb{Q}$ twice; the above theorem works even with intersecting sets.

(ii) The set $\mathbb{A}$ of all algebraic numbers is countable. It is enough to show that the set of integer polynomials is countable, since each polynomial has a finite amount of roots and then $\mathbb{A}$ is a countable union of finite sets. Now, to show that the set of integer polynomials is countable, it is enough to show that for each degree d it is countable, since it is a countable union of all polynomials of degree d (again using the above theorem). To specify a polynomial of degree d you must name its coefficients, so this set injects into $\mathbb{Z}^{d+1}$, so we must just show that $\mathbb{Z}^{d+1}$ is countable (not a bijection since the first term of the polynomial must be nonzero). We know that $\mathbb{Z}^n$ is countable because we can inductively show that $\mathbb{Z}^2, \mathbb{Z}^3, \mathbb{Z}^4, \ldots$ are countable inductively.

7.4. Uncountable sets

Definition. A set is uncountable if there is no way to count the set.

Theorem. $\mathbb{R}$ is uncountable.

Proof (Cantor's Diagonal Argument). We will show that $(0,1)$ is uncountable, then clearly $\mathbb{R}$ is uncountable. Suppose $(0,1)$ is countable. Then given a sequence $r_1, r_2, \ldots$ in $(0,1)$, we just need to find some number $s \in (0,1)$ not contained within this sequence. For each r_n, we have a decimal expansion $r_n = 0.r_{n1}r_{n2}r_{n3}\ldots$. Let us now write all of these numbers in a matrix-style form:

$$\begin{aligned} r_1 &= 0.r_{11}r_{12}r_{13}\ldots \\ r_2 &= 0.r_{21}r_{22}r_{23}\ldots \\ r_3 &= 0.r_{31}r_{32}r_{33}\ldots \\ &\vdots \end{aligned}$$

We just need to construct some number s that is not in this list. So, let us simply make sure that for any given r value, there is at least one digit that does not match. The easiest way to construct such a number is

$$s = 0.s_1s_2s_3\ldots$$

where $s_1 \neq r_{11}$, $s_2 \neq r_{22}$, $s_3 \neq r_{33}$ and so on. We can pick any numbers we like according to these constraints, but we should avoid picking digits 0 and 9 since $0.1000\cdots = 0.0999\ldots$ for example, which could cause unnecessary ambiguity. Then $s \neq r_1, s \neq r_2, \ldots$ since there is at least one mismatched digit in the expansion for each r_i; they differ in decimal digit i. So $\mathbb{R}$ is uncountable. □

This is another proof that transcendental numbers exist. $\mathbb{R}$ is uncountable and $\mathbb{A}$ is countable, so there exists a transcendental number. Indeed, 'most' numbers are transcendental, i.e. $\mathbb{R} \setminus \mathbb{A}$ is uncountable (because if $\mathbb{R} \setminus \mathbb{A}$ were countable, then $\mathbb{R}$ would be $(\mathbb{R} \setminus \mathbb{A}) \cup \mathbb{A}$ which is a finite union of countable sets #).

Theorem. The power set $\mathcal{P}(\mathbb{N})$ is uncountable.

Proof. Suppose the subsets of $\mathbb{N}$ are listed as $S_1, S_2, S_3, \ldots$ then we want to construct another set S that is not equal to any of the other sets S_i. So for each set S_i, we must ensure that S and S_i differ for at least one value. An easy way to do this is to include the number i in the subset if S_i does not contain the number, and to exclude i if $i \in S_i$. Then S differs from S_i at position i. This is the same logic as the diagonal argument above. We have:

$$S = \{n \in \mathbb{N} : n \notin S_n\}$$

So S is not on the list $S_1, S_2, S_3, \ldots$ no matter what way we choose to list the elements, so $\mathcal{P}(\mathbb{N})$ is uncountable. □

Remark. Alternatively, we could just inject $(0, 1)$ into $\mathcal{P}(\mathbb{N})$. For example, consider $x \in (0, 1)$ represented as $0.x_1x_2x_3x_4 \ldots$ in binary where the $x_1, x_2, \ldots$ are zero or one (not ending with an infinite amount of 1s). We can convert this x into a subset of $\mathbb{N}$ by considering the set $\{n \in \mathbb{N} : x_n = 1\}$. Then the uncountability follows.

In fact, our proof of this theorem shows the following.

Theorem. For any set X, there is no bijection from X to the power set $\mathcal{P}(X)$.

For example, $\mathbb{R}$ does not biject with $\mathcal{P}(\mathbb{R})$. The proof in fact will show that there is no surjection from X to its power set; i.e. the power set is 'larger' than X.

Proof. Given any function $f : X \to \mathcal{P}(X)$, we will show f is not surjective. Let $S = \{x \in X : x \notin f(x)\}$. Then S does not belong to the image of f because they differ at element x; for all x we have $S \neq f(x)$. □

Remark. Note that:

(i) This is similar in some sense to Russell's paradox.

(ii) This theorem gives another proof that there is no universal set $\mathcal{E}$, since its power set $\mathcal{P}(\mathcal{E}) \subseteq \mathcal{E}$. But of course, there is always a surjection from a set to a subset. This is a contradiction.

Example. Let $A_i, i \in I$ be a family of open, pairwise disjoint intervals. Must this family be countable? Note that it is not as simple as just listing from left to right, for example consider

$$\left(\frac{1}{2}, 1\right), \left(\frac{1}{3}, \frac{1}{2}\right), \left(\frac{1}{4}, \frac{1}{3}\right), \ldots, (-1, 0)$$

Then the leftmost interval is $(-1, 0)$, but there is no 'next interval' just after it. Also consider

$$\left(0, \frac{1}{2}\right), \left(\frac{1}{2}, \frac{2}{3}\right), \left(\frac{2}{3}, \frac{3}{4}\right), \ldots, (1, 2)$$

Then we can list the first infinitely many intervals, but we will never reach $(1, 2)$. The answer turns out to be true; the family is countable. Here are two important proofs.

Proof 1. Each interval A_i contains a rational number a_i. The rationals $\mathbb{Q}$ are countable. So let us just list the a_i. The family is therefore countable. □

Proof 2. $\{i \in I : A_i \text{ has length } \leq 1\}$ is certainly countable, since it injects into $\mathbb{Z}$ (here, as all A_i contain some integer). Further, $\left\{i \in I : A_i \text{ has length } \leq \frac{1}{2}\right\}$ is countable for the same reason. Essentially, for all n, $\left\{i \in I : A_i \text{ has length } \leq \frac{1}{n}\right\}$ is countable. We have written all the intervals as a countable union (over n) of countable sets. □

To summarise, if we want to show a set X is uncountable:

(i) Run a diagonal argument; or

(ii) Inject an uncountable set into X

To show a set X is countable:

(i) List all the elements (usually fiddly); or

(ii) Inject X into $\mathbb{N}$ (or another countable set); or

(iii) Express X as a countable union of countable sets (usually the best); or

(iv) If X is 'in' or 'near' $\mathbb{R}$, consider $\mathbb{Q}$ (see Proof 2 above).

7.5. Comparing sizes of sets

Intuitively, we might think that:

- 'A bijects with B' means 'A has the same size as B'.
- 'A injects into B' means 'A is at most as large as B'.
- 'A surjects onto B' means 'A is at least as large as B'.

Of course, these analogies break down where B is zero, since there are no functions from A to B in this case. For these to make sense, we require (for $A, B \neq \emptyset$) 'A injects into B' to be true if and only if 'B surjects onto A', and vice versa.

- In the forward direction, we are given an injection $f : A \to B$. Pick some point a_0 in A, and define a surjective function $g : B \to A$ given by

$$b \mapsto \begin{cases} a & \text{if } \exists!\ a \in A, f(a) = b \\ a_0 & \text{otherwise} \end{cases}$$

  Since the mapping f is injective, the first case will always provide a unique value of a.

- Proving the converse, we are given a surjection $g : B \to A$. For each a in A, we have some $a' \in B$ with $g(a') = a$ since g is a surjection. Let $f(a) = a'$ for each $a \in A$, and f is injective.

7.6. Schröder–Bernstein theorem

Further, we must also have that if 'A is at most as large as B' and 'B is at most as large as A', then they must be the same size. Otherwise this size intuition would not make sense.

Theorem (Schröder–Bernstein Theorem)**.** If $f : A \to B$ and $g : B \to A$ are injections, then there exists a bijection $h : A \to B$.

Proof. For $a \in A$, we will write $g^{-1}(a)$ to denote the unique $b \in B$ such that $g(b) = a$, if such a b exists (and similarly for $f^{-1}(b)$). The 'ancestor sequence' of $a \in A$ is

$$g^{-1}(a), f^{-1}g^{-1}(a), g^{-1}f^{-1}g^{-1}(a), \ldots$$

which may terminate. So for any ancestor, after undergoing the relevant function f or g repeatedly, we will end up at a. There are three possible behaviours:

- Let A_0 be the subset of A such that the ancestor sequence stops at even time, i.e. the last ancestor is in A;
- Let A_1 be the subset of A such that the ancestor sequence stops at odd time, i.e. the last ancestor is in B; and
- Let A_∞ be the subset of A such that the ancestor sequence does not terminate.

We specify 0 to be even, i.e. if $a \in A$ has no ancestor $g^{-1}(a)$, then $a \in A_0$. We define similar subsets of B: B_0, B_1, B_∞. Now:

- $f : A \to B$ is a bijection between A_0 and B_1. Clearly if some element a has an even number of ancestors, the ancestors of $f(a)$ are exactly a and all of its ancestors, i.e. an odd number. It is surjective because every element in B_1 has an inverse $f^{-1}(b) \in A_0$ by construction.
- $g : B \to A$ is a bijection between B_0 and A_1 due to the same argument.
- f (or g, both functions work for this proof) bijects A_∞ and B_∞. It is surjective because for every element $b \in B$, it has some ancestor $f^{-1}(b) \in A_\infty$.

So the function $h : A \to B$ is given by

$$h(a) = \begin{cases} f(a) & \text{if } a \in A_0 \\ g^{-1}(a) & \text{if } a \in A_1 \\ f(a) & \text{if } a \in A_\infty \end{cases}$$

is a bijection. □

Let us consider an example of this theorem in action. Do $[0, 1]$ and $[0, 1] \cup [2, 3]$ biject? All we need is to find an injection both ways.

- Let $f : [0, 1] \to [0, 1] \cup [2, 3]$ be the identity map $f(x) = x$.

- Let $g : [0,1] \cup [2,3] \to [0,1]$ be given by $g(x) = x/3$.

It would also be nice to have that, for any sets A and B, either A injects into B or B injects into A. Then we can create a total ordering, rather than a partial ordering; we can compare any two sets. This is proven to be true in the Part II course Logic and Set Theory.

7.7. Arbitrarily large sets

We have the sets

$$\mathbb{N}, \mathcal{P}(\mathbb{N}), \mathcal{P}(\mathcal{P}(\mathbb{N})), \dots, \mathcal{P}^k(\mathbb{N}), \dots$$

Does every set X inject into one of those? It seems like this might be true, but the set

$$X = \mathbb{N} \cup \mathcal{P}(\mathbb{N}) \cup \mathcal{P}(\mathcal{P}(\mathbb{N})) \cup \dots$$

is a counterexample. Let us continue further with this approach.

$$X' = X \cup \mathcal{P}(X) \cup \mathcal{P}(\mathcal{P}(X)) \cup \dots$$

$$X'' = X' \cup \mathcal{P}(X') \cup \mathcal{P}(\mathcal{P}(X')) \cup \dots$$

and so on. Now, does every set inject into one of these sets? No, consider

$$Y = X \cup X' \cup X'' \cup X''' \cup \dots$$

We can keep going forever. So we can't construct a set that all sets inject into.

7.8. What happens next?

This is the end of the Numbers and Sets course. Here are a few of the courses that feed from this course.

- Factorisation is taken further in the IB Groups, Rings and Modules course.
- Fermat's Little Theorem, squares modulo p etc. are taken further in II Number Theory.
- The analysis chapter is extended by IA Analysis.
- Countability and sizes of sets are taken further in the II Logic and Set Theory course.

II. Differential Equations

Lectured in Michaelmas 2020 by Dr. J. R. Taylor

A differential equation is an equation involving one or more unknown functions and their derivatives. These equations arise in many fields of study, such as physics and biology. In this course, we explore many different ways to solve some common types of differential equations.

In many cases, it is not possible to solve differential equations, so it is important to classify various cases that we can solve, and explore them in depth. Heuristically, a differential equation is often easier to solve if it involves fewer variables, and if the derivatives involved have a lower order. First, we will study differential equations in only one variable: the 'ordinary' differential equations. Towards the end of the course, we study 'partial' differential equations, which can involve more than one variable.

Contents

1. Differentiation 55
1.1. Basic definitions 55
1.2. Rules for differentiation 56
1.3. Order of magnitude 56
1.4. Equation of a tangent 58
1.5. Taylor series 58
1.6. L'Hôpital's rule 59
2. Integration 60
2.1. Definition of integration 60
2.2. Fundamental theorem of calculus 61
2.3. Integration techniques 61
3. Multivariate functions 63
3.1. Partial derivatives 63
3.2. Multivariate chain rule 63
3.3. Change of variables 65
3.4. Implicit differentiation 66
3.5. Differentiating an integral with respect to a parameter 67
4. Linear ordinary differential equations 68
4.1. Eigenfunctions 68
4.2. Solving first order ODEs 68
4.3. Constant forcing 68
4.4. Eigenfunction forcing 69
4.5. Non-constant coefficients 70
5. Discrete equations 71
5.1. Numerical integration 71
5.2. Series solutions 71
5.3. Nonlinear first order ODEs 72
6. Isoclines and solution curves 74
6.1. Solution curves 74
6.2. Isoclines 74
6.3. Fixed points and perturbation analysis 75
6.4. Autonomous differential equations 76
7. Phase portraits 78
7.1. Phase portraits 78
7.2. Fixed points in discrete equations 79
7.3. Logistic map 80

8. Higher order linear ODEs . . . 81
8.1. Linear 2nd order ODEs with constant coefficients . . . 81
8.2. Eigenfunctions for 2nd order ODEs . . . 81
8.3. Detuning . . . 82
8.4. Reduction of order . . . 83
8.5. Solution space . . . 83
8.6. Initial conditions . . . 84
8.7. The fundamental matrix and the Wrońskian . . . 85
8.8. Abel's theorem . . . 86
8.9. Equidimensional equations . . . 88
9. Forced second order ODEs . . . 89
9.1. Guesswork . . . 89
9.2. Variation of parameters . . . 89
9.3. Forced oscillating systems: transients and damping . . . 90
9.4. Sinusoidal forcing . . . 91
9.5. Resonance in undamped systems . . . 93
9.6. Impulses and point forces . . . 94
10. Impulse forcing . . . 96
10.1. Dirac δ function . . . 96
10.2. Heaviside step function . . . 96
10.3. Ramp function . . . 97
10.4. Delta function forcing . . . 97
10.5. Heaviside function forcing . . . 99
11. Discrete equations and the method of Frobenius . . . 100
11.1. Higher order discrete equations . . . 100
11.2. Fibonacci sequence . . . 100
11.3. Method of Frobenius . . . 101
11.4. Fuch's theorem . . . 102
11.5. Special cases of indicial equation . . . 106
12. Multivariate calculus . . . 110
12.1. Gradient vector . . . 110
12.2. Stationary points . . . 111
12.3. Taylor series for multivariate functions . . . 111
12.4. Classifying stationary points . . . 112
12.5. Signature of Hessian . . . 113
12.6. Contours near stationary points . . . 114
13. Systems of ODEs . . . 115
13.1. Systems of linear ODEs . . . 115
13.2. Matrix methods . . . 115
13.3. Decoupling ODEs . . . 116

13.4. Phase portraits 117
13.5. Nonlinear systems of ODEs 118
13.6. Lotka–Volterra equations 118
13.7. First order wave equation and method of characteristics 119
14. More PDEs 121
14.1. Second order wave equation 121
14.2. Derivation of diffusion equation 122
14.3. Solving the diffusion equation 123

1. Differentiation

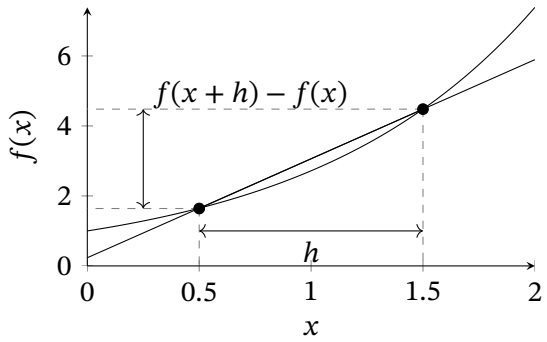
1.1. Basic definitions

Definition (Differential Equation)**.** A differential equation (DE) is an equation involving derivatives of a function or several functions.

Definition (Limit, informally)**.** If $\lim_{x \to x_0} f(x) = A$, then $f(x)$ can be made arbitrarily close to A by making x sufficiently close to x_0.

Note that the definition of the limit does not specify behaviour of $f(x)$ at $x = x_0$; it is perfectly possible that $f(x_0)$ is undefined, or that it is some number not equal to A. Examples of this behaviour would be $1/x$ (undefined at 0), or the Dirac δ function (infinite at 0).

Definition (One-Sided Limit)**.** A left limit is notated $\lim_{x \to x_0^-}$. It requires that the value A represented by the limit is computed by setting x to values smaller than x_0. Analogously, a right limit is notated $\lim_{x \to x_0^+}$. In calculating this limit, x must be greater than x_0.



Definition (Derivative)**.** We can use the definitions of limits to define the derivative of a function $f(x)$ with respect to its argument (in this case, x):

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{1.1}$$

Pictorially, we can see that the definition of the derivative is basically the slope of the line between two points that approach arbitrarily close to each other. In this example, x is 0.5, and h is 1.

Note that for the derivative to exist at a point x, we require that

$$\lim_{h \to 0^-} \frac{f(x+h) - f(x)}{h} = \lim_{h \to 0^+} \frac{f(x+h) - f(x)}{h}$$

This excludes, for example, the derivative of $|x|$ at $x = 0$, as this would have two conflicting answers (-1 and 1).

There are multiple ways of representing derivatives of functions. Here, we show the derivative of $f(x)$ in multiple notation systems:

- $\frac{\mathrm{d}f}{\mathrm{d}x}$: Leibniz notation
- $f'(x)$: Lagrange notation
- $\dot{f}(x)$: Newton notation

For sufficiently smooth functions (meaning that the derivative is valid at each step), we can define derivatives recursively:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\mathrm{d}f}{\mathrm{d}x}\right) = \frac{\mathrm{d}^2 f}{\mathrm{d}x^2} = f''(x) = \ddot{f}(x)$$

1.2. Rules for differentiation

Definition (Chain Rule)**.** Consider a function $f(x) = F(g(x))$. The derivative of $f(x)$ can be written

$$\frac{\mathrm{d}f}{\mathrm{d}x} = F'(g(x)) \cdot g'(x) = \frac{\mathrm{d}F}{\mathrm{d}g}\frac{\mathrm{d}g}{\mathrm{d}x} \tag{1.2}$$

Definition (Product Rule)**.** Consider a function $f(x) = u(x)v(x)$. The derivative of $f(x)$ can be written

$$\frac{\mathrm{d}f}{\mathrm{d}x} = u'(x)v(x) + u(x)v'(x) = u'v + uv' \tag{1.3}$$

Definition (Leibniz' Rule)**.** Consider a function $f(x) = u(x)v(x)$. Recursive derivatives of $f(x)$ can be written

$$\begin{aligned} f &= uv \\ f' &= u'v + uv' \\ f'' &= u''v + 2u'v' + uv'' \\ f''' &= u'''v + 3u''v' + 3u'v'' + uv''' \end{aligned} \tag{1.4}$$

This is analogous to Pascal's triangle and the binomial expansion. The coefficients are

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}$$

1.3. Order of magnitude

The goal of 'order of magnitude' functions is to compare the size of functions in the vicinity of certain points.

Definition (Little o)**.** Given functions $f(x)$ and $g(x)$ such that

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = 0 \tag{1.5}$$

we can say that $f(x) = o(g(x))$ as $x \to x_0$.

This is essentially saying that the function $f(x)$ is much 'smaller' than $g(x)$ as we approach the point x_0. For example, $x^2 = o(x)$ as $x \to 0$, because x^2 becomes vanishingly small compared to x near zero.

Definition (Big O: x_0 finite)**.** Assume we have two functions $f(x)$ and $g(x)$, and a finite number x_0 where we are comparing the functions. If we can find two positive constants M and δ such that

$$|f(x)| \leq M|g(x)| \quad (\forall x, |x - x_0| < \delta) \tag{1.6}$$

then $f(x) = O(g(x))$ as $x \to x_0$.

Informally, the function f can be *bounded by* g in a specific area around the point x_0.

Unlike little o notation, there is no requirement that $f(x)$ becomes vanishingly small compared to $g(x)$, just that it is smaller. Therefore, $x^2 \neq o(x^2)$ but $x^2 = O(x^2)$ (both as $x \to 0$).

Some examples:

- $x^2 = O(x)$ as $x \to 0$. Take $M = 1$, $\delta = 1$.
- $x \neq O(x^2)$ as $x \to 0$. This is because for any value of x smaller than $1/M$, the value of $g(x)$ is Mx^2 which is smaller than x.
- $x^2 = O(x^2)$ as $x \to 0$. Take $M = 1$, and choose an arbitrary δ.

By convention, we usually pick the most restrictive M and δ possible.

Definition (Big O: x_0 infinite)**.** Assume we have two functions $f(x)$ and $g(x)$, and we want to compare the functions' behaviours at infinity. If we can find two positive constants M and x_1 such that

$$|f(x)| \leq M|g(x)| \quad (\forall x > x_1) \tag{1.7}$$

then $f(x) = O(g(x))$ as $x \to \infty$.

This is basically the same as the previous definition—but obviously we can't pick a value slightly less than infinity to test, so we just provide a lower bound on x where the condition holds true.

For example, $2x^3 + 4x + 12 = O(x^3)$ as $x \to \infty$. This is because the function is a cubic, so can be bounded by a cubic as it shoots off to infinity. We can take, for example, $M = 3$ and $x_1 = 3$. Note that we can't just pick $M = 2$ even though asymptotically the function is close to $2x^3$; there is a value added to the $2x^3$ so we'd need to pick a slightly larger number to guarantee that Equation (1.7) is satisfied.

1.4. Equation of a tangent

We can use little o notation to construct the equation of a tangent to a function $f(x)$ at a given x value, x_0. This is the start of the formula for the Taylor series of f at x_0.

First, notice that $o(g(x))/g(x)$ is zero, as $o(g(x))$ is vanishingly small compared to $g(x)$ near the convergence point.

Using Equation (1.1), we can (informally) deduce:

$$\begin{aligned} \left.\frac{\mathrm{d}f}{\mathrm{d}x}\right|_{x=x_0} &= \frac{f(x_0+h)-f(x_0)}{h} \\ &= \frac{f(x_0+h)-f(x_0)}{h} + \frac{o(h)}{h} \\ \therefore f(x_0+h) &= f(x_0) + \left.\frac{\mathrm{d}f}{\mathrm{d}x}\right|_{x=x_0} h + o(h) \end{aligned}$$

If we now take $x = x_0 + h$; $y = f(x)$; $y_0 = f(x_0)$, we have

$$y = y_0 + \left.\frac{\mathrm{d}f}{\mathrm{d}x}\right|_{x=x_0} (x - x_0) + o(h)$$

This is the equation of the tangent to the curve at x_0 if $o(h) = 0$, and this is start of the equation for the Taylor series.

1.5. Taylor series

Suppose that we want to approximate a function $f(x)$ using a polynomial of order n.

$$f(x) \approx \underbrace{a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n}_{\equiv p_n(x)}$$

By assuming that the equality holds, we may set $x = 0$ to get the value of a_0. By differentiating the left and right hand sides k times, we can evaluate both sides at $x = 0$ to get the value of a_k. Therefore, term a_k is equivalent to $f^{(k)}(0)/k!$

$$f(x) \approx p_n(x) = f(0) + xf'(0) + \frac{x^2}{2} f''(0) + \cdots + \frac{x^n}{n!} f^{(n)}(0)$$

Alternatively, repeating the process at x_0, we get the formula for the Taylor polynomial of degree n of $f(x)$:

$$f(x) \approx p_n(x) = f(x_0) + (x - x_0) f'(x_0) + \frac{(x-x_0)^2}{2} f''(x_0) + \cdots + \frac{(x-x_0)^n}{n!} f^{(n)}(x_0)$$

We can write

$$f(x) = p_n(x) + E_n \tag{1.8}$$

where E_n is the error at term n. Recall that $f(x+h) = f(x) + hf'(x) + o(h)$ as $h \to 0$. We can generalise this, provided that the first n derivatives of $f(x)$ exist.

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \cdots + \frac{h^n}{n!}f^{(n)}(x) + o(h^n) \tag{1.9}$$

Comparing Equations (1.8) and (1.9), we see that:

$$E_n = o(h^n)$$

Theorem (Taylor's Theorem). $E_n = O(h^{n+1})$ as $h \to 0$ provided that $f^{(n+1)}(x)$ exists.

Note that the big O notation in Taylor's Theorem is a stronger statement than the little o notation above. For example, $h^{n+a} = o(h^n)$ as $h \to 0\ \forall a \in (0,1)$ since $\lim_{h\to 0} \frac{h^{n+a}}{h^n} = \lim_{h\to 0} h^a = 0$. However, $h^{n+a} \neq O(h^{n+1})$ as $h \to 0$ for $a \in (0,1)$ because we can't bound h^{n+a} using h^{n+1} everywhere in the vicinity of 0.

1.6. L'Hôpital's rule

Let $f(x)$ and $g(x)$ be differentiable functions at $x = x_0$, and that $\lim_{x\to x_0} f(x) = f(x_0) = 0$ and similarly for $g(x)$. L'Hôpital's Rule states that

$$\lim_{x\to x_0} \frac{f(x)}{g(x)} = \lim_{x\to x_0} \frac{f'(x)}{g'(x)} \text{ if } g'(x_0) \neq 0$$

Proof. As $x \to x_0$:

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + o(x - x_0)$$
$$g(x) = g(x_0) + (x - x_0)g'(x_0) + o(x - x_0)$$

But we know that $f(x_0) = g(x_0) = 0$ therefore

$$\frac{f(x)}{g(x)} = \frac{f'(x_0) + \frac{o(x-x_0)}{x-x_0}}{g'(x_0) + \frac{o(x-x_0)}{x-x_0}}$$

By the definition of little o, $o(h)/h$ tends to zero, so

$$\frac{f(x)}{g(x)} = \frac{f'(x)}{g'(x)}$$

□

Note that l'Hôpital's rule can be applied recursively, using higher-order derivatives. For example, consider $f(x) = 3\sin x - \sin 3x$; $g(x) = 2x - \sin 2x$. The limit approaches 3 as $x \to 0$.

2. Integration

2.1. Definition of integration

We use a Riemann sum to approximate the area under a sufficiently well-behaved function $f(x)$ on the real numbers.

$$\sum_{n=0}^{N-1} f(x_n)\Delta x \tag{2.1}$$

where $\Delta x = (b-a)/N$, and $x_n = a + n\Delta x$. How close is (2.1) to the area under $f(x)$ for large N? Consider a specific rectangle in the Riemann sum by fixing n. The area under the curve in the nth rectangle and the area of the rectangle itself differ by a value we denote here as ε. By computing ε's order of magnitude, we can show how much the total error deviates by.

Theorem (Mean Value Theorem)**.** For a continuous function $f(x)$,

$$\int_{x_n}^{x_{n+1}} f(x)\,\mathrm{d}x = f(x_c)\cdot(x_{n+1}-x_n) \tag{2.2}$$

for some $x_c \in (x_n, n_{n+1})$.

We use the Taylor Series of $f(x)$ at x_n to compute a value for x_c.

$$f(x_c) = f(x_n) + O(x_c - x_n)$$

as $x_c - x_n \to 0$. Since $x_n < x_c < x_{n+1}$, which implies $|x_{n+1}-x_n| > |x_c - x_n|$, we can make the statement that

$$f(x_c) = f(x_n) + O(x_{n+1} - x_n)$$

as $x_{n+1} - x_n \to 0$. Thus, by (2.2)

$$\int_{x_n}^{x_{n+1}} f(x)\,\mathrm{d}x = [f(x_n) + O(x_{n+1}-x_n)]\,(x_{n+1}-x_n)$$

By defining $\Delta x = x_{n+1} - x_n$, we have

$$\int_{x_n}^{x_{n+1}} f(x)\,\mathrm{d}x = \Delta x f(x_n) + O(\Delta x^2) \tag{2.3}$$

By rearranging, we can compute ε:

$$\varepsilon = \left|\int_{x_n}^{x_{n+1}} f(x)\,\mathrm{d}x - \Delta x f(x_n)\right| = O(\Delta x^2)$$

Therefore it follows that

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{\Delta x\to 0}\left[\left(\sum_{n=0}^{N-1} f(x_n)\Delta x\right) + O(N\Delta x^2)\right]$$

Note that $O(N\Delta x^2) = O((\frac{b-a}{N})^2 \cdot N) = O(1/N)$, so

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{N\to\infty}\left[\left(\sum_{n=0}^{N-1} f(x_n)\Delta x\right) + O(1/N)\right]$$

Which gives our final result of

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{N\to\infty}\sum_{n=0}^{N-1} f(x_n)\Delta x \tag{2.4}$$

2.2. Fundamental theorem of calculus

Let $F(x) = \int_a^x f(t)\,\mathrm{d}t$. From the definition of the derivative, we have

$$\begin{aligned}\frac{\mathrm{d}F}{\mathrm{d}x} &= \lim_{h\to 0}\frac{1}{h}\left[F(x+h) - F(x)\right]\\ &= \lim_{h\to 0}\frac{1}{h}\left[\int_a^{x+h} f(t)\,\mathrm{d}t - \int_a^x f(t)\,\mathrm{d}t\right]\\ &= \lim_{h\to 0}\frac{1}{h}\int_x^{x+h} f(t)\,\mathrm{d}t\end{aligned}$$

Using (2.4):

$$\begin{aligned}&= \lim_{h\to 0}\frac{1}{h}\left[hf(x) + O(h^2)\right]\\ &= \lim_{h\to 0}\left[f(x) + O(h)\right]\\ &= f(x)\end{aligned}$$

Therefore:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[\int_a^x f(t)\,\mathrm{d}t\right] = f(x) \tag{2.5}$$

2.3. Integration techniques

Three particularly important methods of integration are:

- u-substitution,
- trigonometric substitutions, and
- integration by parts.

Of particular note is the trigonometric substitution method, since it can be difficult to work out exactly which substitution will yield the result. A table is provided below.

Identity	Term in Integrand	Substitution
$\cos^2\theta + \sin^2\theta = 1$	$\sqrt{1-x^2}$	$x = \sin\theta$
$1 + \tan^2\theta = \sec^2\theta$	$1 + x^2$	$x = \tan\theta$
$\cosh^2 u - \sinh^2 u = 1$	$\sqrt{x^2-1}$	$x = \cosh u$
$\cosh^2 u - \sinh^2 u = 1$	$\sqrt{x^2+1}$	$x = \sinh u$
$1 - \tanh^2 u = \operatorname{sech}^2 u$	$1 - x^2$	$x = \tanh u$

3. Multivariate functions

3.1. Partial derivatives

We define the partial derivative of a two-valued function $f(x, y)$ with respect to x (for example) by:

$$\left.\frac{\partial f}{\partial x}\right|_y = \lim_{\delta x \to 0} \frac{f(x + \delta x, y) - f(x, y)}{\delta x} \tag{3.1}$$

For example, if $f(x, y) = x^2 + y^3 + e^{xy^2}$, we have

$$\left.\frac{\partial f}{\partial x}\right|_y = 2x + y^2 e^{xy^2}$$
$$\left.\frac{\partial^2 f}{\partial x^2}\right|_y = 2 + y^4 e^{xy^2}$$

We can also define 'cross-derivatives' by differentiating successively with respect to different variables, for example

$$\left.\frac{\partial}{\partial y}\left(\left.\frac{\partial f}{\partial x}\right|_y\right)\right|_x = 2ye^{xy^2} + 2xy^3 e^{xy^2}$$

The order of computation of cross-derivatives is irrelevant, provided that the required derivatives all exist.

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x}\frac{\partial f}{\partial y} = \frac{\partial}{\partial y}\frac{\partial f}{\partial x} = \frac{\partial^2 f}{\partial y \partial x} \tag{3.2}$$

We use a subscript shorthand to denote partial differentiation. Where the point of evaluation of the derivative is not stated, it is implied to be fixed. For example:

$$\left.\frac{\partial f}{\partial x}\right|_y = \frac{\partial f}{\partial x} = f_x$$

However, with a function $f(x, y, z)$:

$$\left.\frac{\partial f}{\partial x}\right|_{yz} \neq \left.\frac{\partial f}{\partial x}\right|_y$$

because z is not fixed.

3.2. Multivariate chain rule

In this section, all use of o notation is defined to be where all required δ values approach 0. We define the differential of a two-valued function $f(x, y)$ to be

$$\delta f = f(x + \delta x, y + \delta y) - f(x, y) \tag{3.3}$$

We can evaluate this differential by rewriting (3.3) as

$$\begin{aligned}\delta f = f(x+\delta x, y+\delta y) - f(x+\delta x, y) + \\ f(x+\delta x, y) - f(x, y)\end{aligned}$$

We can move from (x, y) to $(x+\delta x, y+\delta y)$ along the path $(x, y) \to (x+\delta x, y) \to (x+\delta x, y+\delta y)$. If we move in this way, then we only need to worry about derivatives in the directions of our axes. From here on in the derivation, the first line will always represent the path segment in the y direction, and the second line will represent the path segment in the x direction.

Now that we've separated the differential into these two axes, we can use Taylor series, treating each line as a single-valued function, to expand each of these path segments along the matching axis.

$$\begin{aligned}\delta f = f(x+\delta x, y) + \delta y \frac{\partial f}{\partial y}(x+\delta x, y) + o(\delta y) - f(x+\delta x, y) + \\ f(x, y) + \delta x \frac{\partial f}{\partial x}(x, y) + o(\delta x) - f(x, y)\end{aligned}$$

We can now cancel the beginning and ending points of each segment of the path, leaving

$$\begin{aligned}\delta f = \delta y \frac{\partial f}{\partial y}(x+\delta x, y) \ + o(\delta y) + \\ \delta x \frac{\partial f}{\partial x}(x, y) + o(\delta x)\end{aligned}$$

We can reduce the remaining $x+\delta x$ term to simply an x term by performing another Taylor expansion.

$$\begin{aligned}\delta f = \delta y \left[\frac{\partial f}{\partial y}(x, y) + \delta x \frac{\partial^2 f}{\partial y^2}(x, y) + o(\delta x) \right] + o(\delta y) + \\ \delta x \frac{\partial f}{\partial x}(x, y) + o(\delta x)\end{aligned}$$

Expanding out this bracket leaves

$$\begin{aligned}\delta f = \delta y \frac{\partial f}{\partial y}(x, y) + \delta x \delta y \frac{\partial^2 f}{\partial y^2}(x, y) + o(\delta x \delta y) + o(\delta y) + \\ \delta x \frac{\partial f}{\partial x}(x, y) + o(\delta x)\end{aligned}$$

We will now change the meanings of each line. Now, we will group terms by factors.

$$\begin{aligned}\delta f = \delta x \frac{\partial f}{\partial x}(x, y) + o(\delta x) + \\ \delta y \frac{\partial f}{\partial y}(x, y) + o(\delta y) + \\ \delta x \delta y \frac{\partial^2 f}{\partial y^2}(x, y) + o(\delta x \delta y)\end{aligned}$$

Because $o(h)$ is significantly smaller than h, we can eliminate all the o terms.

$$\begin{aligned}\delta f = \delta x \frac{\partial f}{\partial x}(x,y) + \\ \delta y \frac{\partial f}{\partial y}(x,y) + \\ \delta x \delta y \frac{\partial^2 f}{\partial y^2}(x,y)\end{aligned}$$

Finally, we can eliminate the $\delta x \delta y$ term because it is vanishingly small as they tend to zero.

$$\delta f = \delta x \frac{\partial f}{\partial x}(x,y) + \delta y \frac{\partial f}{\partial y}(x,y) \tag{3.4}$$

This is the differential form of the multivariate chain rule. We can take the result of this equation in the limit to create the infinitesimal form:

$$\mathrm{d}f = \mathrm{d}x \, \frac{\partial f}{\partial x}(x,y) + \mathrm{d}y \, \frac{\partial f}{\partial y}(x,y) \tag{3.5}$$

By integrating (3.5), we get

$$\int \mathrm{d}f = \int \frac{\partial f}{\partial x} \, \mathrm{d}x + \int \frac{\partial f}{\partial y} \, \mathrm{d}y$$

In definite integral form, we can write

$$\begin{aligned}f(x_2 - x_1, y_2 - y_1) &= \int_{x_1}^{x_2} \frac{\partial f}{\partial x}(x, y_1) \, \mathrm{d}x + \int_{y_1}^{y_2} \frac{\partial f}{\partial y}(x_2, y) \, \mathrm{d}y \\ &= \int_{y_1}^{y_2} \frac{\partial f}{\partial y}(x_1, y) \, \mathrm{d}y + \int_{x_1}^{x_2} \frac{\partial f}{\partial x}(x, y_2) \, \mathrm{d}x \\ &\neq \int_{x_1}^{x_2} \frac{\partial f}{\partial x}(x, y_1) \, \mathrm{d}x + \int_{y_1}^{y_2} \frac{\partial f}{\partial y}(x_1, y) \, \mathrm{d}y\end{aligned}$$

Note that the first two examples of a right hand side go along the paths $(x_1, y_1) \to (x_2, y_1) \to (x_2, y_2)$ and $(x_1, y_1) \to (x_1, y_2) \to (x_2, y_2)$ by performing the integrals. However, the last example does not follow a path from (x_1, y_1) to (x_2, y_2), so it is invalid.

3.3. Change of variables

We can transform derivatives into different coordinate systems to make problems easier to solve. For example, let $f(x, y)$ be some function with a Cartesian coordinate input. We can rewrite it in terms of polar coordinates (r, θ). First, rewrite f as:

$$f(x(r,\theta), y(r,\theta))$$

then we can write the derivatives.

$$\frac{\partial f}{\partial r} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial r}$$

We can do similar evaluations for $\frac{\partial f}{\partial \theta}$, for example.

3.4. Implicit differentiation

Consider some surface defined by $f(x, y, z) = c$. Then f implicitly defines functions such as $z(x, y)$ (provided the function is well-behaved). We can find, for example, $\left.\frac{\partial z}{\partial x}\right|_y$ by using the multivariate chain rule in three dimensions.

$$\left.\frac{\partial f}{\partial x}\right|_y = \left.\frac{\partial f}{\partial x}\right|_{yz} \underbrace{\left.\frac{\partial x}{\partial x}\right|_y}_{=1} + \left.\frac{\partial f}{\partial y}\right|_{xz} \underbrace{\left.\frac{\partial y}{\partial x}\right|_y}_{=0} + \left.\frac{\partial f}{\partial z}\right|_{xy} \left.\frac{\partial z}{\partial x}\right|_y$$

Note that the $\frac{\partial y}{\partial x}$ term is zero because we hold y to be fixed. Simplifying, we get

$$\left.\frac{\partial f}{\partial x}\right|_y = \left.\frac{\partial f}{\partial x}\right|_{yz} + \left.\frac{\partial f}{\partial z}\right|_{xy} \left.\frac{\partial z}{\partial x}\right|_y$$

The left hand side is zero because on the surface $z(x, y)$, f is always equivalent to c so there is never any δf. The $\left.\frac{\partial f}{\partial x}\right|_{yz}$ term, however, is not zero in general because we are not going across the $z(x, y)$ surface—just parallel to the x axis, because we fixed both y and z. Hence,

$$\left.\frac{\partial z}{\partial x}\right|_y = \frac{-\left.\frac{\partial f}{\partial x}\right|_{yz}}{\left.\frac{\partial f}{\partial z}\right|_{xy}}$$

The reciprocal rule for derivatives applies also to partial derivatives so long as the same variables are held fixed. For example, given the function $f(x(r, \theta), y(r, \theta))$, we have

$$\left.\frac{\partial r}{\partial x}\right|_y = \frac{1}{\left.\frac{\partial x}{\partial r}\right|_y}$$

But

$$\frac{\partial r}{\partial x} \neq \frac{1}{\frac{\partial x}{\partial r}}$$

because the left hand side holds y constant and the right hand side holds θ constant.

3.5. Differentiating an integral with respect to a parameter

Consider a family of function $f(x;\alpha)$ where α is some parameter. We can say that α parametrises f. An example of a parametrised function is the logarithm; $f(x;\alpha) = \log_\alpha x$. We define

$$I(\alpha) = \int_{a(\alpha)}^{b(\alpha)} f(x;\alpha)\ \mathrm{d}x$$

So, what is $\frac{\mathrm{d}I}{\mathrm{d}\alpha}$? By definition, we have

$$\begin{aligned}
\frac{\mathrm{d}I}{\mathrm{d}\alpha} &= \lim_{\delta\alpha\to 0} \frac{I(\alpha+\delta\alpha) - I(\alpha)}{\delta\alpha} \\
&= \lim_{\delta\alpha\to 0} \frac{1}{\delta\alpha}\left[\int_{a(\alpha+\delta\alpha)}^{b(\alpha+\delta\alpha)} f(x;\alpha+\delta\alpha)\ \mathrm{d}x - \int_{a(\alpha)}^{b(\alpha)} f(x;\alpha)\ \mathrm{d}x\right] \\
&= \lim_{\delta\alpha\to 0} \frac{1}{\delta\alpha}\left[\int_{a(\alpha)}^{b(\alpha)} f(x;\alpha+\delta\alpha) - f(x;\alpha)\ \mathrm{d}x - \int_{a(\alpha)}^{a(\alpha+\delta)} f(x;\alpha+\delta\alpha)\ \mathrm{d}x + \int_{b(\alpha)}^{b(\alpha+\delta)} f(x;\alpha+\delta\alpha)\ \mathrm{d}x\right] \\
&= \int_{a(\alpha)}^{b(\alpha)} \lim_{\delta\alpha\to 0} \frac{f(x;\alpha+\delta\alpha) - f(x;\alpha)}{\delta\alpha}\ \mathrm{d}x - f(a;\alpha) \lim_{\delta\alpha\to 0} \frac{a(\alpha+\delta\alpha) - a(\alpha)}{\delta\alpha} + f(b;\alpha) \lim_{\delta\alpha\to 0} \frac{b(\alpha+\delta\alpha) - b(\alpha)}{\delta\alpha}
\end{aligned}$$

Therefore:

$$\frac{\mathrm{d}I}{\mathrm{d}\alpha} = \frac{\mathrm{d}}{\mathrm{d}\alpha}\int_{a(\alpha)}^{b(\alpha)} f(x;\alpha)\ \mathrm{d}x = \int_{a(\alpha)}^{b(\alpha)} \frac{\partial f}{\partial \alpha}\ \mathrm{d}x + f(b;\alpha)\frac{\mathrm{d}b}{\mathrm{d}\alpha} - f(a;\alpha)\frac{\mathrm{d}a}{\mathrm{d}\alpha}$$

4. Linear ordinary differential equations

4.1. Eigenfunctions

Definition. The eigenfunction of an operator is a function that is unchanged by the action of the operator (except for a multiplicative scaling).

From this definition, we can see that $e^{\lambda x}$ is the eigenfunction of the differential operator. The eigenvalue of this function is λ, as this is the scaling factor.

(i) Any linear homogeneous ODE with constant coefficients has solutions in the form $e^{\lambda x}$. For example, in the equation $5y' - 3y = 0$ we can try a solution of the form $y = Ae^{\lambda x}$, and we get $5\lambda - 3 = 0$. This equation is known as the characteristic equation.

(ii) Any solution to a linear homogeneous ODE can be scaled to create more solutions. In particular, $y = 0$ is a solution.

(iii) An nth order linear ODE has n linearly independent solutions. In the case of constant coefficient equations, this follows from the Fundamental Theorem of Algebra. However, the proof of this is outside the scope of this course. This implies that the above example has only one solution: $y = Ae^{3x/5}$.

(iv) An nth order ODE requires n initial/boundary conditions to create a particular solution.

4.2. Solving first order ODEs

To solve a differential equation, we can use the following technique to break it apart into two smaller functions:

$$y = y_p + y_c$$

The function y_p is called the particular integral; it is simply any solution the original equation. Normally this does not have any arbitrary constants in it. The other function y_c is the complementary function. This is a solution to the equivalent homogeneous equation, which is formed by setting the right hand side (the side without the dependent variable) to zero. This is generally easier to solve using the exponential function.

By adding the two together, we get the general solution. Alternatively, once we have computed the particular integral, we can simply substitute the equation $y = y_p + y_c$ into the original differential equation to get a new equation in terms of y_c.

Note that we refer to terms that do not depend on the dependent variable as 'forcing functions'.

4.3. Constant forcing

If the equation is linear, has constant coefficients and a constant on the right hand side, we can set $y_p' = 0$. For example, in the equation $5y' - 3y = 10$, we can set $y' = 0$ to get

$y_p = -10/3$.

Now we can insert this general solution into the differential equation. Note that all terms with y_p, along with the right hand side, cancel out because it is a solution. This leaves $5y_c' - 3y = 0$. We can solve this normally (using methods such as trying $Ae^{\lambda x}$ or just directly solving the characteristic equation) to give $y_c = Ae^{-3x/5}$.

Combining the results, we get $y = Ae^{3x/5} - 10/3$.

4.4. Eigenfunction forcing

If the equation has a e^{kt} term as the only forcing function (where the independent variable here is t), we can solve it in a similar way. Here is an example question involving this concept.

> In a sample of rock, isotope A decays into isotope B at a rate proportional to a, the number of nuclei of A, while B decays into isotope C at a rate proportional to b, the number of nuclei of B. Find $b(t)$.

We can formulate an equation as follows:

$$\begin{aligned} \dot{a} = -k_a a \implies a &= a_0 e^{-k_a t} \\ \dot{b} &= k_a a - k_b b \\ \therefore \dot{b} + k_b b &= k_a a_0 e^{-k_a t} \end{aligned}$$

So we have a linear first order ODE with an eigenfunction as the forcing function. We can guess that the particular integral is of the form $b_p = \lambda e^{-k_a t}$.

$$\begin{aligned} -k_a \lambda e^{-k_a t} + k_b \lambda e^{-k_a t} &= k_a a_0 e^{-k_a t} \\ \lambda(k_b - k_a) &= k_a a_0 \\ \therefore \lambda &= \frac{k_a}{k_b - k_a} a_0 \end{aligned}$$

We can form the complementary function by solving:

$$\begin{aligned} \dot{b_c} + k_b b_c &= 0 \\ \therefore b_c &= Ae^{-k_b t} \end{aligned}$$

So combining everything, we have

$$b = \frac{k_a}{k_b - k_a} a_0 e^{-k_a t} + Ae^{-k_b t}$$

In this instance, there is a special property that if $b = 0$ at $t = 0$, then we can divide $b(t)/a(t)$ and completely eliminate a_0, thus letting us calculate the age of a rock without knowing the original amount of isotope A at all.

4.5. Non-constant coefficients

If we have a differential equation in standard form, i.e.

$$y' + p(x)y = f(x)$$

we can multiply the equation by an integrating factor μ to solve it. Ideally, we want the derivative of μ to be $\mu p(x)$ so that the equation becomes

$$\mu y' + \mu p(x)y = \mu y' + \mu' y = (\mu y)' = \mu f(x)$$

So therefore $\mu = e^{\int p(x)\,\mathrm{d}x}$.

5. Discrete equations

A discrete equation (for our purposes) is an equation involving a function that is evaluated at a discrete set of points.

5.1. Numerical integration

We can consider a discrete representation of $y(x)$; let $x_1 \mapsto y_1, x_2 \mapsto y_2$ etc. We can approximate the derivative with

$$\left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{x_n} \approxeq \frac{y_{n+1} - y_n}{h}$$

This is called the 'Forward Euler' approximation of the derivative. It isn't the best, but it is asymptotically equal. We can solve the differential equation $5y' - 3y = 0$ as follows:

$$5\frac{y_{n+1} - y_n}{h} - 3y_n = 0$$

This is known as a difference equation. We can transform this into a recurrence relation as follows:

$$y_{n+1} = \left(1 + \frac{3}{5}h\right) y_n$$

We can apply this iteratively, to get

$$\begin{aligned} y_{n+1} &= \left(1 + \frac{3}{5}h\right) y_n \\ &= \left(1 + \frac{3}{5}h\right)^2 y_{n-1} \\ &= \left(1 + \frac{3}{5}h\right)^n y_0 \end{aligned}$$

So in the limit, this approaches the desired solution. Note that due to the approximation we used for the derivative, for finite n the solution we get will be less than the actual answer.

5.2. Series solutions

Series solutions are a powerful tool for solving ordinary differential equations. We can express the solution in terms of an infinite power series, i.e. we let

$$y(x) = \sum_{n=0}^{\infty} a_n x^n$$

Let us try this on our original differential equation, $5y' - 3y = 0$. We have:

$$y(x) = \sum_{n=0}^{\infty} a_n x^n \qquad\qquad y'(x) = \sum_{n=0}^{\infty} n a_n x^{n-1}$$

Multiplying by x to eliminate the power of $n-1$, we have

$$xy(x) = \sum_{n=0}^{\infty} a_n x^{n+1} \qquad xy'(x) = \sum_{n=0}^{\infty} n a_n x^n$$

Matching the limits of the sums and powers of x:

$$xy(x) = \sum_{n=1}^{\infty} a_{n-1} x^n \qquad xy'(x) = \sum_{n=1}^{\infty} n a_n x^n$$

We can now combine this into one equation.

$$5y' - 3y = 0 \implies 5\sum_{n=1}^{\infty} n a_n x^n - 3\sum_{n=1}^{\infty} a_{n-1} x^n = 0 \implies \sum_{n=1}^{\infty} [5nax^n - 3a_{n-1}x^n] = 0$$

Note that this holds for all x, so we can remove the sum and the x^n term, and solve generically for a_n.

$$\begin{aligned} 5na_n - 3a_{n-1} &= 0 \\ \implies a_n &= \frac{3}{5n} a_{n-1} \\ &= \left(\frac{3}{5}\right)^2 \frac{1}{n(n-1)} a_{n-2} \\ &= \left(\frac{3}{5}\right)^n \frac{1}{n!} a_0 \end{aligned}$$

We now have an explicit equation for y as a power series.

5.3. Nonlinear first order ODEs

Let us consider the equation

$$Q(x,y)\frac{\mathrm{d}y}{\mathrm{d}x} + P(x,y) = 0$$

If it can be written in the form

$$q(y)\,\mathrm{d}y = p(x)\,\mathrm{d}x$$

then by integrating both sides we can find a solution. This is known as a separable equation.

Alternatively, if $Q(x,y)\,\mathrm{d}y + P(x,y)\,\mathrm{d}x$ is an exact differential of some multivariate function $f(x,y)$, then we call this an exact equation. Specifically, due to the multivariate chain rule, we can get

$$\mathrm{d}f = \frac{\partial f}{\partial x}\,\mathrm{d}x + \frac{\partial f}{\partial y}\,\mathrm{d}y$$

So we want $P(x,y) = \frac{\partial f}{\partial x}$ and $Q(x,y) = \frac{\partial f}{\partial y}$. We can exploit cross derivatives to check whether this is truly is an exact equation without having to integrate both P and Q.

$$\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial^2 f}{\partial x \partial y}$$
$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}$$

This is the key condition to check for an exact equation. More specifically, if $\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}$ throughout a simply connected domain $\mathcal{D}$, then $P\,\mathrm{d}x + Q\,\mathrm{d}y$ is an exact differential of a single valued function $f(x,y)$ in $\mathcal{D}$. A simply connected domain is essentially a domain without holes.

We can find f by integrating P and Q, since they are the partial derivatives of f. As an example, let us solve

$$6y(y-x)\frac{\mathrm{d}y}{\mathrm{d}x} + (2x - 3y^2) = 0$$

So here, $P = 2x - 3y^2$ and $Q = 6y(y-x)$. We can check that indeed

$$\frac{\partial P}{\partial y} = -6y \qquad\qquad \frac{\partial Q}{\partial x} = -6y$$

So we have an exact equation as required. Now, we have

$$\left.\frac{\partial f}{\partial x}\right|_y = P = 2x - 3y^2$$
$$\implies f = x^2 - 3xy^2 + h(y)$$

where h is a constant with respect to x, so it must be some function of y. We can differentiate our new definition for f with respect to y, and substitute back into what we know for Q.

$$\left.\frac{\partial f}{\partial y}\right|_x = -6xy + \frac{\mathrm{d}h}{\mathrm{d}y}$$

But also, from the definition of Q,

$$\left.\frac{\partial f}{\partial y}\right|_x = Q = 6y(y-x)$$

So by comparing the two things which we know are equal, we get $\frac{\mathrm{d}h}{\mathrm{d}y} = 6y^2$ so $h = 2y^3 + c$. We plug this back into our value for f, leaving

$$f = x^2 - 3xy^2 + 2y^3 + c$$

So our general solution is

$$x^2 - 3xy^2 + 2y^3 = d$$

6. Isoclines and solution curves

6.1. Solution curves

Nonlinear differential equations are not guaranteed to have closed form solutions. However, we can analyse the behaviour of such an equation without actually having to solve the equation. In this lecture, we consider only equations of the form

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \dot{y} = f(y,t)$$

Each initial condition to this function will generate a different solution curve. Note that these curves may not cross. Suppose that two curves did cross at some point (y,t). Then $\frac{\mathrm{d}y}{\mathrm{d}t}\Big|_y$ would have two different values; the gradient of each curve would have to be different. But $y(t)$ is a single-valued function, so the derivative is also single-valued. So the solution curves can never cross.

Let us consider an example which we can, in fact, solve directly.

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \dot{y} = f(t) = t(1-y^2)$$

This is separable, and we may solve the equation to give

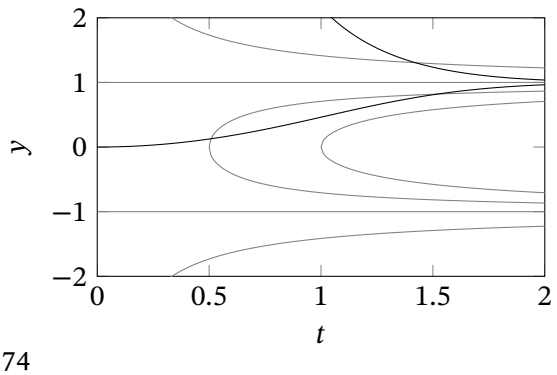
$$y = \frac{A - e^{-t^2}}{A + e^{-t^2}}$$

This general solution produces a family of solution curves parametrised by A. Can we sketch and describe these solutions without using this explicit solution for $y(t)$?

6.2. Isoclines

An isocline is a curve along which $f = \dot{y}$ is constant. To draw these isoclines, we need to work out when f takes certain values.

$$f = 0 \text{ for } y = \pm 1, t = 0$$
$$f < 0 \text{ for } y > 1, y < -1$$
$$f > 0 \text{ for } -1 < y < 1$$



Let us now draw some such isoclines on a graph, here drawn in grey. On the outermost two lines, the value of f, and hence the derivative, is -1. On the lines in the centre, the value of f is 1 and 0.5, both of which are drawn so that it is easier to imagine the infinite set of isoclines. The two horizontal lines at 1 and -1 have $f = 0$. So any

solution curve that passes through these isoclines must have this gradient at the moment it intersects the line. We can therefore visually interpolate what the gradient should be in between these known points.

Two such solution curves are drawn on this graph; the one intersecting zero has $A = 1$ in the solution for y, and the one above it has $A = -1$. Note how, as they intersect the isoclines in red, they have exactly the gradient defined by the isocline. Particularly, the lower solution curve intersects the same isocline twice, and therefore has this exact gradient at two distinct points—we observe these points as the intersection points between the solution curve and the isocline.

Note also that the solutions $y = 1$ and $y = -1$ lie on these isoclines for all t. This is because the isoclines specify that the function has zero gradient on such a straight line, so it makes sense that the function and isocline coincide.

6.3. Fixed points and perturbation analysis

Points such that y is fixed for all t are called fixed points, or equilibrium points. In our example above, $y = 1$ and $y = -1$ are examples of fixed points. Note that the solutions above seemed to gravitate towards $y = 1$ over time; we call such a fixed point 'stable' because any slight perturbation from the value will return back to the fixed point over time. The same is not, however, true for the -1 fixed point. It is considered 'unstable' as any small perturbation will cause y to drift further and further away from -1. We can analyse this more rigorously using perturbation analysis.

Let $y = a$ be a fixed point of $\dot{y} = f(y, t)$ such that $f(a, t) = 0$. Then, consider some small perturbation ε from this fixed point. Now, $y = a + \varepsilon(t)$. By setting the initial value of $\varepsilon(0)$ to some arbitrarily small amount, we want to see the behaviour of $\varepsilon(t)$ as t tends to infinity. This way, if $\varepsilon(t)$ goes to zero, then y will tend towards the fixed point a, so the point is stable. If $\varepsilon(t)$ goes to any other value, then y does not tend to a, so the point is unstable.

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \frac{\mathrm{d}\varepsilon}{\mathrm{d}t} = f(a + \varepsilon, t)$$

Expanding $f(a + \varepsilon, t)$ as a multivariate Taylor series around (a, t), we have

$$= \underbrace{f(a, t)}_{= 0 \text{ by definition}} + \varepsilon \frac{\partial f}{\partial y}(a, t) + O(\varepsilon^2)$$

as ε tends to zero. So for small ε, we have

$$\frac{\mathrm{d}\varepsilon}{\mathrm{d}t} \approx \varepsilon \frac{\partial f}{\partial y}(a, t)$$

which is a linear ordinary differential equation for ε in terms of t, as $\frac{\partial f}{\partial y}(a,t)$ is an expression purely in terms of a and t. If (as $t \to \infty$) ε tends to zero then the point is stable, otherwise ε will tend to $\pm\infty$ and the point is considered unstable. This does not imply that y itself tends to $\pm\infty$, just that the $O(n^2)$ term now becomes important because ε does not tend to zero.

Note that if $\frac{\partial f}{\partial y} = 0$, then we will need to consider the next term in the Taylor expansion, and so on, to make sure that we have an equation that lets us compute ε. In this case, however, the equation for ε will be nonlinear, as we need to consider the ε^2 term, or the ε^3 term, or so on.

In our example, we can deduce that $\frac{\partial f}{\partial y} = -2yt$, so we have:

- $(y = 1)$

$$\begin{aligned} \dot{\varepsilon} &= -2(1)t\varepsilon \\ &= -2t\varepsilon \\ \therefore \varepsilon &= \varepsilon_0 e^{-t^2} \\ \lim_{t\to\infty} \varepsilon_0 e^{-t^2} &= 0 \end{aligned}$$

  so this point is stable.

- $(y = 1)$

$$\begin{aligned} \dot{\varepsilon} &= -2(-1)t\varepsilon \\ &= 2t\varepsilon \\ \therefore \varepsilon &= \varepsilon_0 e^{t^2} \\ \lim_{t\to\infty} \varepsilon_0 e^{t^2} &= \pm\infty \end{aligned}$$

  so this point is unstable.

6.4. Autonomous differential equations

A special case of this is that of autonomous equations, which are defined to be differential equations of the form $\dot{y} = f(y)$. Specifically, the derivative of y does not depend on t. Therefore, near a fixed point $y = a$, we have:

$$\begin{aligned} y &= a + \varepsilon(t) \\ &\therefore \\ dot\varepsilon &= \varepsilon \frac{\mathrm{d}f}{\mathrm{d}y}(a) = \varepsilon k \end{aligned}$$

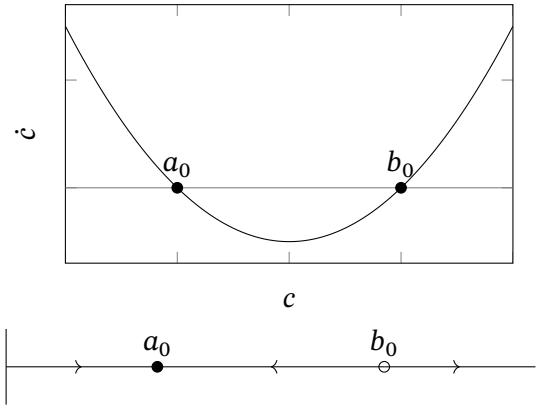
where k is the constant value $\frac{\mathrm{d}f}{\mathrm{d}y}(a)$. Note that we can use normal derivatives in place of partial derivatives because f depends only on y. So the solution is

$$\varepsilon = \varepsilon_0 e^{kt}$$

So, if $k = f'(a) < 0$ then the point is stable, and if $k = f'(a) > 0$ then the point is unstable. This special case is useful, but it is probably only worth memorising the general case to avoid confusion, since it is simple to derive as needed.

7. Phase portraits

7.1. Phase portraits



Another way to analyse solutions to a differential equation is using a geometrical representation of the solution, called a phase portrait. For example,

$$\mathrm{NaOH} + \mathrm{HCl} \to \mathrm{H_2O} + \mathrm{NaCl}$$

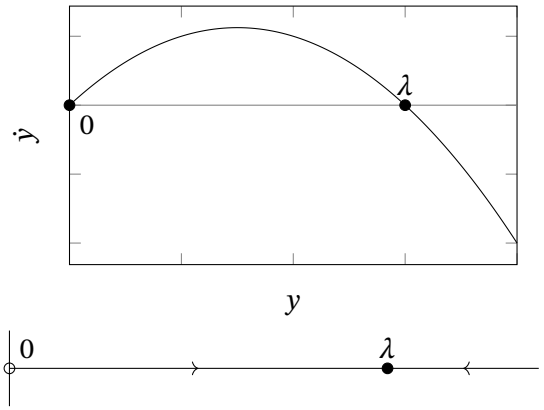
where the amount of molecules of sodium hydroxide is given by $a(t)$, the amount of molecules of hydrochloric acid is given by $b(t)$, the amount of molecules of water is given by $c(t)$, and the amount of molecules of sodium chloride is given by $d(t)$. We can model this using the equation $\frac{\mathrm{d}c}{\mathrm{d}t} = \lambda ab$. As atoms are conserved, we have $a = a_0 - c$ and $b = b_0 = c$. Then:

$$\frac{\mathrm{d}c}{\mathrm{d}t} = \lambda(a_0 - c)(b_0 - c)$$

This is an autonomous nonlinear first order ordinary differential equation. We can create a phase portrait by mapping out $\frac{\mathrm{d}c}{\mathrm{d}t}$ as a function of c, as shown in the first diagram here, which is known as a 2D phase portrait. The second diagram, known as a 1D phase portrait, shows similar information but helps us see the behaviour of fixed points—essentially the arrows point in the direction of motion of c; if $\dot{c}$ is positive then the arrows point to the right, if $\dot{c}$ is negative they point to the left.

Another example is a population model. Let $y(t)$ denote the population. Let αy denote the birth rate, and βy be the death rate. Then, we can model this using a linear model by:

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \alpha y - \beta y \quad \therefore y = y_0 e^{(\alpha-\beta)t}$$



If $\alpha > \beta$ then we have exponential growth; if $\alpha < \beta$ then we have exponential decay. This is an unrealistic model, so we can use a nonlinear model to increase accuracy.

$$\frac{\mathrm{d}y}{\mathrm{d}t} = (\alpha - \beta)y - \gamma y^2$$

When y is sufficiently large, the γ term becomes more relevant; here, the γy^2 term models the increased death rate at high populations. Equivalently, we can write

$$\dot{y} = ry\left(1 - \frac{y}{\lambda}\right)$$

7.2. Fixed points in discrete equations

Consider a first order discrete (or difference) equation of the form

$$x_{n+1} = f(x_n)$$

We define the fixed points of the equation to be any value of x_n where $x_{n+1} = x_n$ or equivalently $f(x_n) = x_n$. We can analyse fixed points' stability just like we can with differential equations, by using perturbation analysis. Let x_f denote a fixed point, and then we will perturb this by a small quantity ε.

$$f(x_f + \varepsilon) = \underbrace{f(x_f)}_{=x_f} + \varepsilon \left.\frac{\mathrm{d}f}{\mathrm{d}x}\right|_{x_f} + O(\varepsilon^2)$$

If we let $x_n = x_f + \varepsilon$, then

$$x_{n+1} \approx f(x_n) = f(x_f + \varepsilon) = x_f + \varepsilon \left.\frac{\mathrm{d}f}{\mathrm{d}x}\right|_{x_f}$$

x_f is stable if $\left|\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right|_{x_f}\right| < 1$, and unstable if this value is greater than 1. This is because if the value is less than 1, x_{n+1} is closer to x_f than x_n was.

7.3. Logistic map

This is an extended example of analysis of discrete equations. Let x_n be the population at generation n. Then, we use the model

$$\frac{x_{n+1} - x_n}{\Delta t} = \lambda x_n - \gamma x_n^2$$

We could contrast this with a nonlinear ordinary differential equation; the left hand side of this equation is analogous to $\frac{\mathrm{d}x}{\mathrm{d}t}$. Alternatively, grouping all x_n terms on the right hand side, we have

$$x_{n+1} = (\lambda \Delta t + 1)x_n - \gamma \Delta t x_n^2$$

We will actually use a slightly simplified model for this, by unifying the γ and λ terms as follows:

$$x_{n+1} = r x_n (1 - x_n) = f(x_n)$$

This is known as the 'logistic map'. We will analyse the fixed points of this equation by solving $f(x_n) = x_n$. We have two solutions, $x_n = 0$ and $x_n = 1 - \frac{1}{r}$. We can analyse their stability using perturbation analysis as before. By letting $f(x) = rx(1 - x)$, thus removing the n index, we have

$$\frac{\mathrm{d}f}{\mathrm{d}x} = r(1 - 2x)$$

At $x_n = 0$, $\frac{\mathrm{d}f}{\mathrm{d}x} = r$. When $0 < r < 1$, the point is stable because the next point produced by the perturbation analysis is closer to the fixed point. If $r > 1$ then the point is unstable.

At $x_n = 1 - \frac{1}{r}$, $\frac{\mathrm{d}f}{\mathrm{d}x} = 2 - r$. For $0 < r < 1$, the value of x_n is greater than 1, so it is unphysical so we discard it. For $r < 1 < 3$, the point is stable. When $r > 3$, the point is unstable.

8. Higher order linear ODEs

8.1. Linear 2nd order ODEs with constant coefficients

The general form of an equation of this type is

$$ay'' + by' + cy = f(x)$$

To solve equations like this, we are going to exploit two facts: the linearity of the differential operator together with the principle of superposition. From the definition of the derivative, we have

$$\frac{\mathrm{d}}{\mathrm{d}x}(y_1 + y_2) = y_1' + y_2'$$

And similarly,

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}(y_1 + y_2) = y_1'' + y_2''$$

For a linear differential operator D built from a linear combination of derivatives, for example

$$D = \left[a\frac{\mathrm{d}^2}{\mathrm{d}x^2} + b\frac{\mathrm{d}}{\mathrm{d}x} + c\right]$$

it then follows that

$$D(y_1 + y_2) = D(y_1) + D(y_2)$$

We will then solve the above general equation in three steps.

(i) Find the complementary functions y_1 and y_2 which satisfy the equivalent homogeneous equation $ay'' + by' + cy = 0$.

(ii) Find a particular integral y_p which solves the original equation.

(iii) If y_1 and y_2 are linearly independent, then $y_1 + y_p$ and $y_2 + y_p$ are each linearly independent solutions, which follows from the fact that $D(y_1) = D(y_2) = 0$ and $D(y_p) = f(x)$.

8.2. Eigenfunctions for 2nd order ODEs

$e^{\lambda x}$ is the eigenfunction of $\frac{\mathrm{d}}{\mathrm{d}x}$, and it is also the eigenfunction of $\frac{\mathrm{d}^2}{\mathrm{d}x^2}$, but with eigenvalue λ^2. More generally, it is the eigenfunction of $\frac{\mathrm{d}^n}{\mathrm{d}x^n}$ with eigenvalue λ^n. In fact, $e^{\lambda x}$ is the eigenfunction of any linear differential operator D. The equation $ay'' + by' + cy = 0$ can be written

$$\underbrace{\left[a\frac{\mathrm{d}^2}{\mathrm{d}x^2} + b\frac{\mathrm{d}}{\mathrm{d}x} + c\right]}_{\equiv D} y = 0$$

Therefore, solutions to this take the form

$$y_c = Ae^{\lambda x}$$

and by substituting, we have

$$a\lambda^2 + b\lambda + c = 0$$

This is known as the characteristic (or auxiliary) equation. From the fundamental theorem of algebra, this must have two real or complex solutions. Now, let λ_1, λ_2 be these roots.

In the case that $\lambda_1 \neq \lambda_2$, $y_1 = Ae^{\lambda_1 x}; y_2 = Be^{\lambda_2 x}$. In this case, the two are linearly independent and complete; they form a basis of solution space. Therefore any other solution to this differential equation can be written as a linear combination of y_1 and y_2. In general, $y_c = Ae^{\lambda_1 x} + Be^{\lambda_2 x}$.

8.3. Detuning

In the case that $\lambda_1 = \lambda_2$, this is known as a degenerate case as we have repeated eigenvalues; y_1 and y_2 are linearly dependent and not complete. Let us take as an example the differential equation $y'' - 4y' + 4y = 0$. We try $y_c = e^{2x}$ as $\lambda = 2$ in this case. We will consider a slightly modified ('detuned') equation to rectify the degeneracy.

$$y'' - 4y' + (4 - \varepsilon^2)y = 0 \text{ where } \varepsilon \ll 1$$

Again we will try $y_c = e^{\lambda x}$, giving

$$\lambda^2 - 4\lambda + (4 - \varepsilon^2) = 0$$

So we have $\lambda = 2 \pm \varepsilon$. The complementary function therefore is $y_c = Ae^{(2+\varepsilon)x} + Be^{(2-\varepsilon)x} = e^{2x}\left(Ae^{\varepsilon x} + Be^{-\varepsilon x}\right)$. We will expand this in a Taylor series for small ε, giving

$$y_c = e^{2x}\left[(A + B) + \varepsilon x(A - B) + O(\varepsilon^2)\right]$$

and by taking the limit, we have

$$\lim_{\varepsilon \to 0} y_c \approx e^{2x}\left[(A + B) + \varepsilon x(A - B)\right]$$

Now consider applying initial conditions to y_c at $x = 0$.

$$y_c\Big|_{x=0} = C \quad y_c'\Big|_{x=0} = D$$

and therefore

$$C = A + B; \quad D = 2C + \varepsilon(A - B)$$

hence

$$A + B = O(1); \quad A - B = O\left(\frac{1}{\varepsilon}\right)$$

in order that D is a constant. Now, let $\alpha = A + B; \beta = \varepsilon(A - B)$, so that we can get constants of $O(1)$ magnitude. Hence,

$$\lim_{\varepsilon \to 0} y_c = e^{2x}\left[\alpha + \beta x\right]$$

In general, if $y_1(x)$ is a degenerate complementary function for linear ODEs with constant coefficients, then $y_2 = xy_1$ is a linearly independent complementary function.

8.4. Reduction of order

Consider a homogeneous second-order linear ODE with non-constant coefficients. The general form of such an equation is

$$y'' + p(x)y' + q(x)y = 0 \tag{8.1}$$

Our objective is to use one solution to this equation (here denoted y_1) to find the other solution y_2. The general idea is to look for a solution of the form

$$y_2(x) = v(x)y_1(x) \tag{8.2}$$

First, note that

$$\begin{aligned} y_2' &= v'y_1 + vy_1' \\ y_2'' &= v''y_1 + 2v'y_1' + vy_1'' \end{aligned}$$

If y_2 is a solution to (8.1), then

$$y_2'' + p(x)y_2' + q(x)y_2 = 0$$

We can use (8.2) and collect terms, to get

$$v \cdot \underbrace{(y_1'' + py_1' + qy_1)}_{0 \text{ since } y_1 \text{ is a solution to (8.1)}} + v' \cdot (2y_1' + py_1) + v'' \cdot y_1 = 0$$

Hence

$$v' \cdot (2y_1' + py_1) + v'' \cdot y_1 = 0$$

This is a first order differential equation for $v'(x)$. Let $u = v'$. Then

$$u'y_1 + u(2y_1' + py_1) = 0$$

This is a separable first order ODE for $u(x)$. So we can solve for $u(x)$ and deduce $v(x)$ by integration.

8.5. Solution space

An nth order linear ODE written

$$p(x)y^{(n)} + q(x)y^{(n-1)} + \cdots + r(x)y = f(x)$$

can be used to write $y^{(n)}(x)$ in terms of lower derivatives of y. For example, the oscillations of a mass on a spring in a damped system can be modelled as

$$m\ddot{y} = -ky - L\dot{y}$$

Therefore the state of the system can be described by an n-dimensional solution vector

$$\mathbf{Y}(x) \equiv \begin{pmatrix} y(x) \\ y'(x) \\ \vdots \\ y^{(n-1)}(x) \end{pmatrix} \tag{8.3}$$

For example, an undamped oscillator modelled by $y'' + 4y = 0$ has solutions

$$y_1 = \cos 2x; \quad y_2 = \sin 2x$$

and has derivatives

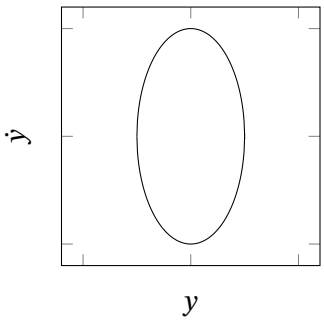
$$y_1' = -2\sin 2x; \quad y_2' = 2\cos 2x$$

and therefore two solution vectors are

$$\mathbf{Y}_1(x) = \begin{pmatrix} y_1 \\ y_1' \end{pmatrix} = \begin{pmatrix} \cos 2x \\ -2\sin 2x \end{pmatrix}$$

and

$$\mathbf{Y}_2(x) = \begin{pmatrix} y_2 \\ y_2' \end{pmatrix} = \begin{pmatrix} \sin 2x \\ 2\cos 2x \end{pmatrix}$$



We can plot the paths of these two solutions using a two-dimensional phase portrait. In this case, both solutions follow an elliptical path. Since $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are linearly independent for all x, any point in solution space (y, y') can be written as a linear combination of these solutions.

Solutions $y_1, y_2, \cdots, y_n$ are linearly independent for any ODE if their solution vectors $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n$ are linearly independent. A set of n linearly independent solution vectors forms a basis for the solution space of an nth order ODE.

8.6. Initial conditions

Consider initial conditions for a second order homogeneous ODE.

$$y(0) = a, \quad y'(0) = b$$

If the general solution is

$$y(x) = Ay_1(x) + By_2(x)$$

then we have the following linear system of equations

$$Ay_1(0) + By_2(0) = a$$
$$Ay_1'(0) + By_2'(0) = b$$

which is a system of two equations for two unknowns. Or alternatively,

$$\underbrace{\begin{pmatrix} y_1(0) & y_2(0) \\ y_1'(0) & y_2'(0) \end{pmatrix}}_{\equiv M} \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}$$

Unique solutions for A and B exist if $\det M \neq 0$.

8.7. The fundamental matrix and the Wrońskian

The fundamental matrix is a matrix formed by placing solution vector $\mathbf{Y}_i$ in the ith column. The Wrońskian, denoted $W(x)$, is the determinant of the fundamental matrix.

$$W(x) \equiv \begin{vmatrix} \vdots & \vdots & & \vdots \\ \mathbf{Y}_1 & \mathbf{Y}_2 & \cdots & \mathbf{Y}_n \\ \vdots & \vdots & & \vdots \end{vmatrix} = \begin{vmatrix} y_1 & y_2 & \cdots & y_n \\ y_1' & y_2' & \cdots & y_n' \\ \vdots & \vdots & \ddots & \vdots \\ y_1^{(n-1)} & y_2^{(n-1)} & \cdots & y_n^{(n-1)} \end{vmatrix}$$

For a second order ODE:

$$W(x) = \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix} = y_1 y_2' - y_2 y_1' \tag{8.4}$$

The solution vectors are linearly independent if $W(x) \neq 0$. This is a convenient test for the linear independence of two solution vectors. In our example above, we had

$$W(x) = \begin{vmatrix} \cos 2x & \sin 2x \\ -2\sin 2x & 2\cos 2x \end{vmatrix} = 2\cos^2 2x + 2\sin^2 2x = 2 \neq 0$$

So the solution vectors are linearly independent for all x.

If $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are linearly dependent, then $W(x) = 0$. Suppose that a third solution $y(x)$ is a linear combination of $y_1(x)$ and $y_2(x)$. Then the solution vectors $\mathbf{Y}, \mathbf{Y}_1, \mathbf{Y}_2$ are a linearly dependent set. Hence

$$\begin{vmatrix} y & y_1 & y_2 \\ y' & y_1' & y_2' \\ y'' & y_1'' & y_2'' \end{vmatrix} = 0$$

For $y_1 = \cos 2x$ and $y_2 = \sin 2x$, we can deduce the original differential equation that produced these solutions by solving for y.

$$\begin{vmatrix} y & \cos 2x & \sin 2x \\ y' & -2\sin 2x & 2\cos 2x \\ y'' & -4\cos 2x & -4\sin 2x \end{vmatrix} = 0$$

$$\implies y(8\sin^2 2x + 8\cos^2 2x)$$

$$-y'(-4\cos 2x \sin 2x + 4\cos 2x \sin 2x)$$

$$+y''(2\cos^2 2x + 2\sin^2 2x) = 0$$

$$\implies y'' + 4y = 0$$

Note that if $W(x) = 0$, this does not necessarily imply linear dependence.

8.8. Abel's theorem

Consider a second order homogeneous ODE:

$$y'' + p(x)y' + q(x)y = 0$$

Theorem (Abel's Theorem)**.** If $p(x)$ and $q(x)$ are continuous on an interval I, then the Wrońskian $W(x)$ is either zero or nonzero for all $x \in I$.

Proof. Let y_1, y_2 be solutions to the equation. Then

$$y_2(y_1'' + p(x)y_1' + q(x)y_1) = 0 \tag{8.5}$$
$$y_1(y_2'' + p(x)y_2' + q(x)y_2) = 0 \tag{8.6}$$

Now, calculating (8.6) − (8.5), we get

$$(y_1y_2'' - y_2y_1'') + p(x)(y_1y_2' - y_2y_1') = 0 \tag{8.7}$$

As we are solving a second order equation, $W(x) = y_1y_2' - y_2y_1'$ and therefore

$$\frac{\mathrm{d}W}{\mathrm{d}x} = y_1y_2'' + y_1'y_2' - y_2'y_1' - y_2y_1'' = y_1y_2'' - y_2y_1''$$

Note that these are the coefficients in (8.7). We have therefore

$$W' + pW = 0 \tag{8.8}$$

Then by separating variables:

$$\frac{\mathrm{d}W}{W} = -p(x)\,\mathrm{d}x$$
$$\int_{x_0}^{x} \frac{\mathrm{d}W}{W} = -\int_{x_0}^{x} p(u)\,\mathrm{d}u$$
$$W(x) = W(x_0)e^{-\int_{x_0}^{x} p(u)\mathrm{d}u}$$

This last equation is known as Abel's Identity, and is very important. Since $p(x)$ is continuous on I with $x \in I$, it is bounded and therefore integrable. Therefore $e^{-\int_{x_0}^{x} p(u)\mathrm{d}u} \neq 0$. It follows that if $W(x_0) = 0$ then $W(x) = 0$ for all x. Likewise, if $W(x_0) \neq 0$, then $W(x) \neq 0$ for all x (on the interval). □

Corollary. If $p(x) = 0$, then $W = W_0$ which is a constant.

Note that we can use this to find $W(x)$ without actually solving the differential equation itself. For example, Bessel's Equation

$$x^2y'' + xy' + (x^2 - n^2)y = 0$$

has no closed form solutions, but the Wrońskian can be calculated be rewriting it as

$$y'' + \frac{1}{x}y' + \frac{x^2 - n^2}{x^2}y = 0$$

and by Abel's Identity,

$$\begin{aligned} W(x) &= W_0 e^{-\int_{x_0}^{x} \frac{1}{u}\mathrm{d}u} \\ &= W_0 e^{-\ln x} \\ &= \frac{W_0}{x} \end{aligned}$$

We can find a second solution y_2 given a solution y_1 using a reduction of order method, but we can also use Abel's Identity.

$$y_1 y_2' - y_2 y_1' = W_0 e^{-\int_{x_0}^{x} p(u)\mathrm{d}u}$$

This is a first order ODE for y_2 which we can now solve:

$$\frac{y_1 y_2' - y_2 y_1'}{y_1^2} = \frac{W_0}{y_1^2} e^{-\int_{x_0}^{x} p(u)\mathrm{d}u}$$

The left hand side is exactly the quotient rule, giving

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{y_2}{y_1} = \frac{W_0}{y_1^2} e^{-\int_{x_0}^{x} p(u)\mathrm{d}u}$$

which can be solved to give y_2 as a function of y_1 and W.

We can use Abel's theorem in higher dimensions. Any linear nth order ODE can be written

$$\mathbf{Y}' + A(x)\mathbf{Y} = 0$$

where A is a matrix; this converts an nth order ODE into a system of n first order ODEs. This will be discussed later in the course. It can be shown that this generalisation of Abel's Identity

$$W' + \mathrm{tr}(A)W = 0$$

holds, and hence

$$W' = W_0 e^{-\int_{x_0}^{x} \mathrm{tr}(A)\mathrm{d}u}$$

and Abel's theorem holds. This is shown on example sheet 3, Question 7.

8.9. Equidimensional equations

An ODE is equidimensional if the differential operator is unaffected by a multiplicative scaling. For example, rescaling

$$x \mapsto X = \alpha x$$

where $\alpha \in \mathbb{R}$. The general form for a second order equidimensional equation is

$$ax^2y'' + bxy' + cy = f(x) \tag{8.9}$$

where a, b, c are constant. Note, $\frac{\mathrm{d}}{\mathrm{d}X} = \frac{1}{\alpha}\frac{\mathrm{d}}{\mathrm{d}x}$, and $\frac{\mathrm{d}^2}{\mathrm{d}X^2} = \frac{1}{\alpha^2}\frac{\mathrm{d}}{\mathrm{d}x^2}$, so plugging this into (8.9) gives

$$aX^2\frac{\mathrm{d}^2y}{\mathrm{d}X^2} + bX\frac{\mathrm{d}y}{\mathrm{d}X} + cy = f\left(\frac{X}{\alpha}\right)$$

The left hand side was unaffected by this rescaling, so the equation is equidimensional.

There are two main methods for solving equidimensional equations.

(i) Note that $y = x^k$ is an eigenfunction of the differential operator $x\frac{\mathrm{d}}{\mathrm{d}x}$. Inspired by this, to solve (8.9) we will look for solutions of the form $y = x^k$, so we have

$$ak(k-1) + bk + c = 0$$

We can simply solve this quadratic for two roots k_1 and k_2. If $k_1 \neq k_2$, then the complementary function is

$$y_c = Ax^{k_1} + Bx^{k_2}$$

(ii) If $k_1 = k_2$, then the substitution $z = \ln x$ turns (8.9) into an equation with constant coefficients.

$$a\frac{\mathrm{d}^2y}{\mathrm{d}z^2} + (b-a)\frac{\mathrm{d}y}{\mathrm{d}z} + cy = f(e^z)$$

Because this has constant coefficients, our complementary functions will be of the form $y = e^{\lambda z}$, which can be solved as usual.

$$y_c = Ae^{\lambda_1 z} + Be^{\lambda_2 z} = Ax^{\lambda_1} + Bx^{\lambda_2}$$

which is the same form as above. In this form, it is easier to see that if the two solutions λ_1, λ_2 are the same, then

$$y_c = Ae^{\lambda_1 z} + Bze^{\lambda_1 z} = Ax^{k_1} + Bx^{k_1}\ln x$$

9. Forced second order ODEs

We want to find methods for finding the particular integral of forced second order ODEs.

9.1. Guesswork

Here, $f(x)$ is the forcing term.

Form of $f(x)$	Form of $y_p(x)$
e^{mx}	Ae^{mx}
$\sin(kx)$ or $\cos(kx)$	$A\sin(kx) + B\cos(kx)$
x^n or an nth degree polynomial	$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$

In general:

(i) Insert our guess into the equation;

(ii) Equate coefficients of functions;

(iii) Solve for unknown coefficients.

9.2. Variation of parameters

This is a method for finding the particular integral y_p given complementary functions y_1, y_2, which are assumed to be linearly independent, with solution vectors

$$\mathbf{Y}_1 = \begin{pmatrix} y_1 \\ y_1' \end{pmatrix}; \quad \mathbf{Y}_2 = \begin{pmatrix} y_2 \\ y_2' \end{pmatrix}$$

Suppose that the solution vector for y_p satisfies

$$\mathbf{Y}_p = \begin{pmatrix} y_p \\ y_p' \end{pmatrix} = u(x)\mathbf{Y}_1 + v(x)\mathbf{Y}_2 \tag{9.1}$$

This is not a linear combination of $\mathbf{Y}_1$ and $\mathbf{Y}_2$, but we can treat it as a linear combination at a fixed x point (just as a way to visualise it). We want to find equations for $u(x)$ and $v(x)$. By comparing components of the vectors on the left and right, we have

$$y_p = uy_1 + vy_2 \tag{a}$$
$$y_p' = uy_1' + vy_2' \tag{b}$$

So therefore,

$$\frac{\mathrm{d}}{\mathrm{d}x}(\text{a}) \implies y_p' = u'y_1 + uy_1' + v'y_2 + vy_2' \tag{c}$$
$$(\text{c}) - (\text{b}) \implies u'y_1 + v'y_2 = 0 \tag{d}$$

Now:

$$\frac{\mathrm{d}}{\mathrm{d}x}(\mathrm{b}) \implies y_p'' = uy_1'' + u'y_1' + v'y_2' + vy_2'' \tag{e}$$

If $y_p'' + p(x)y_p' + q(x)y_p = f(x)$, then

$$(\mathrm{e}) + p(\mathrm{b}) + q(\mathrm{a}) = f(x)$$

But also, y_1 and y_2 satisfy the differential equation

$$y_c'' + py_c' + qy_c = 0 \quad y_c \in \{y_1, y_2\}$$

After we substitute in and cancel a lot of terms, we have

$$u'y_1' + v'y_2' = f(x) \tag{f}$$

Combining (d) and (f), we can deduce u and v.

$$\underbrace{\begin{pmatrix} y_1 & y_2 \\ y_1' & y_2' \end{pmatrix}}_{\text{fundamental matrix}} \begin{pmatrix} u' \\ v' \end{pmatrix} = \begin{pmatrix} 0 \\ f \end{pmatrix}$$

So as long as y_1 and y_2 are linearly independent, i.e. the Wrońskian is nonzero, we can write an equation for u' and v'.

$$\begin{pmatrix} u' \\ v' \end{pmatrix} = \frac{1}{W(x)} \begin{pmatrix} y_2' & -y_2 \\ y_1' & y_1 \end{pmatrix} \begin{pmatrix} 0 \\ f \end{pmatrix}$$

or explicitly,

$$u' = \frac{-y_2 f}{W}$$
$$v' = \frac{y_1 f}{W}$$

and therefore

$$y_p = y_2 \int^x \frac{y_1(t)f(t)}{W(t)}\,\mathrm{d}t - y_1 \int^x \frac{y_2(t)f(t)}{W(t)}\,\mathrm{d}t$$

9.3. Forced oscillating systems: transients and damping

Many physical systems have a restoring force and damping (e.g. friction). For example, the suspension of a car could be modelled with $y(t)$, where y is the height of the wheel, given by

$$M\ddot{y} = F(t) - \underbrace{ky}_{\text{spring}} - \underbrace{L\dot{y}}_{\text{damper}}$$

In standard form, we have

$$\ddot{y} + \frac{L}{M}\dot{y} + \frac{k}{M}y = \frac{F(t)}{M}$$

Let $\tau = \sqrt{\frac{k}{M}}t$. Then we can rewrite this equation using a single parameter:

$$y'' + 2Ky' + y = f(\tau)$$

where $y' = \frac{\mathrm{d}y}{\mathrm{d}\tau}$, $K = \frac{L}{2\sqrt{kM}}$, $f = \frac{F}{k}$. Our unforced system is described here by one parameter K.

In the case of the unforced response (also known as the free or natural response), we have $f = 0$, so

$$y'' + 2Ky' + y = 0$$

Solving by the characteristic equation, we see

$$\lambda = -K \pm \sqrt{K^2 - 1}$$

There are a number of cases here.

(i) ($K < 1$) This produces a decaying oscillation, known as 'underdamped'. λ_1, λ_2 are both complex, and therefore

$$y = e^{-K\tau}\left[A\sin\left(\sqrt{1-K^2}\tau\right) + B\cos\left(\sqrt{1-K^2}\tau\right)\right]$$

The period is $\frac{2\pi}{\sqrt{1-K^2}}$. As K tends to 1, the period tends to ∞.

(ii) ($K = 1$) This is the degenerate case, $\lambda_1 = \lambda_2 = -K$. We can use detuning to deduce

$$y = e^{-K\tau}(A + B\tau)$$

(iii) ($K > 1$) Here, we have two negative real roots λ_1, λ_2; this situation is known as 'overdamped'.

$$y = Ae^{\lambda_1\tau} + Be^{\lambda_2\tau}$$

Note that the unforced response decays in all cases.

9.4. Sinusoidal forcing

Let

$$\ddot{y} + \mu\dot{y} + \omega_0^2 y = \sin\omega t$$

Now let us guess $y_p = A\sin\omega t + B\cos\omega t$. Equating coefficients of $\sin\omega t$ gives

$$-A\omega^2 - B\mu\omega + \omega_0^2 A = 1 \tag{a}$$

and equating $\cos\omega t$ gives

$$-B\omega^2 + A\mu\omega + \omega_0^2 B = 0 \tag{b}$$

Then:

$$\text{(b)} \implies A = B\frac{\omega^2 - \omega_0^2}{\mu\omega}$$

$$\text{(a)} \implies A(\omega_0^2 - \omega^2) = 1 + B\mu\omega$$

So

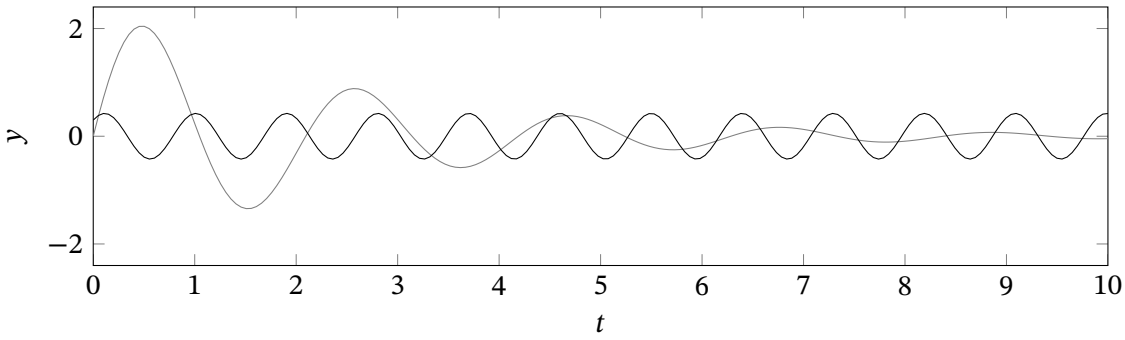
$$A = \frac{\omega_0^2 - \omega^2}{(\omega_0^2 - \omega^2)^2 + \mu^2\omega^2}$$

$$B = \frac{-\mu\omega}{(\omega_0^2 - \omega^2)^2 + \mu^2\omega^2}$$

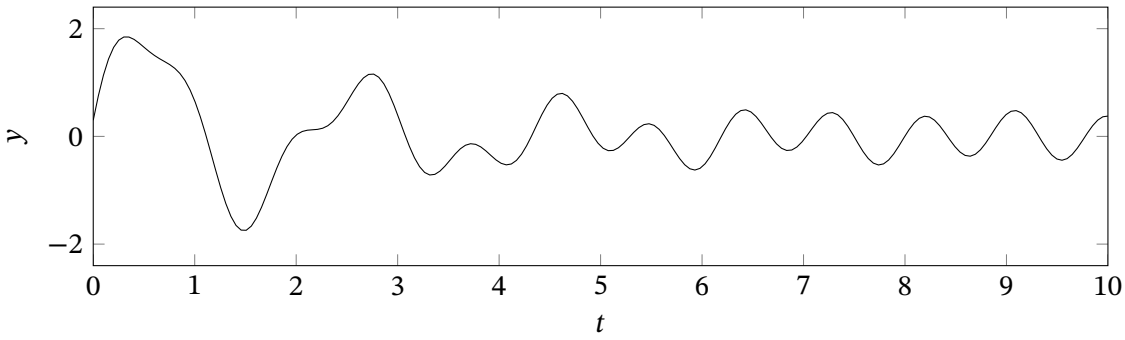
Altogether, we have

$$y_p = \frac{1}{(\omega_0^2 - \omega^2)^2 + \mu^2\omega^2}\left[(\omega_0^2 - \omega^2)\sin\omega t - \mu\omega\cos\omega t\right]$$

Drawing an example of this kind of particular integral (with the complementary function in grey), we can see the following:



And adding both together to form a particular solution gives:



Let us make a few comments about these forced oscillations.

- The complementary function gives us the transient (short-term) response to the initial conditions.
- The particular integral gives the long-term response to the forcing term.

- In some sense, the system 'forgets' about the initial conditions over time due to the damping term.

9.5. Resonance in undamped systems

What happens if $\omega = \omega_0$? If $\mu \neq 0$ (i.e. it is a damped system), then

$$\lim_{\omega \to \omega_0} y_p = \frac{-\cos \omega_0 t}{\mu \omega_0}$$

This is a finite amplitude oscillation. Note that the amplitude increases with decreasing μ, so clearly this solution has a problem at $\mu = 0$. To work with this, we'll let

$$\ddot{y} + \omega_0^2 y = \sin \omega_0 t$$

We will use detuning to get solutions for this equation. Consider instead

$$\ddot{y} + \omega_0^2 y = \sin \omega t$$

where $\omega \neq \omega_0$. We will guess that the particular integral is of the form $y_p = C \sin \omega t$ since by inspection there cannot be any cosine terms.

$$C(-\omega^2 + \omega_0^2) = 1$$

$$\therefore y_p = \frac{1}{\omega_0^2 - \omega^2} \sin \omega t$$

As the system is linear in y and its derivatives, we can freely add some multiple of the complementary function and it will remain a solution.

$$y_p = \frac{1}{\omega_0^2 - \omega^2} \sin \omega t + A \sin \omega_0 t$$

Now let us pick $A = \frac{-1}{\omega_0^2 - \omega^2}$, so

$$y_p = \frac{\sin \omega t - \sin \omega_0 t}{\omega_0^2 - \omega^2}$$

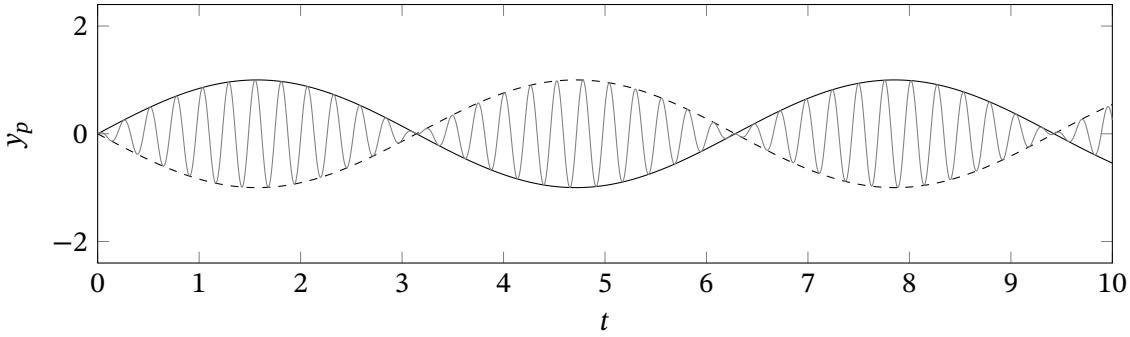
Rewriting this using angle addition and subtraction identities:

$$y_p = \frac{2}{\omega_0^2 - \omega^2} \left[\cos\left(\frac{\omega + \omega_0}{2} t \right) \sin\left(\frac{\omega - \omega_0}{2} t \right) \right]$$

For convenience, let $\Delta\omega \equiv \omega_0 - \omega$, and therefore $\frac{\omega + \omega_0}{2} = w_0 - \frac{1}{2}\Delta\omega$.

$$y_p = \frac{-2}{\Delta\omega(\omega_0 + \omega)} \left[\cos\left(\left(\omega_0 - \frac{\Delta\omega}{2} \right) t \right) \sin \frac{\Delta\omega t}{2} \right]$$

In the following diagram, y_p is drawn in grey, with the sine term acting as an envelope for the higher-frequency cosine term. The phenomenon visible here is known as 'beating', as an oscillator with a fundamental frequency slightly different to the forcing frequency will begin oscillating then stop, and repeat this cycle.



As we reduce $\Delta\omega$ to zero, we have

$$\lim_{\Delta\omega\to 0} \sin\left(\frac{\Delta\omega}{2}t\right) \approx \frac{\Delta\omega}{2}t$$

So

$$\begin{aligned}\lim_{\Delta\omega\to 0} &\approx \frac{-2}{\Delta\omega(\omega_0+\omega_0)}\cos(\omega_0 t)\left(\frac{\Delta\omega}{2}t\right)\\ &\approx \frac{-2t}{\omega_0}\cos\omega_0 t\end{aligned}$$

This is linear growth in amplitude over time. This increase is unbounded an in undamped system. Note that y_p takes the form of one of the complementary functions multiplied by the independent variable.

9.6. Impulses and point forces

Consider a system that experiences a sudden force, for example a car's suspension when driving over a speed bump. Let us define y to be the displacement from the undisturbed height of the suspension. Let the car's mass be M. In a small finite window ε around some time T, the excess force F (the forcing term) on the system is greater than zero. As ε tends to zero, the force becomes a sudden impulse. Let us model this using the equation

$$M\ddot{y} = F(t) - ky - L\dot{y}$$

We can see that before time T, $y = 0$. After this point, there is some kind of oscillation. Note that the value of y is always continuous (otherwise this would violate many laws of physics), but the derivative is not necessarily continuous at the point T. Let us integrate the equation

above in time from $T-\varepsilon$ to $T+\varepsilon$.

$$\lim_{\varepsilon\to 0}\left[M[\dot{y}]_{T-\varepsilon}^{T+\varepsilon} = \int_{T-\varepsilon}^{T+\varepsilon} F(t)\,\mathrm{d}t - k\underbrace{\int_{T-\varepsilon}^{T+\varepsilon} y\,\mathrm{d}t}_{\substack{0 \text{ if } y \text{ is finite}}} - L\underbrace{[y]_{T-\varepsilon}^{T+\varepsilon}}_{\substack{0 \text{ if } y \text{ is continuous}}} \right] \tag{9.2}$$

We now can define the impulse I to be

$$I = \lim_{\varepsilon\to 0}\int_{T-\varepsilon}^{T+\varepsilon} F(t)\,\mathrm{d}t$$

Hence

$$(9.2) \implies I = \lim_{\varepsilon\to 0} M[\dot{y}]_{T-\varepsilon}^{T+\varepsilon}$$

So if the impulse is nonzero, the velocity $\dot{y}$ experiences a sudden change, so it is discontinuous at T. The value of this sudden change in velocity depends on the integral of the force.

10. Impulse forcing

10.1. Dirac δ function

First let us consider a family of functions $D(t;\varepsilon)$ defined by

$$\lim_{\varepsilon\to 0} D(t;\varepsilon) = 0; \quad \forall t \neq 0$$

$$\int_{-\infty}^{\infty} D(t;\varepsilon)\,\mathrm{d}t = 1$$

For example,

$$D(t;\varepsilon) = \frac{1}{\varepsilon\sqrt{\pi}} e^{\frac{-t^2}{\varepsilon^2}}$$

We can now define the Dirac delta function $\delta(t) = \lim_{\varepsilon\to 0} D(t;\varepsilon)$. It has a number of interesting properties:

- $\delta(x) = 0$ for all nonzero x
- $\int_{-\infty}^{\infty} \delta(t)\,\mathrm{d}t = 1$
- (sampling property) For a continuous function $g(x)$:

$$\int_{-\infty}^{\infty} g(x)\delta(x)\,\mathrm{d}x = g(0)\int_{-\infty}^{\infty} \delta(x)\,\mathrm{d}x = g(0)$$

  And more generally,

$$\int_a^b g(x)\delta(x-x_0)\,\mathrm{d}x = \begin{cases} g(x_0) & a \le x_0 \le b \\ 0 & \text{otherwise} \end{cases}$$

10.2. Heaviside step function

Exploiting the definition of the δ function, we will define the Heaviside step function $H(x)$ by

$$H(x) \equiv \int_{-\infty}^{x} \delta(t)\,\mathrm{d}t$$

Here are some of its properties:

- $H(x) = 0$ for $x < 0$
- $H(x) = 1$ for $x > 0$
- $H(0)$ is undefined

10.3. Ramp function

We define the ramp function $r(x)$ by

$$r(x) \equiv \int_{-\infty}^{x} H(t)\,\mathrm{d}t$$

This function is shaped like a ramp:

$$r(x) = \begin{cases} 0 & x < 0 \\ x & x \geq 0 \end{cases}$$

These functions get 'smoother' the more times we integrate.

10.4. Delta function forcing

Consider a linear second order ODE of the form

$$y'' + p(x)y' + q(x)y = \delta(x) \tag{10.1}$$

The key principle is that the highest order derivative 'inherits' the level of discontinuity from the forcing term, since if any other derivative were to contain the discontinuous function, then the next higher derivative would only be more discontinuous. So, y'' behaves somewhat like δ. Here, we will denote this $y'' \sim \delta$—this is extremely non-standard notation, however.

Now, since $\delta(x) = 0$ for all nonzero x, then

$$y'' + py' + qy = 0 \text{ for } x < 0, x > 0$$

We will essentially have two solutions for y; one for before the impulse and one after. We need to combine these together somehow to create the resultant y solution. But this leaves four constants of integration, so surely we can't solve it. Luckily, y satisfies certain 'jump conditions' (the analogous concept to initial conditions in this context):

- $y(x)$ is continuous at $x = 0$ because $y'' \sim \delta \implies y' \sim H \implies y \sim r$. More precisely:

$$\lim_{\varepsilon \to 0} [y]_{x=-\varepsilon}^{x=\varepsilon} = 0$$

- $y'(x)$ is has a jump of 1 at $x = 0$ because $y'' \sim \delta \implies y' \sim H$. Again we can formulate this intuition more precisely by integrating (10.1) around a small window ε:

$$\lim_{\varepsilon \to 0} \int_{-\varepsilon}^{\varepsilon} y'' + p(x)y' + q(x)y\,\mathrm{d}x = \lim_{\varepsilon \to 0} \int_{-\varepsilon}^{\varepsilon} \delta(x)\,\mathrm{d}x$$
$$\lim_{\varepsilon \to 0} [y']_{x=-\varepsilon}^{x=\varepsilon} = 1$$

As a more concrete example, let us solve

$$y'' - y = 3\delta\left(x - \frac{\pi}{2}\right)$$

where

$$x = 0, x = \pi \implies y = 0$$

First, let us solve the interval $0 \le x < \frac{\pi}{2}$.

$$\begin{aligned} y'' - y &= 0 \\ y &= Ae^x + Be^{-x} \\ \text{or } y &= A\sinh x + B\cosh x \text{ redefining } A, B \\ (y = 0 \text{ at } x = 0) &\implies y = A\sinh x \end{aligned}$$

Similarly, between $\frac{\pi}{2} < x \le \pi$ (since the equation is invariant under the transformation $x \mapsto \pi - x$):

$$y = C\sinh(\pi - x)$$

Now, we can apply two jump conditions to solve for these constants:

- Integrating the differential equation over a small region:

$$\lim_{\varepsilon \to 0} [y]_{x=\frac{\pi}{2}-\varepsilon}^{x=\frac{\pi}{2}+\varepsilon} = 3$$

  Hence, taking the derivatives of our two solutions:

$$-A\cosh\frac{\pi}{2} - C\cosh\frac{\pi}{2} = 3$$

- Since the y term is continuous:

$$[y]_{x=\frac{\pi}{2}-\varepsilon}^{x=\frac{\pi}{2}} = 0$$

$$A\sinh\frac{\pi}{2} = C\sinh\frac{\pi}{2} \implies A = C$$

Using both jump conditions, we have $A = C = \frac{-3}{2\cosh\frac{\pi}{2}}$. So our general solution is

$$y = \begin{cases} \frac{-3\sinh x}{2\cosh\frac{\pi}{2}} & 0 \le x < \frac{\pi}{2} \\ \frac{-3\sinh(\pi - x)}{2\cosh\frac{\pi}{2}} & \frac{\pi}{2} < x \le \pi \end{cases}$$

Note that often when working with limits as $\varepsilon \to 0$, we simply elide the limit sign since it is so ubiquitous.

10.5. Heaviside function forcing

Consider

$$y'' + p(x)y' + q(x)y = H(x - x_0) \tag{10.2}$$

Now, $y(x)$ satisfies

$$y'' + py' + qy = 0 \quad (x < x_0) \tag{10.3}$$
$$y'' + py' + qy = 1 \quad (x > x_0) \tag{10.4}$$

Evaluating (10.2) on either side of x_0, we have

$$[y'']_{x_0^-}^{x_0^+} + p(x_0)[y']_{x_0^-}^{x_0^+} + q(x_0)_{x_0^-}^{x_0^+} = 1$$

If $y'' \sim H$ then $y' \sim r$ and then $y \sim \int r$. Hence, y' and y are both continuous. So our jump conditions are

- $[y']_{x_0^-}^{x_0^+} = 0$
- $[y]_{x_0^-}^{x_0^+} = 0$

We can use the two initial or boundary conditions, along with the two jump conditions, to find the four constants in the solutions to (10.3) and (10.4).

11. Discrete equations and the method of Frobenius

11.1. Higher order discrete equations

The general form of an mth order linear discrete equation with constant coefficients is

$$a_m y_{n+m} + a_{m-1} y_{n+m-1} + \cdots + a_1 y_{n+1} + a_0 y_n = f_n \tag{11.1}$$

To solve such an equation, we will exploit some principles used to solve higher order differential equations.

To apply eigenfunction properties, we will define a difference operator $D[y_n] = y_{n+1}$. Then, D has eigenfunction $y_n = k^n$ for k constant, since $D[k^n] = k^{n+1} = k \cdot k^n = k y_n$.

To apply linearity, notice that (11.1) is linear in y, so the general solution $y_n = y_n^{(c)} + y_n^{(p)}$ where $y^{(c)}$ is the complementary function and $y^{(p)}$ is the particular integral.

As an example, let us consider a second order difference equation

$$a_2 y_{n+2} + a_1 y_{n+1} + a_0 y_n = f_n$$

We will first try to solve the homogeneous equation, letting $f = 0$.

$$a_2 y_{n+2} + a_1 y_{n+1} + a_0 y_n = 0$$

We will look for solutions of the form of the eigenfunction: $y_n = k^n$.

$$a_2 k^2 + a_1 k + a_0 = 0$$

This quadratic may be solved to give k_1 and k_2. Then our complementary function is

$$y_n^{(c)} = \begin{cases} A k_1^n + B k_2^n & k_1 \neq k_2 \\ A k^n + B n k^n & k_1 = k_2 = k \end{cases}$$

To solve the particular integral, let us consult this table:

Form of f_n	Form of $y_n^{(p)}$
k^n	Ak^n if $k \neq k_1, k_2$
k_1^n, k_2^n	$Ank_1^n + Bnk_2^n$
n^p	$An^p + Bn^{p-1} + \cdots + Cn + D$

11.2. Fibonacci sequence

The Fibonacci sequence is given by

$$y_n = y_{n-1} + y_{n-2}$$

with initial conditions $y_0 = y_1 = 1$. In standard form, we have

$$y_{n+2} - y_{n+1} - y_n = 0$$

We will look for solutions of the form $y = k^n$. Then

$$k^2 - k - 1 = 0$$

So we have

$$k_1 = \phi = \frac{1+\sqrt{5}}{2}; \quad k_2 = -\phi^{-1} = \frac{1-\sqrt{5}}{2}$$

Solving for the initial conditions gives

$$y_n = \frac{1}{\sqrt{5}}\phi + \frac{1}{\sqrt{5}}\phi^{-1} = \frac{\phi^{n+1} - (-\phi^{-1})^{n+1}}{\sqrt{5}}$$

So we can deduce that

$$\lim_{n\to\infty} \frac{y_{n+1}}{y_n} = \lim_{n\to\infty} \frac{\phi^{n+2} - (-\phi^{-1})^{n+2}}{\phi^{n+1} - (-\phi^{-1})^{n+1}} = \phi$$

11.3. Method of Frobenius

The Method of Frobenius is a way of computing series solutions to linear homogeneous second order ODEs. The general form is

$$p(x)y'' + q(x)y' + r(x)y = 0$$

We will seek a power series expansion about some point $x = x_0$. First, we must classify the point x_0:

- (ordinary point) $x = x_0$ is an ordinary point if the Taylor series of q/p and r/p converge in some region around x_0; i.e. q/p and r/p are analytic.
- (singular point) If x_0 is not ordinary, it is singular. There are two types of singular points:
  - (regular singular point) If the original ODE can be written as

    $$P(x)(x-x_0)^2y'' + Q(x)(x-x_0)y' + R(x)y = 0$$

    and $\frac{Q}{P}$ and $\frac{R}{P}$ are analytic, then $x = x_0$ is a regular singular point. Note that $\frac{Q}{P} = (x-x_0)\frac{q}{p}; \frac{R}{P}(x-x_0)^2\frac{r}{p}$.
  - (irregular singular point) Otherwise, $x = x_0$ is an irregular singular point.

Here are some examples.

(i) $(1-x^2)y'' - 2xy' + 2y = 0$. We have $q/p = \frac{-2x}{1-x^2}$, so $x = \pm 1$ are singular points. But $Q/P = \frac{2x}{1+x}$ which is regular at $x = 1$; a similar argument holds for -1.

(ii) $y'' \sin x + y' \cos x + 2y = 0$. We have $q/p = \cot x$, $r/p = 2\csc x$. So where $x = n\pi$ where $n \in \mathbb{Z}$, we have regular singular points.

(iii) $(1+\sqrt{x})y'' - 2xy' + 2y = 0$. We have $q/p = \frac{-2x}{1+\sqrt{x}}$. Around $x = 0$, the second derivative is undefined, so this is an irregular singular point.

11.4. Fuch's theorem

Theorem. (i) If x_0 is an ordinary point, then there are two linearly independent solutions of the form

$$y = \sum_{n=0}^{\infty} a_n (x - x_0)^n$$

This series is convergent in some region around x_0.

(ii) If x_0 is a regular singular point, then there is at least one solution of the form

$$y = \sum_{n=0}^{\infty} a_n (x - x_0)^{n+\sigma}$$

where σ is real and $a_0 \neq 0$.

Example. Here is an example of case 1.

$$(1-x^2)y'' - 2xy' + 2y = 0 \tag{11.2}$$

We will try to find series solutions about $x_0 = 0$, an ordinary point. We will therefore try solutions of the form

$$y = \sum_{n=0}^{\infty} a_n x^n$$
$$y' = \sum_{n=1}^{\infty} n a_n x^{n-1}$$
$$y'' = \sum_{n=2}^{\infty} n(n-1) a_n x^{n-2}$$

Now, to make all powers of x at least n, we will multiply (11.2) by x^2 for convenience.

$$(1-x^2)x^2 y'' - 2x^3 y' + 2x^2 y = 0$$
$$(1-x^2)x^2 \sum_{n=2}^{\infty} n(n-1) a_n (x-x_0)^{n-2} - 2x^3 \sum_{n=1}^{\infty} n a_n (x-x_0)^{n-1} + 2x^2 \sum_{n=0}^{\infty} a_n (x-x_0)^n = 0$$

$$(1-x^2)\sum_{n=2}^{\infty} n(n-1)a_n x^n - 2x^2\sum_{n=1}^{\infty} na_n x^n + 2x^2\sum_{n=0}^{\infty} a_n x^n = 0$$

$$\sum_{n=2}^{\infty} a_n[n(n-1)(1-x^2)]x^n - 2\sum_{n=1}^{\infty} a_n(nx^2)x^n + 2\sum_{n=0}^{\infty} a_n(x^2)x^n = 0$$

Now, for $n \geq 2$, equating the x^n coefficients we have

$$a_n[n(n-1)] - a_{n-2}[(n-2)(n-3)] - 2a_{n-2}(n-2) + 2a_{n-2} = 0$$

This is a discrete equation. Rewritten in a more standard form, we have

$$n(n-1)a_n = (n^2-3n)a_{n-2}$$

or

$$a_n = \frac{n-3}{n-1}a_{n-2} \tag{11.3}$$

This is known as the recurrence relation. The values of a_0 and a_1 are the unknown constants to be found via initial or boundary conditions. Note that $a_3 = 0$ from (11.3). Therefore, any odd power of x of higher order than x^1 is zero. For even n, we have

$$\begin{aligned}
a_n &= \frac{n-3}{n-1}a_{n-2} \\
a_n &= \frac{n-3}{n-1}\frac{n-5}{n-3}a_{n-4} = \frac{n-5}{n-1}a_{n-4} \\
a_n &= \frac{n-5}{n-1}\frac{n-7}{n-5}a_{n-6} = \frac{n-7}{n-1}a_{n-6} \\
\therefore a_n &= \frac{-1}{n-1}a_0
\end{aligned}$$

Therefore

$$y = a_1 x + a_0\left[1 - x^2 - \frac{x^4}{3} - \frac{x^6}{5} - \frac{x^8}{7} - \ldots\right]$$

Note that

$$\ln(1 \pm x) = \pm x - \frac{x^2}{2} \pm \frac{x^3}{3} - \ldots$$

Therefore

$$\ln\left(\frac{1+x}{1-x}\right) = \ln(1+x) - \ln(1-x) = 2x + 2\frac{x^3}{3} + 2\frac{x^5}{5} + \ldots$$

Hence,

$$y = a_1 x + a_0\left[1 - \frac{x}{2}\ln\left(\frac{1+x}{1-x}\right)\right]$$

Note the behaviour of this function near the singular points of the original differential equation.

Example. Consider the following differential equation:

$$4xy'' + 2(1 - x^2)y' - xy = 0 \tag{11.4}$$

We want to find series solutions about $x = 0$. In this case, $\frac{q}{p}$ is undefined at $x = 0$, so it is a singular point, but it is regular. We will try solutions of the form

$$\begin{aligned} y &= \sum_{n=0}^{\infty} a_n x^{n+\sigma} \\ y' &= \sum_{n=0}^{\infty} (n+\sigma) a_n x^{n+\sigma-1} \\ y'' &= \sum_{n=0}^{\infty} (n+\sigma)(n+\sigma-1) a_n x^{n+\sigma-2} \end{aligned}$$

where $a_0 \neq 0$. For convenience we will multiply (11.4) by x:

$$4x^2y'' + 2(1 - x^2)xy' - x^2y = 0$$

$$4x^2 \sum_{n=0}^{\infty} (n+\sigma)(n+\sigma-1) a_n x^{n+\sigma-2} + 2(1-x^2)x \sum_{n=0}^{\infty} (n+\sigma) a_n x^{n+\sigma-1} - x^2 \sum_{n=0}^{\infty} a_n x^{n+\sigma}$$

Hence,

$$\sum_{n=0}^{\infty} a_n x^{n+\sigma} \left[4(n+\sigma)(n+\sigma-1) + 2\left(1-x^2\right)(n+\sigma) - x^2 \right] = 0 \tag{11.5}$$

We will equate coefficients of $x^{n+\sigma}$ for $n \geq 2$, since here all terms will make some contribution to the coefficient.

$$a_n \left[4(n+\sigma)(n+\sigma-1) + 2(n+\sigma) \right] + a_{n-2} \left[-2(n-2+\sigma) - 1 \right] = 0$$

Therefore,

$$2(n+\sigma)(2n+2\sigma-1)a_n = (2n+2\sigma-3)a_{n-2} \tag{11.6}$$

This is the recurrence relation, which we can use to compute the a_n. A general technique to find σ is to equate the coefficients of the lowest power of x in (11.5). By setting $n = 0$, we can equate coefficients of x^σ, giving

$$a_0(4\sigma(\sigma-1)) + a_0 2\sigma = 0$$

But since $a_0 \neq 0$ in Fuch's Theorem, we have

$$4\sigma(\sigma-1) + 2\sigma = 0$$

So either $\sigma = 0$ or $\sigma = \frac{1}{2}$. We must consider these two cases individually.

- ($\sigma = 0$) Equate coefficients of the lowest powers of x in (11.5).
  - ($n = 0$) The coefficient of x^0 gives
    $$a_0[4(0)(-1)] + a_0[2(0)] = 0$$
    which is true for all a_0. So a_0 is an arbitrary constant.
  - ($n = 1$) The coefficient of x^1 gives
    $$a_1[4(1)(0)] + a_1[2(1)] = 0$$
    so $a_1 = 0$.

  From the recurrence relation (11.6) which is valid for $n \geq 2$, plugging in $\sigma = 0$ gives
  $$2n(2n-1)a_n = (2n-3)a_{n-2} \tag{11.7}$$
  Since $a_1 = 0$, clearly all $a_k = 0$ for odd k. Therefore, using the recurrence relation (11.7) we have
  $$y = a_0\left(1 + \frac{x^2}{4\cdot 3} + \frac{5x^4}{8\cdot 7\cdot 4\cdot 3} + \ldots\right)$$

- ($\sigma = \frac{1}{2}$) This time we will start with the recurrence relation (11.6) with $\sigma = \frac{1}{2}$, relabelling a to b to avoid confusion.
  $$(2n+1)(2n)b_n = (2n-2)b_{n-2} \tag{11.8}$$
  Now let us analyse the coefficients of the lowest powers of x, substituting into (11.5).
  - ($n = 0$) The coefficient of $x^{\frac{1}{2}}$ gives
    $$b_0\left[4\left(\frac{1}{2}\right)\left(\frac{-1}{2}\right)\right] + b_0\left[2\left(\frac{1}{2}\right)\right] = 0$$
    which is true for all b_0. So b_0 is an arbitrary constant.
  - ($n = 1$) The coefficient of $x^{\frac{3}{2}}$ gives
    $$b_1\left[4\left(\frac{3}{2}\right)\left(\frac{1}{2}\right)\right] + b_1\left[2\left(\frac{3}{2}\right)\right] = 0$$
    so $b_1 = 0$.

  As before, all $b_k = 0$ where k is odd. Therefore, using the recurrence relation (11.8), we have
  $$y = b_0 x^{\frac{1}{2}}\left[1 + \frac{x^2}{2\cdot 5} + \frac{3x^4}{2\cdot 5\cdot 4\cdot 9} + \ldots\right]$$

So we have found two linearly independent solutions to the differential equation, given by boundary conditions a_0 and b_0. Note that Fuch's Theorem only specifies that there will be at least one, but we have found two in this case.

11.5. Special cases of indicial equation

Before looking at some examples of the method of Frobenius, we will first look at special cases of the indicial equation provided by Fuch's theorem. Consider an expansion about the point $x = x_0$. Let σ_1, σ_2 be the roots of this equation. There are two cases:

- ($\sigma_1 - \sigma_2 \notin \mathbb{Z}$) We have two linearly independent solutions. So our solution is of the form
$$y = (x - x_0)^{\sigma_1} \sum_{n=0}^{\infty} a_n (x - x_0)^n + (x - x_0)^{\sigma_2} \sum_{n=0}^{\infty} b_n (x - x_0)^n$$
Note that the limit as $x \to x_0$, $y \sim (x - x_0)^{\min(\sigma_1, \sigma_2)}$.
- ($\sigma_1 - \sigma_2 \in \mathbb{Z}$) There is one solution of the form
$$y_1 = (x - x_0)^{\sigma_2} \sum_{n=0}^{\infty} a_n (x - x_0)^n$$
The other solution is of the form
$$y_2 = (x - x_0)^{\sigma_1} \sum_{n=0}^{\infty} b_n (x - x_0)^n + c y_1 \ln(x - x_0)$$
where c may or may not equal zero. If the two solutions are linearly independent without the c term, then $c = 0$. Else, without loss of generality, we can let $c = 1$ since we're dealing with homogeneous equations.
- ($\sigma_1 = \sigma_2 = \sigma$) Here, $c \neq 0$. So our solutions are of the form
$$y_1 = (x - x_0)^{\sigma} \sum_{n=0}^{\infty} a_n (x - x_0)^n$$
$$y_2 = (x - x_0)^{\sigma} \sum_{n=0}^{\infty} b_n (x - x_0)^n + y_1 \ln(x - x_0)$$

Example. Let us solve the equation
$$x^2 y'' - xy = 0 \tag{11.9}$$
where we want series solutions about $x = 0$. Note that this is a regular singular point. We will try solutions of the form
$$y = \sum_{n=0}^{\infty} a_n x^{n+\sigma}$$
Therefore, we have
$$\sum_{n=0}^{\infty} a_n x^{n+\sigma} [(n + \sigma)(n + \sigma - 1) - x] = 0 \tag{11.10}$$

Equating coefficients of $x^{n+\sigma}$ for $n \geq 1$:

$$(n+\sigma)(n+\sigma-1)a_n = a_{n-1} \tag{11.11}$$

By equating the coefficients of the lowest powers of x (here $n = 0$, so we equate coefficients of x^σ), we get an indicial equation for σ:

$$\sigma(\sigma-1)a_0 = 0$$

So either $\sigma = 0$ or $\sigma = 1$, since $a_0 \neq 0$. So the values of σ differ by an integer.

- $(\sigma = 1)$ (11.11) implies that

  $$a_n = \frac{a_{n-1}}{n(n-1)} = \frac{a_0}{(n+1)(n!)^2}$$

  So we have

  $$y_1 = a_0 x\left(1 + \frac{x}{2} + \frac{x^2}{12} + \frac{x^3}{144} + \ldots\right)$$

- $(\sigma = 0)$ (11.11) now gives

  $$n(n-1)b_n = b_{n-1}$$

  Normally we could find b_1 in terms of b_0 using this relation, but this just reduces to $0b_1 = 0$, so we can't deduce it here. When $n = 1$, we can equate coefficients of x in (11.10) (relabelling a to b) to get

  $$b_1(1)(1-1) = 0$$

  So b_1 is arbitrary. Then of course we can find b_2 and so on in terms of smaller b_i values. It turns out that

  $$b_i = a_{i-1}$$

  And therefore $y_2(x)$ is linearly dependent on the previous $y_1(x)$. So we now need to use that logarithmic term to achieve linear independence, so y here is of the form

  $$y_2 = y_1 \ln x + \sum_{x=0}^{\infty} b_n x^n$$

  Why do we have specifically a logarithmic term? We can try the reduction of order method to find the other solution given the existence of y_1. Let $y_2(x) = v(x)y_1(x)$ for some function v. Then we have

  $$x^2(v''y_1 + 2v'y_1') = 0$$

Let $u = v'$, then

$$\begin{aligned} u'y_1 + 2uy_1 &= 0 \\ \frac{u'}{u} &= -2\frac{y_1'}{y_1} \\ \ln u &= \ln\left(y_1^{-2}\right) + \ln B \\ u = v' &= \frac{B}{y_1^2} \\ v' &= \frac{B}{a_0^2 x^2}\left(1 + \frac{x}{2} + \frac{x^2}{12} + \frac{x^3}{144} + \dots\right)^{-2} \end{aligned}$$

Note that the constant of integration gives a constant multiple of y_1, and since the equation is homogeneous the constant does not matter. We will expand this now using the binomial theorem, continually redefining constants since they are arbitrary, to give

$$v' = \frac{B}{a_0^2}\left(\frac{1}{x^2} - \frac{1}{x} + \sum_{n=0}^{\infty} B_n x^n\right)$$

for some constants B_n. Then integrating with respect to x,

$$\begin{aligned} v &= \frac{-B}{a_0^2}\frac{1}{x} - \frac{B}{a_0^2}\ln x + \sum_{n=1}^{\infty} C_n x^n \\ y_2 = vy_1 &= \frac{-B}{a_0} - \frac{B}{2a_0}x + \sum_{n=2}^{\infty} D_n x^n + Cy_1 \ln x \\ &= \sum_{n=0}^{\infty} b_n x^n + cy_1 \ln x \end{aligned}$$

So the appearance of $\ln x$ is natural here.

Example. Let us revisit (11.2).

$$(1 - x^2)y'' - 2xy' + 2y = 0$$

Instead of expanding around $x = 0$, let us now consider expanding around $x = -1$, a singular point. We will redefine the independent variable, let

$$z = 1 + x \implies z(2 - z)y'' - 2(z - 1)y' + 2y = 0$$

Now we will expand around $z = 0$. We know that $z = 0$ is a regular singular point, so we will try solutions of the form

$$y = \sum_{n=0}^{\infty} a_n z^{n+\sigma}; \quad a_0 \neq 0$$

We have

$$\sum_{n=0}^{\infty} a_n z^{n+\sigma-1} \left[(n+\sigma)(n+\sigma-1)(2-z) - 2(n+\sigma)(z-1) + 2z\right] = 0$$

As before, we will equate the coefficients of the lowest power of z (for $n = 0$, these are the coefficients of $z^{\sigma-1}$) to get the indicial equation and recursion relation.

$$2\sigma(\sigma-1)a_0 + 2\sigma a_0 = 0 \implies \sigma^2 = 0$$

So $\sigma = 0$ is a repeated root. Note that we need a term of the form $y_1 \ln(x - x_0)$ in this problem. We will not complete this example here.

12. Multivariate calculus

12.1. Gradient vector

Consider a function $f(x, y)$, and some small displacement $\mathrm{d}\mathbf{s}$. We want to find the rate of change of f in this direction. Recall that the multivariate chain rule tells us that a change in f, given a change in x and y, is given by

$$\begin{aligned}\mathrm{d}f &= \frac{\partial f}{\partial x}\,\mathrm{d}x + \frac{\partial f}{\partial y}\,\mathrm{d}y \\ &= (\mathrm{d}x\,, \mathrm{d}y) \cdot \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) \\ &= \mathrm{d}\mathbf{s} \cdot \nabla f\end{aligned}$$

where $\mathrm{d}\mathbf{s} = (\mathrm{d}x\,, \mathrm{d}y)$; $\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right)$. We call ∇f the 'gradient vector', in this case in Cartesian coordinates. If we let $\mathrm{d}\mathbf{s} = \mathrm{d}s\,\hat{\mathbf{s}}$ where $|\hat{\mathbf{s}}| = 1$, then we can write

$$\mathrm{d}f = \mathrm{d}s\,(\hat{\mathbf{s}} \cdot \nabla f)$$

We define the directional derivative by

$$\frac{\mathrm{d}f}{\mathrm{d}s} = \hat{\mathbf{s}} \cdot \nabla f$$

This is the rate of change of f in the direction given by $\hat{\mathbf{s}}$.

(i) The magnitude of the gradient vector ∇f is the maximum rate of change of $f(x, y)$.

$$|\nabla f| = \max_{\forall\theta}\left(\frac{\mathrm{d}f}{\mathrm{d}s}\right)$$

(ii) The direction of ∇f is the direction in which f increases most rapidly.

$$\left|\frac{\mathrm{d}f}{\mathrm{d}s}\right| = |\nabla f| \cos\theta$$

where θ is the angle between ∇f and $\hat{\mathbf{s}}$, which follows from the definition of the directional derivative.

(iii) If $\mathrm{d}\mathbf{s}$ (and $\hat{\mathbf{s}}$) are parallel to contours of f, then

$$\frac{\mathrm{d}f}{\mathrm{d}s} = \hat{\mathbf{s}} \cdot \nabla f = 0$$

Hence the gradient vector is perpendicular to contours of f, and $|\nabla f|$ is the slope in the 'uphill' direction.

12.2. Stationary points

In general, there is always at least one direction in which the directional derivative is zero, since we can just choose a direction perpendicular to the gradient vector, or equivalently parallel to contours of f. At stationary points, $\frac{\mathrm{d}f}{\mathrm{d}s} = 0$ for all directions, so $\nabla f = \mathbf{0}$. Stationary points may have multiple types:

- Minimum points, where the function is a minimum point in both directions;
- Maximum points, where the function is a maximum point in both directions; and
- Saddle points, where the function is a minimum point in one direction but a maximum point in another direction.

Note:

- Near minima and maxima, the contours of f are elliptical.
- Near a saddle, the contours of f are hyperbolic.
- Contours of f can only cross at saddle points.

12.3. Taylor series for multivariate functions

Let us expand a function $f(x, y)$ around a point $\mathbf{s}_0$, and evaluate it at some point $\mathbf{s}_0 + \delta\mathbf{s}$, where $\delta\mathbf{s} = \delta s\hat{\mathbf{s}}$. The Taylor series expansion in the direction of $\hat{\mathbf{s}}$ is

$$f(s_0 + \delta s) = f(s_0) + \delta s \left.\frac{\mathrm{d}f}{\mathrm{d}s}\right|_{s_0} + \frac{1}{2}(\delta s)^2 \left.\frac{\mathrm{d}^2 f}{\mathrm{d}s^2}\right|_{s_0} + \ldots$$

Further, by the definition of the directional derivative,

$$\frac{\mathrm{d}}{\mathrm{d}s} = \hat{\mathbf{s}} \cdot \nabla$$

Hence

$$\delta s \frac{\mathrm{d}}{\mathrm{d}s} = \delta\mathbf{s} \cdot \nabla$$

Now we can rewrite this Taylor series as follows:

$$f(s_0 + \delta_s) = f(s_0) + (\delta s)(\hat{\mathbf{s}} \cdot \nabla)\, f\Big|_{s_0} + \frac{1}{2}(\delta s)^2(\hat{\mathbf{s}} \cdot \nabla)(\hat{\mathbf{s}} \cdot \nabla)\, f\Big|_{s_0} + \ldots$$

$$f(s_0 + \delta_s) = f(s_0) + \underbrace{(\delta\mathbf{s} \cdot \nabla)\, f\Big|_{s_0}}_{(1)} + \underbrace{\frac{1}{2}(\delta\mathbf{s} \cdot \nabla)(\delta\mathbf{s} \cdot \nabla)\, f\Big|_{s_0}}_{(2)} + \ldots$$

Expressing this in Cartesian coordinates:

$$\mathbf{s}_0 = (x_0, y_0); \quad \delta\mathbf{s} = (\delta x, \delta y); \quad x = x_0 + \delta x; \quad y = y_0 + \delta_y$$

Therefore,

$$(1) = \delta x \frac{\partial f}{\partial x}(x_0, y_0) + \delta y \frac{\partial f}{\partial y}(x_0, y_0)$$

$$\begin{aligned}(2) &= \frac{1}{2}\left(\delta x \frac{\partial}{\partial x} + \delta y + \frac{\partial}{\partial y}\right)\left(\delta x \frac{\partial}{\partial x} + \delta y + \frac{\partial}{\partial y}\right) f\Big|_{x_0,y_0} \\ &= \frac{1}{2}\left(\delta x^2 f_{xx} + \delta x \delta y f_{yx} + \delta y \delta x f_{xy} + \delta y^2 f_{yy}\right)\Big|_{x_0,y_0} \\ &= \frac{1}{2}\begin{pmatrix}\delta x & \delta y\end{pmatrix}\begin{pmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy}\end{pmatrix}\Big|_{x_0,y_0}\begin{pmatrix}\delta x \\ \delta y\end{pmatrix}\end{aligned}$$

The matrix

$$H = \begin{pmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy}\end{pmatrix} = \nabla(\nabla f)$$

as used in the second derivative above, is called the Hessian matrix.

Putting this together, in 2D Cartesian Coordinates, we have

$$\begin{aligned}f(x,y) = f(x_0,y_0) &+ (x-x_0)\, f_x\Big|_{x_0,y_0} + (y-y_0)\, f_y\Big|_{x_0,y_0} \\ &+ \frac{1}{2}\left[(x-x_0)^2\, f_{xx}\Big|_{x_0,y_0} + (y-y_0)^2\, f_{yy}\Big|_{x_0,y_0} + 2(x-x_0)(y-y_0)\, f_{xy}\Big|_{x_0,y_0}\right] + \ldots\end{aligned}$$

And in the general coordinate-independent form:

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \delta\mathbf{x} \cdot \nabla f(\mathbf{x}_0) + \frac{1}{2}\delta\mathbf{x} \cdot [\nabla(\nabla f)]\Big|_{x_0} \cdot \delta\mathbf{x}^\intercal + \ldots$$

12.4. Classifying stationary points

Since $\nabla f = \mathbf{0}$ defines a stationary point, the Taylor series expansion around a stationary point $\mathbf{x} = \mathbf{x}_s$ is

$$f(\mathbf{x}) \approx f(\mathbf{x}_s) + \frac{1}{2}\delta\mathbf{x} \cdot H\Big|_{\mathbf{x}_s} \cdot \delta\mathbf{x}^\intercal$$

So the nature of the stationary point depends on the Hessian matrix H. Consider a function in n-dimensional space

$$f = f(x_1, x_2, \ldots, x_n)$$

Then the n-dimensional Hessian matrix is given by

$$H = \begin{pmatrix} f_{x_1x_1} & f_{x_1x_2} & \cdots & f_{x_1x_n} \\ f_{x_2x_1} & f_{x_2x_2} & \cdots & f_{x_2x_n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{x_nx_1} & f_{x_nx_2} & \cdots & f_{x_nx_n}\end{pmatrix}$$

If all of these derivatives are defined, $f_{x_1x_2} = f_{x_2x_1}$ etc, so $H = H^\intercal$, i.e. H is symmetric, and therefore it can be diagonalised with respect to its principal axes.

$$\delta\mathbf{x} \cdot H \cdot \delta\mathbf{x}^\intercal = \begin{pmatrix} \delta x_1 & \delta x_2 & \cdots & \delta x_n \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} \begin{pmatrix} \delta x_1 \\ \delta x_2 \\ \cdots \\ \delta x_n \end{pmatrix}$$

where the λ_i are eigenvalues of H and the δx_i is the displacement along the principal axis (eigenvector) i. Therefore

$$\delta\mathbf{x} \cdot H \cdot \delta\mathbf{x}^\intercal = \lambda_1 \delta x_1^2 + \lambda_2 \delta x_2^2 + \cdots + \lambda_n \delta x_n^2$$

(i) At a minimum point, $\delta\mathbf{x} \cdot H \cdot \delta\mathbf{x}^\intercal > 0$ for any $\delta\mathbf{x}$ (moving in any direction, we go 'downhill'). So all the $\lambda_i > 0$. So H is positive definite.

(ii) At a maximum point, $\delta\mathbf{x} \cdot H \cdot \delta\mathbf{x}^\intercal < 0$ for any $\delta\mathbf{x}$. So all the $\lambda_i < 0$. H is negative definite.

(iii) At a saddle point, H is indefinite.

12.5. Signature of Hessian

Definition. The signature of H is the pattern of the signs of its subdeterminants.

For a function $f(x_1, x_2, \dots, x_n)$, we want the signs of

$$\underbrace{|f_{x_1x_1}|}_{|H_1|}, \underbrace{\begin{vmatrix} f_{x_1x_1} & f_{x_1x_2} \\ f_{x_2x_1} & f_{x_2x_2} \end{vmatrix}}_{|H_2|}, \dots, \underbrace{\begin{vmatrix} f_{x_1x_1} & f_{x_1x_2} & \cdots & f_{x_1x_n} \\ f_{x_2x_1} & f_{x_2x_2} & \cdots & f_{x_2x_n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{x_nx_1} & f_{x_nx_2} & \cdots & f_{x_nx_n} \end{vmatrix}}_{|H_1|}$$

We know from Vectors and Matrices that if a symmetric matrix H is positive (or negative) definite, then $H_1, H_2, \dots, H_{n-1}$ are positive (or negative) definite. This is known as Sylvester's Criterion. In other words, a minimum (or maximum) point in n-dimensional space is also a minimum (or maximum) in any subspace containing this point. Now let us list the signs of subdeterminants to see the types of signatures.

(i) At a minimum point ($\lambda_i > 0$), the signature is $+, +, +, +, \dots$

(ii) At a maximum point ($\lambda_i < 0$), the signature is $-, +, -, +, \dots$

If $|H| = 0$, we need higher order terms in the Taylor series.

12.6. Contours near stationary points

Consider a coordinate system aligned with the principal axes of the Hessian H in two-dimensional space, so

$$H = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

Let $\delta\mathbf{x} = (\mathbf{x} - \mathbf{x}_s) = (\xi, \eta)$ where $\mathbf{x}_s$ is the stationary point we're considering. In a small region near $\mathbf{x}_s$, the contours of f satisfy

$$f = \text{constant (since } f \text{ is a contour)} \approx f(\mathbf{x}_s) = \frac{1}{2}\delta\mathbf{x} \cdot H \cdot \delta\mathbf{x}^\intercal$$

$$\therefore \lambda_1\xi^2 + \lambda_2\eta^2 \approx \text{constant} \tag{12.1}$$

Near a minimum or maximum point, λ_1 and λ_2 have the same sign. (12.1) implies that the contours of f are elliptical. Near a saddle point, λ_1 and λ_2 have opposite sign so (12.1) shows that the contours of f are hyperbolic. As an example, let us consider

$$f(x, y) = 4x^3 - 12xy + y^2 + 10y + 6$$

Let us first identify the stationary points.

$$f_x = f_y = 0$$

After solving this, we get

$$(x, y) = (1, 1), (5, 25)$$

To get the Hessian matrix:

$$\begin{aligned} f_{xx} &= 24x \\ f_{xy} = f_{yx} &= -12 \\ f_{yy} &= 2 \end{aligned}$$

Now considering the stationary points separately:

- $(1, 1)$:

  $$H = \begin{pmatrix} 24 & -12 \\ -12 & 2 \end{pmatrix} \implies |H_1| = 24; \quad |H| = 48 - 144$$

  The signature is $+, -$, so this is a saddle point.

- $(5, 25)$:

  $$H = \begin{pmatrix} 120 & -12 \\ -12 & 2 \end{pmatrix} \implies |H_1| = 120; \quad |H| = 240 - 144$$

  The signature is $+, +$, so this is a minimum point.

13. Systems of ODEs

13.1. Systems of linear ODEs

Consider two functions $y_1(t), y_2(t)$ which satisfy

$$\dot{y}_1 = ay_1 + by_2 + f_1(t)$$
$$\dot{y}_2 = cy_1 + dy_2 + f_2(t)$$

This is a set of coupled differential equations which we must solve simultaneously. In vector form,

$$\dot{\mathbf{Y}} = M\mathbf{Y} + \mathbf{F}$$

where

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}; \quad M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}; \quad \mathbf{F} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$$

Any nth order differential equation can be written as a system of n first order ODEs. For example, the standard form for a second order linear ODE is

$$\ddot{y} + a\dot{y} + by = f$$

Let $y_1 = y, y_2 = \dot{y}$. Then

$$\mathbf{Y} = \begin{pmatrix} y \\ \dot{y} \end{pmatrix}$$

Hence our two equations are

$$\dot{y}_1 = y_2$$
$$\dot{y}_2 + ay_2 + by_1 = f \implies \dot{y}_2 = -ay_2 - by_1 + f$$

And in vector form,

$$\dot{\mathbf{Y}} = \begin{pmatrix} 0 & 1 \\ -b & -a \end{pmatrix}\mathbf{Y} + \begin{pmatrix} 0 \\ f \end{pmatrix}$$

13.2. Matrix methods

To solve a system of n first-order linear ODEs,

$$\dot{\mathbf{Y}} = M\mathbf{Y} + \mathbf{F} \tag{13.1}$$

we need the following steps.

(i) Write $\mathbf{Y} = \mathbf{Y}_c + \mathbf{Y}_p$ where $\mathbf{Y}_c$ satisfies the homogeneous version of (13.1):

$$\dot{\mathbf{Y}}_c = M\mathbf{Y}_c \tag{13.2}$$

(ii) Seek solutions of $\mathbf{Y}_c$ in the form $\mathbf{v}e^{\lambda t}$.

$$(13.2) \implies \lambda\mathbf{v} = M\mathbf{v}$$

So the vectors $\mathbf{v}$ are the eigenvectors, with eigenvalues λ.

(iii) Find $\mathbf{Y}_p$ based on the form of $\mathbf{F}$.

As a quick example, let us consider

$$\dot{\mathbf{Y}} - \begin{pmatrix} -4 & 24 \\ 1 & -2 \end{pmatrix} \mathbf{Y} = \begin{pmatrix} 4 \\ 1 \end{pmatrix} e^t \tag{13.3}$$

We will try $\mathbf{Y}_c = \mathbf{v}e^{\lambda t}$, and we can compute that the eigenvalues and eigenvectors are

$$\lambda_1 = 2, \mathbf{v}_1 = \begin{pmatrix} 4 \\ 1 \end{pmatrix}; \quad \lambda_2 = -8, \mathbf{v}_2 = \begin{pmatrix} -6 \\ 1 \end{pmatrix}$$

Hence,

$$\mathbf{Y}_c = A \begin{pmatrix} 4 \\ 1 \end{pmatrix} e^{2t} + B \begin{pmatrix} -6 \\ 1 \end{pmatrix} e^{-8t}$$

Now, to solve the particular integral, we will try a $\mathbf{Y}_p$ of the form

$$\mathbf{Y}_p = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} e^t$$

We have

$$\begin{aligned} (13.3) \implies \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} - \begin{pmatrix} -4 & 24 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} &= \begin{pmatrix} 4 \\ 1 \end{pmatrix} \\ I \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} - \begin{pmatrix} -4 & 24 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} &= \begin{pmatrix} 4 \\ 1 \end{pmatrix} \\ \begin{pmatrix} 5 & -24 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} &= \begin{pmatrix} 4 \\ 1 \end{pmatrix} \\ \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} &= \begin{pmatrix} -4 \\ -1 \end{pmatrix} \end{aligned}$$

So the general solution is

$$\mathbf{Y} = A \begin{pmatrix} 4 \\ 1 \end{pmatrix} e^{2t} + B \begin{pmatrix} -6 \\ 1 \end{pmatrix} e^{-8t} - \begin{pmatrix} 4 \\ 1 \end{pmatrix} e^t$$

If the forcing term matches one of the complementary functions, we would try $\mathbf{Y}_p = \mathbf{u}te^{\lambda t}$.

13.3. Decoupling ODEs

From a linear system of n first order ODEs, we can construct n uncoupled nth order ODEs. For the above example (13.3),

$$\dot{y}_1 = -4y_1 + 24y_2 + 4e^t \tag{13.4}$$

$$\dot{y}_2 = y_1 - 2y_2 + e^t \tag{13.5}$$

We will create a linear equation for $\ddot{y}$. First, we will take the derivative of (13.4).

$$\ddot{y}_1 = -4\dot{y}_1 + 24\dot{y}_2 + 4e^t$$

Now we will substitute in (13.5) for $\dot{y}_2$.

$$\ddot{y}_1 = -4\dot{y}_1 + 24(y_1 - 2y_2 + e^t) + 4e^t$$

Now we can substitute back in the original equation (13.4) to remove the y_2 term.

$$\ddot{y}_1 = -4\dot{y}_1 + 24\left(y_1 - \frac{1}{12}(\dot{y}_1 + 4y_1 - 4e^t) + e^t\right) + 4e^t$$

$$\ddot{y}_1 = -4\dot{y}_1 + 24y_1 - 2\dot{y}_1 - 8y_1 + 8e^t + 28e^t$$

$$\ddot{y}_1 + 6\dot{y}_1 - 16y_1 = 36e^t$$

which we can solve as normal. The general solution matches the first component of the general solution vector from above. We can of course construct an analogous equation for y_2, which would match the second component of the solution vector.

13.4. Phase portraits

For some complementary function $\mathbf{Y}_c$ (or equivalently, a solution to a homogeneous system of linear first order ODEs) satisfying

$$\dot{\mathbf{Y}}_c = M\mathbf{Y}_c \tag{13.6}$$

Therefore,

$$\mathbf{Y}_c = A\mathbf{v}_1 e^{\lambda_1 t} + B\mathbf{v}_2 e^{\lambda_2 t}$$

Let us consider three cases.

(i) λ_1, λ_2 real and opposite sign. Without loss of generality, let $\lambda_1 > 0$. The origin here is known as a 'saddle node' as the solution curves for $A = 0$ and $B = 0$ cross at the origin.

(ii) λ_1, λ_2 real and same sign. let us say that without loss of generality that $|\lambda_1| > |\lambda_2|$. If they are both negative, Here the origin is a stable node as all solution curves tend towards it. If λ_1, λ_2 are both positive, The origin here is an unstable node as the curves tend away from it.

(iii) λ_1, λ_2 form a complex conjugate pair. If the real parts are negative, This is a stable spiral; the curves tend towards zero. If the real parts are positive we have an unstable spiral. If the real part is zero, the solution curves are circles. This is known as a centre.

Note that to find the direction of rotations in these phase portraits, we would need to evaluate the system of equations at a given point to find the signs of the derivatives $\dot{y}_1, \dot{y}_2$.

13.5. Nonlinear systems of ODEs

Consider an autonomous system of two nonlinear first order ODEs:

$$\dot{x} = f(x, y) \tag{13.7}$$
$$\dot{y} = g(x, y) \tag{13.8}$$

where 'nonlinear' means that f and g are nonlinear functions of x and y, and where 'autonomous' means that the independent variable t does not explicitly show up in these equations. We will consider the equilibrium points, or fixed points. Let (x_0, y_0) be a fixed point, i.e.

$$\dot{x} = \dot{y} = 0 \implies f(x_0, y_0) = g(x_0, y_0) = 0$$

at this point. We may want to understand the stability of such fixed points by perturbation analysis, like before. Let us consider a small perturbation away from the fixed point.

$$(x, y) = (x_0 + \xi(t), y_0 + \eta(t))$$

We have

$$(13.7) \implies \dot{\xi} = f(x_0 + \xi, y_0 + \eta)$$

We can expand this in a multivariate Taylor series, keeping the first three terms—the constant term and the two linear terms.

$$\begin{aligned}
\dot{\xi} &\approx f(x_0, y_0) + \xi f_x(x_0, y_0) + \eta f_y(x_0, y_0) \\
&= \xi f_x(x_0, y_0) + \eta f_y(x_0, y_0) \\
\dot{\eta} &\approx g(x_0, y_0) + \xi g_x(x_0, y_0) + \eta g_y(x_0, y_0) \\
&= \xi g_x(x_0, y_0) + \eta g_y(x_0, y_0)
\end{aligned}$$

Hence,

$$\begin{pmatrix} \dot{\xi} \\ \dot{\eta} \end{pmatrix} = \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix}\Bigg|_{x_0, y_0} \begin{pmatrix} \xi \\ \eta \end{pmatrix}$$

This is a homogeneous linear system of ODEs. The eigenvalues of M, which we will call λ_1, λ_2, determine the stability and behaviour. This is just the same as the phase portrait analysis above to determine whether the perturbed point tends to the origin or not, and exactly how this movement happens.

13.6. Lotka–Volterra equations

This is a worked example of a coupled set of differential equations, which model a predator-prey system. Let the quantity of prey be represented by x, and the quantity of the predator be y. Then

$$\dot{x} = \alpha x - \beta xy = f(x, y)$$
$$\dot{y} = \delta xy - \gamma y = g(x, y)$$

where $\alpha, \beta, \gamma, \delta$ are positive real constants. We will start by analysing the fixed points, where $\dot{x} = \dot{y} = 0$.

$$\dot{x} = 0 \implies x = 0 \text{ or } y = \frac{\alpha}{\beta}$$

$$\dot{y} = 0 \implies y = 0 \text{ or } x = \frac{\gamma}{\delta}$$

Therefore,

$$(x_0, y_0) = (0,0), \left(\frac{\gamma}{\delta}, \frac{\alpha}{\beta}\right)$$

Using matrix methods,

$$M = \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix} = \begin{pmatrix} \alpha - \beta y & -\beta x \\ \delta y & \delta x - \gamma \end{pmatrix}$$

Now we can analyse the stability of these fixed points by perturbation analysis.

- At the fixed point $(0,0)$, we have

$$\begin{pmatrix} \dot{\xi} \\ \dot{\eta} \end{pmatrix} = \begin{pmatrix} \alpha & 0 \\ 0 & -\gamma \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix}$$

  We can read off the eigenvalues to be α and $-\gamma$. This is a saddle node, since one direction will increase (x) and one will decrease (y).

- At the fixed point $\left(\frac{\gamma}{\delta}, \frac{\alpha}{\beta}\right)$, we have

$$\begin{pmatrix} \dot{\xi} \\ \dot{\eta} \end{pmatrix} = \begin{pmatrix} 0 & -\beta\frac{\gamma}{\delta} \\ \delta\frac{\alpha}{\beta} & 0 \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix}$$

  The characteristic equation is $\chi_M(\lambda) = \lambda^2 + \alpha\lambda = 0$, so $\lambda = \pm\sqrt{-\alpha\gamma}$. Since $\alpha\gamma > 0$, it is more convenient to write $\lambda = \pm i\sqrt{\alpha\gamma}$. Since the real part is zero, this gives a centre node. To work out the direction of rotation, let us consider the x direction,

$$\dot{\xi} = -\beta\frac{\gamma}{\delta}\eta$$

  If $\eta > 0$, then $\dot{\xi} < 0$, so we have anticlockwise rotation.

Now we can sketch a graph taking into account both of these fixed points, visually interpolating the values between them.

13.7. First order wave equation and method of characteristics

We will define a partial differential equation to be a differential equation with multiple independent variables. Here, we will consider three examples, starting with the first order wave equation. Consider a function $y(x,t)$ where

$$\frac{\partial y}{\partial t} - c\frac{\partial y}{\partial x} = 0 \tag{13.9}$$

where c is a constant. We will solve this equation with the method of characteristics. Imagine moving a 'probe' along a path $x(t)$. Then y is a function $y(x(t), t)$, where now the only independent variable is t. Using the multivariate chain rule,

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \frac{\partial y}{\partial t} + \frac{\partial y}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t}$$

Comparing this with (13.9), we note that if $\frac{\mathrm{d}x}{\mathrm{d}t} = -c$, then $\frac{\mathrm{d}y}{\mathrm{d}t} = 0$. So we have found a path along which the 'probe' is at a constant height, i.e. along $x(t) = x_0 - ct$, y is a constant. We can update our graph now showing the 'characteristics' we have just shown to exist. If $y(x, t = 0) = f(x)$, then $y = f(x_0)$ along the characteristics. Hence, our general solution is

$$y = f(x + ct)$$

Let us consider some examples of wave equations f.

(i) (unforced wave equation) Let $y(x, 0) = x^2 - 3$ in (13.9). Then

$$y(x, t) = (x + ct)^2 - 3$$

(ii) (forced wave equation) Let

$$\frac{\partial y}{\partial t} + 5\frac{\partial y}{\partial x} = e^{-t}$$

and

$$y(x, 0) = e^{-x^2}$$

Then along paths with $\frac{\mathrm{d}x}{\mathrm{d}t} = 5$ or $x = x_0 + 5t$,

$$\frac{\mathrm{d}y}{\mathrm{d}t} = e^{-t}$$

So by integration,

$$y = A - e^{-t}$$

along these paths. Applying our initial condition at $t = 0$, our 'probe' is at x_0 and $y(x, 0) = A - 1 = e^{-x_0^2}$. Hence, $A = 1 + e^{-x_0^2}$. So

$$y = 1 + e^{-x_0^2} - e^{-t}$$

along the path given by x_0. Substituting back for a general x, we can create a formula for the general solution of y (not necessarily on a given path):

$$y = 1 + e^{-(x-5t)^2} - e^{-t}$$

14. More PDEs

14.1. Second order wave equation

This equation is typically known as just 'the wave equation', but here we are referring to it as the 'second order' wave equation to distinguish it from the first order equation found in the previous lecture.

$$\frac{\partial^2 y}{\partial t^t} - c^2 \frac{\partial^2 y}{\partial x^2} = 0 \tag{14.1}$$

We will factor out the differential operator:

$$\left(\frac{\partial}{\partial t} - c\frac{\partial}{\partial x}\right)\left(\frac{\partial}{\partial t} + c\frac{\partial}{\partial x}\right) y = 0$$

The two operators commute, hence we have either

$$\left(\frac{\partial}{\partial t} - c\frac{\partial}{\partial x}\right) y = 0; \quad \text{or} \quad \left(\frac{\partial}{\partial t} + c\frac{\partial}{\partial x}\right) y = 0$$

These are both instances of the first order wave equation (13.9).

$$y = f(x + ct); \quad y = g(x - ct)$$

Since (14.1) is linear in y, our general solution is the sum of these two solutions.

$$y = f(x + ct) + g(x - ct)$$

As an example, let us solve

$$y_{tt} - c^2 y_{xx} = 0$$

subject to

$$y = \frac{1}{1 + x^2}; \quad y_t = 0 \quad \text{at } t = 0$$

and further, $y \to 0$ as $x \to \pm\infty$. Our solution is of the form

$$y = f(x + ct) + g(x - ct)$$

We will use the initial conditions to find f, g.

$$f(x) + g(x) = \frac{1}{1 + x^2}$$

$$cf'(x) - cg'(x) = 0$$

The second equation shows that $f' = g'$, or $f = g + A$.

$$2g(x) + A = \frac{1}{1 + x^2}$$

$$g(x) = \frac{1}{2}\left(\frac{1}{1+x^2}\right) - \frac{A}{2}$$
$$f(x) = \frac{1}{2}\left(\frac{1}{1+x^2}\right) + \frac{A}{2}$$

Even though we have a constant of integration A here, since $y = f + g$ the constant vanishes in the general solution. So the constant does not affect the solution and it really is arbitrary. So without loss of generality here we can let $A = 0$. So our solution is

$$y(x,t) = \frac{1}{2}\left[\underbrace{\frac{1}{1+(x+ct)^2}}_{\text{moves left}} + \underbrace{\frac{1}{1+(x-ct)^2}}_{\text{moves right}}\right]$$

14.2. Derivation of diffusion equation

We will consider random walks to derive the diffusion equation. Imagine a particle located at position x at time t. After some change in time Δt, the particle may move to the left or to the right, i.e. $x + \Delta x$ ot $x - \Delta x$. Let $c(x,t)$ be the number of particles at x, t. After a discrete time interval Δt, let

- The probability of moving right one step is p;
- The probability of moving left one step is p; and
- The probability of staying at x is $1 - 2p$.

Considering a large amount of particles,

$$c(x, t+\Delta t) = (1-2p)c(x,t) + p\,(c(x+\Delta x, t) + c(x-\Delta x, t)) \tag{14.2}$$

We will now expand these terms as Taylor series through time and space, for small Δx and Δt. We'll put three terms in the expansion in space since the linear term will cancel when we combine the $+$ and $-$ terms.

$$c(x, t+\Delta t) = c(x,t) + \Delta t \frac{\partial c}{\partial t}(x,t) + O(\Delta t^2)$$

$$c(x \pm \Delta x, t) = c(x,t) \pm \Delta x \frac{\partial c}{\partial x}(x,t) + \frac{\Delta x^2}{2}\frac{\partial^2 c}{\partial x^2}(x,t) + O(\Delta x^3)$$

Now, substituting into (14.2), we have

$$c + \Delta t \frac{\partial c}{\partial t} + O(\Delta t^2) = (1-2p)c + p\left(2c + \Delta x^2 \frac{\partial^2 c}{\partial x^2} + O(\Delta x^3)\right)$$

$$\frac{\partial c}{\partial t} + O(\Delta t) = p\frac{\Delta x^2}{\Delta t}\frac{\partial^2 c}{\partial x^2} + O\left(\frac{\Delta x^3}{\Delta t}\right)$$

We will take the limit as $\Delta x, \Delta t \to 0$ such that $\frac{\Delta x^2}{\Delta t}$ is constant. This will make some things easier. Note that $\frac{\Delta x^3}{\Delta t} = \frac{\Delta x^2}{\Delta t} \cdot \Delta x \to 0$.

$$\frac{\partial x}{\partial t} = \kappa \frac{\partial^2 c}{\partial x^2}; \quad k \equiv \lim_{\Delta x, \Delta t \to 0} p \frac{\Delta x^2}{\Delta t}$$

This is the diffusion equation. Here, κ is the diffusion coefficient.

14.3. Solving the diffusion equation

For example, consider

$$\frac{\partial y}{\partial t} = \kappa \frac{\partial^2 y}{\partial x^2}$$

subject to the initial condition

$$y(x, 0) = \delta(x)$$

where $\delta(x)$ is the Dirac delta function, and where $y \to 0$ as $x \to \pm\infty$. We will convert this PDE into an ODE by constructing a similarity variable

$$\eta \equiv \frac{x^2}{4\kappa t}$$

This form of similarity variable can be motivated by observing units on both sides of the PDE, since κ must have units x^2/t to conserve dimensions. We will seek solutions of the form

$$y = t^{-\alpha} f(\eta)$$

where α, f are to be determined. We will now compute some derivatives:

$$\begin{aligned} y_t &= -\alpha t^{-\alpha-1} f + t^{-\alpha} f_\eta \eta_t \\ y_x &= t^{-\alpha} f_\eta \eta_x \\ y_{xx} &= t^{-\alpha} f_{\eta\eta} (\eta_x)^2 + t^{-\alpha} f_\eta \eta_{xx} \end{aligned}$$

Plugging these into the diffusion equation gives

$$\frac{-\alpha}{t} f + f' \eta_t = \kappa f'' (\eta_x)^2 + \kappa f' \eta_{xx} \tag{14.3}$$

where $f' = f_\eta, f'' = f_{\eta\eta}$.

$$\eta_t = \frac{-x^2}{4\kappa t^2} = \frac{-\eta}{t}$$

$$\eta_x = \frac{2x}{4\kappa t} \implies (\eta_x)^2 = \frac{4x^2}{16\kappa^2 t^2} = \frac{\eta}{\kappa t}$$

$$\eta_{xx} = \frac{2}{4\kappa t}$$

Plugging these results into (14.3) gives

$$\alpha f + f'\eta + f''\eta + \frac{f'}{2} = 0$$

$$\eta \frac{\mathrm{d}}{\mathrm{d}\eta}(f + f') + \frac{1}{2}(f' + 2\alpha f) = 0 \tag{14.4}$$

This is an ODE for $f(\eta)$. We have not yet defined what α is, and it is currently arbitrary, so we can let it be $\frac{1}{2}$ so that it cancels some terms.

$$(14.4) \implies \eta \frac{\mathrm{d}F}{\mathrm{d}\eta} + \frac{F}{2} = 0; \quad F \coloneqq f + f'$$

One solution is that $F = 0$ for all η. This is nontrivial because then $f + f' = 0$. So $f = Ae^{-\eta}$. Then

$$y = At^{-\frac{1}{2}} e^{-\frac{x^2}{4\kappa t}}$$

We can use the delta function initial condition to find A.

$$\delta(x) = \lim_{\varepsilon \to 0} \left[\frac{1}{\varepsilon\sqrt{\pi}} e^{-\frac{x^2}{\varepsilon^2}} \right]$$

So if we let $\varepsilon^2 = 4\kappa t$, then as $t \to 0$, we get $y(x) = \delta(x)$. So

$$\frac{1}{\varepsilon\sqrt{\pi}} = \frac{1}{\sqrt{4\pi\kappa}} t^{-\frac{1}{2}}$$

Hence,

$$A = \frac{1}{\sqrt{4\pi\kappa}}$$

Therefore we have

$$y(x,t) = \frac{1}{\sqrt{4\pi\kappa}} t^{-\frac{1}{2}} e^{-\frac{x^2}{4\kappa t}}$$

III. Groups

Lectured in Michaelmas 2020 by Dr. A. Khukhro

Many mathematical objects have lots of symmetry. To study symmetry in an abstract way, we define the notion of a group. Groups allow us to characterise all of the possible symmetries of an object, so understanding groups allows us to understand symmetry itself. Shapes, numbers, and matrices all give rise to their own groups, which provide insight into how the objects are structured.

As well as studying groups on their own, we also study the ways in which groups can interact. One example is a particular kind of function called a homomorphism, which preserves the structure of the groups in question. The homomorphisms between groups allow us to study each group in more detail.

Contents

1. Axiomatic definition 128
1.1. Intuition with geometry 128
1.2. Definition 128
1.3. Basic properties 129
1.4. Subgroups 130
1.5. Subgroups generated by a subset 132
2. Homomorphisms 134
2.1. Definition and elementary properties 134
2.2. Isomorphisms 135
2.3. Images and kernels 135
3. Types of groups 138
3.1. Direct products of groups 138
3.2. Cyclic groups 139
3.3. Dihedral groups 140
4. Permutation groups 142
4.1. Definition 142
4.2. Cycles 142
4.3. Disjoint cycle decomposition 143
4.4. Products of transpositions 144
5. Möbius transformations 146
5.1. The Möbius group 146
5.2. Properties of the Möbius group 147
6. Cosets and Lagrange's theorem 149
6.1. Cosets 149
6.2. Lagrange's theorem 150
6.3. Groups of small order 152
7. Normal subgroups and quotients 153
7.1. Normal subgroups 153
7.2. Motivation for quotients 155
7.3. Quotients 156
7.4. Examples and properties 157
8. Isomorphism theorems 159
8.1. First isomorphism theorem 159
8.2. Correspondence theorem 160
8.3. Second isomorphism theorem 160
8.4. Third isomorphism theorem 160
8.5. Simple groups 161

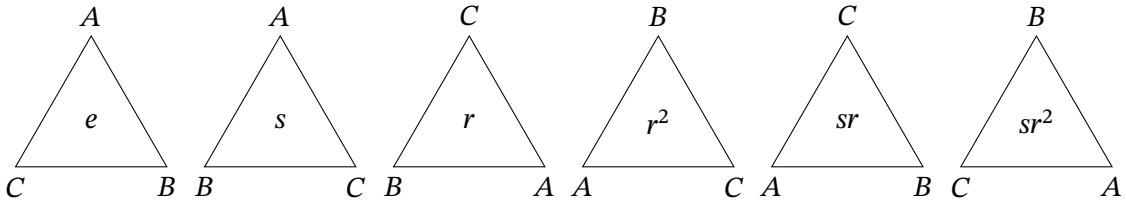
9. Group actions 162
9.1. Definition 162
9.2. Orbits and stabilisers 163
9.3. The Platonic solids 165
9.4. Cauchy's theorem 166
9.5. Left regular action 167
9.6. Cayley's theorem 167
10. Conjugation 169
10.1. Conjugation actions 169
10.2. Conjugation in symmetric groups 171
10.3. Conjugation in alternating groups 172
11. Action of the Möbius group 175
11.1. Introduction 175
11.2. Constructing Möbius maps 176
11.3. Geometric properties of Möbius maps 177
11.4. Cross-ratios 179
12. Matrix groups 181
12.1. Definitions 181
12.2. Matrix encoding of Möbius maps 181
12.3. Actions of matrices on vector spaces 182
12.4. Conjugation action of general linear group 183
12.5. Stabilisers of conjugation action 184
12.6. Geometry of orthogonal groups 185
12.7. Reflections in O_n 186
12.8. Classifying elements of O_2 186
12.9. Classifying elements of O_3 188
13. Groups of order 8 190
13.1. Quaternions 190
13.2. Elements of order 2 190
13.3. Classification of groups of order 8 190

1. Axiomatic definition

1.1. Intuition with geometry

Which is more symmetrical, a scalene triangle or an equilateral triangle? Clearly, the equilateral triangle has more symmetries; you can rotate it 120° or 240°, and you can reflect it across three axes. The scalene triangle has no symmetries that modify the object, but by convention we call the 'do-nothing' operation a symmetry as well.

By a 'symmetry' of an object, we mean something that we can do to it that preserves its structure. In the case of these shapes, we want to preserve the vertices and edges; these symmetries are rotations and reflections. For the equilateral triangle then, what are all the symmetries?



As stated before, we assign the letter e to the identity element. The operation s is a reflection; r is a rotation. By combining these elements, we get the set of elements of the group. Note that order matters: $sr \neq rs$.

1.2. Definition

Definition (Group)**.** A group is a set G together with a way of composing its elements $*$ satisfying $(\forall g, h, k \in G)$:

- (closure) $g * h \in G$
- (identity) $\exists e \in G$ s.t. $e * g = g * e = g$
- (inverses) $\exists g^{-1} \in G$ s.t. $g * g^{-1} = g^{-1} * g = e$
- (associativity) $g * (h * k) = (g * h) * k$

Formally, we might say that a set G with a binary operation $* : G \times G \to G$ is a group if it follows the last three axioms; the first rule is implicit in the function's type.

Here are a few examples of groups.

(i) $G = \{e\}$—this is the 'trivial group'.

(ii) $G = \{$symmetries of the equilateral triangle$\}$; $*$ is defined by: '$g*h$ means doing h then g'.

(iii) $G = (\mathbb{Z}, +)$. This is easy to prove by verifying the axioms.

(iv) $G = (\mathbb{R}, +); (\mathbb{Q}, +); (\mathbb{C}, +)$

(v) $G = (\mathbb{R}^*, \cdot)$ where $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$. Note that $(\mathbb{R}, \cdot)$ is not a group, because $\nexists 0^{-1} \in \mathbb{R}$ s.t. $0^{-1} \cdot 0 = 0 \cdot 0^{-1} = 1$.

(vi) $G = (\mathbb{R}, *)$ where $r * s \coloneqq r + s + 5$.

(vii) $G = (\mathbb{Z}_n, +)$ where $\mathbb{Z}_n = \{0, 1, 2, \cdots, n-1\}$ and addition is done modulo n.

(viii) A vector space with the operation of vector addition is a group.

(ix) $GL_2(\mathbb{R})$ is the set of invertible 2×2 matrices, which is a group with respect to matrix multiplication.

Here are a few non-examples.

(i) $G = (\mathbb{Z}_n, +)$ where addition is not performed modulo n. This group is not closed, e.g. $(n-1) + 2 \notin G$.

(ii) $G = (\mathbb{Z}, \cdot)$ because $\nexists n \in \mathbb{Z}$ s.t. $2 \cdot n = n \cdot 2 = 1$.

(iii) $G = (\mathbb{R}, *)$ where $r * s = r^2 s$ because there is no identity element.

(iv) $G = (\mathbb{N}, *)$ where $n * m \coloneqq |n - m|$ because it is non-associative, e.g. $1 * (2 * 5) = 2$; $(1 * 2) * 5 = 4$.

We use the notation $gh = g \cdot h = g * h$ here to represent the group operation (regardless of the specific operation in question).

1.3. Basic properties

Proposition. Let G be a group. Then,

(i) The identity element e is unique.

(ii) $\forall g \in G$, the inverse g^{-1} is unique.

(iii) $g \cdot h = g \iff h \cdot g = g$

(iv) $g \cdot h = e \iff h \cdot g = e$

(v) $(gh)^{-1} = h^{-1} g^{-1}$

(vi) $(g^{-1})^{-1} = g$

Proof. We prove each case individually.

(i) Assume $\exists e, e'$ which are distinct identity elements. We have $ee' = e$ and $ee' = e'$ by the definition of the inverse so $e = e'$ #

(ii) Suppose h and k are distinct inverses of g. Then $gh = e$ and $gk = e$, so:

$$\begin{aligned} gh &= gk \\ g^{-1}gh &= g^{-1}gk \\ h &= k \ \# \end{aligned}$$

(iii)

$$\begin{aligned} & gh = g \\ \iff & gh = ge \\ \iff & h = e \\ \iff & hg = eg \\ \iff & hg = g \end{aligned}$$

(iv)

$$\begin{aligned} & gh = e \\ \iff & ghg = g \\ \iff & g^{-1}ghg = g^{-1}g \\ \iff & hg = e \end{aligned}$$

(v) $(gh)(h^{-1}g^{-1}) = ghh^{-1}g^{-1} = gg^{-1} = e$

(vi) $g^{-1}g = e$

□

Definition (abelian group)**.** A group G is said to be *abelian* if $\forall a, b, \in G$, $a * b = b * a$.

A common example of an abelian group is the reals under addition. A non-example is the group of invertible 2×2 matrices under matrix multiplication.

Definition. The order of a group G, denoted $|G|$, is the number of elements in the set G. A group G is called a finite group if its order is finite, and it is called an infinite group if its order is infinite.

1.4. Subgroups

Definition. Let $(G, *)$ be a group. A subset $H \subseteq G$ is a subgroup of G if $(H, *)$ is a group. We denote this $H \leq G$.

We must verify each group axiom on a subset to check if it is a subgroup—with the notable exception of the associativity axiom, the property of associativity is inherited by subgroups. Here are some examples of subgroups.

(i) $\{e\}$ is the trivial subgroup

(ii) $G \leq G$

(iii) $(\mathbb{Z}, +) \leq (\mathbb{Q}, +) \leq (\mathbb{R}, +) \leq (\mathbb{C}, +)$

Lemma. Let G be a group. $H \subset G$ is a subgroup of G if and only if H is non-empty and $\forall a, b \in H, ab^{-1} \in H$.

Proof. We prove each axiom.

- (identity) Setting $a = b$ gives $aa^{-1} = e \in H$ as required.
- (inverses) Setting $a = e$, which we know exists from the identity proof above, gives $b^{-1} \in H$.
- (closure) Setting $b = c^{-1}$, we know that $c \in H$, and we can always choose a b such that c is any value we want; and with the property we can see that $ac \in H$ as required by the closure axiom.

□

Proposition. The subgroups of $(\mathbb{Z}, +)$ are precisely the subsets of the form $n\mathbb{Z} \subset Z$ where $n\mathbb{Z} \coloneqq \{nk : k \in \mathbb{Z}\}$.

Proof. First, we know that each $n\mathbb{Z}$ is a subgroup: given any integer $n \in \mathbb{N}$ the axioms hold:

- (closure) given $nk_1, nk_2 \in n\mathbb{Z}$, we have $nk_1 + nk_2 = n(k_1 + k_2) \in n\mathbb{Z}$
- (identity) $e = 0 = n \cdot 0 \in n\mathbb{Z}$
- (inverse) $-nk = n(-k) \in n\mathbb{Z}$

We also prove the converse statement, namely that the only viable subgroups are of the form $(n\mathbb{Z}, +)$. If $H = \{0\}$ then clearly $H = 0\mathbb{Z}$ which is a trivial subgroup. Otherwise, there are some nonzero elements.

There must be at least one positive element in H, since any negative element can be inverted to make a positive one in H. So, let the smallest positive element be n. Since H is a subgroup, it is closed and has inverses. This implies that

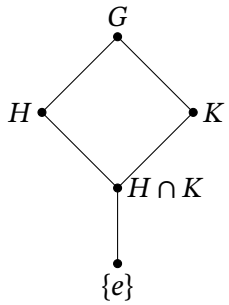
$$n + n + n + \cdots \in H;$$
$$n^{-1} + n^{-1} + n^{-1} + \cdots \in H$$

Therefore $n\mathbb{Z}$ is contained within H. Now, let us show that there are no extra elements. Suppose, for purposes of a contradiction, that $\exists k \in H$ s.t. $k \notin n\mathbb{Z}$. Then, since k is an integer and not a multiple of n, it must lie between two such multiples: $nm < k < n(m + 1)$ where $m \in \mathbb{Z}$. This means that $0 < k - nm < n$ which implies that there is a smaller positive element than n in the set. This is a contradiction, so there are no more elements in the set. □

Proposition. The following statements are true:

- Let H, K be subgroups of G. Then $H \cap K$ is a subgroup of G.
- If $K \leq H$ and $H \leq G$, then $K \leq G$.
- If $K \subseteq H$, $H \leq G$ and $K \leq G$ then $K \leq H$.

We can use a lattice diagram to denote subgroups. Points below other points joined by lines represent subgroups. Let G, H, K be groups and $H \leq G$ and $K \leq G$.



1.5. Subgroups generated by a subset

Definition. Let $X \neq \varnothing$ be a subset of a group G. The subgroup *generated by* X denoted $\langle X \rangle$ is the intersection of all subgroups containing X. Equivalently, $\langle X \rangle$ is the smallest subgroup of G that contains X as a subset. Note that there will always exist some subgroup $\langle X \rangle$ regardless of what X is chosen; a trivial result would be G itself.

We can make a more precise definition of generated groups as follows:

- $\langle X \rangle$ contains e
- $\langle X \rangle$ contains the set X
- $\langle X \rangle$ contains all possible products of X and their inverses

Proposition. Let $X \subseteq G, X \neq \varnothing$. Then $\langle X \rangle$ is the set of elements of G of the form

$$x_1^{\alpha_1} x_2^{\alpha_2} x_3^{\alpha_3} \cdots x_k^{\alpha_k}$$

where $x_i \in X$ (not necessarily distinct), $\alpha_i = \pm 1$, and $k \geq 0$. By convention, the empty product $k = 0$ is defined to be e.

Proof. Let T be the set of such elements of the given form. Clearly, $T \subseteq \langle X \rangle$. Also, T is a subgroup of G, and $X \subseteq T$, so $\langle X \rangle \subseteq T$. Because both $T \subseteq \langle X \rangle$ and $\langle X \rangle \subseteq T$, we have $T = \langle X \rangle$. □

Note that generating sets are not necessarily unique. For example, the group of integers under addition generated by $\langle 1 \rangle$ is equivalent to $\langle 2, 3 \rangle$, both of which are equivalent to $\mathbb{Z}$, for

example. As a discrete example, $\mathbb{Z}_5$ can be generated by any element in the set apart from zero, for example: $\mathbb{Z}_5 = \langle 1 \rangle = \langle 2 \rangle = \langle 3 \rangle = \langle 4 \rangle \neq \langle 0 \rangle$.

2. Homomorphisms

2.1. Definition and elementary properties

Definition. Let $(G, *_G)$, $(H, *_H)$ be groups. A function $\varphi : H \to G$ is a homomorphism if

$$\forall a, b \in H, \quad \varphi(a *_H b) = \varphi(a) *_G \varphi(b)$$

A homomorphism $\varphi : H \to G$ may have the following descriptions:

- injective, if $\varphi(a) = \varphi(b) \implies a = b$;
- surjective, if $\forall g \in G, \exists h \in H$ s.t. $\varphi(h) = g$; and
- bijective, if it is both injective and surjective.

A more intuitive interpretation of the descriptions is:

- A function is injective if the outputs are unique;
- A function is surjective if all outputs are used;
- A function is bijective if there is a one-to-one relation between every element in the input and output sets.

Here are some examples, without proofs.

(i) Given any two groups G and H, $\varphi : H \to G$ defined by $\varphi(h) = e_G$ is a homomorphism.

(ii) The inclusion function $\iota : H \to G$ where $H \leq G$ is an injective homomorphism. The inclusion function is defined as the identity function, simply transferring elements from a subgroup into the supergroup.

(iii) $\varphi : \mathbb{Z} \to \mathbb{Z}_n$ given that $\varphi(k) = k \bmod n$ is a surjective homomorphism.

(iv) $\varphi : (\mathbb{R}, +) \to (\mathbb{R}_{>0}, \cdot)$ where $R_{>0} = \{r \in \mathbb{R} : r > 0\}$ and $\varphi(x) = e^x$ is a bijective homomorphism, otherwise known as an isomorphism.

(v) $\det : GL_2(\mathbb{R}) \to (\mathbb{R}^*, \cdot)$ is a surjective homomorphism.

Proposition. Let $\varphi : H \to G$ be a homomorphism. Then, for all $h \in H$:

(i) $\varphi(e_H) = e_G$

(ii) $\varphi(h^{-1}) = \varphi(h)^{-1}$

(iii) Given another homomorphism $\psi : G \to K$, $\psi \circ \varphi : H \to K$ is a homomorphism.

Proof. We prove each result in order.

(i) Given the identity element of H is e_H and similarly for G,

$$\begin{aligned} \varphi(e_H * e_H) &= \varphi(e_H) * \varphi(e_H) \\ \implies \varphi(e_H) &= \varphi(e_H) * \varphi(e_H) \\ e_G &= \varphi(e_H) \end{aligned}$$

(ii) Consider $\varphi(h) * \varphi(h^{-1}) = \varphi(h * h^{-1}) = \varphi(e_H) = e_G$ which is the defining property of the inverse.

(iii) For all $a, b \in H$:

$$\begin{aligned} (\psi \circ \varphi)(a * b) &= \psi(\varphi(a * b)) \\ &= \psi(\varphi(a) + \varphi(b)) \\ &= \psi(\varphi(a)) + \psi(\varphi(b)) \\ &= (\psi \circ \varphi)(a) + (\psi \circ \varphi)(b) \end{aligned}$$

□

2.2. Isomorphisms

A bijective homomorphism is called an isomorphism. If there exists an isomorphism $\varphi : H \to G$, we say that H is isomorphic to G, or $H \cong G$.

(i) Consider a group G defined as $\{e^{\frac{2\pi ik}{n}} : k \in \mathbb{Z}_n\}$ under multiplication. Then, $(G, \cdot) \cong (\mathbb{Z}_n, +)$ where $\varphi : \mathbb{Z}_n \to G$ is defined as $\varphi(k) = e^{\frac{2\pi ik}{n}}$.

(ii) $\varphi : \mathbb{Z} \to n\mathbb{Z}$ for $n \in \mathbb{N}$ given by $\varphi(k) = nk$. Note that all non-trivial subgroups of $\mathbb{Z}$ are isomorphic to $\mathbb{Z}$.

Proposition. Let $\varphi : H \to G$ be an isomorphism. Then $\varphi^{-1} : G \to H$ is an isomorphism.

Proof. For all $a, b \in G$,

$$\begin{aligned} \varphi^{-1}(a * b) &= \varphi^{-1}\left[\varphi(\varphi^{-1}(a)) * \varphi(\varphi^{-1}(b))\right] \\ &= \varphi^{-1}\left[\varphi(\varphi^{-1}(a) * \varphi^{-1}(b))\right] \\ &= \varphi^{-1}(a) * \varphi^{-1}(b) \end{aligned}$$

So φ^{-1} is a homomorphism. But since φ is bijective, so is φ^{-1}. So φ^{-1} is an isomorphism. □

2.3. Images and kernels

Definition. Let $\varphi : H \to G$ be a homomorphism. Then the image of φ, denoted $\operatorname{Im}\varphi$, is defined as $\{g \in G : g = \varphi(h) \text{ for some } h \in H\}$. The kernel of φ, denoted $\ker\varphi$, is defined as $\{h \in H : \varphi(h) = e_G\}$.

Informally, we can say:

- The image of φ is the set of outputs of φ.
- The kernel of φ is the set of inputs that map to the identity element.

Proposition. $\operatorname{Im}\varphi \leq G$ and $\ker\varphi \leq H$.

Proof. To prove that $\operatorname{Im}\varphi \leq G$, we check the group axioms (apart from associativity, since this is implicit).

- (closure) If $a, b \in \operatorname{Im}\varphi$ then there exist some $x, y \in H$ such that $\varphi(x) = a$ and $\varphi(y) = b$. Therefore, $\varphi(x)\varphi(y) = \varphi(xy)$ which is in the image by definition.
- (identity) $\varphi(e_H) = e_G$
- (inverses) Let $x \in H$ such that $\varphi(x) = a$. Then, because $x^{-1} \in H$, we know that $\varphi(x^{-1}) = \varphi(x)^{-1} \in \operatorname{Im} H$ as required.

Now we prove a similar result for the kernel.

- (closure) If $x, y \in \ker H$ then $\varphi(xy) = \varphi(x)\varphi(y) = e_G e_G = e_G$, which is the requirement for being in the kernel, so $xy \in \ker\varphi$.
- (identity) $\varphi(e_H) = e_G$ so the identity element e_H is in the kernel.
- (inverses) $\varphi(x^{-1}) = \varphi(x)^{-1}$. So if $x \in \ker\varphi$ then $\varphi(x^{-1}) = e_G^{-1} = e_G$ so φ^{-1} is also in the kernel.

□

Here are a few examples of kernels and images of homomorphisms.

(i) If $\varphi : H \to G$ is the trivial homomorphism (mapping every element to the identity) then:
$$\operatorname{Im}\varphi = \{e_G\}; \quad \ker\varphi = H$$

(ii) If $H \leq G$ then the inclusion homomorphism $\iota : H \to G$ has
$$\operatorname{Im}\iota = H; \quad \ker\iota = e_H$$

(iii) $\varphi : \mathbb{Z} \to \mathbb{Z}_n$ where operations are performed modulo n has
$$\operatorname{Im}\varphi = \mathbb{Z}_n; \quad \ker\varphi = n\mathbb{Z}$$

Proposition. Let $\varphi : H \to G$ be a homomorphism. Then

- φ is surjective if and only if $\operatorname{Im}\varphi = G$; and
- φ is injective if and only if $\ker\varphi = \{e_H\}$.

Proof. The first case is trivial. After all, the definition of surjectivity is that all outputs are mapped onto by something, which means that image is equal to this output set. Now, let us prove the injectivity part. We start in the forward direction, then we prove the converse.

Suppose that φ is injective. Then $\varphi(a) = \varphi(b) \implies a = b$. We have that $\varphi(e_H) = e_G$, so e_H must be the only element sent to e_G. Therefore the kernel is simply $\{e_H\}$.

Conversely, suppose that the kernel of φ is simply the identity element. Then, let us suppose there are two elements a, b in H such that $\varphi(a) = \varphi(b)$. Then, $\varphi(ab^{-1}) = \varphi(a)\varphi(b)^{-1} = \varphi(b)\varphi(b)^{-1} = e_G$. Therefore, $ab^{-1} = e_H$, so $a = b$. So φ is injective. □

3. Types of groups

3.1. Direct products of groups

Definition. The direct product of two groups G and H is written $G \times H$, and defined to be $\{(g, h) : g \in G, h \in H\}$, where the group operation is defined by

$$(g_1, h_1) *_{G\times H} (g_2, h_2) = (g_1 *_G g_2, h_1 *_H h_2)$$

We will now prove that this really is a group.

Proof. We prove each axiom.

- (closure) For a pair of elements (g_1, h_1) and (g_2, h_2) in $G\times H$, the product $(g_1 *_G g_2, h_1 *_H h_2)$ is clearly in $G\times H$, because the first entry is in G and the second entry is in H, which is the requirement for being a member of $G \times H$.
- (identity) The element (e_G, e_H) is an identity.
- (inverses) Given an element $(g, h) \in G \times H$, the element (g^{-1}, h^{-1}) satisfies

$$(g^{-1}, h^{-1})(g, h) = (e_G, e_H) = e_{G\times H}$$

- (associativity) Given three elements (g_i, h_i), $i \in \{1, 2, 3\}$, we have

$$\begin{aligned}((g_1, h_1) * (g_2, h_2)) * (g_3, h_3) &= (g_1 * g_2, h_1 * h_2) * (g_3, h_3)\\ &= ((g_1 * g_2) * g_3, (h_1 * h_2) * h_3)\\ &= (g_1 * (g_2 * g_3), h_1 * (h_2 * h_3))\\ &= (g_1, h_1) * (g_2 * g_3, h_2 * h_3)\\ &= (g_1, h_1) * ((g_2, h_2) * (g_3, h_3))\end{aligned}$$

□

$G \times H$ contains subgroups $G \times e_H$ and $e_G \times H$ which are isomorphic to G and H respectively. We name these subgroups simply G and H because they are isomorphic.

Note. In $G \times H$, everything in G commutes with everything in H.

$$\forall g \in G, \forall h \in H, (g, e_H) * (e_G, h) = (e_G, h) * (g, e_H) = (g, h)$$

Theorem (Direct Product Theorem)**.** Let H, K be subgroups of G such that

- $H \cap K = \{e\}$ (the groups intersect only in e)
- $\forall h \in H, \forall k \in K, hk = kh$ (H and K commute in G)
- $\forall g \in G, \exists h \in H, \exists k \in K$ s.t. $g = hk$ ($G = HK$)

Then $G \cong H \times K$.

Proof. Consider $\varphi : H \times K \to G$ where $\varphi((h,k)) = hk$. We now prove that φ is a homomorphism.

$$\begin{aligned}\varphi((h_1,k_1)(h_2,k_2)) &= \varphi((h_1h_2,k_1k_2))\\ &= h_1h_2k_1k_2\\ &= h_1k_1h_2k_2\\ &= \varphi((h_1,k_1))\varphi((h_2,k_2))\end{aligned}$$

Note that by the third property in the theorem we know that φ is surjective. We now prove that φ is also injective.

Suppose that $(h,k) \in \ker\varphi$. Then $\varphi((h,k)) = e_G$ so $hk = e_G$. So $h = k^{-1}$. This means that there is some element that is part of both H and K, for example h. But by the first property in the theorem, this value must be e, so $\ker\varphi = \{e_G\}$, so φ is injective.

φ is an injective, surjective homomorphism, so it is an isomorphism. So G is isomorphic to $H \times K$. □

Now, we can consider direct products in two distinct lenses: a combination of smaller groups to form a large one, or a partition of a large group into two that combine to produce the original.

3.2. Cyclic groups

Definition. Let G be a group, and let $X \subseteq G$ be some subset. If $\langle X\rangle = G$ then X is a generating set of G. We say that G is cyclic if there exists some element a in G such that $\langle a\rangle = G$. a is called a generator of G.

(i) The trivial group $\{e\}$ is generated by its element.

(ii) $(\mathbb{Z},+)$ is a cyclic group generated by $\mathbb{Z} = \langle -1\rangle = \langle 1\rangle$.

(iii) $(\mathbb{Z}_n,+)$, where addition is modulo n, is generated by $\mathbb{Z}_n = \langle k\rangle$ where k and n are coprime.

Theorem. Any cyclic group G is isomorphic to C_n (for some $n \in \mathbb{N}$) or $\mathbb{Z}$.

Proof. Let $G = \langle b\rangle$. Then suppose that there exists some natural number n such that $b^n = e$. We take the smallest such n, and define $\varphi : C_n \to G$ by $\varphi(a^k) = b^k$ where the elements of C_n are e, a, a^2 and so on.

We now show that φ is a homomorphism. For any two elements $a^j, a^k \in C_n$, we have two cases. If $j + k < n$, then $\varphi(a^j \cdot a^k) = \varphi(a^{j+k}) = b^{j+k} = b^j \cdot b^k = \varphi(a^j) \cdot \varphi(a^k)$ as required. Otherwise, $j+k \geq n$, then $\varphi(a^j \cdot a^k) = \varphi(a^{j+k-n}) = b^{j+k}(b^n)^{-1} = b^{j+k} \cdot e = b^{j+k} = b^j \cdot b^k = \varphi(a^j) \cdot \varphi(a^k)$ as required.

Note that φ is bijective:

- $b^n = e \in G$ implies that all elements of G can be written b^k where $0 \leq k < n$, so φ is surjective; and
- Let a^k be an element in the kernel of φ where $0 \leq k < n$. Then $\varphi(a^k) = e \implies b^k = e$. But k must be zero, because any other value would contradict the fact that we chose n to be the smallest number with this property. So the kernel is trivial.

So φ is an isomorphism, and $G \cong C_n$.

If alternatively there exists no n such that $b^n = e$, then we construct $\varphi : \mathbb{Z} \to G$ by $\varphi(k) = b^k$. Then $\varphi(k + m) = b^{k+m} = b^k \cdot b^m = \varphi(k) \cdot \varphi(m)$, so φ is a homomorphism. Clearly φ is surjective because all elements of G can be constructed with powers of b. Now, suppose $m \in \ker \varphi$ where m is nonzero. Then $\varphi(m) = b^m = e$ and $\varphi(-m) = b^{-m} = e$. So one of m and $-m$ is positive, contradicting the fact that there is no such $n > 0$ where $b^n = e$. So the kernel is trivial, so φ is an isomorphism, so $G \cong \mathbb{Z}$. □

Definition. The order of an element $g \in G$ is the smallest $n \in \mathbb{N}$ such that $g^n = e$. We say that $\operatorname{ord} g = n$. If there is no such n, then $\operatorname{ord} g = \infty$.

Note that given some $g \in G$, the subgroup $\langle g \rangle$ is a cyclic group isomorphic to C_n if $\operatorname{ord} g = n$, and isomorphic to $\mathbb{Z}$ if $\operatorname{ord} g = \infty$. So $\operatorname{ord} g = |\langle g \rangle|$.

Proposition. Cyclic groups are abelian.

The proof is trivial.

3.3. Dihedral groups

Definition. The dihedral group D_{2n} is the group of symmetries of a regular n-gon. The group operation is composition of transformations. For example, D_6 is the group of symmetries of a regular triangle.

The elements of a general D_{2n} fall into two categories:

- (rotations) We can rotate the shape around its centre through $\frac{2\pi k}{n}$. There are n such rotations, including the identity element e.
- (reflections) We can reflect the shape across axes through each vertex and the shape's centre. If n is odd, then there are n such symmetries. If n is even, there are $n/2$ such symmetries, but there are a further $n/2$ symmetries through the midpoints of edges and the centre of the shape, leaving a total of n.

Therefore there are (at least) $2n$ elements in D_{2n}. Are these all the elements? To answer this, let us name vertices $v_1, v_2 \cdots v_n$, and let us consider some element g of D_{2n}. There are two characteristics of a rigid symmetry:

- Vertices are mapped to other vertices. So $v_1 \mapsto v_k$ for some $1 \leq k \leq n$.

- Edges are mapped to other edges. So $v_2 \mapsto v_{k+1}$ or v_{k-1} (modulo n). Note that once we define v_1 and v_2, then the location of v_3 is predetermined. Inductively, the entire polygon is pre-determined.

There are n choices for the location of v_1. There are two choices for the location of v_2. So there are only $2n$ elements in D_{2n}. So we have all the elements already. It is also trivial to prove that D_{2n} is a group, simply by verifying the axioms, noting the function composition is always associative.

Note that we can generate D_{2n} using just one rotation and one reflection. Let r be the rotation by $\frac{2\pi}{n}$, and let s be the reflection through v_1 (such that $v_1 \mapsto v_1$). Now,

- r^k gives all possible rotations;
- $r^i s r^{-i}$ gives a reflection through v_{i+1} and the centre;
- $r^{i+1} s r^{-i}$ gives a reflection through the edge joining v_i and v_{i+1}.

These are all three cases, so $D_{2n} = \langle r, s \rangle$.

4. Permutation groups

4.1. Definition

Definition. Given a set X, a permutation of the set is a bijective function $\sigma : X \to X$. The set of all permutations of X is denoted $\operatorname{Sym} X$.

Theorem. $\operatorname{Sym} X$ is a group with respect to composition.

This is provable by checking the group axioms, noting that all bijective functions are invertible, and that function compositions are always associative.

Definition. If $|X| = n$ then S_n is the isomorphism class of $\operatorname{Sym} X$.

Note that $|S_n|$ is $n!$ because the first element has n choices for where to be mapped, the second element has $n - 1$ choices, etc.

4.2. Cycles

We may use a two-row notation for permutations. For example, a permutation $\sigma \in S_3$ such that $\sigma(1) = 2, \sigma(2) = 3, \sigma(3) = 1$ may be written

$$\sigma = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}$$

The columns represent the maps that σ performs: $1 \mapsto 2; 2 \mapsto 3; 3 \mapsto 1$. However, this is quite a clunky, long-winded notation. More often we use a kind of cycle notation, for example

$$\sigma = (1\ 2\ 3)$$

This says that σ represents the cycle $1 \mapsto 2 \mapsto 3 \mapsto 1$. Note that, for example, the cycle (1 2 3) is equivalent to the cycle (2 3 1). We call a cycle like this a 3-cycle because it has 3 elements. So, for example, the cycle $(1\ 2\ 3\ 4\ 5\ 6\ 7) \in S_n$ where $n \geq 7$ is a 7-cycle. Cycles with two elements are called transpositions, and cycles with one element are called singletons.

Since cycles are permutations, we can compose them like this:

$$(1\ 2\ 3\ 4)(3\ 2\ 4) \in S_4$$

We know that the resulting permutation must be a member of S_4 because of the closure axiom. We can deduce what the resulting permutation is in two ways:

- We can find the value of $(1\ 2\ 3\ 4)(3\ 2\ 4)(x)$ for all $1 \leq x \leq n$. This allows us to write the permutation in the two-row notation.

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 3 & 4 \end{pmatrix}$$

- Alternatively, let us begin by just finding $\sigma(1) = 2$. Then, we can find where this result maps to, and so on, until we have a completed cycle. We are guaranteed to form a cycle, as we will prove later. Repeat this cycle generation for all unused numbers, and then you will get a product of cycles, in this case $\sigma = (1\ 2)(3)(4)$.

Note that the inverse of a permutation can be created by swapping the rows. A cycle can be inverted by simply reversing the order of its elements. One more interesting fact is that $D_{2n} \leq S_n$. D_{2n} is a permutation of n vertices with some added constraints (e.g. edges must map to edges), so it makes sense that it would be a subgroup. In particular, $D_6 = S_3$.

Cycles are considered disjoint if no number appears in both.

Lemma. Disjoint cycles commute.

Proof. Let σ, τ be disjoint cycles of S_n. We want to prove that $\sigma\tau(x) = \tau\sigma(x)$ for all $1 \leq x \leq n$. There are three cases:

- If $x \notin \sigma$ and $x \notin \tau$ then immediately $\sigma\tau(x) = x = \tau\sigma(x)$.
- If $x \in \sigma$ but $x \notin \tau$ then because σ is a cycle, $\sigma(x) \in \sigma$; and because the cycles are disjoint, $\sigma(x) \notin \tau$. So $\sigma\tau(x) = \sigma(x) = \tau\sigma(x)$.
- If $x \in \tau$ but $x \notin \sigma$, use the proof above but swap the letters.

□

4.3. Disjoint cycle decomposition

Theorem. Any $\sigma \in S_n$ can be written as a product of disjoint cycles. This is unique up to reordering cycles (and, of course, moving the elements around within a cycle without altering it).

Proof. Let $\sigma \in S_n$. Now consider the infinite list of terms $1, \sigma(1), \sigma^2(1), \sigma^3(1) \cdots$. σ is a bijection from a set to itself so this list will continue infinitely, but there are only n possible items in this set. Therefore, by the pigeonhole principle, there must be two distinct items in that list that are the same. Let us label their indices a and b, such that $\sigma^a(1) = \sigma^b(1)$, and that $a > b$ without loss of generality. Then we can multiply on the right by $\sigma^{-b}(1)$ to get $\sigma^{a-b}(1) = 1$.

Now that we have proven that the number 1 exists multiple times in the list, let us take k to be the smallest positive integer such that $\sigma^k(1) = 1$. Then for $0 \leq l < m < k$, if $\sigma^m(1) = \sigma^l(1)$ then $\sigma^{m-l}(1) = 1$ which contradicts the minimality of k. So the first k numbers on the list are distinct, so $(1\ \sigma(1)\ \sigma^2(1)\ \cdots\ \sigma^{k-1}(1))$ is a cycle.

Repeat this whole process, replacing 1 with different unused values in the set. This will always continue to work because no number that has already appeared can reappear (because σ is a bijection).

Continue until we have exhausted the entire set $\{1, \cdots, n\}$. Then we can multiply together all of the (disjoint) cycles we have generated. Note that each element from $\{1, \cdots, n\}$ must appear exactly once in this product (if it is mapped to itself, it is present as a singleton). It is clear then that this product is equal to σ, because for any input k, no cycles except for the one containing said input (and also, of course, containing the output $\sigma(k)$) will do anything to it.

We can prove the uniqueness of the decomposition by supposing that there might exist two decompositions. Taking any element x in the set $\{1, \cdots, n\}$, we know that σ uniquely defines the cycle that x belongs to. So that means that in both decompositions, the cycle containing x is the same. By repeating this for all x in the set, we can be sure that all cycles are the same, and thus the decompositions in their entirety are the same. Therefore the decomposition is unique. □

The set of cycle lengths of the disjoint cycle decomposition of σ is called the cycle type of σ. For example, $\sigma = (1\ 2\ 3)(5\ 6)$ has cycle type $3, 2$ (or equivalently $2, 3$).

Theorem. The order of $\sigma \in S_n$ is the least common multiple of the cycle lengths in its cycle type.

Proof. The order of a k-cycle is k. Let us decompose σ into a product of disjoint cycles such that $\sigma = \tau_1\tau_2 \cdots \tau_r$. Then $\sigma^m = \tau_1^m\tau_2^m \cdots \tau_r^m$ since disjoint cycles commute.

Let each τ_i be a k_i-cycle. Then if $\sigma^m = e$, $\tau_1^m\tau_2^m \cdots \tau_r^m = e$, and so $\tau_1^m = \tau_2^{-m}\tau_3^{-m} \cdots \tau_r^{-m}$. Note that the right hand side and left hand side permute different elements, so they must both be the identity element e. Repeating this style of argument with every τ shows that $\tau_i^m = e$ so $k_i | m$.

So clearly the lowest common multiple of all of the k_i divides the order of the permutation, $o(\sigma)$. Now, we check that it is actually equal to $o(\sigma)$. Let L be this lowest common multiple. Then $\sigma^L = \tau_1^L\tau_2^L \cdots \tau_r^L = (\tau_1^{k_1})^{L/k_1}(\tau_2^{k_2})^{L/k_2} \cdots (\tau_r^{k_r})^{L/k_r}$. All of these exponents are integers because L is a multiple of each k_i. So we have $e \cdot e \cdots e = e$. So the order of σ is L. □

4.4. Products of transpositions

Proposition. Let $\sigma \in S_n$. Then σ is a product of transpositions.

Proof. It is enough to prove this for just a cycle, then we can use the disjoint cycle decomposition to create a transposition product for the whole σ. We have

$$(a_1\ a_2 \cdots a_n) = (a_1\ a_2)(a_2\ a_3) \cdots (a_{n-1}\ a_n)$$

so the result is immediate. □

Note that this decomposition is not unique in general.

A permutation may be considered even if its transposition decomposition has an even number of terms, or odd otherwise. Note that an even-length cycle has odd parity, and an odd-length cycle has even parity.

Proposition. The parity of a permutation is well-defined, regardless of exactly how you write a permutation.

Proof. Let us denote the amount of cycles in the disjoint cycle decomposition of σ with $\#(\sigma)$. Let $\tau = (c\ d)$. Then the effects of multiplying σ by τ (on the right) have two cases, since it only affects c and d.

- If c and d are in the same cycle in σ, we get the following conversion:

  $$(c\ a_2 \cdots a_{k-1}\ d\ a_{k+1} \cdots a_l) \mapsto (c\ a_{k+1} \cdots a_l)(d\ a_2 \cdots a_{k-1})$$

  So $\#(\sigma\tau) = \#(\sigma) + 1$.

- Otherwise, c and d are in different cycles (possibly singletons) in σ, so we get the following conversion:

  $$(c\ a_2 \cdots a_k)(d\ b_2 \cdots a_l) \mapsto (c\ b_2 \cdots b_l\ d\ a_2 \cdots a_k)$$

  So $\#(\sigma\tau) = \#(\sigma) - 1$.

In either case, parity is flipped. Now, suppose that σ is written as two products of transpositions, where one has m transpositions, and one has n transpositions. Therefore we have $\#(\sigma) \equiv \#(e) + m \bmod 2$, and $\#(\sigma) \equiv \#(e) + n \bmod 2$. But $\#(\sigma)$ is uniquely determined by σ, so both equations match, so $m \equiv n \bmod 2$, so the parity is well-defined. □

Definition. Writing σ as a product of transpositions, the sign of σ is defined as 1 if the permutation is even, and -1 if it is odd.

Note that the function $\operatorname{sign}(\sigma)$ is a homomorphism from S_n to $(\{-1, 1\}, \cdot)$.

Definition. The alternating group A_n is defined as the kernel of the sign homomorphism on S_n. In other words, it is the set of even permutations of S_n.

5. Möbius transformations

5.1. The Möbius group

Möbius groups are an analogous concept to permutation groups, but on the infinite set of the complex numbers. A Möbius transformation f is defined as follows:

$$f : \widehat{\mathbb{C}} \to \widehat{\mathbb{C}}; \quad f(z) = \frac{az+b}{cz+d}; \quad a,b,c,d \in \mathbb{C}; \quad ad-bc \neq 0$$

The reason for the restriction that $ad - bc \neq 0$ is that $ad = bc$ implies that f is a constant value for all z. Note that $\widehat{\mathbb{C}}$ is known as the 'extended complex plane', defined as the complex plane together with a point at infinity, denoted ∞. There are some special points related to Möbius transformations:

- $f\left(\frac{-d}{c}\right)$ is defined to be ∞. This is because the denominator of the fraction would be zero.
- $f(\infty)$ is defined to be $\frac{a}{c}$ if $c \neq 0$. This is because as the length of z increases to infinity, the constant terms b and d vanish. However, if $c = 0$, then the numerator explodes to infinity as the denominator remains constant, so $f(\infty) = \infty$ in this case.

Lemma. Möbius transformations are bijections from $\widehat{\mathbb{C}} \to \widehat{\mathbb{C}}$.

Proof. We can prove this by evaluating $f(f^{-1}(z))$ and $f^{-1}(f(z))$ at all z, taking into account all the special points. The entire proof is not written here, but it suffices to substitute every special point and a generic z into both of these expressions, and show that they equal z in all cases. □

Theorem. The set $\mathcal{M}$ of Möbius maps forms a group under composition of functions.

Proof. We must check each of the group axioms, and we begin with closure. Let $f_1(z) = \frac{a_1z+b_1}{c_1z+d_1}; f_2(z) = \frac{a_2z+b_2}{c_2z+d_2}$. To compose these functions, we first ignore the special points and then check them individually later.

$$\begin{aligned}(f_2 \circ f_1)(z) &= f_2(f_1(z)) \\ &= \frac{a_2\left(\frac{a_1z+b_1}{c_1z+d_1}\right)+b_2}{c_2\left(\frac{a_1z+b_1}{c_1z+d_1}\right)+d_2} \\ &= \frac{(a_1a_2+b_2c_1)z+(a_2b_1+b_2d_1)}{(c_2a_1+d_2c_1)z+(c_2b_1+d_1d_2)} \\ &=: \frac{az+b}{cz+d}\end{aligned}$$

Note that $ad - bc = (a_1a_2 + b_2c_1)(c_2b_1 + d_1d_2) - (a_2b_1 + b_2d_1)(c_2a_1 + d_2c_1) = (a_1d_1 - b_1c_1)(a_2d_2 - b_2c_2)$ which is the product of two nonzero numbers, which is therefore nonzero.

Now we will check all the special points.

$$\begin{aligned}(f_2 \circ f_1)(\infty) &= f_2\left(\frac{a_1}{c_1}\right)\\ &= \frac{a_2\left(\frac{a_1}{c_1}\right)+b_2}{c_2\left(\frac{a_1}{c_1}\right)+d_2}\\ &= \frac{a_1a_2+b_2c_1}{c_2a_1+d_2c_1}\\ &= \frac{a}{c}\\ (f_2 \circ f_1)(\infty) &= f_2(\infty) = \frac{a_2}{c_2}\\ (f_2 \circ f_1)\left(f^{-1}\left(\frac{-d_2}{c_2}\right)\right) &= f_2\left(\frac{-d_2}{c_2}\right)\\ &= \infty\end{aligned}$$

Note that each of these results matches up with our intuitive understanding of infinity in the limit, for instance $(f_2 \circ f_1)(\infty) = \frac{a}{c}$, where naïvely we might assume $(f_2 \circ f_1)(\infty) = \frac{a\cdot\infty+b}{c\cdot\infty+d} = \frac{a}{c}$.

Now we may prove the other group axioms hold for $\mathcal{M}$. Clearly there is an identity element $f(z) = \frac{1z+0}{0z+1}$. We know that there are always inverses because f is a bijection. Finally, we know that all Möbius maps obey the associative law because function composition is always associative. So $\mathcal{M}$ is a group. □

5.2. Properties of the Möbius group

When we are working with Möbius groups, we use the following conventions:

$$\frac{1}{\infty} = 0; \quad \frac{1}{0} = \infty; \quad \frac{a\cdot\infty}{c\cdot\infty} = \frac{a}{c}$$

Firstly, $\mathcal{M}$ is not abelian. For example, let $f_1(z) = z+1$; $f_2(z) = 2z$. Then $(f_2 \circ f_1)(z) = 2z+2$ and $(f_1 \circ f_2)(z) = 2z+1$.

Proposition. Every Möbius transformation can be written as a composition of maps of the following forms:

(i) $f(z) = az$ where $a \neq 0$. This is a dilation or rotation.

(ii) $f(z) = z + b$. This is a translation by b.

(iii) $f(z) = \frac{1}{z}$. This is an inversion.

Proof. Let $f(z) = \frac{az+b}{cz+d}$. Then if $c \neq 0$, $f(z)$ is given by

$$z \xrightarrow{\text{(ii)}} z + \frac{d}{c} \xrightarrow{\text{(iii)}} \frac{1}{z+\frac{d}{c}} \xrightarrow{\text{(i)}} \frac{(-ad+bc)c^{-2}}{z+\frac{d}{c}} \xrightarrow{\text{(ii)}} \frac{a}{c} + \frac{(-ad+bc)c^{-2}}{z+\frac{d}{c}} = \frac{az+b}{cz+d}$$

If $c = 0$, $f(z)$ is given by

$$z \xrightarrow{\text{(i)}} \frac{a}{d}z \xrightarrow{\text{(ii)}} \frac{a}{d}z + \frac{b}{d} = \frac{az+b}{d}$$

□

Note therefore that the set $\mathcal{S}$ of all dilations, rotations, translations and inversions generates $\mathcal{M}$, or in symbolic form, $\langle \mathcal{S} \rangle = \mathcal{M}$.

6. Cosets and Lagrange's theorem

6.1. Cosets

Let H be a subgroup of some group G, and let $g \in G$. Then a set of the form $gH \coloneqq \{gh : h \in H\}$ is called a left coset of H in G. Also, a set of the form $Hg \coloneqq \{hg : h \in H\}$ is called a right coset of H in G. Mostly we use left cosets, but right cosets can be seen in more specific scenarios. Note that the order of group H is the same as the order of the cosets gH and Hg; we can think of gH and Hg as translated copies of H. Note further that gH and Hg are not necessarily groups; in fact in general they are not groups. We now consider some example cosets.

(i) Let $H = \{e\} \leq G$. Then $gH = \{g\}$.

(ii) Let $H = 2\mathbb{Z}$ and let $G = \mathbb{Z}$. Then (where the cosets are written additively):

- $0 + 2\mathbb{Z} = 2\mathbb{Z}$ which is the set of even integers.
- $1 + 2\mathbb{Z} = \{1 + k : k \in 2\mathbb{Z}\}$ which is the set of odd integers.
- $2 + 2\mathbb{Z} = 2\mathbb{Z}$. There are only two distinct cosets of H in G here; every odd integer will create the set of odd integers, and every even integer will create the set of even integers.

(iii) Let $H = \{e, (1\ 2)\}$, and let $G = S_3$. Then, each (left) coset of H in G is given by

- $eH = \{e, (1\ 2)\} = H$
- $(1\ 2)H = \{(1\ 2), e\} = H$
- $(1\ 3)H = \{(1\ 3), (1\ 2\ 3)\}$
- $(1\ 2\ 3)H = \{(1\ 2\ 3), (1\ 3)\}$
- $(2\ 3)H = \{(2\ 3), (1\ 3\ 2)\}$
- $(1\ 3\ 2)H = \{(1\ 3\ 2), (2\ 3)\}$

Note that:

- $eH = H$
- $\forall h \in H, hH = H$ as H is a group and therefore closed under multiplication with h
- $|gH| = |H|$
- $\bigcup_{g \in G} gH = G$, and in this example in particular, each pair of cosets is equal and disjoint to any other pair

6.2. Lagrange's theorem

Definition. We define the index of a subgroup $H \leq G$ in G, written $|G : H|$, to be the number of distinct cosets of H in G.

Theorem (Lagrange's Theorem)**.** Let $H \leq G$ be a subgroup of a finite group G. Then:

(i) $|H| = |gH|$ for any $g \in G$;

(ii) for any $g_1, g_2 \in G$, either $g_1H = g_2H$ or $g_1H \cap g_2H = \emptyset$; and

(iii) $G = \bigcup_{g \in G} gH$

And in particular, $|G| = |G : H| \cdot |H|$.

Proof. We prove each statement independently.

(i) The function $H \to gH$, defined by $h \mapsto gh$, defines a bijection between H and gH, so $|H| = |gH|$.

(ii) Suppose $g_1H \cap g_2H \neq \emptyset$. Then $\exists g \in g_1H \cap g_2H$. So $g = g_1h_1 = g_2h_2$ for some $h_1, h_2 \in H$. So $g_1 = g_2h_2h_1^{-1}$. So for any $h \in H$, we have

$$g_1h = g_2 \underbrace{h_2h_1^{-1}h}_{\in H}$$

So certainly $g_1H \subseteq g_2H$. Employing a symmetric argument for the other way round, we have $g_1H = g_2H$.

(iii) Given some $g \in G$ then $g \in gH$, since $e \in H$. So $G \subseteq \bigcup_{g \in G} gH$. But also, $gH \subseteq G$, so $\bigcup_{g \in G} gH \subseteq G$. So $G = \bigcup_{g \in G} gH$.

So now that we know that G is composed of a union of disjoint cosets, all of which are the same size, we know that $|G|$ is just the number of these cosets multiplied by the size of such a coset, or in other words

$$|G| = |G : H| \cdot |H|$$

□

Note that we could equivalently have used right cosets in place of left cosets. Remember that in general, $gH \neq Hg$, and the set of left cosets is not equal to the set of right cosets.

Proposition. $g_1H = g_2H \iff g_1^{-1}g_2 \in H$.

Proof. We first consider the forwards case. Clearly g_1 is an element of g_1H, as H contains e. Also, g_2 is an element of g_2. So $g_1^{-1}g_2 \in H$. Now for the backwards case. Clearly, g_2H contains the element g_2, as e maps to it. Also, since H contains $g_1^{-1}g_2$, g_1H contains the element $g_1 * (g_1^{-1}g_2) = g_2$. As cosets are either disjoint or equal, and they clearly share the element g_2, then they are equal. □

Note further that $g' \in gH$ implies $g'H = gH$. We may therefore take a single element from each of these distinct cosets, and we will call them $g_1, g_2, \cdots, g_{|G:H|}$. Then

$$G = \bigsqcup_{i=1}^{|G:H|} g_i H$$

where the $\bigsqcup$ symbol denotes a disjoint union of sets. These g_i are called coset representatives of H in G.

Corollary. Let G be a finite group and $g \in G$. Then $(\operatorname{ord} g) \mid |G|$.

Proof. Recall that ord g is defined as the smallest n such that $g^n = e$. We define the subgroup $H \leq G$ as $H = \langle g \rangle$. Then $\operatorname{ord} g = |H|$. By Lagrange's Theorem, we know that $|H| \mid |G|$. □

Corollary. Let G be a finite group, and let $g \in G$. Then $g^{|G|} = e$.

Proof. This follows directly from the previous corollary. $g^{|G|} = g^{n \cdot \operatorname{ord} g}$ for some natural number n, so this simply reduces to e. □

Corollary. Groups of prime order are cyclic, and are generated by any non-identity element.

Proof. Let $|G| = p$, where p is a prime. We will take some $g \in G$, and generate a group from it. By Lagrange's Theorem, $|\langle g \rangle| \mid |G|$, so $|\langle g \rangle|$ is either 1 or p. Now, note that e and g are both elements of $\langle g \rangle$, so if $g \neq e$ then clearly $|\langle g \rangle| > 1$, so $|\langle g \rangle| = p$. □

We can take Lagrange's theorem into the world of number theory, and specifically modular arithmetic, where we are dealing with finite groups. Clearly, $\mathbb{Z}_n$ is a group under addition modulo n, but what happens with multiplication modulo n? Clearly this is not a group—for a start, 0 has no inverse. By removing all elements of the group that have no inverse, we obtain $\mathbb{Z}_n^*$.

Note that for any $x \in \mathbb{Z}_n$, x has a multiplicative inverse if and only if $\operatorname{HCF}(x, n) = 1$, i.e. if x and n are coprime. This follows directly from the fact that we can write 1 as a linear combination of x and n, i.e. $xy + mn = 1$, thus defining y as the multiplicative inverse of x modulo n. From this, it is simple to check that $\mathbb{Z}_n^*$ forms a group under multiplication.

We may also create an equivalent group-theoretic definition of Euler's totient function φ as follows: $\varphi(n) \coloneqq |\mathbb{Z}_n^*|$. We can now use Lagrange's theorem to prove the Fermat–Euler theorem (that is, $\operatorname{HCF}(N, n) = 1 \implies N^{\varphi(n)} \equiv 1 \bmod n$) as follows.

Proof. If N and n are coprime, then there is an element, here denoted a, in $\mathbb{Z}_n$ corresponding to N. So $a^{\varphi(n)} = a^{|\mathbb{Z}_n^*|} = 1$ in $\mathbb{Z}_n$. Since $N = a + kn$, we may expand $N^{\varphi(n)} = a^{\varphi(n)} + n(\cdots) \equiv a^{\varphi(n)} \equiv 1 \bmod n$. □

6.3. Groups of small order

We can completely classify groups of small order; we already know enough to classify all groups up to order 5 using Lagrange's Theorem.

Proposition. If $|G| = 4$, then $G \cong C_4$ or $G \cong C_2 \times C_2$.

Proof. By Lagrange's Theorem, the possible orders of elements of G with $|G| = 4$ are 1 (only the identity), 2 and 4.

- If there is an element g of order 4, then $G = \langle G \rangle$ because $e \neq g \neq g^2 \neq g^3$, so it is cyclic of order 4.
- If there is no such element, then all non-identity elements must have order 2. G is abelian (by question 7 on example sheet 1). Take two distinct elements b, c of order 2. Then:
  - $\langle b \rangle \cap \langle c \rangle = \{e, b\} \cap \{e, c\} = \{e\}$
  - $bc = cb$ as the group is abelian.
  - The element bc is not equal to b or c ($bc = b \implies c = e$ which is an element of order 1). It is also not equal to e because then $b = c^{-1}$ which implies $b = c$. So the remaining element of G is simply bc. So any element in G may be written as the product of an element in $\langle b \rangle$ multiplied by an element in $\langle c \rangle$.

  These are the three conditions of the direct product theorem, so $G = \langle b \rangle \times \langle c \rangle \cong C_2 \times C_2$.

□

Now here is a list the first five smallest groups (we need more tools in order to classify larger groups):

(i) $G = \{e\}$

(ii) $G \cong C_2$ because a group of prime order is cyclic.

(iii) $G \cong C_3$ for the same reason.

(iv) $G \cong C_4$ or $G \cong C_2 \times C_2$ by the proof above.

(v) $G \cong C_5$ because 5 is prime.

7. Normal subgroups and quotients

7.1. Normal subgroups

How and when does it make sense to divide one group by another?

Definition. An subgroup N of a group G is *normal* if $\forall g \in G, gN = Ng$. We write $N \trianglelefteq G$.

The following equivalent definitions hold:

- $\forall g \in G, gN = Ng$
- $\forall g \in G, \forall n \in N, g^{-1}ng \in N$
- $\forall g \in G, g^{-1}Ng = N$

Proof. The first case is the definition. For the second case, clearly (from the first definition) $ng \in gN$. So multiplying on the left by g^{-1}, we have $g^{-1}ng \in N$ as required. For the third case, we can simply multiply the first definition on the left by g^{-1}. Note that these multiplications are distributed over each element in the coset: $a(bC) = \{abc : c \in C\}$. □

(i) $\{e\}$ and G are normal subgroups of G.

(ii) $n\mathbb{Z} \trianglelefteq \mathbb{Z}$. $\forall a \in \mathbb{Z}$, we have $a + n\mathbb{Z} = \{a + nk : k \in \mathbb{Z}\} = \{nk + a : k \in \mathbb{Z}\} = n\mathbb{Z} + a$.

(iii) $A_3 \trianglelefteq S_3$.

- $eA_3 = A_3 = A_3e$
- $(1\ 2\ 3)A_3 = A_3 = A_3(1\ 2\ 3)$
- $(1\ 3\ 2)A_3 = A_3 = A_3(1\ 3\ 2)$
- $(1\ 2)A_3 = \{(1\ 2), (2\ 3), (1\ 3)\} = A_3(1\ 2)$

and so on.

Proposition. (i) Any subgroup of an abelian group is normal.

(ii) Any subgroup of index 2 is normal.

Proof. (i) If G is abelian, then $\forall g \in G, \forall n \in N, g^{-1}ng = n \in N$ which is stronger than required.

(ii) If $H \leq G$ with $|G : H| = 2$, then there are only 2 cosets. $H = eH = He$ is one of the two cosets. Since cosets are disjoint, the other coset must be $G \setminus H$. This is true for both left and right cosets. So the other left and right cosets must be equal, so H is normal.

□

Proposition. If $\varphi : G \to H$ is a homomorphism, then $\ker \varphi \trianglelefteq G$.

Proof. We already know $\ker\varphi$ is a subgroup of G. Now we must check it is normal. Given some $k \in \ker\varphi, g \in G$, we want to show that $g^{-1}kg \in \ker\varphi$. We have $\varphi(g^{-1}kg) = \varphi(g^{-1})\varphi(k)\varphi(g) = \varphi(g^{-1})e\varphi(g) = \varphi(g^{-1}g) = \varphi(e) = e$ so $g^{-1}kg \in \ker\varphi$ as required. □

In fact, we will show later that normal subgroups are exactly kernels of homomorphisms and nothing else.

Here is now a less formal explanation of this theorem and its consequences. Consider some subgroup $K \leq G$. There may be some property P that is true for every element of K and false for every other element of G. Then certainly, for example, given $k_1, k_2 \in K$, we know that k_1k_2 has the same property as it is within K. As another example, let $k \in K$ and let $g \in G \setminus K$. Then kg does not have this property, as $kg \notin K$.

We can encapsulate this behaviour by making a homomorphism from the whole group G to some other group—it *doesn't matter where we end up*, just as long as anything with this particular property maps to the new group's identity element. Let $\varphi : G \to H$, where H is some group that we don't really care about (apart from the identity). This means that any element of K, i.e. any element with property P, is mapped to e_H. By the laws of homomorphisms, any product of $k \in K$ with $g \in G \setminus K$ does not give the identity element, so it does not have this property! This is exactly the behaviour we wanted.

If we can find such a homomorphism, then K is the kernel of this homomorphism. Again, the image of this homomorphism is essentially irrelevant; all we care about is which elements map to the identity. Now, note that by the laws of homomorphisms, given some element $g \in G$ and $k \in K$, $\varphi(g^{-1}kg) = \varphi(g^{-1})\varphi(k)\varphi(g)$. But since k has this desired property, the $\varphi(k)$ term vanishes. So we're left with the identity element. This gives us the result that $g^{-1}kg$ must be an element of K, so it must have property P. This is a definition for a normal subgroup, so K must be normal in order for us to be able to find such a homomorphism φ.

As another small aside, a normal subgroup in this context essentially means this: given some element k with property P, the property is preserved when surrounding k with inverses. This is just a 'translation' of a definition of a normal subgroup: $g^{-1}kg \in K$.

(i) $SL_n(\mathbb{R}) \trianglelefteq GL_n(\mathbb{R})$, where $GL_n(\mathbb{R})$ is the group of invertible matrices of dimension n, and where $SL_n(\mathbb{R})$ is the group of matrices of determinant 1. This is because $\det : GL_n(\mathbb{R}) \to \mathbb{R}^*$, and $SL_n(\mathbb{R}) = \ker(\det)$.

(ii) $A_n \trianglelefteq S_n$ as A_n is the kernel of the sign homomorphism. Alternatively, it is an index 2 subgroup so it must be normal.

(iii) $n\mathbb{Z} \trianglelefteq \mathbb{Z}$ as the kernel of $\varphi : \mathbb{Z} \to \mathbb{Z}_n$, where $\varphi(k) = k \bmod n$, or since $\mathbb{Z}$ is abelian.

With this notion of normal subgroups, we can make some progress into categorising small groups.

Proposition. If $|G| = 6$, then $G \cong C_6$ or $G \cong D_6$.

Proof. By Lagrange's Theorem, the possible element orders are 1 (only the identity), 2, 3, 6.

- If there is an element g of order 6, then $G = \langle g \rangle \cong C_6$.
- Otherwise, (again by question 7 on example sheet 1) there must be an element of the group not of order 2, because if we just had elements of order 2 then $|G|$ would have to be a power of 2. So there is an element r of order 3, so $|\langle r \rangle| = 3$, and by Lagrange's Theorem $|G| = 6 = |G : \langle r \rangle| \cdot |\langle r \rangle|$, so $|G : \langle r \rangle| = 2$. So $\langle r \rangle \trianglelefteq G$. There must also be an element s of order 2, since $|G|$ is even (by question 8 from example sheet 1).

  So, what can $s^{-1}rs$ be? Because $\langle r \rangle$ is normal, then $s^{-1}rs \in \langle r \rangle$. So it is either e, r or r^2.

  - If $s^{-1}rs = e$ then $r = e$ #
  - If $s^{-1}rs = r$ then $sr = rs$, and so sr has order $\mathrm{LCM}(\mathrm{ord}\, s, \mathrm{ord}\, r) = \mathrm{LCM}(2, 3) = 6$ #
  - So $s^{-1}rs = r^2$, then $G = \langle r, s \rangle$ with $r^3 = s^2 = e$ and $sr = r^2 s = r^{-1}s$, which are the defining features of D_6.

□

7.2. Motivation for quotients

Let us consider $n\mathbb{Z} \trianglelefteq \mathbb{Z}$. The cosets are $0 + n\mathbb{Z}$, $1 + n\mathbb{Z}, \cdots, (n-1) + n\mathbb{Z}$. These cosets, although they are subsets of $\mathbb{Z}$, behave a lot like the elements of the group $\mathbb{Z}_n$. For example, if we try to define addition between the cosets:

$$(k + n\mathbb{Z}) + (m + n\mathbb{Z}) \coloneqq (k + m) + n\mathbb{Z}$$

which acts like addition modulo $n\mathbb{Z}$. For a general subgroup $H \leq G$, we could try to do the same.

$$g_1 H \cdot g_2 H \coloneqq g_1 g_2 H$$

But we can write the cosets on the left hand side in many ways, as the representation is dependent on the choice of representative for each coset, so this multiplication may not be well defined. We can guarantee that it is well defined (so that we can turn the set of cosets into a group) by ensuring that

$$g_1' H = g_1 H;\ g_2' H = g_2 H \implies g_1' g_2' H = g_1 g_2 H$$

If $g_1' H = g_1 H;\ g_2' H$, then $g_1' = g_1 h_1$ and $g_2' = g_2 h_2$ for some $h_1 h_2 \in H$. So

$$g_1' g_2' H = g_1 h_1 g_2 \underbrace{h_2 H}_{h_2 H = \{h_2 h : h \in H\} = H}$$

So in order to get $g_1'g_2'H = g_1g_2H$, we need $g_1h_1g_2H = g_1g_2H$ for any elements g_1, g_2, h_1 that we choose. Therefore:

$$\begin{aligned} g_1h_1g_2H &= g_1g_2H \\ g_2^{-1}h_1h_2H &= H \\ \text{or } g_2^{-1}h_1g_2 &\in H \ (\forall g_2 \in G, h_1 \in H) \end{aligned}$$

This is an equivalent condition for the subgroup to be normal.

7.3. Quotients

Proposition. Let $N \trianglelefteq G$. The set of (left) cosets of N in G forms a group under the operation $g_1N \cdot g_2N = g_1g_2N$.

Proof. The group operation is well defined as shown above. We now show the group axioms hold.

- (closure) If g_1N, g_2N are cosets, then g_1g_2N is also a coset.
- (identity) $eN = N$
- (inverses) $(gN)^{-1} = g^{-1}N$
- (associativity) Follows from the associativity of G: $(g_1N \cdot g_2N) \cdot g_3N = g_1g_2N \cdot g_3N = g_1g_2g_3N = g_1N \cdot g_2g_3N = g_1N \cdot (g_2N \cdot g_3N)$

□

Definition. If $N \trianglelefteq G$, the group of (left) cosets of N in G is called the quotient group of G by N, written G/N.

This is a nice way of thinking about quotient groups. Imagine you have a group N of some distinct objects n_1, n_2, n_3 and so on. Imagine lining them all up in a row of length $|N|$. Then the cosets of N in G can be thought of as 'translated copies' of N. For example, let the cosets of N in G be N, g_1N, g_2N and so forth. Now, picture these cosets as copies of N, translated downwards on the page, so that they are like multiple rows, and that therefore there we have a grid containing all elements of G. Now, we have formed a rectangle of area $|G|$ out of $|N|$ columns and c rows, where c is the amount of 'copies' of N. Therefore, $c = \frac{|G|}{|N|}$, as the area of a rectangle is width multiplied by height.

Now, given some element in one of the cosets (i.e. in G) we can do some transformation g to take us to another element. But because we made cosets out of a normal subgroup, multiplying by g is the same as swapping some of the rows, then maybe moving around the order of the elements in each row. It keeps the identity of each row consistent—all elements in a given row are transformed to the same output row. Remember that the word 'row' basically means 'coset'.

This means that we can basically forget about the individual elements in these cosets, all that we really care about is how the rows are swapped with each other under a given transformation. Note, the quotient of 5 in 100 is 20, because there are 20 copies of 5 in 100. So the quotient group of N in G is just all the copies of N in G. The group operation is simply the transformation of rows. If we're talking about G/N, ask the question: 'how do the copies of N in G behave'?

7.4. Examples and properties

(i) The cosets of $n\mathbb{Z}$ in $\mathbb{Z}$ give a group that behaves exactly like $\mathbb{Z}_n$. We write $\mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}_n$. In fact, these are the only quotients of $\mathbb{Z}$, as these are the only subgroups of $\mathbb{Z}$.

(ii) $A_3 \trianglelefteq S_3$ gives S_3/A_3 which has only two elements since $|S_3 : A_3| = 2$, so it is isomorphic to C_2. Note that in general, $|G : N| = \left|G/N\right|$.

(iii) If $G = H \times K$, then both H and K are normal subgroups of G. We have $G/H \cong K$ and $G/K \cong H$

(iv) Consider $N \coloneqq \langle r^2 \rangle \trianglelefteq D_8$. We can check that it is normal by trying $r^{-1}r^2r^{-1} \in N$, and also $s^{-1}r^2s = r^{-2} = r^2 \in N$. Since $\langle r, s \rangle = D_8$, and the generators obey this normal subgroup relation, it follows that $g^{-1}ng$ for all $g \in D_8$. We know $|N| = 2$, so $\left|D_8/N\right| = |D_8 : N| = \frac{|D_8|}{|N|}$ by Lagrange's Theorem. So $\left|D_8/N\right| = 4$. We know that any group of order 4 is isomorphic either to C_4 or $C_2 \times C_2$. We can check that the cosets are $D_8/N = \{N, sN, rN, srN\}$ which does not contain an element of order 4, so it is isomorphic to $C_2 \times C_2$.

We now show a non-example using the subgroup $H \coloneqq \langle (1\ 2) \rangle \leq S_3$ which is not normal, e.g. $(1\ 2\ 3)H \neq H(1\ 2\ 3)$. The cosets are

$$H; \quad (1\ 2\ 3)H = \{(1\ 2\ 3), (1\ 3)\}; \quad (1\ 3\ 2)H = \{(1\ 3\ 2), (2\ 3)\}$$

Attempting a multiplication gives

$$(1\ 2\ 3)H \cdot (1\ 3\ 2)H = (1\ 2\ 3)(1\ 3\ 2)H = H$$

but using a different coset representative,

$$(1\ 3)H \cdot (1\ 3\ 2)H = (1\ 3)(1\ 3\ 2)H = (2\ 3)H \neq H$$

so the multiplication is not well defined so we cannot form the quotient.

- We can check that certain properties are inherited into quotient groups from the original group, such as being abelian and being finite.
- Quotients are not subgroups of the original group. They are associated with the original group in a very different way to subgroups—in general, a coset may not even be isomorphic to a subgroup in the group. The example with direct products above was an example that is not true in general.

- With normality, we need to specify in which group the subgroup is normal. For example, if $K \leq N \leq G$, with $K \trianglelefteq N$. This does not imply that $K \trianglelefteq G$, this would require that $g^{-1}Kg = K$ for all elements g in G, but we only have that $n^{-1}Kn = K$ for all elements n in N, which is a weaker condition. Normality is not transitive—for example, $K \trianglelefteq N \trianglelefteq G$ does not imply $K \trianglelefteq G$.
- However, if $N \leq H \leq G$ and $N \trianglelefteq G$, then the weaker condition $N \trianglelefteq H$ is true.

Theorem. Given $N \trianglelefteq G$, the function $\pi : G \to {}^{G}\!/_{N}$, $\pi(g) = gN$ is a surjective homomorphism called the quotient map. We have $\ker \pi = N$.

Proof. We prove that π is a homomorphism. $\pi(g)\pi(h) = gN \cdot hN = (gh)N = \pi(gh)$ as required. Clearly it is surjective we we can create all possible cosets by applying the π function to a coset representative. Also, $\pi(g) = gN = N$ if and only if $g \in N$, so $\ker \pi = N$. □

Therefore, normal subgroups are exactly kernels of homomorphisms. Using the idea of 'properties' for normal subgroups above, the property in question here is 'belonging to N'. Any element of N is in the coset N, which is the identity coset of ${}^{G}\!/_{N}$. Essentially, the first row of this quotient 'grid' (as described above) is N, which acts as the identity element in the ${}^{G}\!/_{N}$ quotient group.

8. Isomorphism theorems

8.1. First isomorphism theorem

Theorem. Let $\varphi : G \to H$ be a homomorphism. Then $G/\ker\varphi \cong \operatorname{Im}\varphi$.

Proof. Define $\overline{\varphi} : G/\ker\varphi \to \operatorname{Im}\varphi$ using $g\ker\varphi \mapsto \varphi(g)$.

- (well-defined) If $g_1\ker\varphi = g_2\ker\varphi$, then $g_1 = g_2k$, for some $k \in \ker\varphi$. Hence $\overline{\varphi}(g_1\ker\varphi) = \varphi(g_1) = \varphi(g_2k) = \varphi(g_2)\varphi(k) = \varphi(g_2) = \overline{\varphi}(g_2\ker\varphi)$.
- (homomorphism) Let $g, g' \in G$. $\overline{\varphi}(g\ker\varphi \cdot g'\ker\varphi) = \overline{\varphi}(gg'\ker\varphi) = \varphi(gg') = \varphi(g)\varphi(g') = \overline{\varphi}(g\ker\varphi)\cdot\overline{\varphi}(g'\ker\varphi)$.
- (surjective) All elements of $\operatorname{Im}\varphi$ are of the form $\varphi(g)$ for some $g \in G$, so clearly surjective.
- (injective) If $\overline{\varphi}(g\ker\varphi) = e = \varphi(g)$ in $\operatorname{Im}\varphi$ then $g \in \ker\varphi$, so $g\ker\varphi = \ker\varphi$.

□

This is a useful way to understand the first isomorphism theorem. Recall that $G/\ker\varphi$ is really asking the question 'how do the copies of $\ker\varphi$ interact in G'? Well, as φ is a homomorphism, it represents some property that is true for members of a normal subgroup N in G, where $N = \ker\varphi$. Now, we can imagine the grid analogy from before, laying out several copies of N as rows. Let's call the group of these rows K.

Now, multiplying together two rows, i.e. two elements from K, we can apply the homomorphism φ to one of the coset representatives for each row to see how the entire row behaves under φ. We know that all coset representatives give equal results, because each element in a given coset gN can be written as $gn, n \in N$, so $\varphi(gn) = \varphi(g)$. So all elements in the rows behave just like their coset representatives under the homomorphism. Further, all the cosets give different outputs under φ—if they gave the same output they'd have to be part of the same coset. So in some sense, each row represents a distinct output for φ. So the quotient group must be isomorphic to the image of the homomorphism.

Here are some examples.

(i) $\det : GL_2(\mathbb{R}) \to \mathbb{R}^*$, $\operatorname{Im}(\det) = \mathbb{R}^*$, $\ker(\det) = SL_2(\mathbb{R})$. Therefore, $GL_2(\mathbb{R})/SL_2(\mathbb{R}) \cong \mathbb{R}^*$.

(ii) Consider the map $\varphi : \mathbb{R} \to \mathbb{C}^*, \varphi(r) = e^{2\pi ir}$. This is a homomorphism because $\varphi(r+s) = e^{2\pi i(r+s)} = e^{2\pi ir}\cdot e^{2\pi is} = \varphi(r)\cdot\varphi(s)$. The image is the unit circle $|z| = 1$, denoted by S_1; the kernel is $\mathbb{Z}$ as $e^{2\pi iz}$ for some $z \in \mathbb{Z}$, the result is 1. Therefore $\mathbb{R}/\mathbb{Z} = S_1$.

8.2. Correspondence theorem

Now, let's try to understand how subgroups behave inside quotient groups.

Theorem. Let $N \trianglelefteq G$. The subgroups of G/N are in bijective correspondence with subgroups of G containing N.

Proof. Given $N \leq M \leq G$, $N \trianglelefteq G$, then $N \trianglelefteq M$ and clearly $M/N \leq G/N$. Conversely, for every subgroup $H \leq G/N$, we can take the preimage of H under the quotient map $\pi : G \to G/N$, i.e. $\pi^{-1}(H) = \{g \in G : gN \in H\}$. This is a subgroup of G:

- (closure) if $g_1, g_2 \in \pi^{-1}(H)$, then $g_1 g_2 N = g_1 N \cdot g_2 N$ where both elements $g_1 N$ and $g_2 N$ are in H. So $g_1 g_2 N \in H$.
- (identity, inverses easy to check)

$\pi^{-1}(H)$ contains N, since $\forall n \in N, nN = N \in H$. Now we can check that for any $N \leq M \leq G$, $\pi^{-1}(M/N) = M$ and for $H \leq G/N$, $\pi^{-1}(H)/N = H$. So the correspondence is bijective (this satisfies the property that ff^{-1} and $f^{-1}f$ are the identity maps on the relevant sets). □

This correspondence preserves lots of structure: for example, indices, normality, containment. Now, let $N \coloneqq \langle (a^2, b) \rangle$. Note that this is normal because we are in an abelian group. Then, according to the above theorem, the subgroup lattice for $C_4 \times C_2/N$ is bijective with the set of paths on the above lattice that terminate with N (i.e. have N as a subgroup). We took the quotient of a group of order 8 by a group of order 2, so N has order 4, so it must be isomorphic to C_4 (as it has only one subgroup isomorphic to C_2 as can be seen in the lattice, so it cannot be $C_2 \times C_2$).

8.3. Second isomorphism theorem

Let $H \leq G$ and $N \trianglelefteq G$, but $N \not\trianglelefteq H$. We can actually still make a normal subgroup of H by intersecting H with N.

Theorem. Let $H \leq G$ and $N \trianglelefteq G$. Then $H \cap N \trianglelefteq H$ and $H/H \cap N \cong HN/N$.

Proof. When $N \trianglelefteq G, H \leq G$, then $HN = \{hn : h \in H, n \in N\}$ is a subgroup of G, and $HN = \langle H, N \rangle$.

Consider the function $\varphi : H \to HN/N, \varphi(h) \coloneqq hN$. This is a well-defined surjective homomorphism. $\varphi(h) = hN = N \iff h \in N$, but also $h \in H$, so $h \in N \cap H$ is the kernel. So by the First Isomorphism Theorem, $H/N \cap H \cong HN/N$ (note that $HN/N \leq G/N$). □

8.4. Third isomorphism theorem

We noted earlier that normality is preserved inside quotient groups. We can say something analogous about quotients.

Theorem. Let $N \leq M \leq G$ such that $N \trianglelefteq G$ and $M \trianglelefteq G$. Then ${}^{M}\!/_{N} \trianglelefteq {}^{G}\!/_{N}$, and ${}^{G/N}\!/_{M/N} = {}^{G}\!/_{M}$.

Proof. Let us define $\varphi : {}^{G}\!/_{N} \to {}^{G}\!/_{M}$ by $\varphi(gN) = gM$. φ is well defined since $N \leq M$, and it is a surjective homomorphism. $\varphi(gN) = gM = M \iff g \in M$, so its kernel is ${}^{M}\!/_{N}$. By the First Isomorphism Theorem, ${}^{G/N}\!/_{M/N} \cong {}^{G}\!/_{M}$. □

Example. (i) Consider $\mathbb{Z}, H = 3\mathbb{Z}, N = 5\mathbb{Z}$. Then by the Second Isomorphism Theorem, we have

$$H \cap N \trianglelefteq H \implies 15\mathbb{Z} \trianglelefteq 3\mathbb{Z}$$

and, since $HN = \langle H, N\rangle = \mathbb{Z}$ as 3 and 5 are coprime,

$${}^{H}\!/_{H \cap N} \cong {}^{HN}\!/_{N} \implies {}^{3\mathbb{Z}}\!/_{15\mathbb{Z}} \cong {}^{\mathbb{Z}}\!/_{5\mathbb{Z}} \cong \mathbb{Z}_5$$

(ii) Let $C_4 = \langle a\rangle$, $C_2 = \langle b\rangle$, $G = C_4 \times C_2$, $N = \langle (a^2, b)\rangle$, $M = \langle (e, b), (a^2, e)\rangle$. Then $N \leq M \leq G$. By the Third Isomorphism Theorem,

$${}^{(C_4 \times C_2)/N}\!/_{M/N} = {}^{C_4 \times C_2}\!/_{M} = C_2$$

8.5. Simple groups

Definition. A group G is simple if its only normal subgroups are trivial $\{e\}$ and G itself.

- C_p where p is prime is a simple group.
- A_5 is simple. A proof of this will be shown later in the course.

9. Group actions

9.1. Definition

For many of the examples of groups that we have encountered, we have identified elements of that group by their effect on some set, for example the symmetric group S_n permuting the set $\{1, \cdots, n\}$, and the Möbius group being functions $\widehat{\mathbb{C}} \to \widehat{\mathbb{C}}$, and the dihedral group D_{2n} being symmetries of an n-gon. While we can study groups purely algebraically, it can be very useful to see how a group acts on other objects.

Definition. Let G be a group, X be a set. An action of G on X is a function $\alpha : G \times X \to X$, written

$$\alpha(g, x) = \alpha_g(x)$$

satisfying:

- $\alpha_g(x) \in X$ (implied by the function's type)
- $\alpha_e(x) = x;\ \forall x \in X$
- $\alpha_g \circ \alpha_h(x) = \alpha_{gh}(x);\ \forall g, h \in G, \forall x \in X$

We can write $G \curvearrowright X$.

Here are some examples.

(i) Take any G, X and define the trivial action $\alpha_g(x) = x$.

(ii) $S_n \curvearrowright \{1, 2, \cdots, n\}$ by permutation.

(iii) $D_{2n} \curvearrowright \{\text{vertices of a regular } n\text{-gon}\}$, and labelling the vertices as 1 to n, we have $D_{2n} \curvearrowright \{1, 2, \cdots, n\}$.

(iv) $\mathcal{M} \curvearrowright \widehat{\mathbb{C}}$ via Möbius maps.

(v) Symmetries of a cube act on the set of vertices, the set of edges, and even (for example) the set of pairs of opposite faces of the cube.

Examples (i), (ii) show that more than one group can act on a given set. Example (iv) shows that one group can act on many sets. Group actions help us deduce information about the group.

Lemma. $\forall g \in G, \alpha_g : X \to X, x \mapsto \alpha_g(x)$ is a bijection.

Proof. We have that

$$\alpha_g(\alpha_{g^{-1}}(x)) = \alpha_{gg^{-1}}(x) = \alpha_e(x) = x$$

Similarly,

$$\alpha_{g^{-1}}(\alpha_g(x)) = \alpha_{g^{-1}g}(x) = \alpha_e(x) = x$$

So the composition $\alpha_g \circ \alpha_{g^{-1}}$ is the identity on X, and $\alpha_{g^{-1}} \circ \alpha_g$ is also the identity on X, so α_g is a bijection. □

We can also define actions by linking G to $\mathrm{Sym}(X)$.

Proposition. Let G be a group, X a set. Then $\alpha : G \times X \to X$ is an action if and only if the function $\rho : G \to \mathrm{Sym}(X)$ where $\rho(g) = \alpha_g$ is a homomorphism.

Proof. α is an action. By the above lemma, α_g is a bijection from $X \to X$. So $\alpha_g \in \mathrm{Sym}(X)$. Now, we want to show that ρ is a homomorphism. $\rho(gh) = \alpha_{gh}$, and for all $x \in X$, $\alpha_{gh}(x) = \alpha_g \circ \alpha_h(x)$, so $\rho(gh) = \alpha_{gh} = \rho(g) \circ \rho(h)$. So ρ is a homomorphism.

In the other direction, given that ρ is a homomorphism $G \to \mathrm{Sym}(X)$, we can define an action $\alpha : G \times X \to X$ by $\alpha(g, x) = \alpha_g(x) \coloneqq \rho(g)(x)$. α is an action because $\alpha_g \circ \alpha_h = \rho(g)\rho(h) = \rho(gh) = \alpha_{gh}$, and the identity element $\rho(e)$ is the identity element in $\mathrm{Sym}(X)$, so $\alpha_e(x) = \rho(e)(x) = x$ as required. □

Sometimes we write $g(x)$ instead of the more verbose $\alpha_g(x)$.

Definition. The kernel of an action $\alpha : G \times X \to X$ is the kernel of the homomorphism $\rho : G \to \mathrm{Sym}(X)$. These are all the elements of G that preserve every element of X.

Note that $G/\ker\rho \cong \mathrm{Im}\,\rho \leq \mathrm{Sym}(X)$. So in particular, if the kernel is trivial, then $G \leq \mathrm{Sym}(X)$.

(i) D_{2n} acting on the vertices $\{1, \cdots, n\}$ of an n-gon has $\ker\rho = \{e\}$. Every non-trivial element of D_{2n} moves at least one vertex. So $D_{2n} \leq S_n$ by the First Isomorphism Theorem.

(ii) Let G be symmetries of a cube, and consider $X = \{\text{unordered pairs of opposite faces}\}$. Then $|X| = 3$ as there are three unordered pairs of opposite faces. So $\rho : G \to S_3$. Clearly there are symmetries of the cube that realise all the permutations of X, so ρ is surjective. So $G/\ker\rho \cong S_3$. Note that there are clearly non-trivial symmetries (e.g. reflection) that preserve X, so the kernel is non-trivial.

Definition. An action $G \curvearrowright X$ is called faithful if $\ker\rho = \{e\}$.

Then G is isomorphic to a subgroup of $\mathrm{Sym}\,X$ by the First Isomorphism Theorem.

9.2. Orbits and stabilisers

Which elements of X can we 'get to' from a certain $x \in X$ using the action of G?

Definition. Let $G \curvearrowright X$, $x \in X$. The orbit of x is

$$\mathrm{Orb}(x) = G(x) \coloneqq \{g(x) : g \in G\} \subseteq X$$

Which group elements leave a given x unchanged?

Definition. The stabiliser of x is defined by

$$\mathrm{Stab}(x) = G_x \coloneqq \{g \in G : g(x) = x\} \subseteq G$$

Definition. An action is transitive if $\mathrm{Orb}(x) = X$, i.e. we can get to any element from any other element.

As an example, let $G = S_3$. Then we could say, for example, $G \curvearrowright \{1, 2, 3, 4\}$.

- $\mathrm{Orb}(1) = \mathrm{Orb}(2) = \mathrm{Orb}(3) = \{1, 2, 3\}$
- $\mathrm{Orb}(4) = \{4\}$
- $\mathrm{Stab}(1) = \{e, (2\ 3)\}$
- $\mathrm{Stab}(2) = \{e, (1\ 3)\}$
- $\mathrm{Stab}(3) = \{e, (1\ 2)\}$
- $\mathrm{Stab}(4) = G$

Lemma. For any $x \in X$, $\mathrm{Stab}(x) \leq G$.

Proof. Associativity is inherited.

- (closure) g, $h \in \mathrm{Stab}(x)$ implies that $(gh)(x) = g(h(x)) = g(x) = x$ so $gh \in \mathrm{Stab}(x)$.
- (identity) $e(x) = x$ by definition, so $e \in \mathrm{Stab}(x)$.
- (inverses) if $g \in \mathrm{Stab}(x)$ then $g(x) = x$, and therefore $x = g^{-1}(x)$, so $g^{-1} \in \mathrm{Stab}(x)$.

□

Recall from Numbers and Sets: a partition of a set X is a set of subsets of X such that each $x \in X$ belongs to exactly one subset in the partition.

Lemma. Let $G \curvearrowright X$. Then the orbits partition X.

Proof.

- Firstly, for any $x \in X$, $x \in \mathrm{Orb}(x)$. So the union of all orbits is X.
- Suppose that the orbits are not all disjoint. Let $z \in \mathrm{Orb}(x) \cap \mathrm{Orb}(y)$. Then $\exists g_1 \in G$ such that $g_1(x) = z$, and also $\exists g_2 \in G$ such that $g_2(y) = z$, i.e. $y = g_2^{-1}(z)$. So $y = g_2^{-1} g_1(x)$. Thus, for any $g \in G$, $g(y) = g g_2^{-1} g_1(x) \in \mathrm{Orb}(x)$ so $\mathrm{Orb}(y) \subseteq \mathrm{Orb}(x)$. Vice versa, $\mathrm{Orb}(x) \subseteq \mathrm{Orb}(y)$, so $\mathrm{Orb}(x) = \mathrm{Orb}(y)$. Thus orbits are either disjoint or equal.

□

Recall the proof of disjoint cycle notation for $\sigma \in S_n$: we were really finding the orbits in $\{1, 2, \cdots, n\}$ under $\langle \sigma \rangle$, which are disjoint. Note that the sizes of orbits can be different (unlike cosets, where the sizes are always the same).

Theorem (Orbit-Stabiliser Theorem)**.** Let $G \curvearrowright X$, G finite. Then for any $x \in X$,

$$|G| = |\mathrm{Orb}\, x| \cdot |\mathrm{Stab}\, x|$$

Proof. $g(x) = h(x) \iff h^{-1}g(x) = x \iff h^{-1}g \in \mathrm{Stab}(x)$. By a previous result, this statement is true if and only if $g\,\mathrm{Stab}(x) = h\,\mathrm{Stab}(x)$ as cosets. So distinct points in the orbit of x are in bijection with distinct cosets of the stabiliser. So $|\mathrm{Orb}\,x| = |G : \mathrm{Stab}\,x|$ and the result follows. □

In particular, notice that all elements in a given coset $g\,\mathrm{Stab}(x)$ do the same thing to x as g: an element of this coset has the form gh where $h \in \mathrm{Stab}(x)$. Then $gh(x) = g(x)$.

This theorem is very powerful, we can use it for investigating groups further. For example, we can construct another proof that $|D_{2n}| = 2n$ using the Orbit-Stabiliser theorem. D_{2n} acts transitively on $\{1, 2, \cdots, n\}$ so $|\mathrm{Orb}(1)| = n$. $|\mathrm{Stab}(1)| = 2$ because only the identity and the reflection through this point stabilise the point. So $|D_{2n}| = 2n$.

9.3. The Platonic solids

Example (tetrahedron)**.** A tetrahedron has 4 faces (regular, equilateral triangles), 4 vertices, and 6 edges. We will label the vertices $1, 2, 3, 4$. Let G be the group of symmetries of the tetrahedron. Clearly G acts transitively on the vertices (we can get from any vertex to any other through a symmetry). There is no non-trivial symmetry that fixes all the vertices, so $\rho : G \to S_4$ is an injective homomorphism.

$\mathrm{Orb}(1) = \{1, 2, 3, 4\}$ as G is transitive. $\mathrm{Stab}(1) =$ all of the symmetries of the face $\{2, 3, 4\}$, i.e.

$$\mathrm{Stab}(1) = \{e, (2\,3\,4), (2\,4\,3), (2\,3), (3\,4), (2\,4)\} \cong D_6 \cong S_3$$

Then $|G| = |\mathrm{Orb}(1)| \cdot |\mathrm{Stab}(1)| = 4 \cdot 6 = 24 = |S_4|$. Since $G \leq S_4$ and their orders match, $G = S_4$.

Now let G^+ be the subgroup of G formed only of the rotations in G. Again, $\mathrm{Orb}(1) = \{1, 2, 3, 4\}$. Now, $\mathrm{Stab}(1) = \{e, (2\,3\,4), (2\,4\,3)\}$. So $|G^+| = |\mathrm{Orb}(1)| \cdot |\mathrm{Stab}(1)| = 4 \cdot 3 = 12$. Since $G^+ \leq G = S_4$, then we know that $G^+ = A_4$. Indeed, we have all 3-cycles (since these are rotations through vertices), and all elements of the form $(1\,2)(3\,4)$ since these are rotations in the axis through the midpoints of opposite edges.

Example (cube)**.** We label the vertices from 1 to 8 here, and let G be the group of symmetries of the cube acting on the vertices. Clearly the action is transitive, so $|\mathrm{Orb}(1)| = 8$. $\mathrm{Stab}(1) = \{e, r, r^2, s_1, s_2, s_3\}$ where r and r^2 are the rotations through the axis that passes through vertex 1, and where the s_i are the reflections through three planes containing vertex 1. So $|\mathrm{Stab}(1)| = 6$, so $|G| = 48$. We will determine this group completely later on.

Let G^+ be the subgroup of G containing the rotations of G. Then, the action is still transitive, and $|\mathrm{Stab}(1)| = 3$, since we are only looking at the rotations. So $|G^+| = 24$.

Now, to determine this group, let G^+ act on the 4 diagonals in the cube. This gives us a homomorphism $\rho : G^+ \to S_4$. We have all 4-cycles in $\mathrm{Im}\,\rho$, since rotating the cube by quarter turns through the x, y, z axes permute the diagonals in this way. We also have all transpositions (2-cycles) by rotating the cube by a half turn through the plane of two diagonals. In

example sheet 2, we prove that $\langle (1\ 2), (1\ 2\ 3\ 4)\rangle = S_4$, so ρ is surjective. But since the orders match, $G^+ \cong S_4$.

The aforementioned solids are two of the five Platonic solids; the solids in $\mathbb{R}^3$ that have polygonal faces, straight edges and vertices such that their group of symmetries acts transitively on triples (vertex, incident edge, incident face). These are therefore particularly symmetric solids for having this transitive action. The other solids are the octahedron, dodecahedron and icosahedron. The cube and octahedron are 'dual', i.e. they can be inscribed in each other with vertices placed in the centres of faces. The dodecahedron and icosahedron are also dual. Dual solids have the same symmetry groups, so there are only three symmetry groups of Platonic solids.

9.4. Cauchy's theorem

Theorem. Let G be a finite group, p a prime such that $p \mid |G|$. Then G has an element of order p.

Proof. Let $p \mid |G|$. Consider $G^p = G \times G \times \cdots \times G$. This is the group formed of p-tuples of elements of G with coordinate-wise composition. Consider the subset $X \subseteq G^p$, given by

$$X \coloneqq \{(g_1, g_2, \cdots, g_p) \in G^p : g_1 g_2 \cdots g_p = e\}$$

which can be described as 'p-tuples multiplying to e'. Note that if $g \in G$ has order p, then $(g, g, \cdots, g) \in X$; and that if $(g, g, \cdots, g) \in X$ where $g \neq e$, then g has order p.

Now take a cyclic group $C_p = \langle a \rangle$, and let $C_p \curvearrowright X$ by 'cycling':

$$a(g_1, g_2, \cdots, g_p) = (g_2, \cdots, g_p, g_1)$$

This really is an action:

- If $g_1 g_2 \cdots g_p = e$, then $e = g_1^{-1} e g_1 = g_1^{-1} g_1 g_2 \cdots g_p g_1 = g_2 \cdots g_p g_1$ as required. Of course, this applies inductively for any power of a.
- $e(g_1, \cdots, g_p) = (g_1, \cdots, g_p)$ as required.
- $a^k(g_1, \cdots, g_p) = (g_{k+1}, \cdots, g_k) = a \cdot a \cdots a(g_1, \cdots, g_k)$.

Since orbits partition X, the sum of the sizes of the orbits must be $|X|$. We know that $|X| = |G|^{p-1}$, since all choices of g_i are free apart from the last one, which must be the inverse of the product of the other elements. So we have $p-1$ choices of $|G|$ elements, so $|X| = |G|^{p-1}$.

So since $p \mid |G|$, then $p \mid |X|$. By the Orbit-Stabiliser theorem:

$$\left|\mathrm{Orb}((g_1, \cdots, g_p))\right| \cdot \left|\mathrm{Stab}((g_1, \cdots, g_p))\right| = \left|C_p\right| = p$$

So any orbit has size 1 or p, and they sum to $|X| = pk$ for some $k \in \mathbb{N}$. So

$$|X| = pk = \sum_{\text{orbits of size } 1} 1 + \sum_{\text{orbits of size } p} p$$

Clearly, $|\mathrm{Orb}((e, e, \cdots, e))| = 1$. So there must be some other orbits of size 1, so that p divides the amount of orbits of size 1. But orbits of size 1 must be of the form $\mathrm{Orb}((g, g, \cdots, g))$ in order to have the same form under the action of a. So there exists some $g \neq e \in G$ such that $(g, g, \cdots g) \in X$, i.e. $g^p = e$, so $o(g) = p$. □

9.5. Left regular action

Lemma. Let G be a group. G acts on itself by left multiplication. This action is faithful and transitive.

Proof.
- For any $g, x \in G$, $gx \in G$
- $e(x) = e \cdot x = x$
- $(g_1 g_2)x = g_1(g_2 x)$

So it really is an action. It is faithful because $g(x) = gx = x$ implies $g = e$. It is transitive, because given any $x, y \in G$, the action $g = yx^{-1}$ gives $g(x) = y$. □

Definition. This left-multiplication action of a group on itself is known as the left regular action.

9.6. Cayley's theorem

Theorem. Every group is isomorphic to a subgroup of a symmetric group.

Proof. Let $G \curvearrowright G$ by the left regular action. This gives a homomorphism $\rho : G \to \mathrm{Sym}(G)$, with $\ker \rho = \{e\}$ since the action is faithful. So, by the First Isomorphism Theorem, $G/\ker \rho = G \cong \mathrm{Im}\, \rho \leq \mathrm{Sym}(G)$. □

Proposition. Let $H \leq G$. Then G acts on the set of left cosets of H in G by left multiplication, and this action is transitive. (This is called the 'left coset action').

Proof. We check the conditions for actions.

- $g(g_1 H) = g g_1 H$, so $g(g_1 H)$ is a left coset.
- $e(g_1 G) = e g_1 H = g_1 H$
- $(gg')(g_1 H) = g g' g_1 H = g(g'(g_1 H))$

So this is an action. Given two cosets $g_1 H$ and $g_2 H$, the element $(g_1 g_2^{-1})$ acts on $g_2 H$ to give $g_1 H$, so it is transitive. □

Note:

- This is the left regular action if $H = \{e\}$.

- This induces actions of G on its quotient groups G/N.

10. Conjugation

10.1. Conjugation actions

Definition. Given $g, h \in G$, the element hgh^{-1} is the conjugate of g by h.

We should think of conjugate elements as doing the same thing but from different 'points of view'—we change perspective by doing h^{-1}, then do the action g, then reset the perspective back to normal using h.

Here is an example using D_{10}, where the vertices of the regular pentagon are $v_1 \dots v_5$ clockwise. Consider the conjugates s and rsr^{-1}, where s is a reflection through v_1 and the centre, and r is a rotation by $\frac{2\pi}{5}$ clockwise. So rsr^{-1} ends up being just a reflection through v_2 and the centre. So the result of conjugating the reflection by a rotation is still a reflection, just from a different point of view.

Another example is in matrix groups such as $GL_n(\mathbb{R})$ where a conjugate matrix represents the same transformation but with respect to a different basis. This will be covered in more detail later.

As a general principle, conjugate elements can be expected to have similar properties. We will now prove some of these such properties.

Proposition. A group G acts on itself by conjugation.

Proof.
- $g(x) = gxg^{-1} \in G$ for any $g, x \in G$
- $e(x) = exe^{-1} = x$ for any $x \in G$
- $g(h(x)) = ghxh^{-1}g^{-1} = (gh)(x)$

□

Definition. The kernel, orbits and stabilisers have special names:

- The kernel of the conjugation action of G on itself is the centre $Z(G)$:

  $$Z(G) \coloneqq \{g \in G : \forall h \in G,\ ghg^{-1} = h \iff gh = hg\}$$

  In less formal terms, $Z(G)$ is the set of 'elements that commute with everything'.

- An orbit of this action is called a conjugacy class:

  $$\mathrm{ccl}(h) \coloneqq \{ghg^{-1} : g \in G\}$$

  Sometimes this is written $\mathrm{ccl}_G(h)$ to clarify which group we're working on.

- Stabilisers are called centralisers:

  $$C_G(h) \coloneqq \{g \in G : ghg^{-1} = h \iff gh = hg\}$$

  This is the set of 'elements that commute with h'.

Exercise: $Z(G) = \bigcap_{h \in G} C_G(h)$.

Definition. If $H \leq G, g \in G$, then the conjugate of H by g is:

$$gHg^{-1} = \{ghg^{-1} : h \in H\}$$

Proposition. Let $H \leq G, g \in G$. Then gHg^{-1} is also a subgroup of G.

Proof. We check the group axioms.

- (closure) If $gh_1g^{-1}, gh_2g^{-1} \in gHg^{-1}$, then
$$(gh_1g^{-1})(gh_2g^{-1}) = gh_1(g^{-1}g)h_2g^{-1} = g(h_1h_2)g^{-1} \in gHg^{-1}$$
- (identity) $geg^{-1} = e \in gHg^{-1}$
- (inverses) Given $ghg^{-1} \in gHg^{-1}$, the inverse is $gh^{-1}g^{-1}$, which of course is an element of gHg^{-1}.

□

Note that gHg^{-1} is isomorphic to H (proof as exercise).

Proposition. A group G acts by conjugation on the set of its subgroups. The singleton orbits are the normal subgroups.

Proof as exercise. (Recall that $N \trianglelefteq G \iff \forall g \in G,\ gNg^{-1} = N$, which is the same as being stable under conjugation)

Proposition. Normal subgroups are those subgroups that are unions of conjugacy classes. Recall that $\mathrm{ccl}(h) = \{ghg^{-1} : g \in G\}$.

Proof. Let $N \trianglelefteq G$. Then if $h \in N$, then $ghg^{-1} \in N$ for all $g \in G$ because N is a normal subgroup. So $\mathrm{ccl}(h) \subseteq N$. So N is a union of conjugacy classes of its elements;

$$N = \bigcup_{h \in N} \mathrm{ccl}(h)$$

Conversely, if H is a subgroup that is a union of conjugacy classes, then $\forall g \in G, \forall h \in H$, we have $ghg^{-1} \in H$. So $H \trianglelefteq G$. □

As an example, consider $A_3 = \{e, (1\ 2\ 3), (1\ 3\ 2)\} \trianglelefteq S_3$. Now, $A_3 = \{e\} \sqcup \{(1\ 2\ 3), (1\ 3\ 2)\}$. Note that $(1\ 2\ 3), (1\ 3\ 2)$ are conjugates in S_3 but they are not conjugates in A_3.

10.2. Conjugation in symmetric groups

Lemma. Given a k-cycle $(a_1 \dots a_k)$ and $\sigma \in S_n$, we have

$$\sigma(a_1 \dots a_k)\sigma^{-1} = (\sigma(a_1) \dots \sigma(a_k))$$

Proof. Let us apply the left hand side transformation to $\sigma(a_i)$.

$$\sigma(a_1 \dots a_k)\sigma^{-1}\sigma(a_i) = \sigma(a_1 \dots a_k)(a_i) = \sigma(a_{i+1 \bmod k})$$

Now let us consider the effect of the transformation on $\sigma(b)$ for $b \neq a_i$.

$$\sigma(a_1 \dots a_k)\sigma^{-1}\sigma(b) = \sigma(a_1 \dots a_k)(b) = \sigma(b)$$

So these are unchanged. Therefore, the left hand side is equal to the right hand side. □

Proposition. Two elements of S_n are conjugate (in S_n, i.e. via a conjugation by some element in S_n) if and only if they have the same cycle type.

Proof. Two elements that are conjugate will have the same cycle type: given $\sigma \in S_n$, we can write σ as a product of disjoint cycles, say $\sigma = \sigma_1 \dots \sigma_m$. Then if $\rho \in S_n$, $\rho\sigma\rho^{-1} = \rho\sigma_1\rho^{-1}\rho\sigma_2\rho^{-1} \dots \rho\sigma_m\rho^{-1}$ which is a product of the conjugates of the cycles. By the above lemma, the conjugate of a k-cycle is a k-cycle, and because ρ is bijective the $\rho\sigma_i\rho^{-1}$ are all disjoint, so we retain the cycle type of σ under conjugation in S_n.

Conversely, if σ and τ have the same cycle type, then we can write

$$\sigma = (a_1 \dots a_{k_1})(a_{k_1+1} \dots a_{k_2}) \dots$$

$$\tau = (b_1 \dots b_{k_1})(b_{k_1+1} \dots b_{k_2}) \dots$$

in disjoint cycle notation, including singletons. Then all of $\{1, \dots, n\}$ appear in both σ and τ. Then, setting ρ to be defined by $\rho(a_i) = b_i$, which is indeed a permutation, we obtain $\rho\sigma\rho^{-1} = \tau$. □

Let us consider the conjugacy classes of S_4. We can compute the size of C_{S_4} using the orbit-stabiliser theorem; the conjugacy class is the orbit of a particular point under conjugation.

cycle type	example element	size of ccl	size of C_{S_4}	sign
1, 1, 1, 1	e	1	24	+1
2, 1, 1	(1 2)	6	4	−1
2, 2	(1 2)(3 4)	3	8	+1
3, 1	(1 2 3)	8	3	+1
4	(1 2 3 4)	6	4	−1

From this, we can compute all normal subgroups of S_4, since normal subgroups:

- must contain e

- must be a union of conjugacy classes
- must have an order that divides $|S_4| = 24$

To check all possibilities, we will look through all divisors of 24, and check whether we can form a union of conjugacy classes.

- (1) $\{e\}$
- (2) impossible, no conjugacy classes have orders which add to 2
- (3) impossible
- (4) $3 + 1 = 4$ so we have

  $$\{e, (1\,2)(3\,4), (1\,3)(2\,4), (1\,4)(2\,3)\} \cong C_2 \times C_2$$

  This subgroup is often referred to as V_4, the Klein four group.
- (6) impossible
- (8) impossible
- (12) $1 + 3 + 8 = 12$ so we have

  $$\{e, (1\,2)(3\,4), (1\,3)(2\,4), (1\,4)(2\,3), (1\,2\,3), (1\,3\,2), (1\,2\,4), (1\,4\,2), (1\,3\,4), (1\,4\,3), (2\,3\,4), (2\,4\,3)\} = A_4$$

- (24) $S_4 \trianglelefteq S_4$.

So all possible quotients of S_4 are:

- $S_4/\{e\} \cong S_4$
- $S_4/V_4 = \{V_4, (1\,2)V_4, (1\,3)V_4, (2\,3)V_4, (1\,2\,3)V_4, (1\,3\,2)V_4\} \cong S_3$
- $S_4/A_4 \cong C_2$
- $S_4/S_4 \cong \{e\}$

Exercise: repeat with S_5.

10.3. Conjugation in alternating groups

Note that

$$\operatorname{ccl}_{S_n}(\sigma) = \{\tau\sigma\tau^{-1} : \tau \in S_n\}$$
$$\operatorname{ccl}_{A_n}(\sigma) = \{\tau\sigma\tau^{-1} : \tau \in A_n\}$$

So clearly $\operatorname{ccl}_{A_n}(\sigma) \subseteq \operatorname{ccl}_{S_n}(\sigma)$ since $A_n \subseteq S_n$. But elements that are conjugate in S_n may not be conjugate in A_n, for example $(1\,2\,3)$ and $(1\,3\,2)$ in S_3 and A_3.

Some conjugacy classes of S_n are split into smaller conjugacy classes in A_n, since some elements require elements of $S_n \setminus A_n$ to conjugate with each other. By the orbit-stabiliser theorem,

$$|S_n| = |\mathrm{ccl}_{S_n}(\sigma)| \cdot |C_{S_n}(\sigma)|$$
$$|A_n| = |\mathrm{ccl}_{A_n}(\sigma)| \cdot |C_{A_n}(\sigma)|$$

But $|S_n| = 2|A_n|$, and $|\mathrm{ccl}_{S_n}(\sigma)| \geq |\mathrm{ccl}_{A_n}(\sigma)|$. So either:

- $\mathrm{ccl}_{S_n}(\sigma) = \mathrm{ccl}_{A_n}(\sigma)$ and $|C_{S_n}(\sigma)| = 2|C_{A_n}(\sigma)|$, or
- $|\mathrm{ccl}_{S_n}(\sigma)| = 2|\mathrm{ccl}_{A_n}(\sigma)|$ and $C_{S_n}(\sigma) = C_{A_n}(\sigma)$

Definition. When $|\mathrm{ccl}_{S_n}(\sigma)| = 2|\mathrm{ccl}_{A_n}(\sigma)|$, we say that the conjugacy class of σ splits in A_n.

When does a conjugacy class split in A_n?

Proposition. The conjugacy class of $\sigma \in A_n$ splits in A_n if and only if there are no odd permutations that commute with σ.

Proof.

$$|\mathrm{ccl}_{S_n}(\sigma)| = 2|\mathrm{ccl}_{A_n}(\sigma)| \iff C_{S_n}(\sigma) = C_{A_n}(\sigma)$$
$$C_{A_n}(\sigma) = A_n \cap C_{S_n}(\sigma)$$
$$A_n \cap C_{S_n}(\sigma) = C_{S_n}(\sigma) \iff C_{S_n}(\sigma) \text{ contains no odd elements}$$

So no odd permutation is in this centraliser. □

Let us consider an example for conjugacy classes in A_4.

cycle type	example element	odd element in C_{S_4}?	size of ccl_{S_4}	size of ccl_{A_4}
1, 1, 1, 1	e	yes, e.g. (1 2)	1	1
2, 2	(1 2)(3 4)	yes, e.g. (1 2)	3	3
3, 1	(1 2 3)	no	8	two classes of size 4

There is no odd element in $C_{S_4}(1\ 2\ 3)$ because $|C_{S_4}(1\ 2\ 3)| = 3$ and clearly C_{S_4} contains $\langle(1\ 2\ 3)\rangle$, which is a set of 3 elements, so $C_{S_4} = \langle(1\ 2\ 3)\rangle$ which are all even elements.

Let us now consider conjugacy classes in A_5.

cycle type	example element	odd element in C_{S_5}?	size of ccl_{S_5}	size of ccl_{A_5}
1, 1, 1, 1, 1	e	yes, e.g. (1 2)	1	1
2, 2, 1	(1 2)(3 4)	yes, e.g. (1 2)	15	15
3, 1, 1	(1 2 3)	yes, e.g. (4 5)	20	20
5	(1 2 3 4 5)	no	24	two classes of size 12

Lemma. $C_{S_5}(1\ 2\ 3\ 4\ 5) = \langle(1\ 2\ 3\ 4\ 5)\rangle$.

Proof.

$$\left|\mathrm{ccl}_{S_5}(1\,2\,3\,4\,5)\right| = \frac{5 \cdot 4 \cdot 3 \cdot 2}{5} = 24$$

By the orbit-stabiliser theorem,

$$|S_5| = 120 = 24\left|C_{S_5}(1\,2\,3\,4\,5)\right| \implies \left|C_{S_5}(1\,2\,3\,4\,5)\right| = 5$$

Clearly $\langle(1\,2\,3\,4\,5)\rangle \subseteq C_{S_5}(1\,2\,3\,4\,5)$ so $\langle(1\,2\,3\,4\,5)\rangle = C_{S_5}(1\,2\,3\,4\,5)$. Note, this contains only even elements. □

Theorem. A_5 is a simple group.

Proof. Normal subgroups must be unions of conjugacy classes, they must contain e, and their order must divide the order of the group $|A_5| = 60$. The sizes of conjugacy classes we have are $1, 15, 20, 12, 12$ from the example above. The only ways of adding 1 plus some of the other numbers to get a divisor of 60 are

- (1) which can only be the trivial subgroup
- $(1 + 15 + 20 + 12 + 12 = 60)$ which can only be the group itself

So those are the only possible normal subgroups, so it is simple. □

Remark. All A_n for $n \geq 5$ are simple.

11. Action of the Möbius group

11.1. Introduction

We can now study the action of the Möbius group $\mathcal{M}$, which is the group of Möbius maps

$$f : \widehat{\mathbb{C}} \to \widehat{\mathbb{C}}; \quad f(z) = \frac{az+b}{cz+d}; \quad a,b,c,d \in \mathbb{C}; \quad ad-bc \neq 0; \quad \frac{1}{0} = \infty; \quad \frac{1}{\infty} = 0$$

Remark. The above definition defines an action $M \curvearrowright \widehat{\mathbb{C}}$.

Proposition. The action $M \curvearrowright \widehat{\mathbb{C}}$ is faithful (the only elements acting as the identity are the identity), and so $\mathcal{M} \leq \mathrm{Sym}(\widehat{\mathbb{C}})$.

Proof. Consider $\rho : \mathcal{M} \to \mathrm{Sym}(\widehat{\mathbb{C}})$ given by $\rho(f)(z) = f(z)$. Then if $\rho(f) = e_{\mathrm{Sym}(\widehat{\mathbb{C}})}$ (the function $z \mapsto z$) then f is the identity $e_{\mathcal{M}}$. So ρ is injective and the action is faithful. □

Definition. A fixed point of a Möbius map f is a point z such that $f(z) = z$.

Theorem. A Möbius map with at least three fixed points is the identity.

Proof. Let $f(z) = \frac{az+b}{cz+d}$ have at least three fixed points.

- If ∞ is not a fixed point, then the equation $\frac{az+b}{cz+d} = z$ is true for at least three complex numbers. Rewritten,
$$cz^2 + (d-a)z - b = 0$$
By the fundamental theorem of algebra, this can only have at most two distinct roots. So we must have $c = b = 0, d = a$, i.e. $f(z) = z$.
- If ∞ is a fixed point, then $\frac{a\infty+b}{c\infty+d} = \frac{a}{c} = \infty$ so $c = 0$. So for the other two fixed points, $\frac{az+b}{d} = z$ for at least two complex numbers. Rewritten,
$$(a-d)z + b = 0$$
By the fundamental theorem of algebra, this can only have one root. So we must have $a = d, b = 0$, i.e. $f(z) = z$.

□

Corollary. If two Möbius maps coincide on three distinct points in $\widehat{\mathbb{C}}$, then they must be equal.

Proof. Let $f, g \in \mathcal{M}$ be such that $f(z_1) = g(z_1)$, $f(z_2) = g(z_2)$, $f(z_3) = g(z_3)$ for three distinct points $z_1, z_2, z_3 \in \widehat{\mathbb{C}}$. Then $g^{-1}f(z_i) = z_i$ for the same three distinct points. So $g^{-1}f$ is the identity by the theorem above, so $g = f$. □

In less formal words, we can say 'knowing what a Möbius map does to 3 points determines it'.

11.2. Constructing Möbius maps

Theorem. There is a unique Möbius map sending any three distinct points of $\widehat{\mathbb{C}}$ to any three distinct points of $\widehat{\mathbb{C}}$.

Proof. Let the map send distinct points z_1, z_2, z_3 to w_1, w_2, w_3. Suppose first that $w_1 = 0$, $w_2 = 1, w_3 = \infty$. Then

$$f(z) = \frac{(z_2 - z_3)(z - z_1)}{(z_2 - z_1)(z - z_3)}$$

satisfies this requirement. There is a special case if one of the z_i is infinity. Then

$$z_1 = \infty \implies f(z) = \frac{z_2 - z_3}{z - z_3}$$

$$z_2 = \infty \implies f(z) = \frac{z - z_1}{z - z_3}$$

$$z_3 = \infty \implies f(z) = \frac{z - z_1}{z_2 - z_1}$$

Thus we can find a function f_1 sending (z_1, z_2, z_3) to $(0, 1, \infty)$. We can also find a function f_2 sending (w_1, w_2, w_3) to $(0, 1, \infty)$. So surely $f_2^{-1} \circ f_1$ is a map first sending (z_1, z_2, z_3) to $(0, 1, \infty)$, and then from $(0, 1, \infty)$ to (w_1, w_2, w_3), which is the required map. It is unique because of the corollary at the end of the previous section. □

On example sheet 2, it was proven that a conjugate hfh^{-1} of a Möbius map f satisfies:

- $\operatorname{ord}(hfh^{-1}) = \operatorname{ord}(f)$ since $(hfh^{-1})^n = hf^nh^{-1}$
- $f(z) = z \iff hfh^{-1}(h(z)) = h(z)$. In particular, the number of fixed points of a conjugate is the same as that of the original map. The following theorem is a partial converse to this observation.

Theorem. Every non-identity $f \in \mathcal{M}$ has either one or two fixed points.

- If f has one fixed point, then it is conjugate to the map $z \mapsto z + 1$; and
- If f has two fixed points, then it is conjugate to the map $z \mapsto az$ for some $a \in \mathbb{C} \setminus \{0\}$.

Proof. We know that a non-identity element has at most two fixed points, so it suffices to show that it cannot have zero fixed points. If $f(z) = \frac{az+b}{cz+d}$, we can consider the quadratic

$$cz^2 + (d - a)z - b = 0$$

arising from $f(z) = z$. This quadratic must have at least one solution in the complex plane, so in $\mathbb{C}$ there must be at least one fixed point.

- If f has exactly one fixed point z_0, then let us choose some point $z_1 \in \mathbb{C}$ which is not fixed by f. Then the triple $(z_1, f(z_1), z_0)$ are all distinct. So there is some $g \in \mathcal{M}$ such that $(z_1, f(z_1), z_0) \mapsto (0, 1, \infty)$. Now, let us consider gfg^{-1}. We have
  - $0 \mapsto z_1 \mapsto f(z_1) \mapsto 1$
  - $\infty \mapsto z_0 \mapsto z_0 \mapsto \infty$

  So gfg^{-1} has the form $z \mapsto az + 1$ for some complex number a (proof as exercise). If $a \neq 1$ then $\frac{1}{1-a}$ is a fixed point, but this is a contradiction since ∞ can be the only fixed point. So gfg^{-1} has the form $z \mapsto z + 1$, so f is conjugate (via g) to $z \mapsto z + 1$ as required.
- If f has exactly two fixed points z_0 and z_1, then let g be any Möbius map which sends $(z_0, z_1) \mapsto (0, \infty)$. So gfg^{-1} sends:
  - $0 \mapsto z_0 \mapsto z_0 \mapsto 0$
  - $\infty \mapsto z_1 \mapsto z_1 \mapsto \infty$

  So gfg^{-1} fixes zero and infinity. So gfg^{-1} must have the form $z \mapsto az$ where $a = gfg^{-1}(1)$ as required.

□

We can use this to efficiently work out f^n for $f \in \mathcal{M}$. We can quickly see that $gf^ng^{-1} = (gfg^{-1})^n$ will be either

- $z \mapsto z + n$ if f has one fixed point; and
- $z \mapsto a^n z$ if f has two fixed points.

11.3. Geometric properties of Möbius maps

We have seen that the image under $f \in \mathcal{M}$ of three points in $\widehat{\mathbb{C}}$ uniquely determine f. Three points also uniquely define lines and circles in $\widehat{\mathbb{C}}$.

- The equation of a circle with centre $b \in \mathbb{C}$ and radius $r \in \mathbb{R}, r > 0$ is $|z - b| = r$. We can rewrite this as

$$\begin{aligned} & |z-b|^2 - r^2 = 0 \\ \iff & (z-b)\overline{(z-b)} - r^2 = 0 \\ \iff & z\overline{z} - \overline{b}z - b\overline{z} + b\overline{b} - r^2 = 0 \end{aligned} \tag{$*$}$$

- The equation of a straight line in $\mathbb{C}$ is $a\,\mathrm{Re}(z) + b\,\mathrm{Im}(z) = c$, similar to the implicit

form of a straight line in $\mathbb{R}^2$, $ax + by = c$. Expanded, we have

$$\begin{aligned} a\,\mathrm{Re}(z) + b\,\mathrm{Im}(z) &= c \\ a\frac{z+\overline{z}}{2} + b\frac{z-\overline{z}}{2i} &= c \\ \frac{1}{2}\left[a(z+\overline{z}) - bi(z-\overline{z})\right] - c &= 0 \\ \frac{1}{2}\left[z(a-bi) + \overline{z}(a+bi)\right] - c &= 0 \\ \overline{\frac{a+ib}{2}}z + \frac{a+ib}{2}\overline{z} - c &= 0 \end{aligned} \tag{$\dagger$}$$

For a straight line in $\widehat{\mathbb{C}}$, we also consider that ∞ is always on the line. Under a stereographic projection to the Riemann sphere, lines are circles through the north pole (∞).

Both equations ($*$) and ($\dagger$) have the form of the following definition:

Definition. A circle in $\widehat{\mathbb{C}}$ is the set of points satisfying the equation

$$Az\overline{z} + \overline{B}z + B\overline{z} + C = 0$$

where $A, C \in \mathbb{R}$, $B \in \mathbb{C}$, and $|B|^2 > AC$. We consider ∞ to be a solution to this equation if and only if $A = 0$.

Exercise: the set of points satisfying such an equation is always either a circle in $\mathbb{C}$ or a line in $\widehat{\mathbb{C}}$. We call all of these 'circles' in $\widehat{\mathbb{C}}$ by convention, since they're all circles on the Riemann sphere. We should not consider ∞ to be a special point here; it simply 'closes off' any line in $\mathbb{C}$ into a circle in $\widehat{\mathbb{C}}$.

Theorem. Möbius maps preserve circles. In other words, points on a circle in $\widehat{\mathbb{C}}$ are transformed onto points on a (possibly different) circle in $\widehat{\mathbb{C}}$.

Proof. As we saw in a previous section on Möbius maps, maps in $\mathcal{M}$ are generated by

- $z \mapsto az$
- $z \mapsto z + b$
- $z \mapsto \frac{1}{z}$

So it is enough to check that each of these generating maps preserves circles. We will write $S(A, B, C)$ for the circle satisfying

$$Az\overline{z} + \overline{B}z + B\overline{z} + C = 0 \tag{$\clubsuit$}$$

We can check that under a dilation or rotation $z \mapsto az$,

$$S(A, B, C) \mapsto S\left(\frac{A}{\overline{a}a}, \frac{B}{\overline{a}}, C\right)$$

Under a translation $z \mapsto z + b$,

$$S(A, B, C) \mapsto S\left(A, B - Ab, C + Ab\overline{b} - B\overline{b} - \overline{B}b\right)$$

Under an inversion, solutions to (♣) become solutions to

$$Cw\overline{w} + Bw + \overline{B}\overline{w} + A = 0$$

So

$$S(A, B, C) \mapsto S(C, \overline{B}, A)$$

□

Bear in mind when solving various exercises that it is often sufficient to check certain properties apply in the generating set in order to verify that they apply in the general case.

Remark. A circle is determined by three points on it, and a Möbius map is determined by where it sends three points. So in practice, it is easy to find a Möbius map sending a given circle to another given circle.

11.4. Cross-ratios

Recall that given distinct points $z_1, z_2, z_3 \in \widehat{\mathbb{C}}$, we have a unique Möbius map f such that $f(z_1) = 0$, $f(z_2) = 1$, $f(z_3) = \infty$.

Definition. If $z_1, z_2, z_3, z_4 \in \widehat{\mathbb{C}}$ are distinct, then their cross-ratio $[z_1, z_2, z_3, z_4]$ is defined to be $f(z_4)$ where $f \in \mathcal{M}$ is the unique Möbius map f such that $f(z_1) = 0$, $f(z_2) = 1$, $f(z_3) = \infty$.

In particular, $[0, 1, \infty, w] = w$. We have the following formula for computing the cross-ratio.

$$[z_1, z_2, z_3, z_4] = \frac{(z_4 - z_1)(z_2 - z_3)}{(z_2 - z_1)(z_4 - z_3)}$$

with special cases interpreted accordingly where $z_i = \infty$. This result follows from the proof that we can construct a map to send any three distinct points to $0, 1, \infty$. There are in fact 4! different conventions for the cross-ratio, depending on the order of $0, 1, \infty$, so ensure that the correct convention is being used if referring to sources. However, this potential ambiguity is mitigated by the following fact.

Proposition. Double transpositions of the z_i fix the cross-ratio.

Proof. By inspection of the formula, it it clear that this is true. □

Theorem. Möbius maps preserve the cross-ratio. $\forall g \in \mathcal{M}, \forall z_1, z_2, z_3, z_4 \in \widehat{\mathbb{C}}$,

$$[g(z_1), g(z_2), g(z_3), g(z_4)] = [z_1, z_2, z_3, z_4]$$

Proof. Let $f \in \mathcal{M}$ be the unique Möbius map such that

$$f(z_1) = 0; \quad f(z_2) = 1; \quad f(z_3) = \infty$$

so therefore $f(z_4) = [z_1, z_2, z_3, z_4]$. Now, consider $f \circ g^{-1}$:

$$(f \circ g^{-1})g(z_1) = 0; \quad (f \circ g^{-1})g(z_2) = 1; \quad (f \circ g^{-1})g(z_3) = \infty$$

and $f \circ g^{-1}$ is the unique map with this property. So the cross-ratio here is

$$[g(z_1), g(z_2), g(z_3), g(z_4)] = (f \circ g^{-1})g(z_4) = f(z_4)$$

as required. □

Corollary. Four distinct points $z_1, z_2, z_3, z_4 \in \widehat{\mathbb{C}}$ lie on a circle if and only if their cross-ratio is real.

Proof. Let f be the unique Möbius map sending $(z_1, z_2, z_3) \mapsto (0, 1, \infty)$, so that $f(z_4)$ is the required cross-ratio. The circle C passing through z_1, z_2, z_3 is sent by f to the unique circle passing through $0, 1, \infty$, i.e. the real line together with the point at infinity. So z_4 lies on C if and only if $f(z_4)$ lies on $\mathbb{R} \cup \{\infty\}$. But since $f(z_3) = \infty$, $f(z_4) \neq \infty$, so this condition is restricted only to $\mathbb{R}$, excluding a point at infinity. □

12. Matrix groups

12.1. Definitions

We will look at various groups of matrices, their related actions, and study distance-preserving maps on $\mathbb{R}^2$ and $\mathbb{R}^3$. Here are some examples of matrix groups.

- $M_{n\times n}(\mathbb{F})$ is the set of $n \times n$ matrices over the field $\mathbb{F}$.
- $GL_n(\mathbb{F})$ is the set of $n \times n$ matrices over $\mathbb{F}$ which are invertible. This is known as the general linear group over $\mathbb{F}$.
  - $GL_n(\mathbb{F})$ is a group under multiplication.
  - $\det\colon GL_n(\mathbb{F}) \to \mathbb{F}^\times := \mathbb{F} \setminus \{0\}$ is a surjective homomorphism.
  - Given $A \in GL_n(\mathbb{R})$, $A^\intercal$ is the matrix with entries $(A^\intercal)_{ij} = A_{ji}$. It satisfies
    * $(AB)^\intercal = B^\intercal A^\intercal$
    * $(A^{-1})^\intercal = (A^\intercal)^{-1}$
    * $AA^\intercal = I \iff A^\intercal A = I \iff A^\intercal = A^{-1}$
    * $\det A^\intercal = \det A$
- $SL_n(\mathbb{F}) \leq GL_n(\mathbb{F})$ is the kernel of the det homomorphism. This is the special linear group.
- $O_n = O_n(\mathbb{R}) := \{A \in GL_n(\mathbb{R}) : A^\intercal A = I\}$ is the orthogonal group. We can check the group axioms to verify it is a subgroup of $GL_n(\mathbb{R})$.
- $SO_n \leq O_n$ is the kernel of the det homomorphism. This is the special orthogonal group.

Proposition. $\det\colon O_n \to \{\pm 1\}$ is a surjective homomorphism.

Proof. If $A \in O_n$, then $A^\intercal A = I$. So $(\det A)^2 = \det A^\intercal \cdot \det A = \det(A^\intercal A) = \det I = 1$. So $\det A = \pm 1$. It is surjective since $\det I = 1$, and the determinant of the matrix similar to the identity but one of the diagonal entries is -1 has determinant -1. □

12.2. Matrix encoding of Möbius maps

Proposition. The function $\varphi\colon SL_2(\mathbb{C}) \to \mathcal{M}$ mapping

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto f; \quad f(z) = \frac{az+b}{cz+d}$$

is a surjective homomorphism with kernel $\{I, -I\}$.

Proof. Firstly, φ is a homomorphism. If $f_1(z) = \frac{a_1 z+b_1}{c_1 z+d_1}$, $f_2(z) = \frac{a_2 z+b_2}{c_2 z+d_2}$, then we have seen that $f_2(f_1(z))$ can be written in the form $\frac{az+b}{cz+d}$ where

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix} \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix}$$

So

$$\varphi\left(\begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix} \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix}\right) = \varphi \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix} \cdot \varphi \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix}$$

Secondly, φ is surjective. If $\frac{az+b}{cz+d}$ is a Möbius map, then

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in GL_2(\mathbb{C})$$

since $ad - bc \neq 0$. But

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

may not be 1, so we will take D^2 to be this determinant, then we can consider

$$\begin{pmatrix} a/D & b/D \\ c/D & d/D \end{pmatrix}$$

This new matrix has determinant 1 and is equal to the original Möbius map, so we have a matrix in $SL_2(\mathbb{C})$ that maps to any given Möbius map. Finally, we want to find the kernel.

$$\varphi \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \mathrm{id} \in \mathcal{M} \implies \frac{az+b}{cz+d} = z \iff c = d = 0; a = d$$

But since this matrix has determinant 1, $a = d = \pm 1$, and thus $\ker \varphi = \{I, -I\}$. □

Corollary.

$$\mathcal{M} \cong {}^{SL_2(\mathbb{C})}\!/_{\{I,-I\}}$$

Proof. This is an immediate consequence of the first isomorphism theorem. □

The quotient ${}^{SL_2(\mathbb{C})}\!/_{\{I,-I\}}$ is known as the projective special linear group $PSL_2(\mathbb{C})$.

12.3. Actions of matrices on vector spaces

All of the groups defined above act on the corresponding vector spaces. For example, we have $GL_n(\mathbb{F}) \curvearrowright \mathbb{F}^n$. As an example, let $G \leq GL_2(\mathbb{R}) \curvearrowright \mathbb{R}^2$. What are the orbits of this action? Clearly, $\{\mathbf{0}\}$ is a singleton orbit since we are acting by linear maps.

- If $G = GL_2(\mathbb{R})$, G acts transitively on $\mathbb{R}^2 \setminus \{0\}$. We can complete any $\mathbf{v} \neq 0$ to a basis and therefore we have an invertible change of basis matrix sending any basis to any basis. So there are two orbits: $\mathbb{R}^2 \setminus \{0\}$ and $\{0\}$ itself.
- If G is the set of upper triangular matrices given by

$$G = \left\{ \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \in GL_2(\mathbb{R}) \right\} = \left\{ \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} : a, d \neq 0 \right\}$$

  We know that $\mathrm{Orb}(\mathbf{0}) = \{\mathbf{0}\}$. Further:

$$\mathrm{Orb}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \left\{ \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} : \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \in G \right\} = \left\{ \begin{pmatrix} a \\ 0 \end{pmatrix} : a \neq 0 \right\}$$

  We haven't found all of the orbits yet so let us consider another point.

$$\mathrm{Orb}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \left\{ \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} : \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \in G \right\} = \left\{ \begin{pmatrix} b \\ d \end{pmatrix} : d \neq 0 \right\}$$

  We have found all of the orbits since the union gives $\mathbb{R}^2$.

12.4. Conjugation action of general linear group

Recall from Vectors and Matrices: if $\alpha : \mathbb{F}^n \to \mathbb{F}^n$ is a linear map, we can represent α as a matrix A with respect to a basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$. If we choose a different basis $\{\mathbf{f}_1, \ldots, \mathbf{f}_n\}$ then α can also be written as a matrix with respect to this new basis, by the matrix $P^{-1}AP$ where P is the change of basis matrix, defined by

$$\mathbf{f}_j = P_{ij}\mathbf{e}_i$$

This is an example of conjugation.

Proposition. $GL_n(\mathbb{F})$ acts on $M_{n\times n}(\mathbb{F})$ by conjugation. The orbit of a matrix $A \in M_{n\times n}(\mathbb{F})$ is the set of matrices representing the same linear map as A with respect to different bases.

Proof. This is an action:

- $P(A) = PAP^{-1} \in M_{n\times n}(\mathbb{F})$ for any chosen matrix $A \in M_{n\times n}(\mathbb{F})$, $P \in GL_n(\mathbb{F})$
- $I(A) = IAI^{-1} = A$
- $Q(P(A)) = QPAP^{-1}Q^{-1} = (QP)A(QP)^{-1} = (QP)(A)$

As shown in the discussion above, A and B are in the same orbit if and only if $A = PBP^{-1} \iff B = P^{-1}AP$, which is equivalent to this conjugation action. □

Recall from Vectors and Matrices that any matrix in $M_{2\times 2}(\mathbb{C})$ is conjugate to a matrix in Jordan Normal Form, i.e. to one of the following types of matrix:

$$\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}; \quad \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}; \quad \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$$

In the first case, the values λ_1, λ_2 are uniquely determined by the matrix we are trying to conjugate (specifically its eigenvalues). But of course, the order of the eigenvalues is not determined uniquely. Other than this, no two matrices on this list of possible Jordan Normal Forms are conjugate.

- $\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}$ is characterised by having two distinct eigenvalues, a property independent of the chosen basis, so it cannot be conjugate to the others.
- $\begin{pmatrix}\lambda & 0\\ 0 & \lambda\end{pmatrix}$ is only conjugate to itself since it is λI.
- $\begin{pmatrix}\lambda & 1\\ 0 & \lambda\end{pmatrix}$ is characterised by having a repeated eigenvalue λ, but only a one dimensional eigenspace (independent of the basis we choose).

This gives a complete description of the orbits of $GL_n(\mathbb{C}) \curvearrowright M_{n\times n}(\mathbb{C})$.

12.5. Stabilisers of conjugation action

Clearly we have

$$P \in \operatorname{Stab}(A) \iff PAP^{-1} = A \iff PA = AP$$

So if two matrices commute, they stabilise each other. Let us consider the three cases as above.

- For $A = \begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}$:

  $$\begin{pmatrix}a & b\\ c & d\end{pmatrix}\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix} = \begin{pmatrix}\lambda_1 a & \lambda_2 b\\ \lambda_1 c & \lambda_2 d\end{pmatrix}$$
  $$\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}\begin{pmatrix}a & b\\ c & d\end{pmatrix} = \begin{pmatrix}\lambda_1 a & \lambda_1 b\\ \lambda_2 c & \lambda_2 d\end{pmatrix}$$

  So this matrix is in the stabiliser if and only if $b = c = 0$.

  $$\operatorname{Stab}\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix} = \left\{\begin{pmatrix}a & 0\\ 0 & d\end{pmatrix} \in GL_2(\mathbb{C})\right\}$$

- For $A = \begin{pmatrix}\lambda & 0\\ 0 & \lambda\end{pmatrix}$, clearly its stabiliser is $GL_2(\mathbb{C})$ since $A = \lambda I$, and so it commutes with any matrix.
- For $A = \begin{pmatrix}\lambda & 1\\ 0 & \lambda\end{pmatrix}$, the stabiliser is

  $$\operatorname{Stab}\begin{pmatrix}\lambda & 1\\ 0 & \lambda\end{pmatrix} = \left\{\begin{pmatrix}a & b\\ 0 & a\end{pmatrix} \in GL_2(\mathbb{C})\right\}$$

  (Proof as exercise)

12.6. Geometry of orthogonal groups

We will look more closely at the orthogonal group and special orthogonal group, and then focus on symmetries of $\mathbb{R}^2$ and $\mathbb{R}^3$. Let us consider the standard inner product in $\mathbb{R}^n$:

$$\mathbf{x} \cdot \mathbf{y} = x_i y_i = \mathbf{x}^\intercal \mathbf{y}$$

If we consider the columns $\mathbf{p}_1, \dots, \mathbf{p}_n$ of an orthogonal matrix $P \in O_n$, we have

$$(P^\intercal P)_{ij} = \mathbf{p}_i^\intercal \mathbf{p}_j = \mathbf{p}_i \cdot \mathbf{p}_j$$

So since $P \in O_n \iff P^\intercal P = I$, we have

$$\mathbf{p}_i \cdot \mathbf{p}_j = \delta_{ij}$$

Proposition. $P \in O_n$ if and only if the columns of P form an orthonormal basis.

This has been proven by the above discussion. Thinking of $P \in O_n$ as a change of basis matrix, we get the following result.

Proposition. Consider $O_n \curvearrowright M_{n\times n}(\mathbb{R})$ by conjugation. Two matrices are in the same orbit if and only if they represent the same linear map with respect to two orthonormal bases.

Proposition. $P \in O_n$ if and only if $P\mathbf{x} \cdot P\mathbf{y} = \mathbf{x} \cdot \mathbf{y}$, i.e. the matrix preserves the inner product.

Proof. In the forward direction:

$$(P\mathbf{x}) \cdot (P\mathbf{y}) = (P\mathbf{x})^\intercal (P\mathbf{y}) = \mathbf{x}^\intercal P^\intercal P\mathbf{y} = \mathbf{x}^\intercal \mathbf{y} = \mathbf{x} \cdot \mathbf{y}$$

In the backward direction: if $P\mathbf{x} \cdot P\mathbf{y} = \mathbf{x} \cdot \mathbf{y}$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, then taking the standard basis vectors $\mathbf{e}_i, \mathbf{e}_j$ we have

$$P\mathbf{e}_i \cdot P\mathbf{e}_j = \mathbf{e}_i \cdot \mathbf{e}_j = \delta_{ij}$$

So the vectors $P\mathbf{e}_1, \dots, P\mathbf{e}_n$ are orthonormal. These are the columns of P, so $P \in O_n$. □

Corollary. For $P \in O_n$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have

(i) $|P\mathbf{x}| = |\mathbf{x}|$ (P preserves length)

(ii) $P\mathbf{x} \angle P\mathbf{y} = \mathbf{x} \angle \mathbf{y}$ (P preserves angles between vectors)

Proof. (i) Follows from the fact that the inner product is preserved, by taking the inner product of a vector with itself under the transformation.

(ii) Angles are also defined using the inner product,

$$\cos(\mathbf{x} \angle \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}||\mathbf{y}|}$$

Since the inner product and the lengths are preserved, the cosine of the angle is therefore preserved. Since $\cos : [0, \pi] \to [-1, 1]$ is injective, $\mathbf{x} \angle \mathbf{y} = P\mathbf{x} \angle P\mathbf{y}$.

□

12.7. Reflections in O_n

We will consider what the elements of these groups look like when acting upon $\mathbb{R}^n$.

Definition. If $\mathbf{a} \in \mathbb{R}^n$ with $|\mathbf{a}| = 1$, then the reflection in the plane normal to $\mathbf{a}$ is the linear map

$$R_{\mathbf{a}} : \mathbb{R}^n \to \mathbb{R}^n; \quad \mathbf{x} \mapsto \mathbf{x} - 2(\mathbf{x} \cdot \mathbf{a})\mathbf{a}$$

Lemma. $R_{\mathbf{a}}$ lies in O_n.

Proof. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

$$\begin{aligned} R_{\mathbf{a}}(\mathbf{x}) \cdot R_{\mathbf{a}}(\mathbf{y}) &= (\mathbf{x} - 2(\mathbf{x} \cdot \mathbf{a})\mathbf{a}) \cdot (\mathbf{y} - 2(\mathbf{y} \cdot \mathbf{a})\mathbf{a}) \\ &= \mathbf{x} \cdot \mathbf{y} - 2(\mathbf{x} \cdot \mathbf{a})(\mathbf{a} \cdot \mathbf{y}) - 2(\mathbf{y} \cdot \mathbf{a})(\mathbf{x} \cdot \mathbf{a}) + 4(\mathbf{x} \cdot \mathbf{a})(\mathbf{y} \cdot \mathbf{a}) \underbrace{(\mathbf{a} \cdot \mathbf{a})}_{=1} \\ &= \mathbf{x} \cdot \mathbf{y} \end{aligned}$$

So it preserves the inner product, so it is an orthogonal matrix. □

As we might expect, conjugates of reflections by orthogonal matrices are also reflections.

Lemma. Given $P \in O_n$, $PR_{\mathbf{a}}P^{-1} = R_{P\mathbf{a}}$.

Proof. We have

$$\begin{aligned} PR_{\mathbf{a}}P^{-1}(\mathbf{x}) &= P(P^{-1}(\mathbf{x}) - 2(P^{-1}(\mathbf{x}) \cdot \mathbf{a})\mathbf{a}) \\ &= \mathbf{x} - 2(P^{-1}(\mathbf{x}) \cdot \mathbf{a})(P\mathbf{a}) \\ &= \mathbf{x} - 2(P^{\mathsf{T}}(\mathbf{x}) \cdot \mathbf{a})(P\mathbf{a}) \\ &= \mathbf{x} - 2(\mathbf{x}^{\mathsf{T}} P\mathbf{a})(P\mathbf{a}) \\ &= \mathbf{x} - 2(\mathbf{x} \cdot P\mathbf{a})(P\mathbf{a}) \end{aligned}$$

which by inspection is the reflection of $\mathbf{x}$ by the plane with normal $P\mathbf{a}$. □

We know that no reflection matrix can be in SO_n, since this requires the determinant to be $+1$, which is the product of the eigenvalues. The $n - 1$ eigenvectors with eigenvalue $+1$ are $n - 1$ linearly independent vectors spanning the plane, and the single eigenvector with eigenvalue -1 is the normal to the plane. So the determinant is -1.

12.8. Classifying elements of O_2

Theorem. Every element of SO_2 is of the form

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

for some $\theta \in [0, 2\pi)$.

This is an anticlockwise rotation of $\mathbb{R}^2$ about the origin by angle θ. Conversely, every such element lies in SO_2.

Proof. Let

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SO_2$$

We have $A^\intercal A = I$ and $\det A = 1$. So

$$A^\intercal = A^{-1} \implies \begin{pmatrix} a & c \\ b & d \end{pmatrix} = \frac{1}{1}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

So $a = d, b = -c$. Since $ad - bc = 1$, $a^2 + c^2 = 1$. Then we can write $a = \cos\theta$ and $c = \sin\theta$ for a unique $\theta \in [0, 2\pi)$.

Conversely, the determinant of this matrix is 1, and is in O_2, so this element lies in SO_2. □

Theorem. The elements of $O_2 \setminus SO_2$ are the reflections in lines through the origin.

Proof. Let

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in O_2 \setminus SO_2$$

So $A^\intercal A = I$ and $\det A = -1$.

$$A^\intercal = A^{-1} \implies \begin{pmatrix} a & c \\ b & d \end{pmatrix} = \frac{1}{-1}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

So $a = -d, b = c$. Together with $ad - bc = -1$, we have $a^2 + c^2 = 1$. So let $a = \cos\theta$, $c = \sin\theta$ like before, so

$$A = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$$

which can be shown to be a reflection using double angle formulas such that

$$A\begin{pmatrix} \sin\frac{\theta}{2} \\ \cos\frac{\theta}{2} \end{pmatrix} = -\begin{pmatrix} \sin\frac{\theta}{2} \\ \cos\frac{\theta}{2} \end{pmatrix}; \quad A\begin{pmatrix} \cos\frac{\theta}{2} \\ \sin\frac{\theta}{2} \end{pmatrix} = \begin{pmatrix} \cos\frac{\theta}{2} \\ \sin\frac{\theta}{2} \end{pmatrix}$$

So A is a reflection in the plane orthogonal to the vector $\begin{pmatrix} \sin\frac{\theta}{2} \\ \cos\frac{\theta}{2} \end{pmatrix}$. Conversely, any reflection in a line through the origin has this form, so it will be in $O_2 \setminus SO_2$. □

Corollary. Every element of O_2 is the composition of at most two reflections.

Proof. Every element of $O_2 \setminus SO_2$ is a reflection, so this is trivial. If $A \in SO_2$, then we can write

$$A = \underbrace{A\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}}_{\det=-1}\underbrace{\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}}_{\det=-1}$$

So we have expressed A as the product of two reflections. □

12.9. Classifying elements of O_3

Theorem. If $A \in SO_3$, then there exists some unit vector $\mathbf{v} \in \mathbb{R}^3$ with $A\mathbf{v} = \mathbf{v}$, i.e. there exists an eigenvector with eigenvalue 1.

Proof. It is sufficient to show that 1 is an eigenvalue of A, since this guarantees that there is some nonzero eigenvector for this eigenvalue which we can then normalise. This is equivalent to showing that $\det(A - I) = 0$.

$$\begin{aligned}\det(A - I) &= \det(A - AA^\intercal) \\ &= \det(A)\det(I - A^\intercal) \\ &= \det(I - A^\intercal) \\ &= \det((I - A)^\intercal) \\ &= \det(I - A) \\ &= (-1)^3 \det(A - I)\end{aligned}$$

So $2\det(A - I) = 0 \implies \det(A - I) = 0$. □

Corollary. Every element $A \in SO_3$ is conjugate (in SO_3) to a matrix of the form

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix}$$

Proof. By the above theorem, there exists some unit vector $\mathbf{v}_1$ which is an eigenvector of eigenvalue 1. We can extend this vector to an orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ of $\mathbb{R}^3$. Then, for $i = 2, 3$, we have

$$A\mathbf{v}_i \cdot \mathbf{v}_1 = A\mathbf{v}_i \cdot A\mathbf{v}_1 = \mathbf{v}_i \cdot \mathbf{v}_1 = 0$$

So $A\mathbf{v}_2, A\mathbf{v}_3$ lie in the subspace generated by $\mathbf{v}_2, \mathbf{v}_3$, i.e. $\operatorname{span}\{\mathbf{v}_2, \mathbf{v}_3\} = \langle \mathbf{v}_2, \mathbf{v}_3 \rangle$. So A maps this subspace to itself, and we can thus consider the restriction of A to this subspace. The matrix in this new basis will have form

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & a & b \\ 0 & c & d \end{pmatrix}$$

The smaller matrix in the bottom right will still have determinant 1, since we can expand the determinant here by the first row. So A restricted to this subspace is an element of SO_2, so its matrix must be of the form

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

So A has the required form with respect to this new basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$. The change of basis matrix P lies in O_3 since $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is an orthonormal basis. If $P \notin SO_3$, then we can use the basis $\{-\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ instead, which will invert the determinant of P. So in either case $P \in SO_3$. □

This tells us in particular that every element in SO_3 is a rotation about some axis, here $\mathbf{v}_1$.

Corollary. Every element of O_3 is the composition of at most three reflections.

Proof.

- If $A \in SO_3$, then $\exists P \in SO_3$ such that $PAP^{-1} = B$, where B is of the form

$$B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix}$$

  Since this smaller matrix

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

  is a composition of at most two reflections, then B is also a composition of at most two reflections, i.e. $B = B_1B_2$. Since A is a conjugate of B, it is also a composition of at most two reflections, as the conjugate of a reflection is a reflection, and $A = P^{-1}BP = (P^{-1}B_1P)(P^{-1}B_2P)$.

- If $A \in O_3 \setminus SO_3$, then $\det A = -1$ and we can construct

$$A = \underbrace{A\begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{\det=1}\underbrace{\begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{\det=-1}$$

  So the left-hand product lies in SO_3, so it is a composition of at most two reflections. The final element is a reflection in the y–z plane, so the combined product is a composition of at most three reflections.

□

Example (symmetries of the cube, revisited)**.** We can think of symmetry groups of the Platonic solids as subgroups of O_3 by placing the solid at the origin. By question 11 on example sheet 4, we have that $O_3 \cong SO_3 \times C_2$, where C_2 is generated by the map $\mathbf{v} \mapsto -\mathbf{v}$. So if $\mathbf{v} \mapsto -\mathbf{v}$ is a symmetry of our platonic solid, then this group of symmetries will also split as the direct product of $G^+ \times C_2$ where G^+ is the group of rotations (proof as exercise).

So we have that the group of symmetries of the cube is $G^+ \times C_2 \cong S_4 \times C_2$ by the results from earlier.

13. Groups of order 8

13.1. Quaternions

We have already seem all the possibilities of groups of order less than 8. For order 8, we need to first define a new group.

Definition. Consider the subset of matrices of $GL_2(\mathbb{C})$ given by

$$\mathbf{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}; \quad \mathbf{i} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}; \quad \mathbf{j} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}; \quad \mathbf{k} = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}$$

We can form a group from these matrices. The set $\{\pm\mathbf{1}, \pm\mathbf{i}, \pm\mathbf{j}, \pm\mathbf{k}\}$ forms a group with respect to matrix multiplication known as the quaternions, denoted Q_8. The elements therefore satisfy

- $g^4 = \mathbf{1}$
- $(-\mathbf{1})^2 = \mathbf{1}$
- $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -\mathbf{1}$
- $\mathbf{ij} = \mathbf{k}; \mathbf{jk} = \mathbf{i}; \mathbf{ki} = \mathbf{j}$
- $\mathbf{ji} = -\mathbf{k}; \mathbf{kj} = -\mathbf{i}; \mathbf{ik} = -\mathbf{j}$

13.2. Elements of order 2

Lemma. If a finite group has all non-identity elements of order 2, then it is isomorphic to $C_2 \times C_2 \times \cdots \times C_2$.

Proof. By question 7 on example sheet 1, we already know that such a G must be abelian, and that $|G| = 2^n$. If $|G| = 2$, then $G \cong C_2$. If $|G| > 2$, then we can choose some element a_1 of order 2, and then there exists another element $a_2 \notin \langle a_1 \rangle$ of order 2. By the Direct Product Theorem, $\langle a_1, a_2 \rangle \cong \langle a_1 \rangle \times \langle a_2 \rangle$. We can repeat this direct product with elements not in the group to generate the whole group. □

13.3. Classification of groups of order 8

Theorem. A group of order 8 is isomorphic to exactly one of:

- C_8
- $C_4 \times C_2$
- $C_2 \times C_2 \times C_2$
- D_8
- Q_8

Proof. Firstly, the above groups are not isomorphic: $C_8, C_4 \times C_2, C_2 \times C_2 \times C_2$ are all abelian while D_8 and Q_8 are not. The abelian groups can be distinguished by the maximal order of an element. The non-abelian groups can be distinguished by the number of elements of order 2. D_8 has s, r^2, r^2s, while Q_8 only has $-\mathbf{1}$.

Now let G be a group such that $|G| = 8$. If $g \in G$, then $o(g) \mid 8$ by Lagrange's Theorem. So $o(g) = 1, 2, 4, 8$.

- If there is an element of order 8, then $G = \langle g \rangle \cong C_8$.
- If all non-identity elements have order 2, then $G = C_2 \times C_2 \times C_2$ by the above lemma.
- The remaining cases are when there are no elements of order 8, and not all elements are of order 2, so there exists some element h of order 4. Note then that $\langle h \rangle \cong C_4$ and $|G : \langle h \rangle| = 2$, so $\langle h \rangle \trianglelefteq G$. Thus, $g^2 \in \langle h \rangle$ by question 4 on example sheet 3. So $g^2 = e, h, h^2, h^3$.

  Now, consider ghg^{-1}. This must lie in $\langle h \rangle$ since $\langle h \rangle \trianglelefteq G$, and must have order 4 since h does. So $ghg^{-1} = h, h^3$. We will now consider each possible case of g^2 together with each possible case of ghg^{-1}.

  - If $g^2 = h, h^3$ then $g^4 = h^2 \neq e$ so g has order 8 #. So either $g^2 = e$ or $g = h^2$.
  - If $g^2 = e$:
    * If $ghg^{-1} = h$, then $gh = hg$, so g and h commute. Further, $\langle h \rangle \cap \langle g \rangle = \{e\}$, and $G = \langle h \rangle \cdot \langle g \rangle$. By the Direct Product Theorem, $G \cong \langle h \rangle \times \langle g \rangle = C_4 \times C_2$.
    * If $ghg^{-1} = h^3 = h^{-1}$, then since $g^2 = e$, we recognise that the group is the dihedral group D_8 with $h = r$, $g = s$.
  - If $g^2 = h^2$ (note that this does not necessarily imply that $g = h$), we will have
    * If $ghg^{-1} = h$, then g and h commute, so $(gh)^2 = g^2h^2 = h^2h^2 = e$. So gh has order 2. We can again apply the direct product theorem to $\langle h \rangle \cong C_4$ and $\langle gh \rangle \cong C_2$, and we get $g \cong \langle h \rangle \times \langle g \rangle \cong C_4 \times C_2$ again.
    * If $ghg^{-1} = h^3 = h^{-1}$, then we can define a map

      $$\varphi : G \to Q_8$$

      by

      $$\begin{aligned} e &\mapsto \mathbf{1} & g &\mapsto \mathbf{j} \\ h &\mapsto \mathbf{i} & gh &\mapsto -\mathbf{k} \\ h^2 &\mapsto -\mathbf{1} & gh^2 &\mapsto -\mathbf{j} \\ h^3 &\mapsto -\mathbf{i} & gh^3 &\mapsto \mathbf{k} \end{aligned}$$

      Clearly φ is bijective, and we can check that it is a homomorphism. So it is an isomorphism, so $G \cong Q_8$.

□

Remark. We know that in an abelian group, every subgroup is normal. The converse is not true. Just because every subgroup is normal, this does not mean that the group is abelian. For example Q_8 is an example, where its subgroups are $\langle \mathbf{i} \rangle$, $\langle \mathbf{j} \rangle$, $\langle \mathbf{k} \rangle$ (which are normal since they have index 2), and $\langle -\mathbf{1} \rangle$ which is normal since it commutes with everything.

IV. Vectors and Matrices

Lectured in Michaelmas 2020 by Dr. J. M. Evans

The complex numbers can be viewed as a kind of two-dimensional analogue to the real numbers, with a real coordinate and an imaginary coordinate. Euclidean space is a three-dimensional version of the reals, with three coordinates to represent each point. In this course, we generalise these examples, and study vector spaces which can have any dimension.

Functions between vector spaces that preserve the vector space structure are called linear. Linear maps have many different useful properties. One such property is the determinant: if the determinant is any number except zero, the linear map has an inverse function.

Contents

1. Complex numbers 197
1.1. Definition and basic theorems 197
1.2. Complex valued functions 198
1.3. Transformations and primitives 199
2. Vectors in three dimensions 200
2.1. Vector addition and scalar multiplication 200
2.2. Scalar product 200
2.3. Vector product 201
2.4. Basis vectors 201
2.5. Scalar triple product 202
2.6. Vector triple product 202
2.7. Lines 202
2.8. Planes 203
2.9. Other vector equations 203
3. Index notation and the summation convention 204
3.1. Kronecker δ and Levi-Civita ε 204
3.2. Identities 205
4. Higher dimensional vectors 207
4.1. Multidimensional real space 207
4.2. Cauchy–Schwarz inequality 207
4.3. Triangle inequality 208
4.4. Levi-Civita ε in higher dimensions 208
4.5. General real vector spaces 208
4.6. Inner product spaces 209
4.7. Bases and dimensions 210
4.8. Multidimensional complex space 212
5. Linear maps 214
5.1. Introduction 214
5.2. Rank and nullity 215
5.3. Rotations 216
5.4. Reflections and projections 217
5.5. Dilations 217
5.6. Shears 217
5.7. Matrices 217
5.8. Matrix of a general linear map 220
5.9. Linear combinations 221
5.10. Matrix multiplication 221

5.11. Matrix inverses . . . 222
6. Transpose and Hermitian conjugate . . . 224
6.1. Transpose . . . 224
6.2. Hermitian conjugate . . . 225
6.3. Trace . . . 225
6.4. Orthogonal matrices . . . 226
6.5. Unitary matrices . . . 227
7. Adjugates and alternating forms . . . 228
7.1. Inverses in two dimensions . . . 228
7.2. Three dimensions . . . 228
7.3. Levi-Civita ε in higher dimensions . . . 229
7.4. Properties . . . 230
8. Determinant . . . 232
8.1. Definition . . . 232
8.2. Expanding by rows or columns . . . 233
8.3. Row and column operations . . . 233
8.4. Multiplicative property of determinants . . . 234
8.5. Cofactors and determinants . . . 235
8.6. Adjugates and inverses . . . 236
8.7. Systems of linear equations . . . 237
8.8. Geometrical interpretation of solutions of linear equations . . . 239
9. Properties of matrices . . . 240
9.1. Eigenvalues and eigenvectors . . . 240
9.2. The characteristic polynomial . . . 241
9.3. Eigenspaces and multiplicities . . . 241
9.4. Linear independence of eigenvectors . . . 243
9.5. Diagonalisability . . . 244
9.6. Criteria for diagonalisability . . . 245
9.7. Similarity . . . 246
9.8. Real eigenvalues and orthogonal eigenvectors . . . 247
9.9. Unitary and orthogonal diagonalisation . . . 249
10. Quadratic forms . . . 250
10.1. Simple example . . . 250
10.2. Diagonalising quadratic forms . . . 250
10.3. Hessian matrix as a quadratic form . . . 252
11. Cayley–Hamilton theorem . . . 253
11.1. Matrix polynomials . . . 253
11.2. Proofs of special cases of Cayley–Hamilton theorem . . . 254
11.3. Proof in general case (non-examinable) . . . 254

12. Changing bases 256
12.1. Change of basis formula 256
12.2. Changing bases of vector components 258
12.3. Specialisations of changes of basis 259
12.4. Jordan normal form 260
12.5. Jordan normal forms in n dimensions 261
13. Conics and quadrics 263
13.1. Quadrics in general 263
13.2. Conics as quadrics 264
13.3. Standard forms for conics 264
13.4. Conics as sections of a cone 265
14. Symmetries and transformation groups 266
14.1. Orthogonal transformations and rotations 266
14.2. 2D Minkowski space 266
14.3. Lorentz transformations 267
14.4. Application to special relativity 267

1. Complex numbers

1.1. Definition and basic theorems

We construct the complex numbers from $\mathbb{R}$ by adding an element i such that $i^2 = -1$. By definition, any complex number $z \in \mathbb{C} = x + iy$ where $x, y \in \mathbb{R}$. We use the notation $x = \operatorname{Re} z$ and $y = \operatorname{Im} z$ to query the components of a complex number. The complex numbers contains the set of real numbers, due to the fact that $x = x + i0$. We define the operations of addition and multiplication in familiar ways, which lets us state that $\mathbb{C}$ is a field.

We also define the complex conjugate $\overline{z}$ as negating the imaginary part of z. Trivially we can see facts such as $\overline{(\overline{z})} = z$; $\overline{z + w} = \overline{z} + \overline{w}$ and $\overline{zw} = \overline{z} \cdot \overline{w}$.

The Fundamental Theorem of Algebra states that a polynomial of degree n can be written as a product of n linear factors:

$$c_n z^n + \cdots + c_1 z^1 + c_0 z^0 = c_n(z - \alpha_1)(z - \alpha_2) \cdots (z - \alpha_n) \quad (\text{where } c_i, \alpha_i \in \mathbb{C})$$

We can reformulate this statement as follows: a polynomial of degree n has n solutions α_i, counting repeats. This theorem is not proved in this course.

The modulus of complex numbers z_1, z_2 satisfies:

- (composition) $|z_1 z_2| = |z_1||z_2|$, and
- (triangle inequality) $|z_1 + z_2| \leq |z_1| + |z_2|$

Proof. The composition property is trivial. To prove the triangle inequality, we square both sides and compare.

$$\begin{aligned}
\text{LHS} &= |z_1 + z_2|^2 \\
&= (z_1 + z_2)\overline{(z_1 + z_2)} \\
&= |z_1|^2 + \overline{z_1} z_2 + z_1 \overline{z_2} + |z_2|^2 \\
\text{RHS} &= |z_1|^2 + 2|z_1||z_2| + |z_2|^2
\end{aligned}$$

Note that

$$\begin{aligned}
&\overline{z_1} z_2 + z_1 \overline{z_2} \leq 2|z_1||z_2| \\
\iff &\frac{1}{2}\left(\overline{z_1} z_2 + \overline{\overline{z_1} z_2}\right) \leq |z_1||z_2| \\
\iff &\operatorname{Re}(\overline{z_1} z_2) \leq |\overline{z_1} z_2|
\end{aligned}$$

which is true. □

We can alternatively use the map $z_2 \to z_2 - z_1$ to write the triangle inequality as

$$\begin{aligned} |z_2 - z_1| &\geq |z_2| - |z_1| \\ \text{or } |z_2 - z_1| &\geq |z_1| - |z_2| \\ \therefore |z_2 - z_1| &\geq ||z_2| - |z_1|| \end{aligned}$$

De Moivre's Theorem states that

$$(\cos\theta + i\sin\theta)^n = \cos n\theta + i\sin n\theta \quad (\forall n \in \mathbb{Z})$$

We can prove this using induction for $n \geq 0$. To show the negative case, simply use the positive result and raise it to the power of -1.

1.2. Complex valued functions

For $z \in \mathbb{C}$, we can define:

$$\begin{aligned} \exp z &= \sum_{n=0}^{\infty} \frac{1}{n!} z^n \\ \cos z &= \frac{1}{2}\left(e^{iz} + e^{-iz}\right) \\ \sin z &= \frac{1}{2i}\left(e^{iz} - e^{-iz}\right) \end{aligned}$$

By defining $\log z = w$ s.t. $e^w = z$, we have a complex logarithm function. By expanding the definition, we get that $\log z = \log r + i\theta$ where $r = |z|$ and $\theta = \arg z$. Note that because the argument of a complex number is multi-valued, so is the logarithm.

We can define exponentiation in the general case by defining $z^\alpha = e^{\alpha \log z}$. Depending on the choice of α, we have three cases:

- If $\alpha = p \in \mathbb{Z}$ then the result of z^p is unambiguous because

  $$z^p = e^{p\log z} = e^{p(\log r + i\theta + 2\pi i n)}$$

  which has a factor of $e^{2\pi i p n}$ which is 1.
- For a similar reason, a rational exponent has finitely many values.
- But in the general case, there are infinitely many values.

We can calculate results such as the square root of a complex number, which have two results as you might expect.

Note. We can't use facts like $z^\alpha z^\beta = z^{\alpha+\beta}$ in the complex case because the left and right hand sides both have infinite sets of answers, which may not be the same.

1.3. Transformations and primitives

We can represent a line passing through $x_0 \in \mathbb{C}$ parallel to $w \in \mathbb{C}$ using the formula:

$$z = z_0 + \lambda w \quad (\lambda \in \mathbb{R})$$

We can eliminate the dependency on λ by computing the conjugate of both sides:

$$\overline{z} = \overline{z_0} + \lambda \overline{w}$$
$$\overline{w}z - w\overline{z} = \overline{w}z_0 - w\overline{z_0}$$

We can also write the equation for a circle with centre $c \in \mathbb{C}$ and radius $\rho \in \mathbb{R}$:

$$z = c + \rho e^{i\alpha}$$

or equivalently:

$$|z - c| = \left|\rho e^{i\alpha}\right| = \rho$$

or by squaring both sides:

$$|z|^2 - c\overline{z} - \overline{c}z = \rho^2 - |c|^2$$

2. Vectors in three dimensions

We use the normal Euclidean notions of points, lines, planes, length, angles and so on. By choosing an (arbitrary) origin point O, we may write positions as position vectors with respect to that origin point.

2.1. Vector addition and scalar multiplication

We define vector addition using the shape of a parallelogram with points $\mathbf{0}, \mathbf{a}, \mathbf{a}+\mathbf{b}, \mathbf{b}$. We define scalar multiplication of a vector using the line $\overrightarrow{OA}$ and setting the length to be multiplied by the constant. Note that this vector space is an abelian group under addition.

Definition. $\mathbf{a}$ and $\mathbf{b}$ are defined to be parallel if and only if $\mathbf{a} = \lambda\mathbf{b}$ or $\mathbf{b} = \lambda\mathbf{a}$ for some $\lambda \in \mathbb{R}$. This is denoted $\mathbf{a} \parallel \mathbf{b}$. Note that the vectors may be zero, in particular the zero vector is parallel to all vectors.

Definition. The span of a set of vectors is defined as $\operatorname{span}\{\mathbf{a}, \mathbf{b}, \cdots, \mathbf{c}\} = \{\alpha\mathbf{a} + \beta\mathbf{b} + \cdots + \gamma\mathbf{c} : \alpha, \beta, \gamma \in \mathbb{R}\}$. This is the line/plane/volume etc. containing the vectors. The span has an amount of dimensions at most equal to the amount of vectors in the input set. For example, the span of a set of two vectors may be a point, line or plane containing the vectors.

2.2. Scalar product

Definition. Given two vectors $\mathbf{a}, \mathbf{b}$, let θ be the angle between the two vectors. Then, we define

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos\theta$$

Note that if either of the vectors is zero, θ is undefined. However, the dot product is zero anyway here, so this is irrelevant.

Definition. Two vectors $\mathbf{a}$ and $\mathbf{b}$ are defined to be parallel (or orthogonal) if and only if $\mathbf{a} \cdot \mathbf{b} = 0$. This is denoted $\mathbf{a} \perp \mathbf{b}$. This is true in two cases:

(i) $\cos\theta = 0 \iff \theta = \frac{\pi}{2} \mod \pi$, or

(ii) $\mathbf{a} = 0$ or $\mathbf{b} = 0$.

Therefore, the zero vector is perpendicular to all vectors.

Definition. We can decompose a vector $\mathbf{b}$ into components relative to $\mathbf{a}$:

$$\mathbf{b} = \mathbf{b}_{\parallel} + \mathbf{b}_{\perp}$$

where $\mathbf{b}_{\parallel}$ is the component of $\mathbf{b}$ parallel to $\mathbf{a}$, and $\mathbf{b}_{\perp}$ is the component of $\mathbf{b}$ perpendicular to $\mathbf{a}$. In particular, we have that

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a} \cdot \mathbf{b}_{\parallel}$$

2.3. Vector product

Definition. Given two vectors $\mathbf{a}, \mathbf{b}$, let θ be the angle between the two vectors measured with respect to an arbitrary normal $\hat{\mathbf{n}}$. Then, we define

$$\mathbf{a} \wedge \mathbf{b} = \mathbf{a} \times \mathbf{b} = |\mathbf{a}||\mathbf{b}|\hat{\mathbf{n}} \sin\theta$$

Note that by swapping the sign of $\hat{\mathbf{n}}$, θ changes to $2\pi - \theta$, leaving the result unchanged. There are two degenerate cases:

- θ is undefined if $\mathbf{a}$ or $\mathbf{b}$ is the zero vector, but the result is zero anyway because we multiply by the magnitudes of both vectors.
- $\hat{\mathbf{n}}$ is undefined if $\mathbf{a} \parallel \mathbf{b}$, but here $\sin\theta = 0$ so the result is zero anyway.

We can provide several useful interpretations of the cross product:

- The magnitude of $\mathbf{a} \times \mathbf{b}$ is the vector area of the parallelogram defined by the points $\mathbf{0}, \mathbf{a}, \mathbf{a} + \mathbf{b}, \mathbf{b}$.
- By fixing a vector $\mathbf{a}$, we can consider the plane perpendicular to it. If $\mathbf{x}$ is another vector in the plane, $\mathbf{x} \mapsto \mathbf{a} \times \mathbf{x}$ rotates $\mathbf{x}$ by $\frac{\pi}{2}$ in the plane, scaling it by the magnitude of $\mathbf{a}$.

Note that by resolving a vector $\mathbf{b}$ perpendicular to another vector $\mathbf{a}$, we have that

$$\mathbf{a} \times \mathbf{b} = \mathbf{a} \times \mathbf{b}_\perp$$

A final useful property of the cross product is that since the result is perpendicular to both input vectors, we have

$$\mathbf{a} \cdot (\mathbf{a} \times \mathbf{b}) = \mathbf{b} \cdot (\mathbf{a} \times \mathbf{b}) = 0$$

2.4. Basis vectors

To represent vectors as some collection of numbers, we can choose some basis vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ which are 'orthonormal', i.e. they are unit vectors and pairwise orthogonal. Note that

$$\mathbf{e}_i \cdot \mathbf{e}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

The set $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ is called a basis because any vector can be written uniquely as a linear combination of the basis vectors. Because we have orthonormal basis vectors, we can reduce this to

$$\mathbf{a} = \sum_i \mathbf{a}_i \mathbf{e}_i \implies \mathbf{a}_i = \mathbf{e}_i \cdot \mathbf{a}$$

By representing a vector as a linear combination of basis vectors, it is very easy to evaluate the scalar product algebraically. To calculate the vector product, we first need to define whether $\mathbf{e}_1 \times \mathbf{e}_2 = \mathbf{e}_3$ or $-\mathbf{e}_3$. By convention, we assume that the basis vectors are right-handed, i.e. $\mathbf{e}_1 \times \mathbf{e}_2 = \mathbf{e}_3$. Then we can calculate the formula for the cross product in terms of the vectors' components.

2.5. Scalar triple product

The scalar triple product is the scalar product of one vector with the cross product of two more.

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a}) = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}) = [\mathbf{a}, \mathbf{b}, \mathbf{c}]$$

The result of the scalar triple product is the signed volume of the parallelepiped starting at the origin with axes $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$. We can represent this triple product as the determinant of a matrix:

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \begin{vmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 \\ \mathbf{b}_1 & \mathbf{b}_2 & \mathbf{b}_3 \\ \mathbf{c}_1 & \mathbf{c}_2 & \mathbf{c}_3 \end{vmatrix}$$

If the scalar triple product is greater than zero, then $\mathbf{a}, \mathbf{b}, \mathbf{c}$ is called a right handed set. If it is equal to zero, then the vectors are all coplanar: $\mathbf{c} \in \text{span}\{\mathbf{a}, \mathbf{b}\}$.

2.6. Vector triple product

The vector triple product is the cross product of three vectors. Note that this is non-associative. The proof is covered in the subsequent lecture.

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c}$$

$$(\mathbf{a} \times \mathbf{b}) \times \mathbf{c} = (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{b} \cdot \mathbf{c})\mathbf{a}$$

2.7. Lines

A line through $\mathbf{a}$ parallel to $\mathbf{u}$ is defined by

$$\mathbf{r} = \mathbf{a} + \lambda\mathbf{u}$$

where λ is some real parameter. We can eliminate lambda by using the cross product with $\mathbf{u}$. This will allow us to get a $\mathbf{u} \times \mathbf{u}$ term which will cancel to zero.

$$\mathbf{u} \times \mathbf{r} = \mathbf{u} \times \mathbf{a}$$

Informally, this is saying that $\mathbf{r}$ and $\mathbf{a}$ have the same components perpendicular to $\mathbf{u}$. Note that we can also reverse this process. Consider the equation

$$\mathbf{u} \times \mathbf{r} = \mathbf{c}$$

By using the dot product with $\mathbf{u}$ we can say

$$\mathbf{u} \cdot (\mathbf{u} \times \mathbf{r}) = \mathbf{u} \cdot \mathbf{c}$$

If $\mathbf{u} \cdot \mathbf{c} \neq 0$ then the equation is inconsistent. Otherwise, we can suppose that maybe $\mathbf{r} = \mathbf{u} \times \mathbf{c}$ and use the formula for the vector product to get the left hand side to be $\mathbf{u} \times (\mathbf{u} \times \mathbf{c}) = -|\mathbf{u}|^2 \mathbf{c}$. Therefore, by inspection, $\mathbf{a} = -\frac{1}{|\mathbf{u}|^2}(\mathbf{u} \times \mathbf{c})$ is a solution. Now, note that we can add any multiple of $\mathbf{u}$ to $\mathbf{a}$ and it remains a solution. So the general solution is $\mathbf{r} = \mathbf{a} + \lambda\mathbf{u}$.

2.8. Planes

The general point on a plane that passes through $\mathbf{a}$ and has directions $\mathbf{u}$ and $\mathbf{v}$ is

$$\mathbf{r} = \mathbf{a} + \lambda\mathbf{u} + \mu\mathbf{v}$$

where $\mathbf{u}$ and $\mathbf{v}$ are not parallel, and λ and μ are real parameters. We can do a dot product with $\mathbf{n} = (\mathbf{u} \times \mathbf{v})$ to eliminate both parameters.

$$\mathbf{n} \cdot \mathbf{r} = \kappa$$

where $\kappa = \mathbf{n} \cdot \mathbf{a}$. Note that $|\kappa|/|\mathbf{n}|$ is the perpendicular distance from the origin to the plane.

2.9. Other vector equations

The equation of a sphere is given by a quadratic vector equation in $\mathbf{r}$.

$$\mathbf{r}^2 + \mathbf{r} \cdot \mathbf{a} = k$$

We can complete the square to give

$$\left(\mathbf{r} + \frac{1}{2}\mathbf{a}\right)^2 = \frac{1}{4}\mathbf{a}^2 + k$$

which is clearly a sphere with centre $-\frac{1}{2}\mathbf{a}$ and radius $\left(\frac{1}{4}\mathbf{a}^2 + k\right)^{1/2}$.

Another example of a vector equation is

$$\mathbf{r} + \mathbf{a} \times (\mathbf{b} \times \mathbf{r}) = \mathbf{c} \tag{1}$$

where $\mathbf{a}, \mathbf{b}, \mathbf{c}$ are fixed. We can dot with $\mathbf{a}$ to eliminate the second term:

$$\mathbf{a} \cdot \mathbf{r} = \mathbf{a} \cdot \mathbf{c} \tag{2}$$

Note that using the dot product loses information—this is simply a tool to make deductions; (2) does not contain the full information of (1). Combining (1) and (2), and using the formula for the vector triple product, we get

$$\begin{aligned} &\mathbf{r} + (\mathbf{a} \cdot \mathbf{r})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{r} = \mathbf{c} && (3)\\ \implies\ &\mathbf{r} + (\mathbf{a} \cdot \mathbf{c})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{r} = \mathbf{c} \end{aligned}$$

This eliminates the dependency on $\mathbf{r}$ inside the dot product. Now, we can factorise, leaving

$$(1 - \mathbf{a} \cdot \mathbf{b})\mathbf{r} = \mathbf{c} - (\mathbf{a} \cdot \mathbf{c})\mathbf{b} \tag{4}$$

If $1 - \mathbf{a} \cdot \mathbf{b} \neq 0$ then $\mathbf{r}$ has a single solution, a point. Otherwise, the right hand side must also be zero (otherwise the equation is inconsistent). Therefore, $\mathbf{c} - (\mathbf{a} \cdot \mathbf{c})\mathbf{b} = \mathbf{0}$. We can now combine this expression for $\mathbf{c}$ into (3), eliminating the $(1 - \mathbf{a} \cdot \mathbf{b})$ term, to get

$$(\mathbf{a} \cdot \mathbf{r} - \mathbf{a} \cdot \mathbf{c})\mathbf{b} = \mathbf{0}$$

This shows us that (given that $\mathbf{b}$ is nonzero) the solutions to the equation are given by (2), which is the equation of a plane.

3. Index notation and the summation convention

3.1. Kronecker δ and Levi-Civita ε

The Kronecker δ is defined by

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Then $\mathbf{e}_i \mathbf{e}_j = \delta_{ij}$. We can also use δ to rewrite indices: $\sum_i \delta_{ij} \mathbf{a}_i = \mathbf{a}_j$. So

$$\begin{aligned} \mathbf{a} \cdot \mathbf{b} &= \left(\sum_i \mathbf{a}_i \mathbf{e}_i \right) \cdot \left(\sum_j \mathbf{b}_j \mathbf{e}_j \right) \\ &= \sum_{ij} \mathbf{a}_i \mathbf{b}_j (\mathbf{e}_i \cdot \mathbf{e}_j) \\ &= \sum_{ij} \mathbf{a}_i \mathbf{b}_j \delta_{ij} \\ &= \sum_i \mathbf{a}_i \mathbf{b}_i \end{aligned}$$

The Levi-Civita ε is defined by

$$\varepsilon_{ijk} = \begin{cases} +1 & \text{if } ijk \text{ is an even permutation of } [1,2,3] \\ -1 & \text{if } ijk \text{ is an odd permutation of } [1,2,3] \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\begin{aligned} \varepsilon_{123} = \varepsilon_{231} = \varepsilon_{312} = +1 \\ \varepsilon_{132} = \varepsilon_{321} = \varepsilon_{213} = -1 \end{aligned}$$

and all other permutations of $[1,2,3]$ yield 0. This shows that ε is totally antisymmetric; exchanging any pair of indices changes the sign. We now have:

$$\mathbf{e}_i \times \mathbf{e}_j = \sum_k \varepsilon_{ijk} \mathbf{e}_k$$

And:

$$\begin{aligned} \mathbf{a} \times \mathbf{b} &= \left(\sum_i \mathbf{a}_i \mathbf{e}_i \right) \times \left(\sum_j \mathbf{b}_j \mathbf{e}_j \right) \\ \mathbf{a} \times \mathbf{b} &= \sum_{ij} \mathbf{a}_i \mathbf{b}_j \left(\mathbf{e}_i \times \mathbf{e}_j \right) \\ \mathbf{a} \times \mathbf{b} &= \sum_{ijk} \mathbf{a}_i \mathbf{b}_j \varepsilon_{ijk} \mathbf{e}_k \end{aligned}$$

So the individual terms of the cross product can be written

$$(\mathbf{a} \times \mathbf{b})_k = \sum_{ij} \mathbf{a}_i \mathbf{b}_j \varepsilon_{ijk}$$

We use the 'summation convention' to abbreviate the many summation symbols used throughout linear algebra.

(i) An index which occurs exactly once in some term, called a 'free' index, must appear once in every term in that equation.

(ii) An index which occurs exactly twice in a given term, called a 'repeated', 'contracted', or 'dummy' index, is implicitly summed over.

(iii) No index can occur more than twice in a given term.

3.2. Identities

The most general $\varepsilon\varepsilon$ identity is as follows:

$$\begin{aligned}\varepsilon_{ijk}\varepsilon_{pqr} = \delta_{ip}\delta_{jq}\delta_{kr} - \delta_{jp}\delta_{iq}\delta_{kr} \\ + \delta_{jp}\delta_{kq}\delta_{ir} - \delta_{kp}\delta_{jq}\delta_{ir} \\ + \delta_{kp}\delta_{iq}\delta_{jr} - \delta_{ip}\delta_{kq}\delta_{jr}\end{aligned}$$

This is, however, very verbose and not used often throughout the course. It is provable by noting the total antisymmetry in i, j, k and p, q, r on both sides of the equation implies that both sides agree up to a constant factor. We can check that this factor is 1 by substituting in values such as $i = p = 1$, $j = q = 2$ and $k = r = 3$.

The next most generic form is a very useful identity.

$$\varepsilon_{ijk}\varepsilon_{pqk} = \delta_{ip}\delta_{jq} - \delta_{iq}\delta_{jp}$$

This is essentially the first line of the above identity, noting that $k = r$. We can prove this is true by observing the antisymmetry, and that both sides vanish under $i = j$ or $p = q$. So it suffices to check two cases: $i = p, j = q$ and $i = q, j = p$.

We can now continue making more indices equal to each other to get even more specific identities:

$$\varepsilon_{ijk}\varepsilon_{pjk} = 2\delta_{ip}$$

This is easy to prove by noting that $\delta_{jj} = \sum_j \delta_{jj} = 3$, and using the δ rewrite rule.

Finally, we have

$$\varepsilon_{ijk}\varepsilon_{ijk} = 6$$

No indices are free here, so the values of i, j, k themselves are predetermined by the fact that we are in three-dimensional space.

Using the summation convention (as will now be implied for the remainder of the course), we can prove the vector triple product identity

$$\begin{aligned}
[\mathbf{a} \times (\mathbf{b} \times \mathbf{c})]_i &= \varepsilon_{ijk} \mathbf{a}_j (\mathbf{b} \times \mathbf{c})_k \\
&= \varepsilon_{ijk} \mathbf{a}_j \varepsilon_{pqk} \mathbf{b}_p \mathbf{c}_q \\
&= \varepsilon_{ijk} \varepsilon_{pqk} \mathbf{a}_j \mathbf{b}_p \mathbf{c}_q \\
&= (\delta_{ip}\delta_{jq}) \mathbf{a}_j \mathbf{b}_p \mathbf{c}_q - (\delta_{iq}\delta_{jp}) \mathbf{a}_j \mathbf{b}_p \mathbf{c}_q \\
&= (\mathbf{a} \cdot \mathbf{c}) \mathbf{b}_i - (\mathbf{a} \cdot \mathbf{b}) \mathbf{c}_i
\end{aligned}$$

4. Higher dimensional vectors

4.1. Multidimensional real space

We define multidimensional real space as follows:

$$\mathbb{R}^n = \{\mathbf{x} = (x_1, x_2, \cdots, x_n) : x_i \in \mathbb{R}\}$$

We can define addition and scalar multiplication by mapping these operations over each term in the tuple. Therefore, we have a notion of linear combinations of vectors and hence a concept of parallel vectors. We can say, like before in $\mathbb{R}^3$, that $\mathbf{x} \parallel \mathbf{y}$ if and only if $\mathbf{x} = \lambda\mathbf{y}$ or $\mathbf{y} = \lambda\mathbf{x}$.

We define an operator analogous to the scalar product in $\mathbb{R}^3$. The inner product is defined as $x \cdot y = x_i y_i$. Directly from this definition, we can deduce some properties:

- (symmetric) $\mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x}$
- (bilinear) $(\lambda\mathbf{x} + \lambda'\mathbf{x}') \cdot \mathbf{y} = \lambda\mathbf{x} \cdot \mathbf{y} + \lambda'\mathbf{x}' \cdot \mathbf{y}$
- (positive definite) $\mathbf{x} \cdot \mathbf{x} \geq 0$, and the equality holds if and only if $\mathbf{x} = \mathbf{0}$.

We can define the norm of a vector (similar to the concept of length in three-dimension space), denoted $|\mathbf{x}|$, by $|\mathbf{x}|^2 = \mathbf{x} \cdot \mathbf{x}$. We can now define orthogonality as follows: $\mathbf{x} \perp \mathbf{y} \iff \mathbf{x} \cdot \mathbf{y} = 0$.

We define the standard basis vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ by setting each element of the tuple $\mathbf{e}_i$ to zero apart from the ith element, which is set to one. Also, we redefine the Kronecker δ to be valid in higher-dimensional space. Note that under this definition, the standard basis vectors are orthonormal because $\mathbf{e}_i \cdot \mathbf{e}_j = \delta_{ij}$.

4.2. Cauchy–Schwarz inequality

Proposition. For vectors $\mathbf{x}, \mathbf{y}$ in $\mathbb{R}^n$, $|\mathbf{x} \cdot \mathbf{y}| \leq |\mathbf{x}||\mathbf{y}|$, where the equality is true if and only if $\mathbf{x} \parallel \mathbf{y}$.

Proof. If $\mathbf{y} = \mathbf{0}$, then the result is immediate. So suppose that $\mathbf{y} \neq 0$, then for some $\lambda \in \mathbb{R}$, we have

$$\begin{aligned} |\mathbf{x} - \lambda\mathbf{y}|^2 &= (\mathbf{x} - \lambda\mathbf{y}) \cdot (\mathbf{x} - \lambda\mathbf{y}) \\ &= |\mathbf{x}|^2 - 2\lambda\mathbf{x} \cdot \mathbf{y} + \lambda^2|\mathbf{y}|^2 \geq 0 \end{aligned}$$

As this is a positive real quadratic in λ that is always greater than zero, it has at most one real root. Therefore the discriminant is less than or equal to zero.

$$(-2\mathbf{x} \cdot \mathbf{y})^2 - 4|\mathbf{x}|^2|\mathbf{y}|^2 \leq 0 \implies |\mathbf{x} \cdot \mathbf{y}| \leq |\mathbf{x}||\mathbf{y}|$$

where the equality only holds if $\mathbf{x}$ and $\mathbf{y}$ are parallel (i.e. when $\mathbf{x} - \lambda\mathbf{y}$ equals zero for some λ). □

4.3. Triangle inequality

Following from the Cauchy–Schwarz inequality,

$$\begin{aligned}|\mathbf{x}+\mathbf{y}|^2 &= |\mathbf{x}|^2 + 2(\mathbf{x}\cdot\mathbf{y}) + |\mathbf{y}|^2 \\ &\leq |\mathbf{x}|^2 + 2|\mathbf{x}||\mathbf{y}| + |\mathbf{y}|^2 \\ &= (|\mathbf{x}| + |\mathbf{y}|)^2\end{aligned}$$

where the equality holds under the same conditions as above.

4.4. Levi-Civita ε in higher dimensions

Note that the Levi-Civita ε has three indices in $\mathbb{R}^3$. We can extend this ε to higher and lower dimensions by increasing or reducing the amount of indices. It does not make logical sense to use the same ε without changing the amount of indices to define, for example, a vector product in four-dimensional space, since we would have unused indices. The expression $(\mathbf{x}\times\mathbf{y})_k = \varepsilon_{ijk}\mathbf{a}_i\mathbf{b}_j$ works because there is one free index, k, on the right hand side, so we can use this to calculate the values of each element of the result.

We can, however, use this ε to extend the notion of a scalar triple product to other dimensions, for example two-dimensional space, with $[\mathbf{a},\mathbf{b}] \coloneqq \varepsilon_{ij}\mathbf{a}_i\mathbf{b}_j$. This is the signed area of the parallelogram spanning $\mathbf{a}$ and $\mathbf{b}$.

4.5. General real vector spaces

Vector spaces are not studied axiomatically in this course, but the axioms are given here for completeness. A real (as in, $\mathbb{R}$) vector space V is a set of objects with two operators $+ : V\times V\to V$ and $\cdot : \mathbb{R}\times V\to V$ such that

- $(V,+)$ is an abelian group
- $\lambda(v+w) = \lambda v + \lambda w$
- $(\lambda+\mu)v = \lambda v + \mu v$
- $\lambda(\mu v) = (\lambda\mu)v$
- $1v = v$ (to exclude trivial cases for example $\lambda v = 0$ for all v)

A subspace of a real vector space V is a subset $U\subseteq V$ that is a vector space. Equivalently, if all pairs of vectors $v,w\in U$ satisfy $\lambda v+\mu w\in U$, then U is a subspace of V. Note that the span generated from a set of vectors is a subspace, as it is characterised by this equivalent definition. Also, note that the origin must be part of any subspace, because multiplying a vector by zero must yield the origin.

In some real vector space V, let $\mathbf{v}_1,\mathbf{v}_2\cdots\mathbf{v}_r$ be vectors in V. Now consider the linear relation

$$\lambda_1\mathbf{v}_1 + \lambda_2\mathbf{v}_2 + \cdots + \lambda_r\mathbf{v}_r = 0$$

Then we call the set of vectors a linearly independent set if the only solution is where all λ values are zero. Otherwise, it is a linearly dependent set.

4.6. Inner product spaces

An inner product is an extra structure that we can have on a real vector space V, which is often denoted by angle brackets or parentheses. It can also be characterised by axioms (specifically the ones in Section 6.2). Features like the norm of a vector, and theorems like the Cauchy–Schwarz inequality, follow from these axioms.

For example, let us consider the vector space

$$V = \{f : [0,1] \to \mathbb{R} : f \text{ smooth}; f(0) = f(1) = 0\}$$

We can define the inner product to be

$$f \cdot g = \langle f, g \rangle = \int_0^1 f(x)g(x)\,\mathrm{d}x$$

Then by the Cauchy–Schwarz inequality, we have

$$|\langle f, g\rangle| \leq \|f\| \cdot \|g\|$$
$$\therefore \left| \int_0^1 f(x)g(x)\,\mathrm{d}x \right| \leq \sqrt{\int_0^1 f(x)^2\,\mathrm{d}x}\sqrt{\int_0^1 g(x)^2\,\mathrm{d}x}$$

Lemma. In any real inner product space V, if $\mathbf{v}_1 \cdots v_r \neq \mathbf{0}$ are orthogonal, they are linearly independent.

Proof. If $\sum_i \alpha_i \mathbf{v}_i = 0$, then

$$\left\langle \mathbf{v}_j, \sum_i \alpha_i \mathbf{v}_i \right\rangle = 0$$

And because each vector that is not $\mathbf{v}_j$ is orthogonal to it, those terms cancel, leaving

$$\begin{aligned} \therefore \langle \mathbf{v}_j, \alpha_j \mathbf{v}_j \rangle &= 0 \\ \alpha_j \langle \mathbf{v}_j, \mathbf{v}_j \rangle &= 0 \\ \alpha_j &= 0 \end{aligned}$$

So they are linearly independent. □

4.7. Bases and dimensions

In a vector space V, a basis is a set $\mathcal{B} = \{\mathbf{e}_1 \cdots \mathbf{e}_n\}$ such that

- $\mathcal{B}$ spans V; and
- $\mathcal{B}$ is linearly independent, which implies that the coefficients on these basis vectors are unique for any vector in V, since it is impossible to write one vector in terms of the others

Theorem. If $\{\mathbf{e}_1 \cdots \mathbf{e}_n\}$ and $\{\mathbf{f}_1 \cdots \mathbf{f}_m\}$ are bases for a real vector space V, then $n = m$, which we call the dimension of V.

Proof. This proof is non-examinable (without prompts). We can write each basis vector in terms of the others, since they all span the same vector space. Thus:

$$\mathbf{f}_a = \sum_i A_{ai}\mathbf{e}_i; \quad \mathbf{e}_i = \sum_a B_{ia}\mathbf{f}_a$$

Note that indices i, j span from 1 to n, while a, b span from 1 to m. We can substitute one expression into the other, forming:

$$\mathbf{f}_a = \sum_i A_{ai}\left(\sum_b B_{ib}\mathbf{f}_b\right)$$
$$\mathbf{f}_a = \sum_b \left(\sum_i A_{ai}B_{ib}\right)\mathbf{f}_b$$

Note that we have now written $\mathbf{f}_a$ as a linear combination of $\mathbf{f}_b$ for all valid b. But since they are linearly independent, the coefficient of $\mathbf{f}_b$ must be zero if $a \neq b$, and one of $a = b$. Therefore, we have

$$\delta_{ab} = \sum_i A_{ai}B_{ib}$$

We can make a similar statement about $\mathbf{e}_i$:

$$\delta_{ij} = \sum_a B_{ia}A_{aj} = \sum_a A_{aj}B_{ia}$$

Now, assigning $a = b$ and $i = j$, summing over both, and substituting into our two previous expressions for δ, we have:

$$\sum_{ia} A_{ai}B_{ia} = \sum_a \delta_{aa} = \sum_i \delta_{ii}$$
$$= m \qquad = n$$

□

Note that $\{\mathbf{0}\}$ is a trivial subspace of all vector spaces, and it has dimension zero since it requires a linear combination of no vectors.

Proposition. Let V be a vector space with finite subsets $Y = \{\mathbf{w}_1, \cdots, \mathbf{w}_m\}$ that spans V, and $X = \{\mathbf{u}_1, \cdots, \mathbf{u}_k\}$ that is linearly independent. Let $n = \dim V$. Then:

(i) A basis can be found as a subset of Y by discarding vectors in Y as necessary, and that $n \leq m$.

(ii) X can be extended to a basis by adding in additional vectors from Y as necessary, and that $k \leq n$.

Proof. This proof is non-examinable (without prompts).

(i) If Y is linearly independent, then Y is a basis and $m = n$. Otherwise, Y is not linearly independent. So there exists some linear relation

$$\sum_{i=1}^{m} \lambda_i \mathbf{w}_i = \mathbf{0}$$

where there is some i such that $\lambda_i \neq 0$. Without loss of generality (because the order of elements in Y does not matter) we will reorder Y such that $\mathbf{w}_m \neq 0$. So we have

$$\mathbf{w}_m = \frac{-1}{\lambda_m} \sum_{i=1}^{m-1} \lambda_i \mathbf{w}_i$$

So $\operatorname{span} Y = \operatorname{span}(Y \setminus \{\mathbf{w}_m\})$. We can repeat this process of eliminating vectors from Y until linear independence is achieved. We know that this process will end because Y is a finite set. Clearly, in this case, $n < m$. So for all cases, $n \leq m$.

(ii) If X spans V, then X is a basis and $k = n$. Else, there exists some $u_{k+1} \in V$ that is not in the span of X. Then, we will construct an arbitrary linear relation

$$\sum_{i=1}^{k+1} \mu_i \mathbf{u}_i = \mathbf{0}$$

Note that this implies that $\mu_{k+1} = \mathbf{0}$ because it is not in the span of X, and that $\mu_i = 0$ for all $i \leq k$ because the original X was linearly independent. So we know that all the coefficients are zero, and therefore $X \cup \{u_{k+1}\}$ is linearly independent.

Note that we can always choose this u_{k+1} to be an element of Y because we just need to ensure that $u_{k+1} \notin \operatorname{span} X$. Suppose we cannot choose such a vector in Y. Then $Y \subseteq \operatorname{span} X \implies \operatorname{span} Y \subseteq \operatorname{span} X \implies \operatorname{span} X = V$, which is clearly false because X does not span V. This is a contradiction, so we can always choose such a vector from Y. We can repeat this process of taking vectors from Y and adding them to X until we have a basis. This process will always terminate in a finite amount of steps because we are taking new vectors from a finite set Y. Therefore $k \leq n$, as we are adding vectors (increasing k) until $k = n$.

□

It is perfectly possible to have a vector space that has infinite dimensionality. However, they will be rarely touched upon in this course apart from specific examples, like the following example. Let $V = \{f : [0,1] \to \mathbb{R} : f \text{ smooth}, f(0) = f(1) = 0\}$. Then let $S_n(x) = \sqrt{2}\sin(n\pi x)$ where n is a natural number $1, 2, \cdots$. Clearly, $S_n \in V$ for all n. The inner product of two of these S functions is given by

$$\begin{aligned}\langle S_n, S_m\rangle &= 2\int_0^1 \sin(n\pi x)\sin(m\pi x)\,\mathrm{d}x \\ &= \delta_{mn}\end{aligned}$$

So S_n are orthonormal and therefore linearly independent. So we can continue adding more vectors until it becomes a basis. However, the set of all S_n is already infinite—so V must have infinite dimensionality.

4.8. Multidimensional complex space

We define $\mathbb{C}^n$ by

$$\mathbb{C}^n \coloneqq \{\mathbf{z} = (z_1, z_2, \cdots, z_n) : \forall i, z_i \in \mathbb{C}\}$$

We define addition and scalar multiplication in obvious ways. Note that we have a choice over what the scalars are allowed to be. If we only allow scalars that are real numbers, $\mathbb{C}^n$ can be considered a real vector space with bases $(0, \cdots, 1, \cdots, 0)$ and $(0, \cdots, i, \cdots, 0)$ and dimension $2n$. Alternatively, if we let the scalars be any complex numbers, we don't need to have imaginary bases, thus giving us a complex vector space with bases $(0, \cdots, 1, \cdots, 0)$ and dimension n. We can say that $\mathbb{C}^n$ has dimension $2n$ over $\mathbb{R}$, and dimension n over $\mathbb{C}$. From here on, unless stated otherwise, we treat $\mathbb{C}^n$ to be a complex vector space.

We can define the inner product by

$$\langle \mathbf{z}, \mathbf{w}\rangle \coloneqq \sum_j \overline{z_j} w_j$$

The conjugate over the z terms ensures that the inner product is positive definite. It has these properties, analogous to the properties of the inner product in the real vector space $\mathbb{R}^n$:

- (Hermitian) $\langle \mathbf{z}, \mathbf{w}\rangle = \overline{\langle \mathbf{w}, \mathbf{z}\rangle}$
- (linear/antilinear) $\langle \mathbf{z}, \lambda\mathbf{w} + \lambda'\mathbf{w}'\rangle = \lambda\langle \mathbf{z}, \mathbf{w}\rangle + \lambda'\langle \mathbf{z}, \mathbf{w}'\rangle$ and $\langle \lambda\mathbf{z} + \lambda'\mathbf{z}', w\rangle = \overline{\lambda}\langle \mathbf{z}, \mathbf{w}\rangle + \overline{\lambda'}\langle \mathbf{z}', \mathbf{w}\rangle$
- (positive definite) $\langle \mathbf{z}, \mathbf{z}\rangle = \sum_j |z_j|^2$ which is real and greater than or equal to zero, where the equality holds if and only if $\mathbf{z} = \mathbf{0}$.

We can also define the norm of $\mathbf{z}$ to satisfy $|\mathbf{z}| \geq 0$ and $|\mathbf{z}|^2 = \langle \mathbf{z}, \mathbf{z}\rangle$. Note that the standard basis for $\mathbb{C}^n$ is orthonormal, since the inner product of any two basis vectors $\mathbf{e}_j$ and $\mathbf{e}_k$ is given by δ_{jk}.

Here is an example of the use of the complex inner product on $\mathbb{C}^1 = \mathbb{C}$. Note first that $\langle z, w\rangle = \overline{z}w$. Let $z = a_1 + ia_2$ and $w = b_1 + ib_2$ where $a_1, a_2, b_1, b_2 \in \mathbb{R}$. Then

$$\begin{aligned}\langle z, w\rangle &= \overline{z}w \\ &= (a_1b_1 + a_2b_2) + i(a_1b_2 - a_2b_1) \\ &= (z \cdot w) + i[z, w]\end{aligned}$$

We can therefore use the inner product to compute two different scalar products at the same time.

5. Linear maps

5.1. Introduction

A linear map (or linear transformation) is some operation $T : V \to W$ between vector spaces V and W preserving the core vector space structure (specifically, the linearity). It is defined such that

$$T(\lambda\mathbf{x} + \mu\mathbf{y}) = \lambda T(\mathbf{x}) + \mu T(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in V$ where the scalars λ and μ match up with the scalar field that V and W use (so this could be $\mathbb{R}$ or $\mathbb{C}$ in our examples). Much of the language used for linear maps between vector spaces is analogous to the language used for homomorphisms between groups.

Note that a linear map is completely determined by its action on a basis $\{\mathbf{e}_1, \cdots, \mathbf{e}_n\}$ where $n = \dim V$, since

$$T\left(\sum_i x_i \mathbf{e}_i\right) = \sum_i x_i T(\mathbf{e}_i)$$

We denote $\mathbf{x}' = T(\mathbf{x}) \in W$, and define $\mathbf{x}'$ as the image of x under T. Further, we define

$$\mathrm{Im}(T) = \{\mathbf{x}' \in W : \mathbf{x}' = T(\mathbf{x}) \text{ for some } \mathbf{x} \in V\}$$

to be the image of T, and we define

$$\ker(T) = \{\mathbf{x} \in V : T(\mathbf{x}) = \mathbf{0}\}$$

to be the kernel of T.

Lemma. $\ker T$ is a subspace of V, and $\mathrm{Im}\, T$ is a subspace of W.

Proof. To verify that some subset is a subspace, it suffices to check that it is non-empty, and that it is closed under linear combinations.

$\ker T$ is non-empty because $\mathbf{0} \in \ker T$. For $\mathbf{x}, \mathbf{y} \in \ker T$, we have $T(\lambda\mathbf{x} + \mu\mathbf{y}) = \lambda T(\mathbf{x}) + \mu T(\mathbf{y}) = \mathbf{0} \in \ker T$ as required.

$\mathrm{Im}\, T$ is non-empty because $\mathbf{0} \in \mathrm{Im}\, T$. For $\mathbf{x}, \mathbf{y} \in V$, let $\mathbf{x}' = T(\mathbf{x})$ and $\mathbf{y}' = T(\mathbf{y})$, therefore $\mathbf{x}', \mathbf{y}' \in \mathrm{Im}\, T$. Now, $\lambda\mathbf{x}' + \mu\mathbf{y}' = T(\lambda\mathbf{x} + \mu\mathbf{y})$ so it is closed under linear combinations as required. □

Here are some examples of images and kernels.

(i) The zero linear map $\mathbf{x} \mapsto \mathbf{0}$ has:

$$\mathrm{Im}\, T = \{\mathbf{0}\}$$
$$\ker T = V$$

(ii) The identity linear map $\mathbf{x} \mapsto \mathbf{x}$ has:

$$\begin{aligned} \operatorname{Im} T &= V \\ \ker T &= \{\mathbf{0}\} \end{aligned}$$

(iii) Let $T : \mathbb{R}^3 \to \mathbb{R}^3$, such that

$$\begin{aligned} x_1' &= 3x_1 - x_2 + 5x_3 \\ x_2' &= -x_1 - 2x_3 \\ x_3' &= 2x_1 + x_2 + 3x + 3 \end{aligned}$$

This map has

$$\begin{aligned} \operatorname{Im} T &= \left\{ \lambda \begin{pmatrix} 3 \\ -1 \\ 2 \end{pmatrix} + \mu \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} : \lambda, \mu \in \mathbb{R} \right\} \\ \ker T &= \left\{ \lambda \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix} : \lambda \in \mathbb{R} \right\} \end{aligned}$$

5.2. Rank and nullity

We define the rank of a linear map to be the dimension of its image, and the nullity of a linear map to be the dimension of its kernel.

$$\operatorname{rank} T = \dim \operatorname{Im} T; \quad \operatorname{null} T = \dim \ker T$$

Note that therefore for $T : V \to W$, we have $\operatorname{rank} T \leq \dim W$ and $\ker T \leq \dim V$.

Theorem. For some linear map $T : V \to W$,

$$\operatorname{rank} T + \operatorname{null} T = \dim V$$

Proof. This proof is non-examinable (without prompts). Let $\mathbf{e}_1, \cdots, \mathbf{e}_k$ be a basis for $\ker T$, so $T(\mathbf{e}_i) = \mathbf{0}$ for all valid i. We may extend this basis by adding more vectors $\mathbf{e}_i$ where $k < i \leq n$ until we have a basis for V, where $n = \dim V$. We claim that the set $\mathcal{B} = \{T(\mathbf{e}_{k+1}), \cdots, T(\mathbf{e}_n)\}$ is a basis for $\operatorname{Im} T$. If this is true, then clearly the result follows because $k = \dim \ker T = \operatorname{null} T$ and $n - k = \dim \operatorname{Im} T = \operatorname{rank} T$.

To prove the claim we need to show that $\mathcal{B}$ spans $\operatorname{Im} T$ and that it is a linearly independent set.

- $\mathcal{B}$ spans $\operatorname{Im} T$ because for any $\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{e}_i$, we have

$$T(\mathbf{x}) = \sum_{i=k+1}^{n} x_i T(\mathbf{e}_i) \in \operatorname{span} \mathcal{B}$$

- $\mathcal{B}$ is linearly independent. Consider a general linear combination of basis vectors:

$$\sum_{i=k+1}^{n} \lambda_i T(\mathbf{e}_i) = 0 \implies T\left(\sum_{i=k+1}^{n} \lambda_i \mathbf{e}_i\right) = 0$$

so

$$\sum_{i=k+1}^{n} \lambda_i \mathbf{e}_i \in \ker T$$

Because this is in the kernel, it may be written in terms of the basis vectors of the kernel. So, we have

$$\sum_{i=k+1}^{n} \lambda_i \mathbf{e}_i = \sum_{i=1}^{k} \mu_i \mathbf{e}_i$$

This is a linear relation in terms of all basis vectors of V. So all coefficients are zero.

□

5.3. Rotations

Linear maps are often used to describe geometrical transformations, such as rotations, reflections, projections, dilations and shears. A convenient way to express these maps is by describing where the basis vectors are mapped to. In $\mathbb{R}^2$, we may describe a rotation anticlockwise around the origin by angle θ with

$$\begin{aligned}\mathbf{e}_1 &\mapsto \cos\theta \mathbf{e}_1 + \sin\theta \mathbf{e}_2\\ \mathbf{e}_2 &\mapsto -\sin\theta \mathbf{e}_1 + \cos\theta \mathbf{e}_2\end{aligned}$$

In $\mathbb{R}^3$ we can construct a similar transformation for a rotation around the $\mathbf{e}_3$ axis with

$$\begin{aligned}\mathbf{e}_1 &\mapsto \cos\theta \mathbf{e}_1 + \sin\theta \mathbf{e}_2\\ \mathbf{e}_2 &\mapsto -\sin\theta \mathbf{e}_1 + \cos\theta \mathbf{e}_2\\ \mathbf{e}_3 &\mapsto \mathbf{e}_3\end{aligned}$$

We can extend this to a general rotation in $\mathbb{R}^3$ about an axis given by a unit normal vector $\hat{\mathbf{n}}$. For any vector $\mathbf{x} \in \mathbb{R}^3$ we can resolve parallel and perpendicular to $\hat{\mathbf{n}}$ as follows.

$$\mathbf{x} = \mathbf{x}_\parallel + \mathbf{x}_\perp; \quad \mathbf{x}_\parallel = (\mathbf{x}\cdot\hat{\mathbf{n}})\hat{\mathbf{n}}; \quad \mathbf{x}_\perp = \mathbf{x} - (\mathbf{x}\cdot\hat{\mathbf{n}})\hat{\mathbf{n}}$$

Note that $\hat{\mathbf{n}}$ resembles the $\mathbf{e}_3$ axis here, and $\mathbf{x}_\perp$ resembles the $\mathbf{e}_1$ axis. So we can compute the equivalent of $\mathbf{e}_2$ using the cross product, $\hat{\mathbf{n}} \times \mathbf{x}_\perp = \hat{\mathbf{n}} \times \mathbf{x}$. Now we may define the map with

$$\begin{aligned}\mathbf{x}_\parallel &\mapsto \mathbf{x}_\parallel\\ \mathbf{x}_\perp &\mapsto (\cos\theta)\mathbf{x}_\perp + (\sin\theta)(\hat{\mathbf{n}} \times \mathbf{x})\end{aligned}$$

So all together, we have

$$\mathbf{x} \mapsto (\cos\theta)\mathbf{x} + (1-\cos\theta)(\hat{\mathbf{n}}\cdot\mathbf{x})\hat{\mathbf{n}} + (\sin\theta)(\hat{\mathbf{n}} \times \mathbf{x})$$

5.4. Reflections and projections

For a plane with normal $\hat{\mathbf{n}}$, we define a projection to be

$$\begin{aligned}\mathbf{x}_\parallel &\mapsto \mathbf{0}\\ \mathbf{x}_\perp &\mapsto \mathbf{x}_\perp\\ \mathbf{x} &\mapsto \mathbf{x}_\perp = \mathbf{x} - (\mathbf{x}\cdot\hat{\mathbf{n}})\hat{\mathbf{n}}\end{aligned}$$

and a reflection to be

$$\begin{aligned}\mathbf{x}_\parallel &\mapsto -\mathbf{x}_\parallel\\ \mathbf{x}_\perp &\mapsto \mathbf{x}_\perp\\ \mathbf{x} &\mapsto \mathbf{x}_\perp - \mathbf{x}_\parallel = \mathbf{x} - 2(\mathbf{x}\cdot\hat{\mathbf{n}})\hat{\mathbf{n}}\end{aligned}$$

The same expressions also apply in $\mathbb{R}^2$, where we replace the plane with a line.

5.5. Dilations

Given scale factors $\alpha, \beta, \gamma > 0$, we define a dilation along the axes by

$$\begin{aligned}\mathbf{e}_1 &\mapsto \alpha\mathbf{e}_1\\ \mathbf{e}_2 &\mapsto \beta\mathbf{e}_2\\ \mathbf{e}_3 &\mapsto \gamma\mathbf{e}_3\end{aligned}$$

5.6. Shears

Let $\mathbf{a}, \mathbf{b}$ be orthogonal unit vectors in $\mathbb{R}^3$, i.e. $|\mathbf{a}| = |\mathbf{b}| = \mathbf{0}$ and $\mathbf{a}\cdot\mathbf{b} = 0$, and we define a real parameter λ. A shear is defined as

$$\begin{aligned}\mathbf{x} &\mapsto \mathbf{x}' = \mathbf{x} + \lambda\mathbf{a}(\mathbf{x}\cdot\mathbf{b})\\ \mathbf{a} &\mapsto \mathbf{a}\\ \mathbf{b} &\mapsto \mathbf{b} + \lambda\mathbf{a}\end{aligned}$$

This definition holds equivalently in $\mathbb{R}^2$.

5.7. Matrices

Consider a linear map $T : \mathbb{R}^n \to \mathbb{R}^m$, with standard bases $\{\mathbf{e}_i\} \in \mathbb{R}^n$, $\{\mathbf{f}_a\}, \in \mathbb{R}^m$, and with $T(\mathbf{x}) = \mathbf{x}'$. Let further

$$\mathbf{x} = \sum_i x_i\mathbf{e}_i = \begin{pmatrix} x_1\\ x_2\\ \vdots\\ x_n\end{pmatrix}; \quad x' = \sum_a x'_a\mathbf{f}_a = \begin{pmatrix} x'_1\\ x'_2\\ \vdots\\ x'_m\end{pmatrix}$$

Linearity implies that T is fixed by specifying

$$T(\mathbf{e}_i) = \mathbf{e}'_i = \mathbf{C}_i \in \mathbb{R}^m$$

We take these $\mathbf{C}$ as columns of an $m \times n$ array or matrix M, with rows denoted as $\mathbf{R}_a \in \mathbb{R}^n$.

$$\begin{pmatrix} \uparrow & & \uparrow \\ \mathbf{C}_1 & \cdots & \mathbf{C}_n \\ \downarrow & & \downarrow \end{pmatrix} = M = \begin{pmatrix} \leftarrow & \mathbf{R}_1 & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{R}_m & \rightarrow \end{pmatrix}$$

M has entries $M_{ai} \in \mathbb{R}$, where a labels rows and i labels columns, so

$$(\mathbf{C}_i)_a = M_{ai} = (\mathbf{R}_a)_i$$

The action of T is then given by the matrix M multiplying the vector $\mathbf{x}$ in the following way:

$$\mathbf{x}' = M\mathbf{x}$$

defined by

$$x'_a = M_{ai}x_i$$

or explicitly:

$$\begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_m \end{pmatrix} = \begin{pmatrix} M_{11} & M_{12} & \cdots & M_{1n} \\ M_{21} & M_{22} & \cdots & M_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ M_{m1} & M_{m2} & \cdots & M_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} M_{11}x_1 + M_{12}x_2 + \cdots + M_{1n}x_n \\ M_{21}x_1 + M_{22}x_2 + \cdots + M_{2n}x_n \\ \vdots \\ M_{m1}x_1 + M_{m2}x_2 + \cdots + M_{mn}x_n \end{pmatrix}$$

To check that the matrix multiplication above gives the action of T, we can plug in a generic value $\mathbf{x}$, and we get

$$\mathbf{x}' = T\left(\sum_i x_i \mathbf{e}_i\right) = \sum_i x_i T(\mathbf{e}_i) = \sum_i x_i \mathbf{C}_i$$

and by taking component a of the vector, we have

$$x'_a = \sum_i x_i (\mathbf{C}_i)_a = \sum_i x_i M_{ai}$$

as required. Note also that

$$x'_a = M_{ai}x_i = (\mathbf{R}_a)_i x_i = \mathbf{R}_a \cdot \mathbf{x}$$

We can now regard the properties of T as properties of M (suitably interpreted). For example:

- $\mathrm{Im}(T) = \mathrm{Im}(M) = \mathrm{span}\{\mathbf{C}_1, \cdots, \mathbf{C}_n\}$. In words, the image of a matrix is the span of its columns.

- $\ker(T) = \ker(M) = \{\mathbf{x} : \forall a, \mathbf{R}_a \cdot \mathbf{x} = 0\}$. In some sense, the kernel of M is the subspace perpendicular to all of its rows.

Example. (i) The zero map $\mathbb{R}^n \to \mathbb{R}^m$ corresponds to the zero matrix

$$M = 0 \text{ with } M_{ai} = 0$$

(ii) The identity map $\mathbb{R}^n \to \mathbb{R}^n$ corresponds to the identity (or unit) matrix

$$M = I \text{ with } I_{ij} = \delta_{ij}$$

(iii) The map $\mathbb{R}^3 \to \mathbb{R}^3$ given by $\mathbf{x}' = T(\mathbf{x}) = M\mathbf{x}$ with

$$M = \begin{pmatrix} 3 & 1 & 5 \\ -1 & 0 & -2 \\ 2 & 1 & 3 \end{pmatrix}$$

gives

$$\begin{pmatrix} x_1' \\ x_2' \\ x_3' \end{pmatrix} = \begin{pmatrix} 3x_1 + x_2 + 5x_3 \\ -x_1 - 2x_3 \\ 2x_1 + x_2 + 3x_3 \end{pmatrix}$$

In this case, we may read off the column vectors $\mathbf{C}_a$ from the matrix. Note that since they form a linearly dependent set, we have

$$\mathrm{Im}(T) = \mathrm{Im}(M) = \mathrm{span}\{\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3\} = \mathrm{span}\{\mathbf{C}_1, \mathbf{C}_2\}$$

Here, $\mathbf{R}_2 \times \mathbf{R}_3 = \begin{pmatrix} 2 & -1 & -1 \end{pmatrix}^\intercal = \mathbf{u}$ is actually perpendicular to all rows as they form a linearly dependent set. So

$$\ker(T) = \ker(M) = \{\lambda \mathbf{u}\}$$

(iv) A rotation through θ in $\mathbb{R}^2$ is given by (building from the images of the basis vectors):

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

(v) A dilation $\mathbf{x}' = M\mathbf{x}$ with scale factors α, β, γ along axes in $\mathbb{R}^3$ is given by

$$\begin{pmatrix} \alpha & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & \gamma \end{pmatrix}$$

(vi) A reflection in a plane perpendicular to a unit vector $\hat{\mathbf{n}}$ is given by a matrix H that must have the property that

$$\mathbf{x}' = H\mathbf{x} = \mathbf{x} - 2(\mathbf{x} - \hat{\mathbf{n}})\hat{\mathbf{n}}$$
$$x_i' = x_i - 2x_j n_j n_i = H_{ij} x_j$$

And by comparing coefficients of x_j, and using δ to rewrite x_i using the j index, we have

$$H_{ij} = \delta_{ij} - 2n_i n_j$$

For example, with $\hat{\mathbf{n}} = \frac{1}{\sqrt{3}}\begin{pmatrix}1 & 1 & 1\end{pmatrix}$, then $n_i n_j = \frac{1}{3}$ for all i, j, so

$$H = \frac{1}{3}\begin{pmatrix}1 & -2 & -2\\ -2 & 1 & -2\\ -2 & -2 & 1\end{pmatrix}$$

(vii) A shear is defined by a matrix S such that

$$\mathbf{x}' = S\mathbf{x} = \mathbf{x} + \lambda(\mathbf{b}\cdot\mathbf{x})\mathbf{a}$$

where $\mathbf{a}$, $\mathbf{b}$ are unit vectors with $\mathbf{a} \perp \mathbf{b}$, and where λ is a real scale factor. Therefore:

$$\begin{aligned} x'_i &= x_i + \lambda b_j x_j a_i = S_{ij}x_j \\ \therefore S_{ij} &= \delta_{ij} + \lambda a_i b_j \end{aligned}$$

For example in $\mathbb{R}^2$ with $\mathbf{a} = \begin{pmatrix}1\\0\end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix}0\\1\end{pmatrix}$, we have

$$S = \begin{pmatrix}1 & \lambda\\ 0 & 1\end{pmatrix}$$

(viii) A rotation matrix R in $\mathbb{R}^3$ with axis $\hat{\mathbf{n}}$ and angle θ must satisfy

$$\begin{aligned} \mathbf{x}' &= R\mathbf{x} = (\cos\theta)\mathbf{x} + (1-\cos\theta)(\hat{\mathbf{n}}\cdot\mathbf{x})\hat{\mathbf{n}} + (\sin\theta)(\hat{\mathbf{n}}\times\mathbf{x}) \\ x'_i &= (\cos\theta)x_i + (1-\cos\theta)n_j x_j n_i - (\sin\theta)\varepsilon_{ijk}x_j n_k = R_{ij}x_j \\ \therefore R_{ij} &= \delta_{ij}(\cos\theta) - (1-\cos\theta)n_i n_j - (\sin\theta)\varepsilon_{ijk}n_k \end{aligned}$$

5.8. Matrix of a general linear map

Consider a linear map $T : V \to W$ between general real or complex vector spaces of dimension n, m respectively. We will choose bases $\{\mathbf{e}_i\}$ for V and $\{\mathbf{f}_a\}$ for W. The matrix representing the linear map T with respect to these bases is an $m \times n$ array with entries $M_{ai} \in \mathbb{R}$ or $\mathbb{C}$ as appropriate, defined by

$$T(\mathbf{e}_i) = \sum_a \mathbf{f}_a M_{ai}$$

Then

$$\mathbf{x}' = T(\mathbf{x}) \iff x'_a = \sum_i M_{ai}x_i = M_{ai}x_i$$

where

$$\mathbf{x} = \sum_i x_i \mathbf{e}_i; \quad \mathbf{x}' = \sum_a x_a \mathbf{f}_a$$

Note therefore that (in real vector spaces) given choices of bases $\{\mathbf{e}_i\}$ and $\{\mathbf{f}_a\}$, V is identified with $\mathbb{R}_n$ in the sense that any vector has n real components, and that W is identified with R_m analogously, and that therefore T is identified with an $m \times n$ real matrix M. Note further that entries in column i of M are components of $T(\mathbf{e}_i)$ with respect to basis $\{\mathbf{f}_a\}$.

5.9. Linear combinations

If $T : V \to W$ and $S : V \to W$, between real or complex vector spaces V, W of dimension n, m respectively, are linear, then

$$\alpha T + \beta S : V \to W$$

is also a linear map, where

$$(\alpha T + \beta S)(\mathbf{x}) = \alpha T(\mathbf{x}) + \beta S(\mathbf{x})$$

for any $\mathbf{x} \in V$. So the set of linear maps is a vector space. If M and N are the $m \times N$ matrices for T, S then $\alpha M + \beta N$ is the $m \times n$ matrix for the linear combination above, where

$$(\alpha M + \beta N)_{ai} + \alpha M_{ai} + \beta N_{ai}; \quad a = 1, \cdots, m; \quad i = 1, \cdots, n$$

with respect to the same bases.

5.10. Matrix multiplication

If A is an $m \times n$ matrix with entries A_{ai}, and B is an $n \times p$ matrix with entries B_{ir}, then we define AB to be an $m \times p$ matrix with entries

$$(AB)_{ar} = A_{ai} B_{ir}; \quad a = 1, \cdots, m; \quad i = 1, \cdots, n; \quad r = 1, \cdots, p$$

The product is not defined unless the amount of columns of A matches the number of rows of B.

Matrix multiplication corresponds to composition of linear maps. Consider linear maps:

$$\begin{aligned} S : \mathbb{R}^p \to \mathbb{R}^n;\ S(\mathbf{x}) = B\mathbf{x},\ \mathbf{x} \in \mathbb{R}^p \\ T : \mathbb{R}^n \to \mathbb{R}^m;\ T(\mathbf{x}) = A\mathbf{x},\ \mathbf{x} \in \mathbb{R}^n \\ \implies T \circ S : \mathbb{R}^p \to \mathbb{R}^m;\ (T \circ S)(\mathbf{x}) = (AB)x \end{aligned}$$

since

$$[(AB)\mathbf{x}]_a = (AB)_{ar} x_r$$

and

$$A(B(\mathbf{x})) = A_{ai}(B\mathbf{x})_i = A_{ai} B_{ir} x_r = (AB)_{ar} x_r$$

as required. The definition of matrix multiplication ensures that these answers agree. Of course, this proof works for complex or general vector spaces.

Whenever the products are defined, then for any scalars λ and μ:

- $(\lambda M + \mu N)P = \lambda MP + \mu NP$
- $P(\lambda M + \mu N) = \lambda PM + \mu PN$
- $(MN)P = M(NP)$
- $IM = MI = M$ where $I_{ij} = \delta_{ij}$

We may view matrix multiplication in the following ways.

(i) Regarding a vector $\mathbf{x} \in \mathbb{R}^n$ as a column vector (an $n \times 1$ matrix), then the matrix-vector and matrix-matrix multiplication rules agree.

(ii) Consider the product AB where A is an $m \times n$ matrix and B is an $n \times p$, with columns $\mathbf{C}_r(B) \in \mathbb{R}^n$ and columns $\mathbf{C}_r(AB) \in \mathbb{R}^m$, where $1 \leq r \leq p$. The columns are related by $\mathbf{C}_r(AB) = A\mathbf{C}_r(B)$. Less formally, each column in the right matrix is acted on by the left matrix as if it were a vector, then the resultant vectors are combined into the output matrix.

(iii) In terms of rows and columns,

$$AB = \begin{pmatrix} & \vdots & \\ \leftarrow & \mathbf{R}_n(A) & \rightarrow \\ & \vdots & \end{pmatrix} \begin{pmatrix} \cdots & \uparrow & \cdots \\ & \mathbf{C}_r(B) & \\ & \downarrow & \end{pmatrix}$$

gives

$$\begin{aligned}(AB)_{ar} &= [\mathbf{R}_a(A)]_i \, [\mathbf{C}_r(B)]_i \\ &= \mathbf{R}_a(A) \cdot \mathbf{C}_r(B) \text{ for real matrices, where the } \cdot \text{ is the dot product in } R^n\end{aligned}$$

5.11. Matrix inverses

If A is an $m \times n$ then B, an $n \times m$ matrix, is a left inverse of A if $BA = I$ (the $n \times n$ identity matrix). C is a right inverse of A if $AC = I$ (the $m \times m$ identity matrix). If $m = n$ (A is square), then one of these implies the other; there is no distinction between left and right inverses. We say that $B = C = A^{-1}$, *the* inverse of the matrix A, such that $AA^{-1} = A^{-1}A = I$. Not every matrix has an inverse. If such an inverse exists, A is called invertible, or non-singular.

Consider $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^n$ or $\mathbb{C}^n$, and M is an $n \times n$ matrix. If M^{-1} exists, we can solve the equation $\mathbf{x}' = M\mathbf{x}$ for $\mathbf{x}$, given $\mathbf{x}'$, because we can apply the matrix inverse on the left. For example, where $n = 2$, we have

$$M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}$$

and

$$x_1' = M_{11}x_1 + M_{12}x_2$$
$$x_2' = M_{21}x_1 + M_{22}x_2$$

We can solve these simultaneous equations to construct the general matrix inverse.

$$M_{22}x_1' - M_{12}x_2' = (\det M)x_1$$
$$-M_{21}x_1' + M_{11}x_2' = (\det M)x_2$$

where $\det M = M_{11}M_{22} - M_{12}M_{21}$, called the determinant of the matrix. Where the determinant is nonzero, the matrix inverse

$$M^{-1} = \frac{1}{\det M}\begin{pmatrix} M_{22} & -M_{12} \\ -M_{21} & M_{11} \end{pmatrix}$$

exists. Note that

$$\mathbf{C}_1 = M\mathbf{e}_1 = \begin{pmatrix} M_{11} \\ M_{21} \end{pmatrix}$$
$$\mathbf{C}_2 = M\mathbf{e}_2 = \begin{pmatrix} M_{12} \\ M_{22} \end{pmatrix}$$
$$\iff \det M = [\mathbf{C}_1, \mathbf{C}_2] = [M\mathbf{e}_1, M\mathbf{e}_2] \text{ in } \mathbb{R}^2$$

So the determinant gives the signed factor by which areas are scaled under the action of M. $\det M$ is nonzero if and only if $M\mathbf{e}_1$ and $M\mathbf{e}_2$ are linearly independent, which is true if and only if the image of M has dimension 2, i.e. M has maximal rank. For example, a shear

$$S(\lambda) = \begin{pmatrix} 1 & \lambda \\ 0 & 1 \end{pmatrix}$$

has determinant 1, so areas are preserved. In particular, in this case,

$$S^{-1}(\lambda) = \begin{pmatrix} 1 & -\lambda \\ 0 & 1 \end{pmatrix} = S(-\lambda)$$

As another example, we know that a matrix $R(\theta)$ for a rotation about a fixed axis $\hat{\mathbf{n}}$ through angle θ has formula

$$R(\theta)_{ij}R(-\theta)_{jk} = (\delta_{ij}\cos\theta + (1-\cos\theta)n_in_j - \varepsilon_{ijp}n_p\sin\theta) \times (\delta_{jk}\cos\theta + (1-\cos\theta)n_jn_k + \varepsilon_{jkq}n_q\sin\theta)$$

Expanding out, noting that $n_in_i = 1$ as $\hat{\mathbf{n}}$ is a unit vector, and cancelling:

$$= \delta_{ik}\cos^2\theta + 2\cos\theta(1-\cos\theta)n_in_k + (1-\cos\theta)^2n_in_k - \varepsilon_{ijp}\varepsilon_{jkq}n_pn_q\sin^2\theta$$

By using an $\varepsilon\varepsilon$ identity:

$$= \delta_{ik}\cos^2\theta + (1-\cos^2\theta)n_in_k + \delta_{ik}n_pn_p\sin^2\theta - (\sin^2\theta)n_in_k$$
$$= \delta_{ik}\cos^2\theta + \delta_{ik}n_pn_p\sin^2\theta$$
$$= \delta_{ik}\cos^2\theta + \delta_{ik}\sin^2\theta$$
$$= \delta_{ik}$$

as required.

6. Transpose and Hermitian conjugate

6.1. Transpose

If M is an $m \times n$ (real or complex) matrix, the transpose $M^\intercal$ is an $n \times m$ matrix defined by

$$(M^\intercal)_{ia} = M_{ai}$$

which essentially exchanges rows and columns. Here are some key properties.

- $(\alpha A + \beta B)^\intercal = \alpha A^\intercal + \beta B^\intercal$ for α, β scalars, and A, B both $m \times n$ matrices.
- $(AB)^\intercal = B^\intercal A^\intercal$, where A is $m \times n$ and B is $n \times p$. This is because

$$\begin{aligned}[(AB)^\intercal]_{ra} &= (AB)_{ar} \\ &= A_{ai}B_{ir} \\ &= (A^\intercal)_{ia}(B^\intercal)_{ri} \\ &= (B^\intercal)_{ri}(A^\intercal)_{ia} \\ &= (B^\intercal A^\intercal)_{ra}\end{aligned}$$

- If $\mathbf{x}$ is a column vector (or an $n \times 1$ matrix), $\mathbf{x}^\intercal$ is the equivalent row vector (a $1 \times n$ matrix).
- The inner product in $\mathbb{R}^n$ can therefore be written $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^\intercal \mathbf{y}$. Note that this is not equivalent to $\mathbf{x}\mathbf{y}^\intercal$, which is known as the outer product, which results in a matrix not a scalar.
- If M is $n \times n$ (square) then M is:
  - symmetric iff $M^\intercal = M$, or $M_{ij} = M_{ji}$
  - antisymmetric iff $M^\intercal = -M$, or $M_{ij} = -M_{ji}$
- Any M which is square can be written as a sum of a symmetric and and an antisymmetric part

$$M = S + A \quad \text{where } S = \frac{1}{2}(M + M^\intercal); \quad A = \frac{1}{2}(M - M^\intercal)$$

  as S is symmetric and A is antisymmetric by construction.
- If A is 3×3 and antisymmetric, then we can write

$$A_{ij} = \varepsilon_{ijk} a_k \text{ where } A = \begin{pmatrix} 0 & a_3 & -a_2 \\ -a_3 & 0 & a_1 \\ a_2 & -a_1 & 0 \end{pmatrix}$$

  Then, we have

$$(A\mathbf{x})_i = \varepsilon_{ijk} a_k x_j = (\mathbf{x} \times \mathbf{a})_i$$

6.2. Hermitian conjugate

Let M be an $m \times n$ matrix. Then the Hermitian conjugate (also known as the conjugate transpose) $M^\dagger$ is an $n \times m$ matrix defined by

$$(M^\dagger)_{ia} = \overline{M_{ai}}$$

If M is square, then M is Hermitian if and only if $M^\dagger = M$, or alternatively $M_{ia} = \overline{M_{ai}}$; M is anti-Hermitian if $M^\dagger = -M$, or alternatively $M_{ia} = -\overline{M_{ai}}$. Similarly to above, if $\mathbf{z}$ is a column vector in $\mathbb{C}^n$ (an $n \times 1$ matrix), then the complex inner product is given by $\mathbf{z} \cdot \mathbf{w} = \mathbf{z}^\dagger \mathbf{w}$.

6.3. Trace

For a complex $n \times n$ (square) matrix M, the trace of the matrix, denoted $\mathrm{tr}(M)$, is defined by

$$\mathrm{tr}(M) = M_{ii} = M_{11} + M_{22} + \cdots + M_{nn}$$

It has a number of key properties.

- $\mathrm{tr}(\alpha M + \beta N) = \alpha \,\mathrm{tr}\, M + \beta \,\mathrm{tr}\, N$ where α and β are scalars, and M and N are $n \times n$ matrices.
- $\mathrm{tr}(MN) = \mathrm{tr}(NM)$ where M is $m \times n$ and N is $n \times m$. MN and NM need not have the same dimension, but their traces are identical. We can check this as follows: $\mathrm{tr}(MN) = (MN)_{aa} = M_{ai}N_{ia} = N_{ia}M_{ai} = (NM)_{ii} = \mathrm{tr}(NM)$.
- $\mathrm{tr}(M^\intercal) = \mathrm{tr}(M)$
- $\mathrm{tr}(I) = \delta_{ii} = n$ where n is the dimensionality of the vector space.
- If S is $n \times n$ and symmetric, let

$$\begin{aligned} T &= S - \frac{1}{n}\mathrm{tr}(S)I \\ \text{or } T_{ij} &= S_{ij} - \frac{1}{n}\mathrm{tr}(S)\delta_{ij} \\ \text{then } \mathrm{tr}(T) = T_{ii} &= S_{ii} = \frac{1}{n}\mathrm{tr}(S)\delta_{ii} \\ &= \mathrm{tr}(S) - \frac{1}{n}\mathrm{tr}(S) = 0 \end{aligned}$$

  Then $S = T + \frac{1}{n}\mathrm{tr}(S)I$ where T is traceless and the right hand term $\frac{1}{n}\mathrm{tr}(S)I$ is 'pure trace'.
- If A is $n \times n$ antisymmetric, $\mathrm{tr}(A) = A_{ii} = 0$.

6.4. Orthogonal matrices

A real $n \times n$ matrix U is orthogonal if and only if its transpose is its inverse.

$$U^\intercal U = UU^\intercal = I$$

These conditions can be written

$$U_{ki}U_{kj} = U_{ik}U_{jk} = \delta_{ij}$$

In words, the left hand side says that the columns of U are orthonormal, and the middle part of the equation says that the rows of U are orthonormal.

$$U^\intercal U = \begin{pmatrix} & \vdots & \\ \leftarrow & \mathbf{C}_i & \rightarrow \\ & \vdots & \end{pmatrix}\begin{pmatrix} & \uparrow & \\ \cdots & \mathbf{C}_j & \cdots \\ & \downarrow & \end{pmatrix} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}$$

For example, if $U = R(\theta)$ is a rotation through θ around an axis $\hat{\mathbf{n}}$, then $U^\intercal = R(\theta)^\intercal = R(-\theta) = R(\theta)^{-1} = U^{-1}$. An equivalent definition for orthogonality is: U is orthogonal if and only if it preserves the inner product on $\mathbb{R}^n$.

$$(U\mathbf{x}) \cdot (U\mathbf{y}) = \mathbf{x} \cdot \mathbf{y} \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$

To check equivalence:

$$\begin{aligned} (U\mathbf{x}) \cdot (U\mathbf{y}) &= (U\mathbf{x})^\intercal (U\mathbf{y}) \\ &= (\mathbf{x}^\intercal U^\intercal)(U\mathbf{y}) \\ &= \mathbf{x}^\intercal (U^\intercal U)\mathbf{y} \\ &= \mathbf{x}^\intercal \mathbf{y} \\ &= \mathbf{x} \cdot \mathbf{y} \end{aligned}$$

which is true if and only if $U^\intercal U = I$. Note that in $\mathbb{R}^n$, the columns of U are $U\mathbf{e}_i, \cdots, U\mathbf{e}_n$ so the inner product is preserved when U acts on the standard basis vectors if and only if

$$(U\mathbf{e}_i) \cdot (U\mathbf{e}_j) = \mathbf{e}_i \cdot \mathbf{e}_j = \delta_{ij}$$

i.e. the columns of U are orthonormal.

Let us now try to find a general 2×2 orthogonal matrix. We begin by transforming the basis vectors. $\mathbf{e}_i = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ must be transformed to a unit vector. Therefore, in the most general sense:

$$U\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$$

for some parameter θ. Now, the other basis vector $\mathbf{e}_2$ must be orthogonal to it, and so it must be

$$U\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \pm\begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}$$

So we have two cases:

$$U = R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}; \quad U = H = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$$

where R is a rotation by θ and H is a reflection in $\mathbb{R}^2$, where

$$\hat{\mathbf{n}} = \begin{pmatrix} -\sin\frac{\theta}{2} \\ \cos\frac{\theta}{2} \end{pmatrix}$$

because

$$H_{ij} = \delta_{ij} - 2n_i n_j \therefore H = \begin{pmatrix} 1 - 2\sin^2\frac{\theta}{2} & 2\sin\frac{\theta}{2}\cos\frac{\theta}{2} \\ 2\sin\frac{\theta}{2}\cos\frac{\theta}{2} & 1 - 2\cos^2\frac{\theta}{2} \end{pmatrix}$$

which simplifies as required. Note that $\det R = +1$, but $\det H = -1$.

6.5. Unitary matrices

A complex $n \times n$ matrix U is called unitary if and only if

$$U^\dagger U = UU^\dagger = I$$

Equivalently, U is unitary if and only if it preserves the complex inner product on $\mathbb{C}_n$:

$$\langle U\mathbf{z}, U\mathbf{w}\rangle = \langle \mathbf{z}, \mathbf{w}\rangle \quad \forall \mathbf{z}, \mathbf{w} \in \mathbb{C}^n$$

To check equivalence:

$$\begin{aligned} \langle U\mathbf{z}, U\mathbf{w}\rangle &= (U\mathbf{z})^\dagger (U\mathbf{w}) \\ &= (\mathbf{z}^\dagger U^\dagger)(U\mathbf{w}) \\ &= \mathbf{z}^\dagger (U^\dagger U)\mathbf{w} \\ &= \mathbf{z}^\dagger \mathbf{w} \end{aligned}$$

which of course matches if and only if $U^\dagger U = I$.

7. Adjugates and alternating forms

7.1. Inverses in two dimensions

Consider a linear map $T: \mathbb{R}^n \to \mathbb{R}^n$. If T is invertible (i.e. bijective), then $\ker T = \{\mathbf{0}\}$ as T is injective, and $\operatorname{Im} T = \mathbb{R}^n$ as T is surjective. These conditions are actually equivalent due to the rank-nullity theorem. Conversely, if the conditions hold, then $T(\mathbf{e}_1), T(\mathbf{e}_2), \cdots, T(\mathbf{e}_n)$ must be a basis of the image, so we can just define T^{-1} by defining its actions on the basis vectors $T(\mathbf{e}_1), T(\mathbf{e}_2) \cdots T(\mathbf{e}_n)$, specifically mapping them to the standard basis.

How can we test whether the conditions above hold for a matrix M representing T, and how can we find M^{-1} from M explicitly? For any $n \times n$ matrix M (not necessarily invertible), we will define the adjugate matrix $\widetilde{M}$ and the determinant $\det M$ such that

$$\widetilde{M}M = (\det M)I \tag{$*$}$$

Then if $\det M \neq 0$, M is invertible, where

$$M^{-1} = \frac{1}{\det M}\widetilde{M}$$

From $n = 2$, recall that $(*)$ holds with

$$M = \begin{pmatrix} M_{11} & M_{21} \\ M_{12} & M_{22} \end{pmatrix}; \quad \widetilde{M} = \begin{pmatrix} M_{22} & -M_{21} \\ -M_{12} & M_{11} \end{pmatrix}; \quad \det M = [M\mathbf{e}_1, M\mathbf{e}_2] = \varepsilon_{ij} M_{i1} M_{j2}$$

The determinant in this case is the factor by which areas scale under M. $\det M \neq 0$ if and only if $M\mathbf{e}_1, M\mathbf{e}_2$ are linearly independent.

7.2. Three dimensions

For $n = 3$, we will define similarly

$$\det M = [M\mathbf{e}_1, M\mathbf{e}_2, M\mathbf{e}_3] = \varepsilon_{ijk} M_{i1} M_{j2} M_{k3}$$

We define it like this because this is the factor by which volumes scale under M in three dimensions. So

$$\det M \neq 0 \iff \{M\mathbf{e}_1, M\mathbf{e}_2, M\mathbf{e}_3\} \text{ linearly independent, or } \operatorname{Im} M = \mathbb{R}^3$$

Now we define $\widetilde{M}$ from M using row/column notation.

$$\begin{aligned} \mathbf{R}_1(\widetilde{M}) &= \mathbf{C}_2(M) \times \mathbf{C}_3(M) \\ \mathbf{R}_2(\widetilde{M}) &= \mathbf{C}_3(M) \times \mathbf{C}_1(M) \\ \mathbf{R}_3(\widetilde{M}) &= \mathbf{C}_1(M) \times \mathbf{C}_2(M) \end{aligned}$$

Note that therefore,

$$(\widetilde{M}M)_{ij} = \mathbf{R}_i(\widetilde{M}) \cdot \mathbf{C}_j(M) = \underbrace{(\mathbf{C}_1(M) \times \mathbf{C}_2(M) \cdot \mathbf{C}_3(M))}_{\det M} \delta_{ij}$$

as claimed. For example, let us invert the following matrix.

$$M = \begin{pmatrix} 1 & 3 & 0 \\ 0 & -1 & -2 \\ 4 & 1 & -1 \end{pmatrix}$$

$$\mathbf{C}_2 \times \mathbf{C}_3 = \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} \times \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix} = \begin{pmatrix} -1 \\ 3 \\ 6 \end{pmatrix}$$

$$\mathbf{C}_3 \times \mathbf{C}_1 = \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix} \times \begin{pmatrix} 1 \\ 0 \\ 4 \end{pmatrix} = \begin{pmatrix} 8 \\ -1 \\ -2 \end{pmatrix}$$

$$\mathbf{C}_1 \times \mathbf{C}_2 = \begin{pmatrix} 1 \\ 0 \\ 4 \end{pmatrix} \times \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 11 \\ -1 \end{pmatrix}$$

$$\widetilde{M} = \begin{pmatrix} -1 & 3 & 6 \\ 8 & -1 & -2 \\ 4 & 11 & -1 \end{pmatrix}$$

$$\det M = \mathbf{C}_1 \cdot \mathbf{C}_2 \times \mathbf{C}_3 = 23$$

$$\widetilde{M}M = 23I$$

7.3. Levi-Civita ε in higher dimensions

Recall (from IA Groups):

- A permutation σ on the set $\{1, 2, \cdots, n\}$ is a bijection from the set to itself, specified by an ordered list $\sigma(1), \sigma(2), \cdots, \sigma(n)$.
- Permutations form a group S_n, called the symmetric group of order $n!$
- A transposition $\tau = (p, q)$ where $p \neq q$ is a permutation that swaps p and q.
- Any permutation is a product of of k transpositions, where k is unique modulo 2 for a given σ. In this course, we will write $\varepsilon(\sigma)$ to mean the sign (or signature) of the permutation, $(-1)^k$. σ is even if the sign is 1, and odd if the sign is -1.

The alternating symbol ε in $\mathbb{R}^n$ or $\mathbb{C}^n$ is an n-index object (tensor) defined by

$$\underbrace{\varepsilon_{ij\cdots l}}_{n \text{ indices}} = \begin{cases} +1 & \text{if } i, j \cdots, l \text{ is an even permutation of } 1, 2, \cdots, n \\ -1 & \text{if } i, j \cdots, l \text{ is an odd permutation of } 1, 2, \cdots, n \\ 0 & \text{otherwise, i.e. if any indices take the same value} \end{cases}$$

Thus if σ is any permutation, then

$$\varepsilon_{\sigma(1)\cdots\sigma(n)} = \varepsilon(\sigma)$$

So $\varepsilon_{ij\ldots l}$ is totally antisymmetric and changes sign whenever a pair of indices are exchanged.

Definition. Given vectors $\mathbf{v}_1, \cdots \mathbf{v}_n \in \mathbb{R}^n$ or $\mathbb{C}^n$, the alternating form combines them to give the scalar

$$\begin{aligned}[\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n] &= \varepsilon_{ij\ldots l}(\mathbf{v}_1)_i(\mathbf{v}_2)_j \cdots (\mathbf{v}_n)_l \\ &= \sum_{\sigma \in S_n} \varepsilon(\sigma) \cdot (\mathbf{v}_1)_{\sigma(1)} \cdot (\mathbf{v}_2)_{\sigma(2)} \cdots (\mathbf{v}_n)_{\sigma(n)}\end{aligned}$$

7.4. Properties

(i) The alternating form is multilinear.

$$\begin{aligned}[\mathbf{v}_1, \cdots, \mathbf{v}_{p-1}, \alpha\mathbf{u} + \beta\mathbf{w}, \mathbf{v}_{p+1} \cdots, \mathbf{v}_n] = \alpha[\mathbf{v}_1, \cdots, \mathbf{v}_{p-1}, \mathbf{u}, \mathbf{v}_{p+1} \cdots, \mathbf{v}_n] \\ + \beta[\mathbf{v}_1, \cdots, \mathbf{v}_{p-1}, \mathbf{w}, \mathbf{v}_{p+1} \cdots, \mathbf{v}_n]\end{aligned}$$

(ii) It is totally antisymmetric. $[\mathbf{v}_{\sigma(1)}, \mathbf{v}_{\sigma(2)}, \cdots, \mathbf{v}_{\sigma(n)}] = \varepsilon(\sigma)[\mathbf{v}_1, \cdots, \mathbf{v}_n]$

(iii) Standard basis vectors give a positive result: $[\mathbf{e}_i, \cdots, \mathbf{e}_n] = 1$.

These three properties fix the alternating form completely, and they also imply

(iv) If $\mathbf{v}_p = \mathbf{v}_q$ where $p \neq q$, then

$$[\mathbf{v}_1, \cdots, \mathbf{v}_p, \cdots, \mathbf{v}_q, \cdots, \mathbf{v}_n] = 0$$

(v) If $\mathbf{v}_p$ can be written as a non-trivial linear combination of the other vectors, then

$$[\mathbf{v}_1, \cdots, \mathbf{v}_p, \cdots, \mathbf{v}_n] = 0$$

Property (iv) follows from property (ii), where we swap $\mathbf{v}_p$ and $\mathbf{v}_q$. Property (v) follows from substituting the linear combination representation of $\mathbf{v}_p$ into the alternating form expression, the using properties (i) and (iv). To justify (ii) above, it suffices to check a transposition $\tau = (p\ q)$ where (without loss of generality) $p < q$, then since transpositions generate all permutations the result follows.

$$\begin{aligned}&[\mathbf{v}_1, \cdots, \mathbf{v}_{p-1}, \mathbf{v}_q, \mathbf{v}_{p+1}, \cdots, \mathbf{v}_{q-1}, \mathbf{v}_p, \mathbf{v}_{q+1}, \cdots, \mathbf{v}_n] \\ &= \sum_\sigma \varepsilon(\sigma)(\mathbf{v}_1)_{\sigma(1)} \cdots (\mathbf{v}_{p-1})_{\sigma(p-1)}(\mathbf{v}_q)_{\sigma(p)}(\mathbf{v}_{p+1})_{\sigma(p+1)} \\ &\qquad \cdots (\mathbf{v}_{q-1})_{\sigma(q-1)}(\mathbf{v}_p)_{\sigma(q)}(\mathbf{v}_{q+1})_{\sigma(q+1)} \\ &= \sum_\sigma \varepsilon(\sigma)(\mathbf{v}_1)_{\sigma'(1)} \cdots (\mathbf{v}_{p-1})_{\sigma'(p-1)}(\mathbf{v}_q)_{\sigma'(q)}(\mathbf{v}_{p+1})_{\sigma'(p+1)} \\ &\qquad \cdots (\mathbf{v}_{q-1})_{\sigma'(q-1)}(\mathbf{v}_p)_{\sigma'(p)}(\mathbf{v}_{q+1})_{\sigma'(q+1)}\end{aligned}$$

where $\sigma' = \sigma\tau$

$$\begin{aligned}&= -\sum_{\sigma'} \varepsilon(\sigma')(\mathbf{v}_1)_{\sigma'(1)} \cdots (\mathbf{v}_{p-1})_{\sigma'(p-1)}(\mathbf{v}_p)_{\sigma'(p)}(\mathbf{v}_{p+1})_{\sigma'(p+1)} \\ &\qquad \cdots (\mathbf{v}_{q-1})_{\sigma'(q-1)}(\mathbf{v}_q)_{\sigma'(q)}(\mathbf{v}_{q+1})_{\sigma'(q+1)} \\ &= -[\mathbf{v}_1, \cdots, \mathbf{v}_{p-1}, \mathbf{v}_p, \mathbf{v}_{p+1}, \cdots, \mathbf{v}_{q-1}, \mathbf{v}_q, \mathbf{v}_{q+1}, \cdots, \mathbf{v}_n]\end{aligned}$$

as required.

Proposition. $[\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n] \neq 0$ if and only if $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$ are linearly independent.

Proof. To show the forward implication, let us suppose that they are not linearly independent and use property (v). Then we can express some $\mathbf{v}_p$ as a linear combination of the others. Then $[\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n] = 0$.

To show the other direction, note that $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_3$ means that they span, and if they span then each of the standard basis vectors $\mathbf{e}_i$ can be written as a linear combination of the $\mathbf{v}$ vectors, i.e. $\mathbf{e}_i = U_{ai}\mathbf{v}_a$. Then

$$\begin{aligned}[\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n] &= [U_{a1}\mathbf{v}_a, U_{b2}\mathbf{v}_b, \cdots, U_{cn}\mathbf{v}_c] \\ &= U_{a1}U_{b2}\cdots U_{cn}[\mathbf{v}_a, \mathbf{v}_b, \cdots, \mathbf{v}_c] \\ &= U_{a1}U_{b2}\cdots U_{cn}\varepsilon_{ab\cdots c}[\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n]\end{aligned}$$

By definition, the left hand side is $+1$, so $[\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n]$ is nonzero. □

As an example of these ideas, let

$$\mathbf{v}_1 = \begin{pmatrix} i \\ 0 \\ 0 \\ 2 \end{pmatrix}; \quad \mathbf{v}_2 = \begin{pmatrix} 0 \\ 0 \\ 5i \\ 0 \end{pmatrix}; \quad \mathbf{v}_3 = \begin{pmatrix} 3 \\ 2i \\ 0 \\ 0 \end{pmatrix}; \quad \mathbf{v}_4 = \begin{pmatrix} 0 \\ 0 \\ i \\ 1 \end{pmatrix}; \quad \text{where } \mathbf{v}_j \in \mathbb{C}_4$$

Then

$$\begin{aligned}[\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4] &= 5i[\mathbf{v}_1, \mathbf{e}_3, \mathbf{v}_3, \mathbf{v}_4] \\ &= 5i[i\mathbf{e}_1 + 2\mathbf{e}_4, \mathbf{e}_3, 3\mathbf{e}_1 + 2i\mathbf{e}_2, -i\mathbf{e}_3 + \mathbf{e}_4]\end{aligned}$$

By multilinearity, we can eliminate all $\mathbf{e}_3$ terms not in the second position because they will cancel with it, giving

$$= 5i[i\mathbf{e}_1 + 2\mathbf{e}_4, \mathbf{e}_3, 3\mathbf{e}_1 + 2i\mathbf{e}_2, \mathbf{e}_4]$$

And likewise with $\mathbf{e}_4$:

$$= 5i[i\mathbf{e}_1, \mathbf{e}_3, 3\mathbf{e}_1 + 2i\mathbf{e}_2, \mathbf{e}_4]$$

And again with $\mathbf{e}_1$:

$$\begin{aligned}&= 5i[i\mathbf{e}_1, \mathbf{e}_3, 2i\mathbf{e}_2, \mathbf{e}_4] \\ &= 5i \cdot 2i \cdot i[\mathbf{e}_1, \mathbf{e}_3, \mathbf{e}_2, \mathbf{e}_4] \\ &= 10i[\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4] \\ &= 10i\end{aligned}$$

8. Determinant

8.1. Definition

For an $n \times n$ matrix M with columns $\mathbf{C}_a = M\mathbf{e}_a$, then the determinant $\det(M) = |M| \in \mathbb{R}$ or $\mathbb{C}$ is given by any of the following equivalent definitions.

$$\begin{aligned}\det M &= [\mathbf{C}_1, \mathbf{C}_2, \cdots, \mathbf{C}_n] \\ &= [M\mathbf{e}_1, M\mathbf{e}_2, \cdots, M\mathbf{e}_n] \\ &= \varepsilon_{ij\ldots l} M_{i1} M_{j2} \cdots M_{ln} \\ &= \sum_\sigma \varepsilon(\sigma) M_{\sigma(1)1} M_{\sigma(2)2} \cdots M_{\sigma(n)n}\end{aligned}$$

Here are some examples.

(i) $n = 2$

$$\det M = \sum_\sigma M_{\sigma(1)1} M_{\sigma(2)2} = \begin{vmatrix} M_{11} & M_{21} \\ M_{12} & M_{22} \end{vmatrix} = M_{11}M_{22} - M_{12}M_{21}$$

(ii) M diagonal, i.e. $M_{ij} = 0$ for $i \neq j$

$$M = \begin{pmatrix} M_{11} & 0 & \cdots & 0 \\ 0 & M_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & M_{nn} \end{pmatrix} \implies \det M = M_{11}M_{22}\cdots M_{nn}$$

(iii) Let M be $n \times n$, A be $(n-1) \times (n-1)$, where

$$M = \left(\begin{array}{c|c} A & 0 \\ \hline 0 & 1 \end{array}\right)$$

We call M a matrix 'in block form'. So $M_{ni} = M_{in} = 0$ if $i \neq n$. So we can restrict the permutation σ to only transmuting the first $(n-1)$ terms, i.e. $\sigma(n) = n$. So $\det M = \det A$.

Proposition. If $\mathbf{R}_a$ are the rows of M, $\det M$ is given by

$$\begin{aligned}\det M &= [\mathbf{R}_1, \mathbf{R}_2, \cdots, \mathbf{R}_n] \\ &= \varepsilon_{ij\ldots l} M_{1i} M_{2j} \cdots M_{nl} \\ &= \sum_\sigma \varepsilon(\sigma) M_{1\sigma(1)} M_{2\sigma(2)} \cdots M_{n\sigma(n)}\end{aligned}$$

i.e. $\det M = \det M^\intercal$.

Proof. Recall that $(\mathbf{C}_a)_i = M_{ia} = (\mathbf{R}_i)_a$. We need to show that one of these definitions is equivalent to one of the previous definitions, then all other equivalent definitions follow. We use the Σ definition by considering the product $M_{1\sigma(1)}M_{2\sigma(2)}\cdots M_{n\sigma(n)}$. We may rewrite this product in a different order: $M_{\rho(1)1}M_{\rho(2)2}\cdots M_{\rho(n)n}$. Then $\rho = \sigma^{-1}$. But then $\varepsilon(\sigma) = \varepsilon(\rho)$, and a sum over σ is equivalent to a sum over ρ. □

8.2. Expanding by rows or columns

For an $n \times n$ matrix M with entries M_{ia}, we define the minor M^{ia} to be the $(n-1) \times (n-1)$ determinant of the matrix obtained by deleting row i and column a from M.

Proposition. The determinant of a generic $n \times n$ matrix M is given by

$$\begin{aligned} \det M &= \sum_i (-1)^{i+a} M_{ia} M^{ia} \text{ for a fixed } a \\ &= \sum_a (-1)^{i+a} M_{ia} M^{ia} \text{ for a fixed } i \end{aligned}$$

This process is known as expanding by row i or by column a. As an example, let us take the following 4×4 complex matrix

$$M = \begin{pmatrix} i & 0 & 3 & 0 \\ 0 & 0 & 2i & 0 \\ 0 & 5i & 0 & -i \\ 2 & 0 & 0 & 1 \end{pmatrix}$$

Then, the determinant is given by (expanding by row 3)

$$\begin{aligned} \det M &= -5i \begin{vmatrix} i & 3 & 0 \\ 0 & 2i & 0 \\ 2 & 0 & 1 \end{vmatrix} + i \begin{vmatrix} i & 0 & 3 \\ 0 & 0 & 2i \\ 2 & 0 & 0 \end{vmatrix} \\ &= -5i \left[i \begin{vmatrix} 2i & 0 \\ 0 & 1 \end{vmatrix} - 3 \begin{vmatrix} 0 & 0 \\ 2 & 1 \end{vmatrix} \right] + i \left[-2i \begin{vmatrix} i & 0 \\ 2 & 0 \end{vmatrix} \right] \\ &= -5i[i \cdot 2i - 3 \cdot 0] + i[-2i \cdot 0] \\ &= -5i[-2] + i[0] \\ &= 10i \end{aligned}$$

8.3. Row and column operations

Consider the following consequences of the properties of the determinant:

- (row and column scaling) If $\mathbf{R}_i \mapsto \lambda \mathbf{R}_i$ for a fixed i, or $\mathbf{C}_a \mapsto \lambda \mathbf{C}_a$, then $\det M \mapsto \lambda \det M$ by multilinearity. If we scale all rows or columns, then $M \mapsto \lambda M$, so $\det M \mapsto \lambda^n \det M$ where M is an $n \times n$ matrix.
- (row and column operations) If $\mathbf{R}_i \mapsto \mathbf{R}_i + \lambda \mathbf{R}_j$ where $i \neq j$ (or the corresponding conversion with columns), then $\det M \mapsto \det M$.
- (row and column exchanges) If we swap $\mathbf{R}_i$ and $\mathbf{R}_j$ (or two columns), then $\det M \mapsto -\det M$.

For example, let us find the determinant of matrix A, where

$$A = \begin{pmatrix} 1 & 1 & a \\ a & 1 & 1 \\ 1 & a & 1 \end{pmatrix}; \quad a \in \mathbb{C}$$

Then:

$$\det A = \begin{vmatrix} 1 & 1 & a \\ a & 1 & 1 \\ 1 & a & 1 \end{vmatrix}$$

$$\mathbf{C}_1 \mapsto \mathbf{C}_1 - \mathbf{C}_3 : \quad \det A = \begin{vmatrix} 1-a & 1 & a \\ a-1 & 1 & 1 \\ 0 & a & 1 \end{vmatrix}$$

$$\det A = (1-a)\begin{vmatrix} 1 & 1 & a \\ -1 & 1 & 1 \\ 0 & a & 1 \end{vmatrix}$$

$$\mathbf{C}_2 \mapsto \mathbf{C}_2 - \mathbf{C}_3 : \quad \det A = (1-a)\begin{vmatrix} 1 & 1-a & a \\ -1 & 0 & 1 \\ 0 & a-1 & 1 \end{vmatrix}$$

$$\det A = (1-a)^2\begin{vmatrix} 1 & 1 & a \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{vmatrix}$$

$$\mathbf{R}_1 \mapsto \mathbf{R}_1 + \mathbf{R}_2 + \mathbf{R}_3 : \quad \det A = (1-a)^2\begin{vmatrix} 0 & 0 & a+2 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{vmatrix}$$

$$\det A = (1-a)^2(a+2)\begin{vmatrix} -1 & 0 \\ 0 & -1 \end{vmatrix} = (1-a)^2(a+2)$$

8.4. Multiplicative property of determinants

Theorem. For $n \times n$ matrices M, N, $\det(MN) = \det M \cdot \det N$.

We can prove this using the following elaboration on the definition of the determinant:

Lemma.

$$\varepsilon_{i_1 i_2 \cdots i_n} M_{i_1 a_1} M_{i_2 a_2} \cdots M_{i_n a_n} = (\det M)\varepsilon_{a_1 a_2 \cdots a_n}$$

Proof. The left hand side and right hand side are each totally antisymmetric (alternating) in $a_1, a_2, \cdots, a_n$, so they must be related by a constant of proportionality. To fix the constant, we can simply consider taking $a_i = i$ and the result follows. □

Now, we prove the above theorem.

Proof. Using the lemma above:

$$\begin{aligned}\det MN &= \varepsilon_{i_1 i_2 \cdots i_n}(MN)_{i_1 1}(MN)_{i_2 2}\cdots(MN)_{i_n n} \\ &= \varepsilon_{i_1 i_2 \cdots i_n} M_{i_1 k_1} M_{i_2 k_2} \cdots M_{i_n k_n} N_{k_1 1} N_{k_2 2} \cdots N_{k_n n} \\ &= (\det M)\varepsilon_{a_1 a_2 \cdots a_n} N_{k_1 1} N_{k_2 2} \cdots N_{k_n n} \\ &= (\det M)(\det N)\end{aligned}$$

as required. □

Note the following consequences.

(i) $M^{-1}M = I \implies \det\left(M^{-1}\right)\det(M) = \det I = 1$. Therefore, $\det\left(M^{-1}\right) = (\det M)^{-1}$, so $\det M$ must be nonzero for M to be invertible.

(ii) For R real and orthogonal, $R^\intercal R = I \implies \det(R^\intercal)\det(R) = 1$. But $\det(R^\intercal) = \det R$, so $(\det R)^2 = 1$, so $\det R = \pm 1$.

(iii) For U complex and unitary, $U^\dagger U = I \implies \det\left(U^\dagger\right)\det(U) = 1$. But since $U^\dagger = \overline{U^\intercal}$, we have $\overline{\det U}\det U = 1$, so $|(\det U)^2| = 1$, so $|\det U| = 1$.

8.5. Cofactors and determinants

Consider a column of some $n \times n$ matrix M, written in the form

$$\mathbf{C}_a = \sum_i M_{ia}\mathbf{e}_i$$

$$\begin{aligned}\implies \det M &= [\mathbf{C}_1, \cdots, \mathbf{C}_a, \cdots, \mathbf{C}_n] \\ &= [\mathbf{C}_1, \cdots, \mathbf{C}_{a-1}, \sum_i M_{ia}\mathbf{e}_i, \mathbf{C}_{a+1}, \cdots, \mathbf{C}_n] \\ &= \sum_i M_{ia}\Delta_{ia}\end{aligned}$$

where

$$\begin{aligned}\Delta_{ia} &= [\mathbf{C}_1, \cdots, \mathbf{C}_{a-1}, \mathbf{e}_i, \mathbf{C}_{a+1}, \cdots, \mathbf{C}_n] \\ &= \begin{vmatrix} & & & 0 & & & \\ & A & & \vdots & & B & \\ & & & 0 & & & \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ & & & 0 & & & \\ & C & & \vdots & & D & \\ & & & 0 & & & \end{vmatrix}\end{aligned}$$

where the zero entries in the rows arise from antisymmetry, giving

$$\begin{aligned} &= \underbrace{(-1)^{n-a}}_{\text{amount of column transpositions}} \cdot \underbrace{(-1)^{n-i}}_{\text{amount of row transpositions}} \begin{vmatrix} A & B \\ C & D \end{vmatrix} \\ &= (-1)^{i+a} M^{ia} \end{aligned}$$

where M^{ia} is the minor in this position; the determinant of the matrix with this particular row and column removed. We call Δ_{ia} the cofactor.

$$\det M = \sum_i M_{ia}\Delta_{ia} = \sum_i (-1)^{i+a} M_{ia} M^{ia}$$

Similarly, by considering rows,

$$\det M = \sum_a M_{ia}\Delta_{ia} = \sum_a (-1)^{i+a} M_{ia} M^{ia}$$

8.6. Adjugates and inverses

Reasoning as above, consider $\mathbf{C}_b = \sum_i M_{ib}\mathbf{e}_i$. Then,

$$[\mathbf{C}_1, \cdots, \mathbf{C}_{a-1}, \mathbf{C}_b, \mathbf{C}_{a+1}, \cdots, \mathbf{C}_n] = \sum_i M_{ib}\Delta_{ia}$$

If $a = b$ then clearly this is $\det M$. Otherwise, $\mathbf{C}_b$ is equal to one of the other columns, so $\sum_i M_{ib}\Delta_{ia} = 0$.

$$\sum_i M_{ib}\Delta_{ia} = (\det M)\delta_{ab}$$

Similarly,

$$\sum_a M_{ja}\Delta_{ia} = (\det M)\delta_{ij}$$

Now, let Δ be the matrix of cofactors (i.e. entries Δ_{ia}), and we define the adjugate $\widetilde{M} = \Delta^\intercal$. Then

$$\Delta_{ia} M_{ib} = \widetilde{M}_{ai} M_{ib} = (\widetilde{M}M)_{ab} = (\det M)\delta_{ab}$$

Therefore,

$$\widetilde{M}M = (\det M) I$$

We can reach this result similarly considering the other index. Hence, if $\det M \neq 0$ then $M^{-1} = \frac{1}{\det M}\widetilde{M}$.

8.7. Systems of linear equations

Consider a system of n linear equations in n unknowns x_i written in matrix-vector form:

$$A\mathbf{x} = \mathbf{b}, \quad \mathbf{x}, \mathbf{b} \in \mathbb{R}^n,$$

where A is an $n \times n$ matrix. There are three possibilities:

(i) $\det A \neq 0 \implies A^{-1}$ exists so there is a unique solution $\mathbf{x} = A^{-1}\mathbf{b}$

(ii) $\det A = 0$ and $b \notin \operatorname{Im} A$ means that there is no solution

(iii) $\det A = 0$ and $b \in \operatorname{Im} A$ means that there are infinitely many solutions of the form

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{u}$$

where $\mathbf{u} \in \ker A$ and $\mathbf{x}_0$ is a particular solution

A solution therefore exists if and only if $A\mathbf{x}_0 = \mathbf{b}$ for some $\mathbf{x}_0$, which is true if and only if $\mathbf{b} \in \operatorname{Im} A$. Then $\mathbf{x}$ is also a solution if and only if $\mathbf{u} = \mathbf{x} - \mathbf{x}_0$ satisfies

$$A\mathbf{u} = \mathbf{0}$$

This equation is known as the equivalent homogeneous problem. Now, $\det A \neq 0 \iff \operatorname{Im} A = \mathbb{R}^n \iff \ker A = \{\mathbf{0}\}$. So in case (i), there is always a unique solution for any $\mathbf{b}$. But $\det A = 0 \iff \operatorname{rank}(A) < n \iff \operatorname{null} A > 0$. Then either $b \notin \operatorname{Im} A$ as in case (ii), or $b \in \operatorname{Im} A$ as in case (iii).

If $\mathbf{u}_1, \ldots, \mathbf{u}_k$ is a basis for $\ker A$, then the general solution to the homogeneous problem is some linear combination of these basis vectors, i.e.

$$\mathbf{u} = \sum_{i=1}^{k} \lambda_i \mathbf{u}_i, \quad k = \operatorname{null} A$$

This is similar to the complementary function and particular integral technique used to solve linear differential equations.

For example, in $A\mathbf{x} = \mathbf{b}$, let

$$A = \begin{pmatrix} 1 & 1 & a \\ a & 1 & 1 \\ 1 & a & 1 \end{pmatrix}; \quad \mathbf{b} = \begin{pmatrix} 1 \\ c \\ 1 \end{pmatrix}; \quad a, c \in \mathbb{R}$$

We have previously found that $\det A = (a-1)^2(a+2)$. So the cases are:

- $(a \neq 1, a \neq -2)$ $\det A \neq 0$ and A^{-1} exists; we previously found this to be

$$A^{-1} = \frac{1}{(1-a)(2+a)} \begin{pmatrix} 1 & 1+a & 1 \\ 1 & 1 & -1-a \\ -1-a & 1 & 1 \end{pmatrix}$$

For these values of a, there is a unique solution for any c, demonstrating case (i) above:

$$\mathbf{x} = A^{-1}\mathbf{b} = \frac{1}{(1-a)(2+a)}\begin{pmatrix} 2-c-ca \\ c-a \\ c-a \end{pmatrix}$$

Geometrically, this solution is simply a point.

- $(a = 1)$ In this case, the matrix is simply

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \implies \operatorname{Im} A = \operatorname{span}\left\{\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\right\} = \left\{\lambda\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\right\}; \quad \ker A = \operatorname{span}\left\{\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}\right\}$$

Note that $\mathbf{b} \in \operatorname{Im} A$ if and only if $c = 1$, where a particular solution is

$$\mathbf{x}_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

So the general solution is given by

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{u} = \begin{pmatrix} 1-\lambda-\mu \\ \lambda \\ \mu \end{pmatrix}$$

In summary, for $a = 1$, $c = 1$ we have case (iii). Geometrically this is a plane. For $a = 1, c \neq 1$, we have case (ii) where there are no solutions.

- $(a = -2)$ The matrix becomes

$$A = \begin{pmatrix} 1 & 1 & -2 \\ -2 & 1 & 1 \\ 1 & -2 & 1 \end{pmatrix} \implies \operatorname{Im} A = \operatorname{span}\left\{\begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}\right\}; \quad \ker A = \left\{\lambda\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\right\}$$

Now, $\mathbf{b} \in \operatorname{Im} A$ if and only if $c = -2$, the particular solution is

$$\mathbf{x}_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

The general solution is therefore

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{u} = \begin{pmatrix} 1+\lambda \\ \lambda \\ \lambda \end{pmatrix}$$

In summary, for $a = -2$ and $c = -2$ we have case (iii). Geometrically this is a line. For $a = -2, c \neq -2$, we have case (ii) where there are no solutions.

8.8. Geometrical interpretation of solutions of linear equations

Let $\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3$ be the rows of the 3×3 matrix A. Then the rows represent the normals of planes. This is clear by expanding the matrix multiplication of the homogeneous form:

$$\begin{aligned} A\mathbf{u} = \mathbf{0} \iff \mathbf{R}_1 \cdot \mathbf{u} &= 0 \\ \mathbf{R}_2 \cdot \mathbf{u} &= 0 \\ \mathbf{R}_3 \cdot \mathbf{u} &= 0 \end{aligned}$$

So the solution of the homogeneous problem (i.e. finding the general solution) amounts to determining where the planes intersect.

- ($\operatorname{rank} A = 3$) The rows are linearly independent, so the three planes' normals are linearly independent and the planes intersect at $\mathbf{0}$ only.
- ($\operatorname{rank} A = 2$) The normals span a plane, so the planes intersect in a line.
- ($\operatorname{rank} A = 1$) The normals are parallel and therefore the planes coincide.
- ($\operatorname{rank} A = 0$) The normals are all zero, so any vector in $\mathbb{R}^3$ solves the equation.

Now, let us consider instead the original problem $A\mathbf{x} = \mathbf{b}$:

$$\begin{aligned} A\mathbf{b} = \mathbf{0} \iff \mathbf{R}_1 \cdot \mathbf{u} &= b_1 \\ \mathbf{R}_2 \cdot \mathbf{u} &= b_2 \\ \mathbf{R}_3 \cdot \mathbf{u} &= b_3 \end{aligned}$$

The planes still have normals $\mathbf{R}_i$ as before, but they do not necessarily pass through the origin.

- ($\operatorname{rank} A = 3$) The planes' normals are linearly independent and the planes intersect at a point; this is the unique solution.
- ($\operatorname{rank} A < 3$) The existence of a solution depends on the value of $\mathbf{b}$.
  - ($\operatorname{rank} A = 2$) The planes may intersect in a line as before, but they may instead form a sheaf (the planes pairwise intersect in lines but they do not as a triple), or two planes could be parallel and not intersect each other at all.
  - ($\operatorname{rank} A = 1$) The normals are parallel, so the planes may coincide or they might be parallel. There is no solution unless all three planes coincide.

9. Properties of matrices

9.1. Eigenvalues and eigenvectors

For a linear map $T : V \to V$, a vector $\mathbf{v} \in V$ with $\mathbf{v} \neq 0$ is called an eigenvector of T with eigenvalue λ if $T(\mathbf{v}) = \lambda\mathbf{v}$. If $V = \mathbb{R}^n$ or $\mathbb{C}^n$, and T is given by an $n \times n$ matrix A, then

$$A\mathbf{v} = \lambda v \iff (A - \lambda I)\mathbf{v} = \mathbf{0}$$

and for a given λ, this holds for some $\mathbf{v} \neq 0$ if and only if

$$\det(A - \lambda I) = 0$$

This is called the characteristic equation for A. So λ is an eigenvalue if and only if it is a root of the characteristic polynomial

$$\chi_A(t) = \det(A - tI) = \begin{vmatrix} A_{11} - t & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} - t & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} - t \end{vmatrix}$$

We can look for eigenvalues as roots of the characteristic polynomial or characteristic equation, and then determine the corresponding eigenvectors once we've deduced what the possibilities are. Here are a few examples.

(i) $V = \mathbb{C}^2$:

$$A = \begin{pmatrix} 2 & i \\ -i & 2 \end{pmatrix} \implies \det(A - \lambda I) = (2 - \lambda)^2 - 1 = 0$$

So we have $(2 - \lambda)^2 = 1$ so $\lambda = 1$ or 3.

- $(\lambda = 1)$

$$(A - I)\mathbf{v} = \begin{pmatrix} 1 & i \\ -i & 1 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \mathbf{0} \implies \mathbf{v} = \alpha\begin{pmatrix} 1 \\ i \end{pmatrix}$$

  for any $\alpha \neq 0$.

- $(\lambda = 3)$

$$(A - 3I)\mathbf{v} = \begin{pmatrix} -1 & i \\ -i & -1 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \mathbf{0} \implies \mathbf{v} = \beta\begin{pmatrix} 1 \\ -i \end{pmatrix}$$

  for any $\beta \neq 0$.

(ii) $V = \mathbb{R}^2$:

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \implies \det(A - \lambda I) = (1 - \lambda)^2 = 0$$

So $\lambda = 1$ only, a repeated root.

$$(A - I)\mathbf{v} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \mathbf{0} \implies \mathbf{v} = \alpha\begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

for any $\alpha \neq 0$. There is only one (linearly independent) eigenvector here.

(iii) $V = \mathbb{R}^2$ or $\mathbb{C}^2$:

$$U = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \implies \chi_U(t) = \det(U - tI) = t^2 - 2t\cos\theta + 1$$

The eigenvalues λ are $e^{\pm i\theta}$. The eigenvectors are

$$\mathbf{v} = \alpha \begin{pmatrix} 1 \\ \mp i \end{pmatrix}; \quad \alpha \neq 0$$

So there are no real eigenvalues or eigenvectors except when $\theta = n\pi$.

(iv) $V = \mathbb{C}^n$:

$$A = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} \implies \chi_A(t) = \det(A - tI) = (\lambda_1 - t)(\lambda_2 - t)(\lambda_3 - t)\dots(\lambda_n - t)$$

So the eigenvalues are all the λ_i, and the eigenvectors are $\mathbf{v} = \alpha\mathbf{e}_i$ ($\alpha \neq 0$) for each i.

9.2. The characteristic polynomial

For an $n \times n$ matrix A, the characteristic polynomial $\chi_A(t)$ has degree n:

$$\chi_A(t) = \sum_{j=0}^{n} c_j t^j = (-1)^n(t - \lambda_1)\dots(t - \lambda_n)$$

(i) There exists at least one eigenvalue (solution to χ_A), due to the fundamental theorem of algebra, or n roots counted with multiplicity.

(ii) $\operatorname{tr}(A) = A_{ii} = \sum_{i=1}^{n} \lambda_i$, the sum of the eigenvalues. Compare terms of degree $n-1$ in t, and from the determinant we get

$$(-t)^{n-1}A_{11} + (-t)^{n-1}A_{22} + \cdots + (-t)^{n-1}A_{nn}$$

The overall sign matches with the expansion of $(-1)^n(t - \lambda_1)(t - \lambda_2)\dots(t - \lambda_n)$.

(iii) $\det(A) = \chi_A(0) = \prod_{i=1}^{n} \lambda_i$, the product of the eigenvalues.

(iv) If A is real, then the coefficients c_i in the characteristic polynomial are real, so $\chi_A(\lambda) = 0 \iff \chi_A(\overline{\lambda}) = 0$. So the non-real roots occur in conjugate pairs if A is real.

9.3. Eigenspaces and multiplicities

For an eigenvalue λ of a matrix A, we define the eigenspace

$$E_\lambda = \{\mathbf{v} : A\mathbf{v} = \lambda\mathbf{v}\} = \ker(A - \lambda I)$$

All nonzero vectors in this space are eigenvectors. The geometric multiplicity is

$$m_\lambda = \dim E_\lambda = \text{null}(A - \lambda I)$$

equivalent to the number of linearly independent eigenvectors with the given eigenvalue λ. The algebraic multiplicity is

$$M_\lambda = \text{the multiplicity of } \lambda \text{ as a root of } \chi_A(t)$$

i.e. $\chi_A(t) = (t-\lambda)^{M_t} f(t)$, where $f(\lambda) \neq 0$.

Proposition. $M_\lambda \geq m_\lambda$ (and $m_\lambda \geq 1$ since λ is an eigenvalue). The proof of this proposition is delayed until the next section where we will then have the tools to prove it.

Here are some examples.

(i)
$$A = \begin{pmatrix} -2 & 2 & -3 \\ 2 & 1 & -6 \\ -1 & -2 & 0 \end{pmatrix} \implies \chi_A(t) = \det(A - tI) = (5-t)(t+3)^2$$

So $\lambda = 5, -3$. $M_5 = 1, M_{-3} = 2$. We will now find the eigenspaces.

- $(\lambda = 5)$
$$E_5 = \left\{ \alpha \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} \right\}$$
- $(\lambda = -3)$
$$E_{-3} = \left\{ \alpha \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} + \beta \begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix} \right\}$$

Note that to compute the eigenvectors, we just need to solve the equation $(A-\lambda I)\mathbf{x} = \mathbf{0}$. In the case of $\lambda = -3$, for example, we then have

$$\begin{pmatrix} 1 & 2 & -3 \\ 2 & 4 & -6 \\ -1 & -2 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \mathbf{0}$$

We can use the first line of the matrix to get a linear combination for x_1, x_2, x_3, specifically $x_1 + 2x_2 = 3x_3 = 0$, so we can eliminate one of the variables (here, x_1) to get

$$\mathbf{x} = \begin{pmatrix} -2x_2 + 3x_3 \\ x_2 \\ x_3 \end{pmatrix} = \mathbf{0}$$

Now, $\dim E_5 = m_5 = 1 = M_5$. Similarly, $\dim E_{-3} = m_{-3} = 2 = M_{-3}$.

(ii)
$$A = \begin{pmatrix} -3 & -1 & 1 \\ -1 & -3 & 1 \\ -2 & -2 & 0 \end{pmatrix} \implies \chi_A(t) = \det(A - tI) = -(t+2)^3$$

We have a root $\lambda = -2$ with $M_{-2} = 3$. To find the eigenspace, we will look for solutions of:
$$(A + 2I)\mathbf{x} = \begin{pmatrix} -1 & -1 & 1 \\ -1 & -1 & 1 \\ -2 & -2 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \mathbf{0} \implies \mathbf{x} = \begin{pmatrix} -x_2 + x_3 \\ x_2 \\ x_3 \end{pmatrix}$$

So
$$E_{-2} = \left\{ \alpha \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + \beta \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \right\}$$

Further, $m_{-2} = 2 < 3 = M_{-2}$.

(iii) A reflection in a plane through the origin with unit normal $\hat{\mathbf{n}}$ satisfies
$$H\hat{\mathbf{n}} = -\hat{\mathbf{n}}; \quad \forall \mathbf{u} \perp \hat{\mathbf{n}}, H\mathbf{u} = \mathbf{u}$$

The eigenvalues are therefore ± 1 and $E_{-1} = \{\alpha\hat{\mathbf{n}}\}$, and $E_1 = \{\mathbf{x} : \mathbf{x} \cdot \hat{\mathbf{n}} = 0\}$. The multiplicities are given by $M_{-1} = m_{-1} = 1, M_1 = m_1 = 2$.

(iv) A rotation about an axis $\hat{\mathbf{n}}$ through angle θ in $\mathbb{R}^3$ satisfies
$$R\hat{\mathbf{n}} = \hat{\mathbf{n}}$$

So the axis of rotation is the eigenvector with eigenvalue 1. There are no other real eigenvalues unless $\theta = n\pi$. The rotation restricted to the plane perpendicular to $\hat{\mathbf{n}}$ has eigenvalues $e^{\pm i\theta}$ as shown above.

9.4. Linear independence of eigenvectors

Proposition. Let $\mathbf{v}_1, \dots, \mathbf{v}_r$ be eigenvectors of an $n \times n$ matrix A with eigenvalues $\lambda_1, \dots, \lambda_r$. If the eigenvalues are distinct, then the eigenvectors are linearly independent.

Proof. Note that if we take some linear combination $\mathbf{w} = \sum_{j=1}^{r} \alpha_j \mathbf{v}_j$, then $(A - \lambda I)\mathbf{w} = \sum_{j=1}^{r} \alpha_j(\lambda_j - \lambda)\mathbf{v}_j$. Here are two methods for getting this proof.

(i) Suppose the eigenvectors are linearly dependent, so there exist linear combinations $\mathbf{w} = \mathbf{0}$ where some α are nonzero. Let p be the amount of nonzero α values. So, $2 \le p \le r$. Now, pick such a $\mathbf{w}$ for which p is least. Without loss of generality, let α_1 be one of the nonzero coefficients. Then
$$(A - \lambda_1 I)\mathbf{w} = \sum_{j=2}^{r} \alpha_j(\lambda_j - \lambda_1)\mathbf{v}_j = \mathbf{0}$$

This is a linear relation with $p - 1$ nonzero coefficients #.

(ii) Alternatively, given a linear relation $\mathbf{w} = \mathbf{0}$,

$$\prod_{j\neq k}(A - \lambda_j I)\mathbf{w} = \alpha_k \prod_{j\neq k}(\lambda_k - \lambda_j)\mathbf{v}_k = \mathbf{0}$$

for some fixed k. So $\alpha_k = 0$. So the eigenvectors are linearly independent as claimed.

□

Corollary. With conditions as in the proposition above, let $\mathcal{B}_{\lambda_i}$ be a basis for the eigenspace E_{λ_i}. Then $\mathcal{B} = \mathcal{B}_{\lambda_1} \cup \mathcal{B}_{\lambda_2} \cup \cdots \cup \mathcal{B}_{\lambda_r}$ is linearly independent.

Proof. Consider a general linear combination of all these vectors, it has the form

$$\mathbf{w} = \mathbf{w}_1 + \mathbf{w}_2 + \cdots + \mathbf{w}_r$$

where each $\mathbf{w}_i \in E_i$. Applying the same arguments as in the proposition, we find that

$$\mathbf{w} = 0 \implies \forall i\, \mathbf{w}_i = 0$$

So each $\mathbf{w}_i$ is the trivial linear combination of elements of $\mathcal{B}_{\lambda_i}$ and the result follows. □

9.5. Diagonalisability

Proposition. For an $n \times n$ matrix A acting on $V = \mathbb{R}^n$ or $\mathbb{C}^n$, the following conditions are equivalent:

(i) there exists a basis of eigenvectors of A for V, named $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ which $A\mathbf{v}_i = \lambda_i \mathbf{v}_i$ for each i; and

(ii) there exists an $n \times n$ invertible matrix P with the property that

$$P^{-1}AP = D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

If either of these conditions hold, then A is diagonalisable.

Proof. Note that for any matrix P, AP has columns $A\mathbf{C}_i(P)$, and PD has columns $\lambda_i \mathbf{C}_i(P)$. Then (i) and (ii) are related by choosing $\mathbf{v}_i = \mathbf{C}_i(P)$. Then $P^{-1}AP = D \iff AP = PD \iff A\mathbf{v}_i = \lambda_i \mathbf{v}_i$.

In essence, given a basis of eigenvectors as in (i), the relation above defines P, and if the eigenvectors are linearly independent then P is invertible. Conversely, given a matrix P as in (ii), its columns are a basis of eigenvectors. □

Let's try some examples.

(i) Let

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \implies E_1 = \left\{ \alpha \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\}$$

This is a single eigenvalue $\lambda = 1$ with one linearly independent eigenvector. So there is no basis of eigenvectors for $\mathbb{R}^2$ or $\mathbb{C}^2$, so A is not diagonalisable.

(ii) Let

$$U = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \implies E_{e^{i\theta}} = \left\{ \alpha \begin{pmatrix} 1 \\ -i \end{pmatrix} \right\}; \quad E_{e^{-i\theta}} = \left\{ \beta \begin{pmatrix} 1 \\ i \end{pmatrix} \right\}$$

which are two linearly independent complex eigenvectors. So,

$$P = \begin{pmatrix} 1 & 1 \\ -i & i \end{pmatrix}; \quad P^{-1} = \frac{1}{2} \begin{pmatrix} 1 & i \\ 1 & -i \end{pmatrix}; \quad P^{-1}UP = \begin{pmatrix} e^{i\theta} & 0 \\ 0 & e^{i\theta} \end{pmatrix}$$

So U is diagonalisable over $\mathbb{C}^2$ but not over $\mathbb{R}^2$.

9.6. Criteria for diagonalisability

Proposition. Consider an $n \times n$ matrix A.

(i) A is diagonalisable if it has n distinct eigenvalues (sufficient condition).

(ii) A is diagonalisable if and only if for every eigenvalue λ, $M_\lambda = m_\lambda$ (necessary and sufficient condition).

Proof. Use the proposition and corollary above.

(i) If we have n distinct eigenvalues, then we have n linearly independent eigenvectors. Hence they form a basis.

(ii) If λ_i are all the distinct eigenvalues, then $\mathcal{B}_{\lambda_1} \cup \cdots \cup \mathcal{B}_{\lambda_r}$ are linearly independent. The number of elements in this new basis is $\sum_i m_{\lambda_i} = \sum_i M_{\lambda_i} = n$ which is the degree of the characteristic polynomial. So we have a basis.

Note that case (i) is just a specialisation of case (ii) where both multiplicities are 1. □

Let us consider some examples.

(i) Let

$$A = \begin{pmatrix} -2 & 2 & -3 \\ 2 & 1 & -6 \\ -1 & -2 & 0 \end{pmatrix} \implies \lambda = 5, -3; \quad M_5 = m_5 = 1; \quad M_{-3} = m_{-3} = 2$$

So A is diagonalisable by case (ii) above, and moreover

$$P = \begin{pmatrix} 1 & -2 & 3 \\ 2 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}; \quad P^{-1} = \frac{1}{8} \begin{pmatrix} 1 & 2 & -3 \\ -2 & 4 & 6 \\ 1 & 2 & 5 \end{pmatrix} \implies P^{-1}AP = \begin{pmatrix} 5 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & -3 \end{pmatrix}$$

(ii) Let

$$A = \begin{pmatrix} -3 & -1 & 1 \\ -1 & -3 & 1 \\ -2 & 2 & 0 \end{pmatrix} \implies \lambda = -2; \quad M_{-2} = 3 > m_{-2} = 2$$

So A is not diagonalisable. As a check, if it were diagonalisable, then there would be some matrix P such that $P^{-1}AP = -2I \implies A = P(-2I)P^{-1} = -2I$ #.

9.7. Similarity

Matrices A and B (both $n \times n$) are similar if $B = P^{-1}AP$ for some invertible $n \times n$ matrix P. This is an equivalence relation.

Proposition. If A and B are similar, then

(i) $\operatorname{tr} B = \operatorname{tr} A$

(ii) $\det B = \det A$

(iii) $\chi_B = \chi_A$

Proof. (i)

$$\begin{aligned} \operatorname{tr} B &= \operatorname{tr}\left(P^{-1}AP\right) \\ &= \operatorname{tr}\left(APP^{-1}\right) \\ &= \operatorname{tr} A \end{aligned}$$

(ii)

$$\begin{aligned} \det B &= \det\left(P^{-1}AP\right) \\ &= \det P^{-1} \det A \det P \\ &= \det A \end{aligned}$$

(iii)

$$\begin{aligned} \det(B - tI) &= \det\left(P^{-1}AP - tI\right) \\ &= \det\left(P^{-1}AP - tP^{-1}P\right) \\ &= \det\left(P^{-1}(A - tI)P\right) \\ &= \det P^{-1} \det(A - tI) \det P \\ &= \det(A - tI) \end{aligned}$$

□

9.8. Real eigenvalues and orthogonal eigenvectors

Recall that an $n \times n$ matrix A is hermitian if and only if $A^\dagger = \overline{A}^\intercal = A$, or $\overline{A_{ij}} = A_{ji}$. If A is real, then it is hermitian if and only if it is symmetric. The complex inner product for $\mathbf{v}, \mathbf{w} \in \mathbb{C}^n$ is $\mathbf{v}^\dagger \mathbf{w} = \sum_i \overline{v_i} w_i$, and for $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, this reduces to the dot product in $\mathbb{R}^n$, $\mathbf{v}^\intercal \mathbf{w}$.

Here is a key observation. If A is hermitian, then

$$(A\mathbf{v})^\dagger \mathbf{w} = \mathbf{v}^\dagger (A\mathbf{w})$$

Theorem. For an $n \times n$ matrix A that is hermitian:

(i) Every eigenvalue λ is real;

(ii) Eigenvectors $\mathbf{v}, \mathbf{w}$ with different eigenvalues λ, μ respectively, are orthogonal, i.e. $\mathbf{v}^\dagger \mathbf{w} = 0$; and

(iii) If A is real and symmetric, then for each eigenvalue λ we can choose a real eigenvector, and part (ii) becomes $\mathbf{v} \cdot \mathbf{w} = 0$.

Proof. (i) Using the observation above with $\mathbf{v} = \mathbf{w}$ where $\mathbf{v}$ is any eigenvector with eigenvalue λ, we get

$$\begin{aligned} \mathbf{v}^\dagger (A\mathbf{v}) &= (A\mathbf{v})^\dagger \mathbf{v} \\ \mathbf{v}^\dagger (\lambda \mathbf{v}) &= (\lambda \mathbf{v})^\dagger \mathbf{v} \\ \lambda \mathbf{v}^\dagger (\mathbf{v}) &= \overline{\lambda} (\mathbf{v})^\dagger \mathbf{v} \end{aligned}$$

As $\mathbf{v}$ is an eigenvector, it is nonzero, so $\mathbf{v}^\dagger \mathbf{v} \neq 0$, so

$$\lambda = \overline{\lambda}$$

(ii) Using the same observation,

$$\begin{aligned} \mathbf{v}^\dagger (A\mathbf{w}) &= (A\mathbf{v})^\dagger \mathbf{w} \\ \mathbf{v}^\dagger (\mu \mathbf{w}) &= (\lambda \mathbf{v})^\dagger \mathbf{w} \\ \mu \mathbf{v}^\dagger \mathbf{w} &= \lambda \mathbf{v}^\dagger \mathbf{w} \end{aligned}$$

Since $\lambda \neq \mu$, $\mathbf{v}^\dagger \mathbf{w} = 0$, so the eigenvectors are orthogonal.

(iii) Given $A\mathbf{v} = \lambda \mathbf{v}$ with $\mathbf{v} \in \mathbb{C}^n$ but A is real, let

$$\mathbf{v} = \mathbf{u} + i\mathbf{u}'; \quad \mathbf{u}, \mathbf{u}' \in \mathbb{R}^n$$

Since $\mathbf{v}$ is an eigenvector, and this is a linear equation, we have

$$A\mathbf{u} = \lambda \mathbf{u}; \quad A\mathbf{u}' = \lambda \mathbf{u}'$$

So $\mathbf{u}$ and $\mathbf{u}'$ are eigenvectors. $\mathbf{v} \neq 0$ implies that at least one of $\mathbf{u}$ and $\mathbf{u}'$ are nonzero, so there is at least one real eigenvector with this eigenvalue.

□

Case (ii) is a stronger claim for hermitian matrices than just showing that eigenvectors are linearly independent. Furthermore, previously we considered bases $\mathcal{B}_\lambda$ for each eigenspace E_λ, and it is now natural to choose bases $\mathcal{B}_\lambda$ to be orthonormal when we are considering hermitian matrices. Here are some examples.

(i) Let

$$A = \begin{pmatrix} 2 & i \\ -i & 2 \end{pmatrix}; \quad A^\dagger = A; \quad \lambda = 1, 3; \quad \mathbf{u}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ i \end{pmatrix}; \quad \mathbf{u}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -i \end{pmatrix}$$

We have chosen coefficients for the vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ such that they are unit vectors. As shown above, they are then orthonormal. We know that having distinct eigenvalues means that a matrix is diagonalisable. So let us set

$$P = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix} \implies P^{-1}AP = D = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$$

Since the eigenvectors are orthonormal, so are the columns of P, so $P^{-1} = P^\dagger$ (i.e. P is unitary).

(ii) Let

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

A is real and symmetric, with eigenvalues $\lambda = -1, 2$ with $M_{-1} = 2, M_2 = 1$. Further,

$$E_{-1} = \text{span}\{\mathbf{w}_1, \mathbf{w}_2\}; \quad \mathbf{w}_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}; \quad \mathbf{w}_2 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$$

So $m_{-1} = 2$, and the matrix is diagonalisable. Let us choose an orthonormal basis for E_{-1} by taking

$$\mathbf{u}_1 = \frac{1}{|\mathbf{w}_1|}\mathbf{w}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$$

and we can consider

$$\mathbf{w}_2' = \mathbf{w}_2 - (\mathbf{u}_1 \cdot \mathbf{w}_2)\mathbf{u}_1 = \begin{pmatrix} 1/2 \\ 1/2 \\ -1 \end{pmatrix}$$

so that $\mathbf{w}_2'$ is orthogonal to $\mathbf{u}_1$ by construction. We can then normalise this vector to get

$$\mathbf{u}_2 = \frac{1}{|\mathbf{w}_2'|}\mathbf{w}_2' = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}$$

and therefore

$$\mathcal{B}_{-1} = \{\mathbf{u}_1, \mathbf{u}_2\}$$

is an orthonormal basis. For E_2, let us choose $\mathcal{B}_2 = \{\mathbf{u}_3\}$ where

$$\mathbf{u}_3 = \frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix}$$

Together,

$$\mathcal{B} = \left\{\frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\\0\end{pmatrix}, \frac{1}{\sqrt{6}}\begin{pmatrix}1\\1\\-2\end{pmatrix}, \frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix}\right\}$$

is an orthonormal basis for $\mathbb{R}^3$. Let P be the matrix with columns $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$, then $P^{-1}AP = D$ as required. Since we have chosen an orthonormal basis, P is orthogonal, so $P^\intercal AP = D$.

9.9. Unitary and orthogonal diagonalisation

Theorem. Any $n \times n$ hermitian matrix A is diagonalisable.

(i) There exists a basis of eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_n \in \mathbb{C}^n$ with $A\mathbf{u}_i = \lambda\mathbf{u}_i$; equivalently

(ii) There exists an $n \times n$ invertible matrix P with $P^{-1}AP = D$ where D is the matrix with eigenvalues on the diagonal, where the columns of P are the eigenvectors $\mathbf{u}_i$.

In addition, the eigenvectors $\mathbf{u}_i$ can be chosen to be orthonormal, so

$$\mathbf{u}_i^\dagger \mathbf{u}_j = \delta_{ij}$$

or equivalently, the matrix P can be chosen to be unitary,

$$P^\dagger = P^{-1} \implies P^\dagger AP = D$$

In the special case that the matrix A is real, the eigenvectors can be chosen to be real, and so

$$\mathbf{u}^\intercal \mathbf{u}_j = \mathbf{u}_i \cdot \mathbf{u}_j = \delta_{ij}$$

so P is orthogonal, so

$$P^\intercal = P^{-1} \implies P^\intercal AP = D$$

10. Quadratic forms

10.1. Simple example

Consider a function $\mathcal{F}: \mathbb{R}^2 \to \mathbb{R}$ defined by

$$\mathcal{F}(\mathbf{x}) = 2x_1^2 - 4x_1x_2 + 5x_2^2$$

This can be simplified by writing

$$\mathcal{F}(\mathbf{x}) = x_1'^2 + 6x_2'^2$$

where

$$x_1' = \frac{1}{\sqrt{5}}(2x_1 + x_2); \quad x_2' = \frac{1}{\sqrt{5}}(-x_1 + 2x_2)$$

This can be found by writing $\mathcal{F}(\mathbf{x}) = \mathbf{x}^\intercal A\mathbf{x}$ where

$$A = \begin{pmatrix} 2 & -2 \\ -2 & 5 \end{pmatrix}$$

by inspection from the original equation, and then diagonalising A. We find the eigenvalues to be $\lambda = 1, 6$, with eigenvectors

$$\frac{1}{\sqrt{5}}\begin{pmatrix} 2 \\ 1 \end{pmatrix}; \quad \frac{1}{\sqrt{5}}\begin{pmatrix} -1 \\ 2 \end{pmatrix}$$

10.2. Diagonalising quadratic forms

In general, a quadratic form is a function $\mathcal{F}: \mathbb{R}^n \to \mathbb{R}$ given by

$$\mathcal{F}(\mathbf{x}) = \mathbf{x}^\intercal A\mathbf{x} \implies \mathcal{F}(\mathbf{x})_{ij} = x_i A_{ij} x_j$$

where A is a real symmetric $n \times n$ matrix. Any antisymmetric part of A would not contribute to the result, so there is no loss of generality under this restriction. From the section above, we know we can write $P^\intercal AP = D$ where D is a diagonal matrix containing the eigenvalues, and P is constructed from the eigenvectors, with orthonormal columns $\mathbf{u}_i$. Setting $\mathbf{x}' = P^\intercal \mathbf{x}$, or equivalently $\mathbf{x} = P\mathbf{x}'$, we have

$$\begin{aligned} \mathcal{F}(\mathbf{x}) &= \mathbf{x}^\intercal A\mathbf{x} \\ &= (P\mathbf{x}')^\intercal A(P\mathbf{x}') \\ &= (\mathbf{x}')^\intercal P^\intercal AP\mathbf{x}' \\ &= (\mathbf{x}')^\intercal D\mathbf{x}' \\ &= \sum_i \lambda_i x_i'^2 = \lambda_1 x_1'^2 + \lambda_2 x_2'^2 + \ldots \end{aligned}$$

We say that $\mathcal{F}$ has been diagonalised. Now, note that

$$\begin{aligned}\mathbf{x}' &= x_1'\mathbf{e}_1 + \cdots + x_n'\mathbf{e}_n \\ \mathbf{x} &= x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n \\ &= x_1'\mathbf{u}_1 + \cdots + x_n'\mathbf{u}_n\end{aligned}$$

where the $\mathbf{e}_i$ are the standard basis vectors, since

$$\mathbf{x}_i' = \mathbf{u}_i \cdot \mathbf{x} \iff \mathbf{x}' = P^\intercal \mathbf{x}$$

Hence the $\mathbf{x}_i'$ can be regarded as coordinates with respect to a new set of axes defined by the orthonormal eigenvector basis, known as the principal axes of the quadratic form. They are related to the standard axes (given by basis vectors $\mathbf{e}_i$) by the orthogonal transformation P.

Example (two dimensions)**.** Consider $\mathcal{F}(\mathbf{x}) = \mathbf{x}^\intercal A\mathbf{x}$ with

$$A = \begin{pmatrix}\alpha & \beta \\ \beta & \alpha\end{pmatrix}$$

The eigenvalues are $\lambda = \alpha + \beta, \alpha - \beta$ and

$$\mathbf{u}_1 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix}; \quad \mathbf{u}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix}-1\\1\end{pmatrix}$$

So in terms of the standard basis vectors,

$$\mathcal{F}(\mathbf{x}) = \alpha x_1^2 + 2\beta x_1 x_2 + \alpha x_2^2$$

And in terms of our new basis vectors,

$$\mathcal{F}(\mathbf{x}) = (\alpha + \beta)x_1'^2 + (\alpha - \beta)x_2'^2$$

where

$$\begin{aligned}\mathbf{x}_1' &= \mathbf{u}_1 \cdot \mathbf{x} = \frac{1}{\sqrt{2}}(x_1 + x_2) \\ \mathbf{x}_2' &= \mathbf{u}_2 \cdot \mathbf{x} = \frac{1}{\sqrt{2}}(-x_1 + x_2)\end{aligned}$$

Taking for example $\alpha = \frac{3}{2}, \beta = \frac{-1}{2}$, we have $\lambda_1 = 1, \lambda_2 = 2$. If we choose $\mathcal{F} = 1$, this represents an ellipse in our new coordinate system:

$$x_1'^2 + 2x_2'^2 = 1$$

If instead we chose $\alpha = \frac{-1}{2}, \beta = \frac{3}{2}$. We now have $\lambda_1 = 1, \lambda_2 = -2$. The locus at $\mathcal{F} = 1$ gives a hyperbola:

$$x_1'^2 - 2x_2'^2 = 1$$

Example (three dimensions). In $\mathbb{R}^3$, note that if $\lambda_1, \lambda_2, \lambda_3$ are all strictly positive, then $\mathcal{F} = 1$ gives an ellipsoid. This is analogous to the $\mathbb{R}^2$ case above.

Let us consider an example. Earlier, we found that the eigenvalues of the matrix A where

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

are $\lambda_1 = \lambda_2 = -1, \lambda_3 = 2$, where

$$\mathbf{u}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}; \quad \mathbf{u}_2 = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}; \quad \mathbf{u}_3 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Then

$$\begin{aligned} \mathcal{F}(\mathbf{x}) &= 2x_1x_2 + 2x_2x_3 + 2x_3x_1 \\ &= -x_1'^2 - x_2'^2 + 2x_3'^2 \end{aligned}$$

Now, $\mathcal{F} = 1$ corresponds to

$$2x_3'^2 = 1 + x_1'^2 + x_2'^2$$

So we can more clearly see that this is a hyperboloid of two sheets in $\mathbb{R}^3$ with rotational symmetry between the x_1 and x_2 axes. Further, $\mathcal{F} = -1$ corresponds to

$$1 + 2x_3'^2 = x_1'^2 + x_2'^2$$

Here, this is a hyperboloid of one sheet since for any fixed x_3 coordinate, it defines a circle in the x_1 and x_2 axes.

10.3. Hessian matrix as a quadratic form

Consider a smooth function $f : \mathbb{R}^n \to \mathbb{R}$ with a stationary point at $\mathbf{x} = \mathbf{a}$, i.e. $\frac{\partial f}{\partial x_i} = 0$ at $\mathbf{x} = \mathbf{a}$. By Taylor's theorem,

$$f(\mathbf{a} + \mathbf{h}) + f(\mathbf{a}) + \mathcal{F}(\mathbf{h}) + O(|\mathbf{h}|^3)$$

where $\mathcal{F}$ is a quadratic form with

$$A_{ij} = \frac{1}{2} \frac{\partial^2 f}{\partial x_i \partial x_j}$$

all evaluated at $\mathbf{x} = \mathbf{a}$. Note that this A is half of the Hessian matrix, and that the linear term vanishes since we are at a stationary point. Rewriting this $\mathbf{h}$ in terms of the eigenvectors of A (the principal axes), we have

$$\mathcal{F} = \lambda_1 h_1'^2 + \lambda_2 h_2'^2 + \cdots + \lambda_n h_n'^2$$

So clearly if $\lambda_i > 0$ for all i, then f has a minimum at $\mathbf{x} = \mathbf{a}$. If $\lambda_i < 0$ for all I, then f has a maximum at $\mathbf{x} = \mathbf{a}$. Otherwise, it has a saddle point. Note that it is often sufficient to consider the trace and determinant of A, since $\operatorname{tr} A = \lambda_1 + \lambda_2$ and $\det A = \lambda_1\lambda_2$.

11. Cayley–Hamilton theorem

11.1. Matrix polynomials

If A is an $n \times n$ complex matrix and

$$p(t) = c_0 + c_1 t + c_2^2 + \cdots + c_k t^k$$

is a polynomial, then

$$p(A) = c_0 I + c_1 A + c_2 A^2 + \cdots + c_k A^k$$

We can also define power series on matrices (subject to convergence). For example, the exponential series which always converges:

$$\exp(A) = I + A + \frac{1}{2}A^2 + \cdots + \frac{1}{r!}A^r + \ldots$$

For a diagonal matrix, polynomials and power series can be computed easily since the power of a diagonal matrix just involves raising its diagonal elements to said power. Therefore,

$$D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} \implies p(D) = \begin{pmatrix} p(\lambda_1) & 0 & \cdots & 0 \\ 0 & p(\lambda_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p(\lambda_n) \end{pmatrix}$$

Therefore,

$$\exp(D) = \begin{pmatrix} e^{\lambda_1} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_n} \end{pmatrix}$$

If $B = P^{-1}AP$ (similar to A) where P is an $n \times n$ invertible matrix, then

$$B^r = P^{-1}A^r P$$

Therefore,

$$p(B) = p(P^{-1}AP) = P^{-1}p(A)P$$

Of special interest is the characteristic polynomial,

$$\chi_A(t) = \det(A - tI) = c_0 + c_1 t + c_2 t^2 + \cdots + c_n t^n$$

where $c_0 = \det A$, and $c_n = (-1)^n$.

Theorem (Cayley–Hamilton Theorem)**.**

$$\chi_A(A) = c_0 I + c_1 A + c_2 A^2 + \cdots + c_n A^n = 0$$

Less formally, a matrix satisfies its own characteristic equation.

Remark. We can find an expression for the matrix inverse.

$$-c_0 I = A(c_1 + c_2 A + \cdots + c_n A^{n-1})$$

If $c_0 = \det A \neq 0$, then

$$A^{-1} = \frac{-1}{c_0}(c_1 + c_2 A + \cdots + c_n A^{n-1})$$

11.2. Proofs of special cases of Cayley–Hamilton theorem

Proof for a 2×2 matrix. Let A be a general 2×2 matrix.

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \implies \chi_A(t) = t^2 - (a+d)t + (ad - bc)$$

We can check the theorem by substitution.

$$\chi_A(A) = A^2 - (a+d)A - (ad-bc)I$$

This is shown on the last example sheet. □

Proof for diagonalisable $n \times n$ matrices. Consider A with eigenvalues λ_i, and an invertible matrix P such that $P^{-1}AP = D$, where D is diagonal.

$$\chi_A(D) = \begin{pmatrix} \chi_A(\lambda_1) & 0 & \cdots & 0 \\ 0 & \chi_A(\lambda_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \chi_A(\lambda_n) \end{pmatrix} = 0$$

since the λ_i are eigenvalues. Then

$$\chi_A(A) = \chi_A(PDP^{-1}) = P\chi_A(D)P^{-1} = 0$$

□

11.3. Proof in general case (non-examinable)

Proof. Let $M = A - tI$. Then $\det M = \det(A - tI) = \chi_A(t) = \sum_{r=0} c_r t^r$. We can construct the adjugate matrix.

$$\widetilde{M} = \sum_{r=0}^{n-1} B_r t^r$$

Therefore,

$$\begin{aligned} \widetilde{M}M = (\det M)I &= \left(\sum_{r=0}^{n-1} B_r t^r\right)(A - tI) \\ &= B_0A + (B_1A - B_0)t + (B_2A - B_1)t^2 + \cdots + (B_{n-1}A - B_{n-2})t^{n-1} - B_{n-1}t \end{aligned}$$

Now by comparing coefficients,

$$\begin{aligned} C_0I &= B_0A \\ C_1I &= B_1A - B_0 \\ &\vdots \\ C_{n-1}I &= B_{n-1}A - B_{n-2} \\ C_nI &= -B_{n-1} \end{aligned}$$

Summing all of these coefficients, multiplying by the relevant powers,

$$\begin{aligned}
&C_0 I + C_1 A + C_2 A^2 + \cdots + C_n A^n \\
&= B_0 A + (B_1 A^2 - B_0 A) + (B_2 A^3 - B_1 A^2) + \cdots + (B_{n-1} A^n - B_{n-2} A^{n-1}) - B_{n-1} A^n \\
&= 0
\end{aligned}$$

□

12. Changing bases

12.1. Change of basis formula

Recall that given a linear map $T : V \to W$ where V and W are real or complex vector spaces, and choices of bases $\{\mathbf{e}_i\}$ for $i = 1, \dots, n$ and $\{\mathbf{f}_a\}$ for $a = 1, \dots, m$, then the $m \times n$ matrix A with respect to these bases is defined by

$$T(\mathbf{e}_i) = \sum_a \mathbf{f}_a A_{ai}$$

So the entries in column i of A are the components of $T(\mathbf{e}_i)$ with respect to the basis $\{\mathbf{f}_a\}$. This is chosen to ensure that the statement $\mathbf{y} = T(\mathbf{x})$ is equivalent to the statement that $y_a = A_{ai}x_i$, where $\mathbf{y} = \sum_a y_a \mathbf{f}_a$ and $\mathbf{x} = \sum_i x_i \mathbf{e}_i$. This equivalence holds since

$$T\left(\sum_i x_i \mathbf{e}_i\right) = \sum_i x_i T(\mathbf{e}_i) = \sum_i x_i \left(\sum_a \mathbf{f}_a A_{ai}\right) = \sum_a \underbrace{\left(\sum_i A_{ai} x_i\right)}_{y_a} \mathbf{f}_a$$

as required. For the same linear map T, there is a different matrix representation A' with respect to different bases $\{\mathbf{e}'_i\}$ and $\{\mathbf{f}'_a\}$. To relate A with A', we need to understand how the new bases relate to the original bases. The change of base matrices P ($n \times n$) and Q ($m \times m$) are defined by

$$\mathbf{e}'_i = \sum_j \mathbf{e}_j P_{ji}; \quad \mathbf{f}'_a = \sum_b \mathbf{f}_b Q_{ba}$$

The entries in column i of P are the components of the new basis $\mathbf{e}'_i$ in terms of the old basis vectors $\{\mathbf{e}_j\}$, and similarly for Q. Note, P and Q are invertible, and in the relation above we could exchange the roles of $\{\mathbf{e}_i\}$ and $\{\mathbf{e}'_i\}$ by replacing P with P^{-1}, and similarly for Q.

Proposition (Change of base formula for a linear map)**.** With the definitions above,

$$A' = Q^{-1}AP$$

First we will consider an example, then we will construct a proof. Let $n = 2, m = 3$, and

$$T(\mathbf{e}_1) = \mathbf{f}_1 + 2\mathbf{f}_2 - \mathbf{f}_3 = \sum_a \mathbf{f}_a A_{a1}$$
$$T(\mathbf{e}_2) = -\mathbf{f}_1 + 2\mathbf{f}_2 + \mathbf{f}_3 = \sum_a \mathbf{f}_a A_{a2}$$

Therefore,

$$A = \begin{pmatrix} 1 & -1 \\ 2 & 2 \\ -1 & 1 \end{pmatrix}$$

Consider a new basis for V, given by

$$\mathbf{e}'_1 = \mathbf{e}_1 - \mathbf{e}_2 = \sum_i \mathbf{e}_i P_{i1}$$
$$\mathbf{e}'_2 = \mathbf{e}_1 + \mathbf{e}_2 = \sum_i \mathbf{e}_i P_{i2}$$

$$P = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

Consider further a new basis for W, given by

$$\mathbf{f}'_1 = \mathbf{f}_1 - \mathbf{f}_3 = \sum_a \mathbf{f}_a Q_{a1}$$
$$\mathbf{f}'_2 = \mathbf{f}_2 = \sum_a \mathbf{f}_a Q_{a2}$$
$$\mathbf{f}'_3 = \mathbf{f}_1 + \mathbf{f}_3 = \sum_a \mathbf{f}_a Q_{a3}$$

$$Q = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}$$

From the change of base formula,

$$\begin{aligned} A' &= Q^{-1}AP \\ &= \begin{pmatrix} 1/2 & 0 & -1/2 \\ 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 2 & 2 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 2 & 0 \\ 0 & 4 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

Now checking this result directly,

$$T(\mathbf{e}'_1) = 2\mathbf{f}_1 - 2\mathbf{f}_3 = 2\mathbf{f}'_1$$
$$T(\mathbf{e}'_2) = 4\mathbf{f}_2 = 4\mathbf{f}'_2$$

which matches the content of the matrix as required. Now, let us prove the proposition in general.

Proof.

$$\begin{aligned} T(\mathbf{e}'_i) &= T\left(\sum_j \mathbf{e}_j P_{ji}\right) \\ &= \sum_j T(\mathbf{e}_j) P_{ji} \\ &= \sum_j \left(\sum_a \mathbf{f}_a A_{aj}\right) P_{ji} \\ &= \sum_{ja} \mathbf{f}_a A_{aj} P_{ji} \end{aligned}$$

But on the other hand,

$$\begin{aligned} T(\mathbf{e}'_i) &= \sum_b \mathbf{f}'_b A'_{bi} \\ &= \sum_b \left(\sum_a \mathbf{f}_a Q_{ab}\right) A'_{bi} \\ &= \sum_{ab} \mathbf{f}_a Q_{ab} A'_{bi} \end{aligned}$$

which is a sum over the same set of basis vectors, so we may equate coefficients of $\mathbf{f}_a$.

$$\begin{aligned} \sum_j A_{aj} P_{ji} &= \sum_b Q_{ab} A'_{bi} \\ (AP)_{ai} &= (QA')_{ai} \end{aligned}$$

Therefore

$$AP = QA' \implies A' = Q^{-1}AP$$

as required. □

12.2. Changing bases of vector components

Here is another way to arrive at the formula $A' = Q^{-1}AP$. Consider changes in vector components

$$\begin{aligned} \mathbf{x} &= \sum_i x_i \mathbf{e}_i = \sum_j x'_j \mathbf{e}'_j \\ &= \sum_i \left(\sum_j P_{ij} x'_j\right) \mathbf{e}_i \\ \implies x_i &= P_{ij} x'_j \end{aligned}$$

We will write

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}; \quad X' = \begin{pmatrix} x'_1 \\ \vdots \\ x'_n \end{pmatrix}$$

Then $X = PX'$ or $X' = P^{-1}X$. Similarly,

$$\mathbf{y} = \sum_a y_a \mathbf{f}_a = \sum_b y'_b \mathbf{f}'_b$$
$$\implies y_a = Q_{ab} y'_b$$

Then $Y = QY'$ or $Y' = Q^{-1}Y$. So the matrices are defined to ensure that

$$Y = AX; \quad Y' = A'X'$$

Therefore,

$$QY' = APX' \implies Y' = (Q^{-1}AP)X' \implies A' = Q^{-1}AP$$

12.3. Specialisations of changes of basis

Now, let us consider some special cases (in increasing order of specialisation).

(i) Let $V = W$ with $\mathbf{e}_i = \mathbf{f}_i$ and $\mathbf{e}'_i = \mathbf{f}'_i$. So $P = Q$ and the change of basis is

$$A' = P^{-1}AP$$

Matrices representing the same linear map but with respect to different bases are similar. Conversely, if A, A' are similar, then we can construct an invertible change of basis matrix P which relates them, so they can be regarded as representing the same linear map. In an earlier section we noted that $\mathrm{tr}(A') = \mathrm{tr}(A)$, $\det(A') = \det(A)$ and $\chi_A(t) = \chi_{A'}(t)$. so these are intrinsic properties of the linear map, not just the particular matrix we choose to represent it.

(ii) Let $V = W = \mathbb{R}^n$ or $\mathbb{C}^n$ where $\mathbf{e}_i$ is the standard basis, with respect to which, T has matrix A. If there exists a basis of eigenvectors, $\mathbf{e}'_i = \mathbf{v}_i$ with $A\mathbf{v}_i = \lambda_i \mathbf{v}_i$. Then

$$A' = P^{-1}AP = D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

and

$$\mathbf{v}_i = \sum_k \mathbf{e}_j P_{ji}$$

so the eigenvectors are the columns of P.

(iii) Let A be hermitian, i.e. $A^\dagger = A$, then we always have a basis of orthonormal eigenvectors $\mathbf{e}'_i = \mathbf{u}_i$. Then the relations in (ii) apply, and P is unitary, $P^\dagger = P^{-1}$.

12.4. Jordan normal form

Also known as the (Jordan) Canonical Form, this result classifies $n \times n$ complex matrices up to similarity.

Proposition. Any 2×2 complex matrix A is similar to one of the following:

(i) $A' = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ with $\lambda_1 \neq \lambda_2$, so $\chi_A(t) = (t - \lambda_1)(t - \lambda_2)$.

(ii) $A' = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$, so $\chi_A(t) = (t - \lambda)^2$.

(iii) $A' = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$, so $\chi_A(t) = (t - \lambda)^2$ as in case (ii).

Proof. $\chi_A(t)$ has two roots over $\mathbb{C}$.

(i) For distinct roots λ_1, λ_2, we have $M_{\lambda_1} = m_{\lambda_1} = M_{\lambda_2} = m_{\lambda_2} = 1$. So the eigenvectors $\mathbf{v}_1, \mathbf{v}_2$ provide a basis. Hence $A' = P^{-1}AP$ with the eigenvectors as the columns of P.

(ii) For a repeated root λ with $M_\lambda = m_\lambda = 2$, the same argument applies.

(iii) For a repeated root λ with $M_\lambda = 2$, $m_\lambda = 1$, we do not have a basis of eigenvectors so we cannot diagonalise the matrix. We only have one linearly independent eigenvector, which we will call $\mathbf{v}$. Let $\mathbf{w}$ be any other vector such that $\{\mathbf{v}, \mathbf{w}\}$ are linearly independent. Then

$$\begin{aligned} A\mathbf{v} &= \lambda\mathbf{v} \\ A\mathbf{w} &= \alpha\mathbf{v} + \beta\mathbf{w} \end{aligned}$$

The matrix representing this linear map with respect to the basis vectors $\{\mathbf{v}, \mathbf{w}\}$ is therefore

$$\begin{pmatrix} \lambda & \alpha \\ 0 & \beta \end{pmatrix}$$

Let us solve for some of these unknowns. We know that the characteristic polynomial of this matrix must be $(t-\lambda)^2$, so $\beta = \lambda$. Also, $\alpha \neq 0$, otherwise we have case (ii) above. So now we can set $\mathbf{u} = \alpha\mathbf{v}$, so

$$\begin{aligned} A(\alpha\mathbf{v}) &= \lambda(\alpha\mathbf{v}) \\ A\mathbf{w} &= \alpha\mathbf{v} + \beta\mathbf{w} \end{aligned}$$

So with respect to the basis $\{\mathbf{u}, \mathbf{w}\}$ we get the matrix A to be

$$A' = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$$

□

Alternative Proof. Here is an alternative approach for case (iii). If A has characteristic polynomial

$$\chi_A(t) = (t-\lambda)^2$$

but $A \neq \lambda I$, then there exists some vector $\mathbf{w}$ for which $\mathbf{u} = (A-\lambda I)\mathbf{w} \neq \mathbf{0}$. So $(A-\lambda I)\mathbf{u} = (A-\lambda I)^2\mathbf{w} = \mathbf{0}$ by the Cayley–Hamilton theorem. So

$$A\mathbf{u} = \lambda\mathbf{u}$$
$$A\mathbf{w} = \mathbf{u} + \lambda\mathbf{w}$$

So with basis $\{\mathbf{u}, \mathbf{w}\}$ we have the matrix

$$A' = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$$

□

Here is a concrete example using this alternative proof method.

$$A = \begin{pmatrix} 1 & 4 \\ -1 & 5 \end{pmatrix} \implies \chi_A(t) = (t-3)^2$$

So

$$A - 3I = \begin{pmatrix} -2 & 4 \\ -1 & 2 \end{pmatrix}$$

We will choose $\mathbf{w} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and we find $\mathbf{u} = (A-3I)\mathbf{w} = \begin{pmatrix} -2 \\ -1 \end{pmatrix}$. $\mathbf{w}$ is not an eigenvector, as required for the construction. By the reasoning in the alternative argument above, $\mathbf{u}$ is an eigenvector by construction.

$$A\mathbf{u} = 3\mathbf{u}$$
$$A\mathbf{w} = \mathbf{u} + 3\mathbf{w}$$

So

$$P = \begin{pmatrix} -2 & 1 \\ -1 & 0 \end{pmatrix} \implies P^{-1} = \begin{pmatrix} 0 & -1 \\ 1 & -2 \end{pmatrix}$$

and we can check that

$$P^{-1}AP = \begin{pmatrix} 3 & 1 \\ 0 & 3 \end{pmatrix} = A'$$

12.5. Jordan normal forms in n dimensions

To extend the arguments above to larger matrices, consider the $n \times n$ matrix

$$N = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

When applied to the standard basis vectors in $\mathbb{C}^n$, the action of this matrix sends $\mathbf{e}_n \mapsto \mathbf{e}_{n-1} \mapsto \cdots \mapsto \mathbf{e}_1 \mapsto \mathbf{0}$. This is consistent with the property that $N^n = 0$. The kernel of this matrix has dimension 1. Now consider the matrix $J = \lambda I + N$, as follows:

$$N = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ 0 & 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda \end{pmatrix}$$

This matrix has

$$\chi_J(t) = (\lambda - t)^n$$

with $M_\lambda = n$ and $m_\lambda = 1$, since the kernel of $J - \lambda I = N$ has dimension 1 as before. The general result is as follows.

Theorem. Any $n \times n$ complex matrix A is similar to a matrix of the form

$$A' = \left(\begin{array}{c|c|c|c} J_{n_1}(\lambda_1) & 0 & \cdots & 0 \\ \hline 0 & J_{n_2}(\lambda_2) & \cdots & 0 \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline 0 & 0 & \cdots & J_{n_r}(\lambda_r) \end{array}\right)$$

where each diagonal block is a Jordan block $J_{n_r}(\lambda_r)$ which is an $n_r \times n_r$ matrix J with eigenvalue λ_r. $\lambda_1, \ldots, \lambda_r$ are eigenvalues of A and A', and the same eigenvalue may appear in different blocks. Further, $n_1 + n_2 + \cdots + n_r = n$ so we end up with an $n \times n$ matrix. A is diagonalisable if and only if A' consists entirely of 1×1 blocks. The expression above is the Jordan Normal Form.

The proof is non-examinable and depends on the Part IB courses Linear Algebra, and Groups, Rings and Modules, so is not included here.

13. Conics and quadrics

13.1. Quadrics in general

A quadric in $\mathbb{R}^n$ is a hypersurface defined by an equation of the form

$$Q(\mathbf{x}) = \mathbf{x}^\intercal A\mathbf{x} + \mathbf{b}^\intercal\mathbf{x} + c = 0$$

for some nonzero, symmetric, real $n \times n$ matrix A, $b \in \mathbb{R}^n$, $c \in \mathbb{R}$. In components,

$$Q(\mathbf{x}) = A_{ij}x_ix_j + b_ix_i + c = 0$$

We will classify solutions for $\mathbf{x}$ up to geometrical equivalence, so we will not distinguish between solutions here which are related by isometries in $\mathbb{R}^n$ (distance-preserving maps, i.e. translations and orthogonal transformations about the origin).

Note that A is invertible if and only if it has no zero eigenvalues. In this case, we can complete the square in the equation $Q(\mathbf{x}) = 0$ by setting $\mathbf{y} = \mathbf{x} + \frac{1}{2}A^{-1}\mathbf{b}$. This is essentially a translation isometry, moving the origin to $\frac{1}{2}A^{-1}\mathbf{b}$.

$$\begin{aligned}\mathbf{y}^\intercal A\mathbf{y} &= (\mathbf{x} + \frac{1}{2}A^{-1}\mathbf{b})^\intercal A(\mathbf{x} + \frac{1}{2}A^{-1}\mathbf{b}) \\ &= (\mathbf{x}^\intercal + \frac{1}{2}\mathbf{b}^\intercal(A^{-1})^\intercal)A(\mathbf{x} + \frac{1}{2}A^{-1}\mathbf{b}) \\ &= \mathbf{x}^\intercal A\mathbf{x} + \mathbf{b}^\intercal\mathbf{x} + \frac{1}{4}\mathbf{b}^\intercal A^{-1}\mathbf{b}\end{aligned}$$

since $(A^\intercal)^{-1} = (A^{-1})^\intercal$. Then,

$$Q(\mathbf{x}) = 0 \iff \mathcal{F}(\mathbf{y}) = k$$

with

$$\mathcal{F}(\mathbf{y}) = \mathbf{y}^\intercal A\mathbf{y}$$

which is a quadratic form with respect to a new origin $\mathbf{y} = \mathbf{0}$, and where $k = \frac{1}{4}\mathbf{b}^\intercal A^{-1}\mathbf{b} - c$. Now we can diagonalise $\mathcal{F}$ as in the above section, in particular, orthonormal eigenvectors give the principal axes, and the eigenvalues of A and the value of k determine the geometrical nature of the solution of the quadric. In $\mathbb{R}^3$, the geometrical possibilities are (as we saw before):

(i) eigenvalues positive, k positive gives an ellipsoid;

(ii) eigenvalues different signs, k nonzero gives a hyperboloid

If A has one or more zero eigenvalues, then the analysis we have just provided changes, since we can no longer construct such a $\mathbf{y}$ vector, since A^{-1} does not exist. The simplest standard form of Q may have both linear and quadratic terms.

13.2. Conics as quadrics

Quadrics in $\mathbb{R}^2$ are curves called conics. Let us first consider the case where $\det A \neq 0$. By completing the square and diagonalising A, we get a standard form

$$\lambda_1 {x_1'}^2 + \lambda_2 {x_2'}^2 = k$$

The variables x_i' correspond to the principal axes and the new origin. We have the following cases.

- $(\lambda_1, \lambda_2 > 0)$ This is an ellipse for $k > 0$, and a point for $k = 0$. There are no solutions for $k < 0$.
- $(\lambda_1 > 0, \lambda_2 < 0)$ This gives a hyperbola for $k > 0$, and a hyperbola in the other axis if $k < 0$. If $k = 0$, this is a pair of lines. For instance, ${x_1'}^2 - {x_2'}^2 = 0 \implies (x_1' - x_2')(x_1' + x_2') = 0$.

If $\det A = 0$, then there is exactly one zero eigenvalue since $A \neq 0$. Then:

- $(\lambda_1 > 0, \lambda_2 = 0)$ We will diagonalise A in the original expression for the quadric. This gives

  $$\lambda_1 {x_1'}^2 + b_1' x_1' + b_2' x_2' + c = 0$$

  This is a new equation in the coordinate system defined by A's principal axes. Completing the square here in the x_1' term, we have

  $$\lambda_1 {x_1''}^2 + b_2' x_2' + c' = 0$$

  where $x_1'' = x_1' + \frac{1}{2\lambda_1} b_1'$, and $c' = c - \frac{{b_1'}^2}{4\lambda_1^2}$. If $b_2' = 0$, then x_2 can take any value; and we get a pair of lines if $c' < 0$, a single line if $c' = 0$, and no solutions if $c' > 0$. Otherwise, $b_2' \neq 0$, and the equation becomes

  $$\lambda_1 {x_1''}^2 + b_2' x_2'' = 0$$

  where $x_2'' = x_2' + \frac{1}{b_2'} c'$, and clearly this equation is a parabola.

All changes of coordinates correspond to translations (shifts of the origin) or orthogonal transformations, both of which preserve distance and angles.

13.3. Standard forms for conics

The general forms of conics can be written in terms of lengths a, b (the semi-major and semi-minor axes), or equivalently a length scale ℓ and a dimensionless eccentricity constant e.

- First, let us consider Cartesian coordinates. The formulas are:

conic	formula	eccentricity	foci
ellipse	$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$	$b^2 = a^2(1-e^2)$, and $e < 1$	$x = \pm ae$
parabola	$y^2 = 4ax$	one quadratic term vanishes, $e = 1$	$x = +a$
hyperbola	$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$	$b^2 = a^2(e^2-1)$, and $e > 1$	$x = \pm ae$

- Polar coordinates are a convenient alternative to Cartesian coordinates. In this coordinate system, we set the origin to be at a focus. Then, the formulas are

$$r = \frac{\ell}{1 + e\cos\theta}$$

  - For the ellipse, $e < 1$ and $\ell = a(1-e^2)$;
  - For the parabola, $e = 1$ and $\ell = 2a$; and
  - For the hyperbola, $e > 1$ and $\ell = a(e^2-1)$. There is only one branch for the hyperbola given by this polar form.

13.4. Conics as sections of a cone

The equation for a cone in $\mathbb{R}^3$ given by an apex $\mathbf{c}$, an axis $\hat{\mathbf{n}}$, and an angle $\alpha < \frac{\pi}{2}$, is

$$(\mathbf{x} - \mathbf{c}) \cdot \hat{\mathbf{n}} = |\mathbf{x} - \mathbf{c}| \cos\alpha$$

Less formally, the angle of $\mathbf{x}$ away from $\hat{\mathbf{n}}$ must be α. By squaring this equation, we can essentially define two cones which stretch out infinitely far and meet at the centre point $\mathbf{c}$.

$$((\mathbf{x} - \mathbf{c}) \cdot \hat{\mathbf{n}})^2 = |\mathbf{x} - \mathbf{c}|^2 \cos^2\alpha$$

Let us choose a set of coordinate axes so that our equations end up slightly easier. Let $\mathbf{c} = c\mathbf{e}_3$, $\hat{\mathbf{n}} = \cos\beta\mathbf{e}_1 - \sin\beta\mathbf{e}_3$. Then essentially the cone starts at $(0,0,c)$ and points 'downwards' in the $\mathbf{e}_1$–$\mathbf{e}_3$ plane. Then the conic section is the intersection of this cone with the $\mathbf{e}_1$–$\mathbf{e}_2$ plane, i.e. $x_3 = 0$.

$$(x_1\cos\beta - c\sin\beta)^2 = (x_1^2 + x_2^2 + c^2)\cos^2\alpha$$

$$\iff (\cos^2\alpha - \cos^2\beta)x_1^2 + (\cos^2\alpha)x_2^2 + 2x_1 c\sin\beta\cos\beta = \text{const.}$$

Now we can compare the signs of the x_1^2 and x_2^2 terms. Clearly the x_2^2 term is always positive, so we consider the sign of the x_1^2 term.

- If $\cos^2\alpha > \cos^2\beta$ (i.e. $\alpha < \beta$), then we have an ellipse.
- If $\cos^2\alpha = \cos^2\beta$ (i.e. $\alpha = \beta$), then we have a parabola.
- If $\cos^2\alpha < \cos^2\beta$ (i.e. $\alpha > \beta$), then we have a hyperbola.

14. Symmetries and transformation groups

14.1. Orthogonal transformations and rotations

We know that if a matrix R is orthogonal, we have $R^\intercal R = I \iff (R\mathbf{x})\cdot(R\mathbf{y}) = \mathbf{x}\cdot\mathbf{y} \iff$ the rows or columns are orthonormal. The set of $n \times n$ matrices R forms the orthogonal group $O_n = O(n)$. If $R \in O(n)$ then $\det R = \pm 1$. $SO_n = SO(n)$ is the special orthogonal group, which is the subgroup of $O(n)$ defined by $\det R = 1$. If some matrix R is an element of $O(n)$, then R preserves the modulus of n-dimensional volume. If $R \in SO(n)$, then R preserves not only the modulus but also the sign of such a volume.

$SO(n)$ consists precisely of all rotations in $\mathbb{R}^n$. $O(n) \setminus SO(n)$ consists of all reflections. For some specific $H \in O(n) \setminus SO(n)$, any element of $O(n)$ can be written as a product of H with some element in $SO(n)$, i.e. R or RH with $R \in SO(n)$. For example, if n is odd, we can choose $H = -I$.

Now, we can consider the transformation $x_i' = R_{ij}x_j$ under two distinct points of view.

- (active) The rotation R acts on the vector $\mathbf{x}$ and yields a new vector $\mathbf{x}'$. The x_i' are components of the transformed vector in terms of the standard basis vectors.
- (passive) The x_i' are components of the same vector $\mathbf{x}$ but with respect to new orthonormal basis vectors $\mathbf{u}_i$. In general, $\mathbf{x} = \sum_i x_i\mathbf{e}_i = \sum_i x_i'\mathbf{u}_i$ which is true where $\mathbf{u}_i = \sum_j R_{ij}\mathbf{e}_j = \sum_j \mathbf{e}_j P_{ji}$. So $P = R^{-1} = R^\intercal$ where P is the change of basis matrix.

14.2. 2D Minkowski space

Consider a new 'inner product' on $\mathbb{R}^2$ given by

$$(\mathbf{x},\mathbf{y}) = \mathbf{x}^\intercal J\mathbf{y}; \quad J = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}$$

$$\therefore \left(\begin{pmatrix}x_0\\x_1\end{pmatrix},\begin{pmatrix}y_0\\y_1\end{pmatrix}\right) = x_0y_0 - x_1y_1$$

We start indexing these vectors from zero, not one. Here are some important properties.

- This 'inner product' is not positive definite. In fact, $(\mathbf{x},\mathbf{x}) = x_0^2 - x_1^2$. (This is a quadratic form for $\mathbf{x}$ with eigenvalues ± 1.)
- It is bilinear and symmetric.
- Defining $\mathbf{e}_0 = \begin{pmatrix}1\\0\end{pmatrix}$ and $\mathbf{e}_1 = \begin{pmatrix}0\\1\end{pmatrix}$, they obey

  $$(\mathbf{e}_0,\mathbf{e}_0) = -(\mathbf{e}_1,\mathbf{e}_1) = 1; \quad (\mathbf{e}_0,\mathbf{e}_1) = 0$$

  This is similar to orthonormality, in this generalised sense.

This inner product is known as the Minkowski metric on $\mathbb{R}^2$. $\mathbb{R}^2$ with this metric is called Minkowski space.

14.3. Lorentz transformations

Let us consider a matrix

$$M = \begin{pmatrix} M_{00} & M_{01} \\ M_{10} & M_{11} \end{pmatrix}$$

giving a map $\mathbb{R}^2 \to \mathbb{R}^2$; this preserves the Minkowski metric if and only if $(M\mathbf{x}, M\mathbf{y}) = (\mathbf{x}, \mathbf{y})$ for any vectors $\mathbf{x}, \mathbf{y}$. Expanded, this condition is

$$(M\mathbf{x})^\intercal J(M\mathbf{y}) = \mathbf{x}^\intercal M^\intercal JM\mathbf{y} = \mathbf{x}^\intercal J\mathbf{y}$$

$$\implies M^\intercal JM = J$$

The set of such matrices form a group. Also, $\det M = \pm 1$ for the same reason as before. Furthermore, $|M_{00}|^2 \geq 1$, so either $M_{00} \geq 1$ or $M_{00} \leq -1$. The subgroup with $\det M = +1$ and $M_{00} \geq 1$ is known as the Lorentz group.

Let us find the general form of M, by using the fact that the columns $M\mathbf{e}_0$ and $M\mathbf{e}_i$ are orthonormal with respect to the Minkowski metric.

$$(M\mathbf{e}_0, M\mathbf{e}_0) = M_{00}^2 - M_{10}^2 = (\mathbf{e}_0, \mathbf{e}_0) = 1 \quad (\text{hence } |M_{00}|^2 \geq 1)$$

Taking $M_{00} \geq 1$, we can write

$$M\mathbf{e}_0 = \begin{pmatrix} \cosh\theta \\ \sinh\theta \end{pmatrix}$$

for some real value θ. For the other column,

$$(M\mathbf{e}_0, M\mathbf{e}_1) = 0;\ (M\mathbf{e}_1, M\mathbf{e}_1) = -1 \implies M\mathbf{e}_1 = \pm\begin{pmatrix} \sinh\theta \\ \cosh\theta \end{pmatrix}$$

The sign is fixed to be positive by the condition that $\det M = +1$.

$$M = \begin{pmatrix} \cosh\theta & \sinh\theta \\ \sinh\theta & \cosh\theta \end{pmatrix}$$

The curves defined by $(\mathbf{x}, \mathbf{x}) = k$ where k is a constant are hyperbolas. This is analogous to how the curves defined by $\mathbf{x}\cdot\mathbf{x} = k$ are circles. So applying M to any vector on a given branch of a hyperbola, the resultant vector remains on the hyperbola. Note that these matrices obey the rule $M(\theta_1)M(\theta_2) = M(\theta_1 + \theta_2)$. This confirms that they form a group.

14.4. Application to special relativity

Let

$$M(\theta) = \gamma(v)\begin{pmatrix} 1 & v \\ v & 1 \end{pmatrix};\quad v = \tanh\theta;\quad \gamma = (1 - v^2)^{-\frac{1}{2}}$$

Here, v lies in the range $-1 < v < 1$. We will rename x_0 to be t, which is now our time coordinate. x_1 will just be written x, our one-dimensional space coordinate. Then,

$$\mathbf{x}' = M\mathbf{x} \iff \begin{cases} t' & = \gamma \cdot (t + vx) \\ x' & = \gamma \cdot (x + vt) \end{cases}$$

This is a Lorentz transformation, or 'boost', relating the time and space coordinates for observers moving with relative velocity v in Special Relativity, in units where the speed of light c is taken to be 1. The γ factor in the Lorentz transformation gives rise to time dilation and length contraction effects. The group property $M(\theta_3) = M(\theta_1)M(\theta_2)$ with $\theta_3 = \theta_1 + \theta_2$ corresponds to the velocities

$$v_i = \tanh \theta_i \implies v_3 = \frac{v_1 + v_2}{1 + v_1 v_2}$$

This is consistent with the fact that all velocities are less than the speed of light, 1.

V. Dynamics and Relativity

Lectured in Lent 2021 by Prof. P. H. Haynes

In the first part of this course, we study the classical laws of motion. We apply physical laws to study various phenomena such as gravity, friction, and orbits. Many such laws take the form of differential equations, and by solving these equations we can compute things like trajectories of particles.

In the second part, we study special relativity. We explore things like time dilation and the twin paradox, and how the laws of physics seem to change when particles are travelling very close to the speed of light.

Contents

1. Basic definitions and Newton's laws 273
1.1. Basic concepts 273
1.2. Newton's laws of motion 273
1.3. Boosts 274
1.4. Galilean transformations 274
1.5. Newton's second law 275
1.6. Gravitational force 275
1.7. Electromagnetic force 276
2. Dimensional analysis 277
2.1. Choice of units 277
2.2. Scaling and dimensional independence 277
3. Forces and potential energy 279
3.1. Forces 279
3.2. More general potentials 280
3.3. Equilibrium points 281
3.4. Force and potential in three spatial dimensions 283
4. Gravitational and electromagnetic forces 285
4.1. Gravity 285
4.2. Gravitational and inertial mass 286
4.3. One-dimensional approximation to gravity 286
4.4. Escape velocity 286
4.5. Electromagnetism 287
5. Friction 288
5.1. Dry friction 288
5.2. Fluid drag 288
5.3. Work done by friction 289
5.4. Projectiles experiencing linear drag 289
6. Angular motion and orbits 291
6.1. Angular momentum 291
6.2. Orbits 291
6.3. Central forces 291
6.4. Polar coordinates in the plane 292
6.5. Circular motion 292
7. Orbits and stability 294
7.1. Motion in a central force field 294
7.2. Orbits under gravity 295
7.3. Stability of circular orbits 295

7.4. The orbit equation . . . 296
7.5. The Kepler problem . . . 297
7.6. Energy and eccentricity . . . 298
7.7. Kepler's laws of planetary motion . . . 299
7.8. Rutherford scattering . . . 299
8. Rotating frames . . . 301
8.1. Introduction . . . 301
8.2. The centrifugal force . . . 302
8.3. The Coriolis force . . . 303
8.4. Dropping a particle in a rotating frame . . . 304
9. Systems of particles . . . 306
9.1. Basic setup . . . 306
9.2. Centre of mass . . . 306
9.3. Motion relative to the centre of mass . . . 307
9.4. Angular momentum . . . 307
9.5. Energy . . . 309
10. Applications of orbits . . . 310
10.1. Two body problem . . . 310
10.2. Variable mass problems and the rocket problem . . . 311
11. Rigid bodies . . . 312
11.1. Definition . . . 312
11.2. Recap of angular velocity . . . 312
11.3. Moment of inertia for a rigid body . . . 312
11.4. Calculating moments of inertia . . . 314
11.5. Results on moments of inertia . . . 315
11.6. General motion of a rigid body . . . 316
11.7. Simple pendulum . . . 319
11.8. Comparison of sliding and rolling . . . 319
11.9. Transition from sliding to rolling . . . 321
12. Special relativity . . . 322
12.1. Introduction and postulates . . . 322
12.2. Lorentz transformations . . . 322
12.3. General properties of Lorentz transformation . . . 324
13. Space-time diagrams, simultaneity and causality . . . 325
13.1. Space-time diagrams . . . 325
13.2. Comparing velocities . . . 326
13.3. Simultaneity . . . 326
13.4. Causality . . . 326
13.5. Time dilation . . . 327
13.6. The twin paradox . . . 327

13.7. Length contraction 330
14. Geometry of spacetime 332
14.1. Invariant interval 332
14.2. Signs of the invariant interval 332
14.3. The Lorentz group 333
14.4. Rapidity 334
15. Relativistic physics 335
15.1. Proper time 335
15.2. 4-velocity 335
15.3. Transformation of velocities 336
15.4. Energy-momentum 4-vector 337
15.5. Massless particles 338
15.6. Newton's second law 338
15.7. Special relativity with particle physics 339
15.8. Particle decay 339
15.9. Higgs to photon decay 340
15.10. Particle scattering 340
15.11. Particle creation 342

1. Basic definitions and Newton's laws

1.1. Basic concepts

Definition. A particle is an object which has negligible size. It therefore does not have an alignment or rotation. It has a finite mass $m > 0$, and perhaps an electric charge q (which may be positive or negative). The position of the particle is described by a position vector $\mathbf{r}(t)$ or $\mathbf{x}(t)$, with respect to an origin O.

Definition. The Cartesian components of this vector $\mathbf{r}(t)$ are given by (x, y, z), where $\mathbf{r} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$, with $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$ orthonormal. The choice of coordinate axes defines a frame of reference S.

Definition. The velocity of a particle is $\mathbf{u}(t) = \dot{\mathbf{r}} = \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{r}(t)$. The velocity is tangential to the path, or *trajectory*, of the particle.

Definition. The momentum of a particle is $\mathbf{p} = m\mathbf{u}$.

Definition. The acceleration of a particle is $\mathbf{a} = \dot{\mathbf{u}} = \ddot{\mathbf{r}}$.

Note. The time derivative of $\mathbf{u}(t)$, for example, is defined using the limit definition:

$$\dot{\mathbf{u}}(t) = \lim_{h \to 0} \frac{\mathbf{u}(t+h) - \mathbf{u}(t)}{h}$$

with $\mathbf{u} \to \mathbf{u}_0$ if and only if $|\mathbf{u} - \mathbf{u}_0| \to 0$. With Cartesian basis vectors, we can evaluate derivatives componentwise, bringing the differential operator inside each vector component.

The derivatives of scalar and vector functions interoperate as expected. Suppose we have a scalar function $f(t)$ and vector functions $\mathbf{g}(t), \mathbf{h}(t)$, then for example we have

$$\frac{\mathrm{d}}{\mathrm{d}t}(f\mathbf{g}) = \frac{\mathrm{d}f}{\mathrm{d}t}\mathbf{g} + f\frac{\mathrm{d}\mathbf{g}}{\mathrm{d}t}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{g} \cdot \mathbf{h}) = \frac{\mathrm{d}\mathbf{g}}{\mathrm{d}t} \cdot \mathbf{h} + \mathbf{g} \cdot \frac{\mathrm{d}\mathbf{h}}{\mathrm{d}t}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{g} \times \mathbf{h}) = \frac{\mathrm{d}\mathbf{g}}{\mathrm{d}t} \times \mathbf{h} + \mathbf{g} \times \frac{\mathrm{d}\mathbf{h}}{\mathrm{d}t}$$

Take note of the ordering of the terms involving $\mathbf{g}$ and $\mathbf{h}$ when using the vector product.

1.2. Newton's laws of motion

(i) (Galileo's Law of Inertia) There exist inertial frames of reference in which a particle remains at rest or moves in a straight line at constant speed (i.e. at constant velocity), unless it is acted on by a force.

(ii) In an inertial frame of reference, the rate of change of momentum of a particle is equal to the force acting on it.

(iii) To every action, there is an equal and opposite reaction. The forces exerted between two particles are equal in magnitude and opposite in direction.

Note that the second law is a statement about vectors. All of these statements that we have made about particles can also be extended to finite bodies, composed of many particles.

1.3. Boosts

In an inertial frame, the acceleration of a particle is zero if the force acting on the particle is zero.

$$\ddot{\mathbf{r}} = \mathbf{0} \iff \mathbf{F} = \mathbf{0}$$

There is no unique inertial frame of reference. If S is an inertial frame, then any other frame S' moving at constant velocity relative to S is also an inertial frame. For example, suppose that S' is moving at speed v in the x direction. Then here

$$x' = x - vt; \quad y' = y; \quad z' = z; \quad t' = t$$

and we can generalise this to S' moving in an arbitrary direction relative to S, i.e.

$$\mathbf{r}' = \mathbf{r} - \mathbf{v}t$$

where $\mathbf{v}$ is the velocity of S' relative to S. This type of transformation is known as a 'boost'. For a particle with position vector $\mathbf{r}(t)$ in S (and position vector $\mathbf{r}'(t)$ in S'), we can compute the velocity $\mathbf{u} = \dot{\mathbf{r}}$ and acceleration $\mathbf{a} = \ddot{\mathbf{r}}$ as follows:

$$\mathbf{u}' = \mathbf{u} - \mathbf{v}; \quad \mathbf{a}' = \mathbf{a}$$

This can be seen by taking the derivative of the 'boost' formula.

1.4. Galilean transformations

A general Galilean Transformation is any transformation that preserves inertial frames. They are combinations of:

- boosts $\mathbf{r}' = \mathbf{r} - \mathbf{v}t$ where $\mathbf{v}$ is constant,
- translations of space (moving the origin) $\mathbf{r}' = \mathbf{r} - \mathbf{r}_0$ where r_0 is constant,
- translations of time $t' = t - t_0$ where t_0 is constant,
- rotations and reflections in space $\mathbf{r}' = R\mathbf{r}$ where R is a constant orthogonal matrix.

This set generates the Galilean group. For any Galilean transformation we have

$$\ddot{\mathbf{r}} = \mathbf{0} \iff (\ddot{\mathbf{r}'}) = \mathbf{0}$$

The principle of Galilean relativity is that the laws of Newtonian physics are the same in all inertial frames. In other words, the laws of physics are always the same:

- at any point in space
- at any point in time
- in any direction
- at any constant velocity

Any set of equations which describe Newtonian physics must preserve this Galilean invariant. This shows that measurement of velocity cannot be absolute, it must be relative to a specific inertial frame of reference—but conversely, measurement of acceleration *is* absolute.

1.5. Newton's second law

For any particle subject to a force $\mathbf{F}$, the momentum $\mathbf{p}$ of the particle satisfies

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = \mathbf{F}$$

where $\mathbf{p} = m\mathbf{u}$. For this part of the course, let us assume that m is constant. Then $\mathbf{F} = \dot{\mathbf{p}} = m\mathbf{a}$. We can interpret this value m as a measure of 'reluctance to accelerate', i.e. its inertia. If $\mathbf{F}$ is specified as a function of $\mathbf{r}, \dot{\mathbf{r}}, t$, then we have a second order differential equation for $\mathbf{r}$. In order to solve this equation, we must provide two initial conditions, such as $\mathbf{r}_0$ and $\dot{\mathbf{r}}_0$ at some initial time t_0. The trajectory of the particle is then determined for all future and past times.

1.6. Gravitational force

Consider two particles, one at $\mathbf{r}_1$ and one at $\mathbf{r}_2$. Newton's law of gravitation states that the gravitational force on $\mathbf{r}_1$ is given by

$$\mathbf{F}_1 = \frac{-Gm_1m_2(\mathbf{r}_1 - \mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|^3}$$

where G is the gravitational constant, and $\mathbf{F}_2$ is given by $-\mathbf{F}_1$. Note that:

- This is known as an inverse square law, since the magnitude of the output is proportional to the inverse of the square of the distance between the particles.
- This is an attractive force, since it is in the direction $\mathbf{r}_2 - \mathbf{r}_1$.
- This obeys Newton's Third Law, since $\mathbf{F}_2 = -\mathbf{F}_1$.
- By inspection, G must have dimension $L^3 \cdot M^{-1} \cdot T^{-2}$, i.e. length cubed over mass over time squared.

1.7. Electromagnetic force

Consider a particle with electric charge q, in the presence of an electric field $\mathbf{E}(\mathbf{r}, t)$ and a magnetic field $\mathbf{B}(\mathbf{r}, t)$. The Lorentz force law states that

$$\mathbf{F}(\mathbf{r}, \dot{\mathbf{r}}, t) = q\,(\mathbf{E} + \dot{\mathbf{r}} \times \mathbf{B})$$

As an example, let $\mathbf{E} = \mathbf{0}$ everywhere, and let $\mathbf{B}$ be a constant vector. Then

$$m\ddot{\mathbf{r}} = q\dot{\mathbf{r}} \times \mathbf{B}$$

We can solve this differential equation for $\mathbf{r}$. Let us choose axes such that $\mathbf{B} = B\hat{\mathbf{z}}$, i.e. $\mathbf{B}$ is in the z direction. Evaluating the cross product, $m\ddot{z} = 0$, so $z = z_0 + ut$ where z_0 and u are constants. Further,

$$m\ddot{x} = qB\dot{y}; \quad m\ddot{y} = -qB\dot{x}$$

For convenience, let us define $\omega = qB/m$, and then

$$x = x_0 - \alpha\cos(\omega(t - t_0)); \quad x = y_0 + \alpha\sin(\omega(t - t_0))$$

This describes circles in the x–y plane, and constant velocity motion in the z direction. This results in a helix in the direction of the magnetic field, clockwise when viewed from the direction of $\mathbf{B}$.

2. Dimensional analysis

2.1. Choice of units

Many problems in dynamics involve three basic dimensional quantities: length, mass and time. These are commonly referred to using the symbols L, M and T, to be generic over the choice of unit system. The dimensions of some quantity x can therefore be expressed in terms of powers of L, M, T. So the dimension of density is $M \cdot L^{-3}$. The dimension of force is $M \cdot L \cdot T^{-2}$.

Only 'power law' functions of these quantities are allowed; we are not allowed to exponentiate a dimensional quantity, for example. This is because $e^L = 1 + L + \frac{1}{2}L^2 + \ldots$ would be comparing a dimensionless constant 1 with some length, and some area, and so forth. This comparison does not make any sense.

We can choose a unit system that is convenient, for example SI units. It defines the metre for L, the kilogram for M and the second for T. So many other physical quantities can be formed from these. For example, the SI unit for the gravitational constant is $\mathrm{m}^3\,\mathrm{kg}^{-1}\,\mathrm{s}^{-2}$. In this unit system, we can say $G = 6.67 \times 10^{-11}\,\mathrm{m}^3\,\mathrm{kg}^{-1}\,\mathrm{s}^{-2}$.

As a general principle, dynamical and physical equations must work for any consistent choice of units. If, however, we used SI units for length, mass and time, but the imperial unit pound-force as the unit for force, the equations would be inconsistent.

2.2. Scaling and dimensional independence

Suppose that a dimensional quantity Y depends on a set of dimensional quantities $X_1, \ldots, X_n$, so the dimension of Y is $L^a M^b T^c$ and the dimension of the X_i are $L^{a_i} M^{b_i} T^{c_i}$.

If $n \leq 3$, then $Y = C \cdot X_1^{p_1} X_2^{p_2} X_3^{p_3}$, and p_1, p_2, p_3 can be found by balancing the dimensions. Hence $a = a_1 p_1 + a_2 p_2 + a_3 p_3$ and so forth for b and c. This yields a unique solution for p_1, p_2, p_3 if these three equations are linearly independent, i.e. if the dimensions of X_1, X_2, X_3 are independent.

If $n > 3$, then this property of dimensional independence does not hold; it is always possible to express one of the four (or more) dimensions in terms of the other three. So let us choose X_1, X_2, X_3 to be dimensionally independent, and then we can incorporate X_4, X_5 and so on as dimensionless quantities:

$$\lambda_1 = \frac{X_4}{X_1^{q_{11}} X_2^{q_{12}} X_3^{q_{13}}}; \quad \lambda_2 = \frac{X_5}{X_1^{q_{21}} X_2^{q_{22}} X_3^{q_{23}}} \cdots$$

where the powers q_{ij} have been chosen such that the λ are dimensionless. Then

$$Y = X_1^{p_1} X_2^{p_2} X_3^{p_3} \cdot C(\lambda_1, \lambda_2, \ldots, \lambda_{n-3})$$

This is known as Bridgman's Theorem.

Example. As an example, let us consider a simple pendulum with a string of length ℓ, released from rest, when the horizontal distance from the end of the pendulum to the rest position is d. How does the period P of the pendulum depend on the four dimensional quantities m, ℓ, d, g?

We know that the dimension of the period is T, time. The dimension of m is M, the dimension of g is $L \cdot T^{-2}$, and the dimensions of ℓ and d are both L. We will form one dimensionless group, since $n = 4$ in this case. A simple way of doing so is letting $\lambda = d/\ell$. So $P = m^{p_1}\ell^{p_2}g^{p_3} \cdot f(d/\ell)$. Comparing units, we have $T = M^{p_1}L^{p_2}(L \cdot T^{-2})^{p_3}$. Solving, we get $p_1 = 0, p_2 = \frac{1}{2}, p_3 = \frac{-1}{2}$. Applying Bridgman's Theorem, we have $P = \sqrt{\ell/g} \cdot f(d/\ell)$. This does not completely specify the formula, but it does provide useful insights. For example, doubling both d and ℓ, $P \mapsto \sqrt{2}P$, since d/ℓ does not change.

Example. Taylor used publicly available data on the fireball's growth over time in order to estimate the energy released in the first atomic explosion. Let $R(t)$ be the radius of the fireball as a function of time, which has dimension L. The time t has dimension T. The density of air ρ has dimension $M \cdot L^{-3}$. The energy of the explosion is E which has dimension $M \cdot L^2 \cdot T^{-2}$. Then, $R = C \cdot t^{\alpha}\rho^{\beta}E^{\gamma}$. By balancing dimensions, we have $\alpha = \frac{2}{5}, \beta = \frac{-1}{5}, \gamma = \frac{1}{5}$. Then, $R(t) = C \cdot t^{\frac{2}{5}}\rho^{\frac{-1}{5}}\gamma^{\frac{1}{5}}$.

Taylor then verified this $\frac{2}{5}$ power law, and estimated the value of E as $\frac{\rho R^5}{C^5 t^2}$. It was observed that $\frac{R^5}{t^2} \sim 6.7 \times 10^{13}\,\mathrm{m^5\,s^{-1}}$, and $\rho \sim 1.25\,\mathrm{kg\,m^{-3}}$. Then if $C \sim 1$ then $E \sim 1 \times 10^{14}$ J, which is approximately 2.4×10^4 t of TNT.

3. Forces and potential energy

3.1. Forces

Consider a particle of mass m at position $x(t)$ in one spatial dimension. Let us consider the action of a force $F(x)$ on the particle, i.e. a force dependent entirely on the position and not the velocity or time. We define the potential energy $V(x)$ by

$$F(x) = -\frac{\mathrm{d}V}{\mathrm{d}x}$$

Hence,

$$V(x) = -\int^{x} F(u)\,\mathrm{d}u$$

The lower limit is unspecified to give an arbitrary constant in $V(x)$. If possible, the constant is usually chosen such that as $|x| \to \infty$, we have $V \to 0$. By Newton's Second Law,

$$m\ddot{x} = -\frac{\mathrm{d}V}{\mathrm{d}x}$$

We define the kinetic energy $T = \frac{1}{2}m\dot{x}^2$. The total energy in the system E is defined as $T + V = \frac{1}{2}m\dot{x}^2 + V(x)$. We will show that total energy is conserved: $\frac{\mathrm{d}E}{\mathrm{d}t} = 0$.

Proof.

$$\begin{aligned}
\frac{\mathrm{d}E}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}m\dot{x}^2 + V(x)\right) \\
&= m\dot{x}\ddot{x} + \frac{\mathrm{d}V}{\mathrm{d}x}\dot{x} \\
&= \dot{x}\left(m\ddot{x} + \frac{\mathrm{d}V}{\mathrm{d}x}\right) \\
&= \dot{x}(0) \\
&= 0
\end{aligned}$$

□

In general, in order to conserve a total energy $\frac{1}{2}m\dot{x}^2 + \Phi$, we require that

$$\dot{x}F = -\frac{\mathrm{d}\phi}{\mathrm{d}t}$$

It is usually the case that there exists no such Φ if F depends on $\dot{x}$ or t.

Example. Let us consider the example of the harmonic oscillator, i.e.

$$F(x) = -kx$$

Then we can construct

$$V(x) = -\int^{x} -ku\,\mathrm{d}u = \int^{x} ku\,\mathrm{d}u = \frac{1}{2}kx^2$$

where we have chosen the arbitrary constant conveniently. Note that we can explicitly solve the second order ordinary differential equation to compute x as a function of t:

$$x(t) = A\cos\omega t + B\sin\omega t; \quad \dot{x}(t) = -\omega A\sin\omega t + \omega B\cos\omega t$$

where $\omega = \sqrt{\frac{k}{m}}$. We can check that energy E is conserved:

$$\begin{aligned} E &= \frac{1}{2}m\dot{x}^2 + \frac{1}{2}kx^2 \\ &= \frac{1}{2}m\left(-\omega A\cos\omega t + \omega B\sin\omega t\right)^2 + \frac{1}{2}k\left(A\sin\omega t + B\cos\omega t\right)^2 \\ &= \frac{1}{2}k(A^2 + B^2) \end{aligned}$$

3.2. More general potentials

Note that conservation of energy is a first integral of Newton's Second Law. In one dimension, conservation of energy gives useful information about a particle's motion that can help in deriving x as a function of t. In the previous example, we verified that conservation of energy holds having already solved the differential equation, but it can often be more useful to consider energy while solving the equation.

$$E = \frac{1}{2}m\dot{x}^2 + V(x)$$

Hence,

$$\dot{x} = \pm\sqrt{\frac{2}{m}(E - V(x))}$$

Therefore,

$$\int_{x_0}^{x} \frac{\mathrm{d}u}{\sqrt{\frac{2}{m}(E - V(u))}} = t - t_0$$

where $x(t_0) = x_0$. This gives t as a function of x; we can invert this function to give x as a function of x. Realistically, this integral is mostly useful to get structural insight rather than actually solving x as a function of time, since it is difficult to do this analytically. As an example, let

$$V(x) = \lambda(x^3 - 3\beta^2 x)$$

where λ, β are positive constants. What happens if we release the particle from rest at $x = x_0$? We can draw the graph of $V(x)$ and imagine the height of the graph as the height of a 'rail' that the particle sits on, acted on under gravity, i.e. the particle 'falls' from higher $V(x)$ to lower $V(x)$, gaining kinetic energy as it falls. Since we start at rest, $E = V(x_0)$ at $t = 0$, and in the subsequent motion $E \leq V(x_0)$. We have a few cases:

(i) ($x_0 < -\beta$) $x_0 = -\beta$ is a maximum point on the graph. The particle will move to the left with $x(t) \to -\infty$ as $t \to \infty$.

(ii) ($-\beta < x_0 < 2\beta$) Note that $V(-\beta) = V(2\beta)$; they are the same height on the graph. Since there is no friction in this model, the particle's motion is confined to the region $-\beta < x < 2\beta$ and will oscillate forever.

(iii) ($2\beta < x_0$) The particle will move to the left, reaching $x = -\beta$, and then will continue to the left, since it has kinetic energy at this point. So $x \to -\infty$ as $t \to \infty$.

We also have special cases on the turning points $\pm\beta$, where the particle does not move. There is another case at $x_0 = 2\beta$: the particle will move to the left, accelerating until $x = \beta$, then decelerating until $x = -\beta$, where it will then stop moving at this maximum point. How long does it take for the particle to move from $x_0 = 2\beta$ to $x = -\beta$, where it rests? We can use the integral above to compute this, letting $t_0 = 0$ and $x(0) = 2\beta$.

$$\int_{x(t)}^{2\beta} \frac{\mathrm{d}\tilde{x}}{\sqrt{\frac{2\lambda}{m}(2\beta^3 - \tilde{x}^3 + 3\beta^2\tilde{x})}} = t$$

$$\int_{x(t)}^{2\beta} \frac{\mathrm{d}\tilde{x}}{\sqrt{\frac{2\lambda}{m}(\tilde{x} + \beta)^2(2\beta - \tilde{x})}} = t$$

$$\int_{x(t)}^{2\beta} \frac{\mathrm{d}\tilde{x}}{(\tilde{x} + \beta)\sqrt{\frac{2\lambda}{m}(2\beta - \tilde{x})}} = t$$

This integral diverges as $\tilde{x} \to -\beta$, so it takes an infinite amount of time to come to rest at this maximum point; specifically it exhibits logarithmic behaviour.

3.3. Equilibrium points

An equilibrium point is defined as a point where the potential is stationary, in other words where the force on the particle is zero. So the particle stays at rest for all time. In the example in the previous lecture, $x = \pm\beta$ were the equilibrium points. We can analyse the motion close to the equilibrium point in order to work out whether the equilibrium point is stable or unstable. Let x_0 be an equilibrium point, so $V'(x_0) = 0$. We can expand $V(x)$ as a series, assuming that $x - x_0$ is small.

$$V(x) = V(x_0) + \frac{1}{2}(x - x_0)^2 V''(x_0) + o((x - x_0)^2)$$

In the neighbourhood of x_0,

$$m\ddot{x} = -V'(x) \approx -(x - x_0)V''(x_0)$$

- If $V''(x_0) > 0$, we have a local minimum of potential, which gives rise to a stable equilibrium point. The equation of motion of a particle near x_0 is a harmonic oscillator. The angular frequency of oscillation is $\omega = \sqrt{\frac{V''(x_0)}{m}}$.

- If $V''(x_0) < 0$, we have a local maximum of potential, which gives rise to an unstable equilibrium point. Any perturbation from this point will cause an increased deviation from the point. The equation of motion near this point is exponential; almost always exponentially increasing rather than decreasing. The growth rate is $\gamma = \sqrt{\frac{-V''(x_0)}{m}}$.

- If $V''(x) = 0$, we must use higher-order terms from the Taylor series in order to determine the behaviour.

Let us consider the example of a simple pendulum with a mass m held by a rigid beam of length ℓ. Let the angle between the beam and the vertical direction be θ. By Newton's second law,

$$F(x = \ell\theta) = m\ell\ddot{\theta} = -mg\sin\theta$$

We can derive an energy equation by using $F(x) = -V'(x)$.

$$V(x = \ell\theta) = -\int_0^{\ell\theta} F(u)\,\mathrm{d}u = -mg\ell\cos\theta$$

The kinetic energy T is given by

$$T = \frac{1}{2}m\ell^2\dot{\theta}^2$$

We can check that $\frac{\mathrm{d}E}{\mathrm{d}t} = 0$ at all t. The stationary points of V are at $\theta = 0$ and $\theta = \pi$ (assuming $0 \le \theta < 2\pi$). The $\theta = 0$ point is stable, since $V''(\theta = 0) > 0$. The $\theta = \pi$ point is unstable. If $-mg\ell < E < mg\ell$, the pendulum will oscillate between two values since it cannot continue spinning in circles. In particular, this oscillation occurs about a position of stable equilibrium. However, if we add additional energy into this system, either $\dot{\theta} > 0$ or $\dot{\theta} < 0$ for all time. It is impossible to have $E < -mg\ell$ since this is the minimum value of the potential.

Now, let us consider the period P of the oscillation of θ after releasing the particle from rest at some initial angle θ_0. Note that the oscillation consists of $\theta_0 \to 0 \to -\theta_0 \to 0 \to \theta_0$. By symmetry, this period is four times the time it takes to go from θ_0 to 0. From the energy

equation, we can deduce

$$\begin{aligned} P &= 4\int_0^{\theta_0} \frac{\mathrm{d}\theta}{\sqrt{\frac{2g\ell}{\ell^2}(\cos\theta - \cos\theta_0)}} \\ &= 4\sqrt{\frac{\ell}{g}}\int_0^{\theta_0} \frac{\mathrm{d}\theta}{\sqrt{2\cos\theta - 2\cos\theta_0}} \\ &= 4\sqrt{\frac{\ell}{g}}F(\theta_0) \end{aligned}$$

where f is notably a function only of θ_0. Recall from the dimensional analysis lecture that

$$P = \sqrt{\frac{l}{g}}H\left(\frac{d}{\ell}\right)$$

noting that d/ℓ and θ both define the initial condition. So we have deduced this unknown function H. This integral is difficult to compute exactly; however, we can compute an approximation when θ_0 (and hence θ) is small.

$$\begin{aligned} F(\theta_0) &= \int_0^{\theta_0} \frac{\mathrm{d}\theta}{\sqrt{\theta_0^2 - \theta^2}} \\ &= \frac{\pi}{2} \end{aligned}$$

which is independent of θ_0. Hence, for small angles,

$$P \approx 2\pi\sqrt{\frac{\ell}{g}}$$

3.4. Force and potential in three spatial dimensions

Consider a particle moving in three spatial dimensions under a force $\mathbf{F}$. Then Newton's second law states

$$m\ddot{\mathbf{r}} = \mathbf{F}$$

We define the kinetic energy by

$$T = \frac{1}{2}|\dot{\mathbf{r}}|^2 = \frac{1}{2}|\mathbf{u}|^2$$

Then

$$\frac{\mathrm{d}T}{\mathrm{d}t} = m\dot{\mathbf{r}}\cdot\ddot{\mathbf{r}} = \mathbf{F}\cdot\dot{\mathbf{r}} = \mathbf{F}\cdot\mathbf{u}$$

This is the rate of working of the force on the particle. Let us consider the total work done by a force on a particle as it moves along a finite curve C from t_1 to t_2. Then the total work

done is the line integral

$$W = \int_{t_1}^{t_2} \mathbf{F} \cdot \mathbf{u} \, \mathrm{d}t = \int_{t_1}^{t_2} \mathbf{F} \cdot \dot{\mathbf{r}} \, \mathrm{d}t = \int_{\mathbf{r}(t_1)}^{\mathbf{r}(t_2)} \mathbf{F} \cdot \mathrm{d}\mathbf{r}$$

Note that we must specify that this integral acts along the curve C, since any other curve could connect the points $\mathbf{r}(t_1)$ and $\mathbf{r}(t_2)$. We can write this integral in terms of coordinates:

$$\int_{\mathbf{r}(t_1)}^{\mathbf{r}(t_2)} F_x \, \mathrm{d}x + F_y \, \mathrm{d}y + F_z \, \mathrm{d}z$$

Now, if force is only a function of the position $\mathbf{r}$, then we say that $\mathbf{F}(\mathbf{r})$ defines a force field. A *conservative* force field is such that

$$\mathbf{F}(\mathbf{r}) = -\nabla V(\mathbf{r})$$

for some function $V(\mathbf{r})$. In component form, this is equivalent to

$$F_i = -\frac{\partial V}{\partial x_i}$$

If the force is conservative, then the energy $E = T + V(\mathbf{r})$ is conserved.

Proof.

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\mathrm{d}T}{\mathrm{d}t} + \frac{\mathrm{d}}{\mathrm{d}t} V(\mathbf{r}) = m\dot{\mathbf{r}} \cdot \ddot{\mathbf{r}} + \nabla V \cdot \dot{\mathbf{r}} = m\dot{\mathbf{r}} \cdot \ddot{\mathbf{r}} - m\ddot{\mathbf{r}} \cdot \dot{\mathbf{r}} = 0$$

□

Let us consider the total work done on the particle under a conservative force. From the properties of the gradient vector,

$$W = \int_C \mathbf{F} \cdot \mathrm{d}\mathbf{r} = -\int_C \nabla V \cdot \mathrm{d}\mathbf{r} = V(\mathbf{r}_1) - V(\mathbf{r}_2)$$

Note that this is dependent only on the end points of the curve; it is irrelevant of the path taken. Hence, if C is closed, then no net work is done by the force. Note that in general, $F(\mathbf{r})$ is not conservative, so in general there is no $V(\mathbf{r})$ such that $\mathbf{F} = -\nabla V$. In fact, $\mathbf{F}(\mathbf{r})$ is conservative if

$$\nabla \times \mathbf{F}(\mathbf{r}) = \mathbf{0}$$

4. Gravitational and electromagnetic forces

4.1. Gravity

The gravitational force experienced by a mass m at position vector $\mathbf{r}$ relative to a mass M is given by

$$\mathbf{F} = \frac{-GMm}{|\mathbf{r}|^3} \cdot \mathbf{r} = \frac{-GMm}{|\mathbf{r}|^2} \cdot \hat{\mathbf{r}}$$

This is a conservative force:

$$\mathbf{F}(\mathbf{r}) = -\nabla V(\mathbf{r}); \quad V(\mathbf{r}) = \frac{-GMm}{r}$$

To remove the factor of m, we define the 'gravitational potential' Φ_g to be

$$\Phi_g(\mathbf{r}) = \frac{-GM}{r}$$

We further define the gravitational field

$$\mathbf{g}(\mathbf{r}) = -\nabla\Phi_g(\mathbf{r}) = \frac{-GM}{r^2}\hat{\mathbf{r}}$$

Note that this is dependent only on M, and not m. These quantities are related to $\mathbf{F}$ and V by scale factors of m.

$$V(\mathbf{r}) = m\Phi_g(\mathbf{r}); \quad \mathbf{F}(\mathbf{r}) = m\mathbf{g}$$

We can generalise these expressions to define the gravitational potential associated with many point masses M_i for $i = 1, \dots, n$. Then,

$$\Phi_g(\mathbf{r}) = -\sum_{i=1}^{n} \frac{GM_i}{|\mathbf{r} - \mathbf{r}_i|}$$

$$\mathbf{g}(\mathbf{r}) = -\sum_{i=1}^{n} \frac{GM_i}{|\mathbf{r} - \mathbf{r}_i|^3}(\mathbf{r} - \mathbf{r}_i)$$

We can extend this to a continuous mass distribution by generalising the summation into an integral. In particular, for a uniform spherical distribution of mass centred at the origin, we have that outside the sphere

$$\Phi_G(\mathbf{r}) = \frac{-GM}{r}$$

which is equivalent to the formula for a point mass at the origin. So we can represent any spherical distribution of mass as a particle, provided we never consider behaviour inside the sphere.

4.2. Gravitational and inertial mass

Note that in the equations for gravitational force, mass plays two roles.

- Inertial mass: In Newton's second law, $m\ddot{\mathbf{r}} = \mathbf{F}$ shows that the mass encapsulates the resistance to motion
- Gravitational mass: In the law of gravitation, $\mathbf{F} = \frac{-GMm}{r^2}\hat{\mathbf{r}}$, showing the scale factor by which the mass affects the force.

It turns out that these 'masses' are not exactly the same; they differ by a factor of around 1×10^{-12}. In this course, we will consider these masses to be identical since the factor is very small.

4.3. One-dimensional approximation to gravity

Let us consider a one-dimensional approximation. Consider a mass m at some height z above the surface of a planet of mass M and radius R, where $z \ll R$. Using the binomial expansion, the potential is approximated by

$$V(R+z) = \frac{-GMm}{R+z} \approx \frac{-GMm}{R} + \frac{GMmz}{R^2} - \dots$$

The first term in the expansion is a constant, and the second term is mgz where g is a constant. So when $z \ll R$,

$$V(R+z) \approx mgz; \quad g = \frac{GM}{R^2} \approx 9.8\,\mathrm{m\,s^{-2}}$$

4.4. Escape velocity

Consider a particle leaving the surface of a planet of mass M and radius R, starting with velocity v. Can this particle escape the gravitational attraction of the planet, and fly off to infinity? By conservation of energy,

$$E = T + V = \frac{1}{2}mv^2 - \frac{GMm}{r}$$

If $E < 0$, the particle does not have sufficient energy to leave the 'potential well' V. If $E > 0$, the particle can escape to infinity. The critical velocity v_0 at which the particle can escape with lowest energy (the escape velocity) is therefore computed by setting $E = 0$ at $r = R$, i.e.

$$\frac{1}{2}v_0^2 = \frac{GM}{R} \implies v_0 = \sqrt{\frac{2GM}{R}}$$

Note that light has a finite velocity, c. Therefore it must be possible that a mass is large enough that even the speed of light is insufficient for a particle to escape from a given radius. This describes a black hole. Of course, at this point we would need to invoke Einstein's theory of relativity in order to properly describe the behaviour of such an object.

4.5. Electromagnetism

We know that the force $\mathbf{F}$ acting on a particle with charge q is

$$\mathbf{F} = q\mathbf{E} + q\dot{\mathbf{r}} \times \mathbf{B}$$

where $\mathbf{E}, \mathbf{B}$ are functions of $\mathbf{r}$ and t. This is known as the Lorentz force law. Let us first consider time-independent fields $\mathbf{E}(\mathbf{r}), \mathbf{B}(\mathbf{r})$ as a simplification. In this case, we can write

$$\mathbf{E} = -\nabla\Phi_e(\mathbf{r})$$

where Φ_e is the electrostatic potential. The force $q\mathbf{E}$ is therefore conservative. We now prove that for time independent $\mathbf{E}(\mathbf{r})$ and $\mathbf{B}(\mathbf{r})$, $\mathbf{F}$ is conservative.

Proof.

$$\begin{aligned} E &= \frac{1}{2}m|\dot{\mathbf{r}}|^2 + q\Phi_e(\mathbf{r}) \\ \frac{\mathrm{d}E}{\mathrm{d}t} &= m\dot{\mathbf{r}} \cdot \ddot{\mathbf{r}} + q\dot{\mathbf{r}} \cdot \nabla\Phi_e(\mathbf{r}) \\ &= \dot{\mathbf{r}} \cdot (m\ddot{\mathbf{r}} + q\nabla\Phi_e) \\ &= \dot{\mathbf{r}} \cdot (q\mathbf{E} + q\dot{\mathbf{r}} \times \mathbf{B} + q\nabla\Phi_e) \\ &= \dot{\mathbf{r}} \cdot (q\dot{\mathbf{r}} \times \mathbf{B}) \\ &= 0 \end{aligned}$$

since this is a triple product where two of the vectors are parallel. Since $\mathbf{B}$ acts perpendicular to the velocity, it does not do work on the particle. □

Analogously to point masses, we may consider point charges. A particle with charge Q located at the origin generates an electrostatic potential and electric field

$$\Phi_e(\mathbf{r}) = \frac{Q}{4\pi\varepsilon_0 r}; \quad \mathbf{E}(\mathbf{r}) = -\nabla\Phi_e = \frac{Q}{4\pi\varepsilon_0 r^2}\hat{\mathbf{r}}$$

where $\varepsilon_0 = 8.85 \times 10^{-12}\,\mathrm{m^{-3}\,kg^{-1}\,s^2\,C^2}$ is the electric constant. So the force on a particle of charge q located at $\mathbf{r}$ is given by

$$\mathbf{F} = -q\nabla\Phi_e = \frac{Qq}{4\pi\varepsilon_0 r^2}\hat{\mathbf{r}}$$

This is called the Coulomb force. A negative sign is an attractive force; a positive sign is a repulsive force. This can be seen by considering a perturbation from the origin.

5. Friction

Friction is a contact force, unlike the forces we have discussed previously. It is a convenient encapsulation of many complicated molecular phenomena; it is not a fundamental force.

5.1. Dry friction

The friction associated with solid bodies in contact is called 'dry' friction. It has two associated forces: the normal force $\mathbf{N}$ perpendicular to the contact surface, which prevents objects from passing through each other, and the tangential force $\mathbf{F}$ parallel to the contact surface, which resists the relative tangential motion of the bodies in contact. When the two bodies are static, the empirically-derived formula relating the forces is

$$|\mathbf{F}| \leq \mu_s |\mathbf{N}|$$

where μ_s is the coefficient of static friction. If the objects start to move relative to each other, this is kinetic friction. In this case,

$$|\mathbf{F}| = \mu_k |\mathbf{N}|$$

where μ_k is the coefficient of kinetic friction. Generally $\mu_s > \mu_k > 0$.

5.2. Fluid drag

When a solid body moves through a fluid (a liquid or a gas), it experiences a drag force. There are two important equations that model fluid drag. The linear drag formula is

$$\mathbf{F} = -k_1 \mathbf{u}$$

This formula is most relevant to 'small' objects, moving through a viscous fluid. Stokes' drag law for a moving sphere states that

$$k_1 = 6\pi\eta R$$

where η is the viscosity of the fluid, and R is the radius of the sphere. The quadratic drag formula is

$$\mathbf{F} = -k_2 |\mathbf{u}| \mathbf{u}$$

This formula is more relevant to 'large' objects, moving through a less viscous fluid. Of course, $k_1 \neq k_2$ since they have different dimensions. Typically, we have

$$k_2 = \rho_{\text{fluid}} C_D R^2$$

where C_D is the drag coefficient, and R^2 is the size of the cross section.

5.3. Work done by friction

Note that since friction always acts in a direction opposite to a component of motion, the body loses kinetic energy if the fluid (or other body) is assumed to be at rest. The rate of work under a fluid's drag force is

$$\mathbf{F}\cdot\mathbf{u} = \begin{cases} -k_1|\mathbf{u}|^2 & \text{linear drag} \\ -k_2|\mathbf{u}|^3 & \text{quadratic drag} \end{cases}$$

In the latter case, the total work done is proportional to $|\mathbf{u}|^2$ multiplied by the total distance travelled. The fluid gains energy, which may manifest as heat.

5.4. Projectiles experiencing linear drag

Let us consider the example of a projectile moving through the air, under uniform gravity and a linear drag force.

$$m\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = m\mathbf{g} - k\mathbf{u}$$

Solving with an integrating factor, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{u}e^{kt/m}\right) = m\mathbf{g}e^{kt/m}$$

$$\mathbf{u} = \frac{m\mathbf{g}}{k} + \mathbf{C}e^{-kt/m}$$

We can find $\mathbf{C}$ using the initial conditions, say at $t = 0, \mathbf{x} = 0, \mathbf{u} = \mathbf{U}$.

$$\mathbf{u} = \frac{m\mathbf{g}}{k} + \left(\mathbf{U} - \frac{-\mathbf{g}}{k}\right)e^{-kt/m}$$

Then

$$\begin{aligned}\mathbf{x} &= \frac{m\mathbf{g}}{k}t - \frac{m}{k}\left(\mathbf{U} - \frac{-\mathbf{g}}{k}\right)e^{-kt/m} + D \\ &= \frac{m\mathbf{g}}{k}t + \frac{m}{k}\left(\mathbf{U} - \frac{-\mathbf{g}}{k}\right)\left(1 - e^{-kt/m}\right)\end{aligned}$$

Now, considering the components of $\mathbf{x} = (x, y, z)$ and $\mathbf{u} = (u, v, w)$, we can choose

$$\mathbf{U} = (U\cos\theta, 0, U\sin\theta); \quad \mathbf{g} = (0, 0, -g)$$

Then

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} U\cos\theta e^{-kt/m} \\ 0 \\ \left(U\sin\theta + \frac{mg}{k}\right)e^{-kt/m} - \frac{mg}{k} \end{pmatrix}$$

Note that the terminal velocity is $(0, 0, -mg/k)$, achieved on a time scale of m/k (as seen from the exponential term). Further,

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \frac{mU\cos\theta}{k}\left(1 - e^{-kt/m}\right) \\ 0 \\ \frac{m}{k}\left(U\sin\theta + \frac{mg}{k}\right)\left(1 - e^{-kt/m}\right) - \frac{mgt}{k} \end{pmatrix}$$

There exists a range R of this particle, since initially the particle moves upwards, but as time increases the particle begins moving downwards again. R is a function of U, θ, m, k, g. We can construct the dimensionless group $\frac{kU}{mg} = \frac{U/g}{m/k}$, which can be thought of as the gravitational time scale divided by the frictional time scale. Dimensional analysis shows that

$$R = \frac{U^2}{g} F\left(\theta, \frac{kU}{mg}\right)$$

When $\frac{kU}{mg} \ll 1$, this is very small friction.

$$R = \frac{U^2}{g} \cdot 2\sin\theta\cos\theta$$

When $\frac{kU}{mg} \gg 1$, this is very large friction.

$$R = \frac{U^2}{g}\left(\frac{mg}{kU}\cos\theta\right)$$

R is a decreasing function of $\frac{kU}{mg}$.

6. Angular motion and orbits

6.1. Angular momentum

We define the angular momentum for a particle with position vector $\mathbf{r}(t)$, of mass m, moving under the influence of a force $\mathbf{F}$ as

$$\mathbf{L} = \mathbf{r} \times \mathbf{p} = \mathbf{r} \times m\dot{\mathbf{r}}$$

Then

$$\dot{\mathbf{L}} = m\dot{\mathbf{r}} \times \dot{\mathbf{r}} + m\mathbf{r} \times \ddot{\mathbf{r}} = \mathbf{r} \times \mathbf{F} = \mathbf{G}$$

This term $\mathbf{r} \times \mathbf{F} = \mathbf{G}$ is sometimes called the torque or the moment of the force. The values of $\mathbf{L}$ and $\mathbf{G}$ depend on the choice of origin, so we typically refer to the angular momentum about a particular point. If $\mathbf{r} \times \mathbf{F} = \mathbf{0}$, then the angular momentum is conserved. The angular momentum around some suitably chosen point may be constant; this may help with calculations since we are free to choose the origin.

6.2. Orbits

We will begin the topic of orbits by considering the problem of gravitational orbits. Let

$$m\ddot{\mathbf{r}} = -\nabla V(r)$$

This represents a particle moving in a conservative force that is a function only of the radius from the origin. For this problem, we are assuming that the 'central' mass remains fixed at the origin. This is a good approximation if the central mass is significantly larger than m.

6.3. Central forces

We define a central force as a conservative force with the potential $V(r)$ being a function only of the radius from the origin. Consequently,

$$\mathbf{F} = -\nabla V(r) = -\nabla V(|\mathbf{r}|) = -\frac{\mathrm{d}V}{\mathrm{d}r}\hat{\mathbf{r}}$$

Consider the angular momentum $\mathbf{L}$ about the origin, given by

$$\dot{\mathbf{L}} = \mathbf{r} \times \mathbf{F} = \mathbf{r} \times \left(-\frac{\mathrm{d}V}{\mathrm{d}r}\hat{\mathbf{r}}\right) = 0$$

So angular momentum about the origin is conserved for any central force. Further, from the definition of $\mathbf{L}$,

$$\mathbf{L} \cdot \mathbf{r} = 0$$

Hence, the motion of the particle is confined to the plane through the origin, perpendicular to $\mathbf{L}$. This reduces a three-dimensional problem into a two-dimensional problem.

6.4. Polar coordinates in the plane

A convenient choice of coordinates to use is the set of two-dimensional polar coordinates, by choosing the z axis such that the orbit lies in the plane $z = 0$. Then

$$x = r\cos\theta; \quad y = r\sin\theta$$

Then, relative to the Cartesian axes,

$$\mathbf{e}_r = \hat{\mathbf{r}} = \begin{pmatrix}\cos\theta\\ \sin\theta\end{pmatrix}; \quad \mathbf{e}_\theta = \begin{pmatrix}-\sin\theta\\ \cos\theta\end{pmatrix}$$

At any point, $\mathbf{e}_r$, $\mathbf{e}_\theta$ form an orthonormal basis, but the basis can point in different directions for different values of θ. In other words, they form a set of orthonormal curvilinear coordinates. We have

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbf{e}_r = \mathbf{e}_\theta; \quad \frac{\mathrm{d}}{\mathrm{d}\theta}\mathbf{e}_\theta = -\mathbf{e}_r$$

Note that for a moving particle, r and θ are functions of position, and hence functions of time. So we can use the following results:

$$\frac{\mathrm{d}\mathbf{e}_r}{\mathrm{d}t} = \frac{\mathrm{d}\theta}{\mathrm{d}t}\frac{\mathrm{d}\mathbf{e}_r}{\mathrm{d}\theta} = \mathbf{e}_\theta\frac{\mathrm{d}\theta}{\mathrm{d}t}; \quad \frac{\mathrm{d}\mathbf{e}_\theta}{\mathrm{d}t} = \frac{\mathrm{d}\theta}{\mathrm{d}t}\frac{\mathrm{d}\mathbf{e}_\theta}{\mathrm{d}\theta} = -\mathbf{e}_r\frac{\mathrm{d}\theta}{\mathrm{d}t}$$

We can compute expressions for velocity and acceleration in terms of these new coordinates.

$$\begin{aligned}\mathbf{r} &= r\mathbf{e}_r\\ \therefore \dot{\mathbf{r}} &= \dot{r}\mathbf{e}_r + r\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}_r\\ &= \dot{r}\mathbf{e}_r + r\dot{\theta}\mathbf{e}_\theta\end{aligned}$$

So $\dot{r}$ is the radial component of the velocity, and $r\dot{\theta}$ is the angular component of the velocity. Note that $\dot{\theta}$ is the angular velocity. Further:

$$\begin{aligned}\ddot{\mathbf{r}} &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\dot{r}\mathbf{e}_r + r\dot{\theta}\mathbf{e}_\theta\right)\\ &= \ddot{r}\mathbf{e}_r + \dot{r}\dot{\mathbf{e}}_r + \dot{r}\dot{\theta}\mathbf{e}_\theta + r\ddot{\theta}\mathbf{e}_\theta + r\dot{\theta}\dot{\mathbf{e}}_\theta\\ &= \ddot{r}\mathbf{e}_r + \dot{r}\dot{\theta}\mathbf{e}_\theta + \dot{r}\dot{\theta}\mathbf{e}_\theta + r\ddot{\theta}\mathbf{e}_\theta + r\dot{\theta}\left(-\dot{\theta}\mathbf{e}_r\right)\\ &= \left(\ddot{r} - r\dot{\theta}^2\right)\mathbf{e}_r + \left(2\dot{r}\dot{\theta} + r\ddot{\theta}\right)\mathbf{e}_\theta\end{aligned}$$

Again we can read off the radial and angular components of the acceleration.

6.5. Circular motion

Let us consider the example of circular motion with constant angular velocity. Then we can set $r = a$, $\dot{\theta} = \omega$, and let $\dot{r} = \ddot{r} = \ddot{\theta} = 0$. We can find that

$$\dot{\mathbf{r}} = a\omega\mathbf{e}_\theta; \quad \ddot{\mathbf{r}} = -a\omega^2\mathbf{e}_r$$

The acceleration is in the inward radial direction, which constrains the particle to follow a circular path instead of flying off tangentially towards infinity. Therefore, by Newton's second law, there is a constant force in this direction.

7. Orbits and stability

7.1. Motion in a central force field

By Newton's second law, the force in a central force field is given by

$$m\ddot{\mathbf{r}} = \mathbf{F} = -\nabla V = -\frac{\mathrm{d}V}{\mathrm{d}r}\mathbf{e}_r$$

The results from the previous lecture give

$$m\left(\ddot{r} - r\dot{\theta}^2\right)\mathbf{e}_r + m\left(2\dot{r}\dot{\theta} + r\ddot{\theta}\right)\mathbf{e}_\theta = \frac{-\mathrm{d}V}{\mathrm{d}r}\mathbf{e}_r \qquad (*)$$

But the right hand side has no angular component, so $m\left(2\dot{r}\dot{\theta} + r\ddot{\theta}\right) = 0$. Then

$$\frac{m}{r}\frac{\mathrm{d}}{\mathrm{d}t}\left(r^2\dot{\theta}\right) = 0$$

So the quantity $h = r^2\dot{\theta}$, known as the *specific* angular momentum (since it contains no mass component) is constant. Note that the angular momentum $\mathbf{L}$ is given by

$$\mathbf{L} = m\mathbf{r} \times \dot{\mathbf{r}} = mr\mathbf{e}_r \times \left(\dot{r}\mathbf{e}_r + r\dot{\theta}\mathbf{e}_\theta\right) = mr^2\dot{\theta}\mathbf{e}_z$$

Hence the magnitude of the angular momentum is constant. Now, let us consider the radial component in $(*)$.

$$\begin{aligned} m\ddot{r} - mr\dot{\theta}^2 &= -\frac{\mathrm{d}V}{\mathrm{d}r} \\ m\ddot{r} &= -\frac{\mathrm{d}V}{\mathrm{d}r} + \frac{mh^2}{r^3} \\ m\ddot{r} &= -\frac{\mathrm{d}V_{\text{eff}}}{\mathrm{d}r} \end{aligned}$$

where

$$V_{\text{eff}}(r) = V(r) + \frac{mh^2}{2r^2}$$

where V_{eff} is called the effective potential. In other words, the motion of the particle is equivalent to one-dimensional motion under the influence of the effective potential. The energy of the particle is given by

$$T + V(r) = \frac{1}{2}m\left(\dot{r}^2 + r^2\dot{\theta}^2\right) + V(r) = \frac{1}{2}m\dot{r}^2 + \frac{mh^2}{2r^2} + V(r) = \frac{1}{2}m\dot{r}^2 + V_{\text{eff}}(r)$$

which is consistent with our description of the effective potential.

7.2. Orbits under gravity

As an example, let us consider

$$V(r) = \frac{-GMm}{r}; \quad V_{\text{eff}}(r) = \frac{-GMm}{r} + \frac{mh^2}{2r^2}$$

The effective potential has a single minimum point at r_*, and a single root at r_0. In other words, $V'_{\text{eff}}(r_*) = 0$ and $V_{\text{eff}}(r_0) = 0$. We can compute that

$$r_0 = \frac{h^2}{2GM}; \quad r_* = \frac{h^2}{GM}$$

The minimum energy is therefore

$$E_{\min} = \frac{-m(GM)^2}{2h^2}$$

What is the possible motion of the particle? At $E = E_{\min}$, we have $r(t) = r_*$, an equilibrium position. Further, $\dot{\theta} = \frac{h}{r_*^2}$ everywhere. At $E_{\min} < E < 0$, then $r(t)$ oscillates between a minimum point (periapsis/perihelion/perigee) and a maximum point (apoapsis/aphelion/apogee), and $\dot{\theta}$ varies. If $E_{\min} \geq 0$, the particle escapes to infinity. This is sometimes called an unbound orbit.

7.3. Stability of circular orbits

Consider a general potential $V(r)$. Does a circular orbit exist, and is it stable? We will assume that the angular momentum is given and nonzero. For a circular orbit, the radius is a constant value r_*, so $\ddot{r} = 0$ and hence $V'_{\text{eff}}(r_*) = 0$. We know that we have a stable equilibrium if V_{eff} has a minimum at this point. Correspondingly, it is unstable if this is a maximum. So, for instance, it is stable if $V''_{\text{eff}}(r_*) > 0$. Now, let us rewrite these conditions in terms of $V(r)$.

$$V'(r_*) - \frac{mh^2}{r_*^3} = 0; \quad V''(r_*) = V''(r_*) + \frac{3mh^2}{r_*^4} > 0$$

We can combine these to give the condition for stability as

$$V''(r_*) + \frac{3V'(r_*)}{r_*} > 0$$

Now let us consider an example,

$$V(r) = \frac{-km}{r^p}$$

where $p > 0$, $k > 0$. If $p = 1$, this is an example of an inverse square law. We have a circular orbit if

$$\frac{pkm}{r_*^{p+1}} - \frac{mh^2}{r_*^3} = 0$$

Hence,

$$r_*^{p-2} = \frac{pkm}{mh^2} \implies r_* = \left(\frac{pkm}{mh^2}\right)^{\frac{1}{p-2}}$$

So there exists a circular orbit for all h provided $p \neq 2$. Is this a stable orbit?

$$V''(r_*) + \frac{3V'(r_*)}{r_*} = \frac{-kmp(p+1)}{r_*^{p+2}} + \frac{3kmp}{r_*^{p+2}} = \frac{p(2-p)km}{r_*^{p+2}}$$

So this is greater than zero (stable) if $0 < p < 2$ and less than zero (unstable) if $p > 2$.

7.4. The orbit equation

What shape does a non-circular orbit trace out? We could in principle find $r(t)$ by the energy equation

$$E = \frac{1}{2}m\dot{r}^2 + V_{\text{eff}}(r) = \text{constant}$$

Hence

$$t = \pm\sqrt{\frac{m}{2}}\int^r \frac{\mathrm{d}u}{\sqrt{E - V_{\text{eff}}(u)}}$$

Then we can use $r(t)^2\dot{\theta} = h$ to deduce $\theta(t)$. However in practice, this is not useful. An analytic solution is only possible for a small family of effective potential functions. It is somewhat more convenient to find r in terms of θ, not in terms of t. We can write

$$\frac{\mathrm{d}}{\mathrm{d}t} = \frac{\mathrm{d}\theta}{\mathrm{d}t}\frac{\mathrm{d}}{\mathrm{d}\theta} = \frac{h}{r^2}\frac{\mathrm{d}}{\mathrm{d}\theta}$$

Applying this to Newton's second law, we have

$$m\frac{h}{r^2}\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{h}{r^2}\frac{\mathrm{d}}{\mathrm{d}\theta}r\right) - \frac{mh^2}{r^3} = F(r)$$

The $\frac{h}{r^2}\frac{\mathrm{d}}{\mathrm{d}\theta}r$ term suggests using the substitution $u = \frac{1}{r}$. Then

$$mhu^2\frac{\mathrm{d}}{\mathrm{d}\theta}\left(-h\frac{\mathrm{d}u}{\mathrm{d}\theta}\right) - mh^2u^3 = F(u^{-1})$$

$$\frac{\mathrm{d}^2u}{\mathrm{d}\theta^2} + u = \frac{-1}{mh^2u^2}F(u^{-1})$$

This is known as the orbit equation. We can solve this for u as a function of θ.

7.5. The Kepler problem

The Kepler problem is the orbit problem, specialised to the case of a gravitational central force. The force is

$$F(r) = \frac{-mk}{r^2}$$

where the constant k is equivalent to GM. Hence,

$$\frac{\mathrm{d}^2u}{\mathrm{d}\theta^2} + u = \frac{-1}{mh^2u^2} \cdot -mku^2 = \frac{k}{h^2}$$

This gives us a linear equation in u, which is promising for solving this equation. The solution is

$$u = \frac{k}{h^2} + A\cos(\theta - \theta_0)$$

where A and θ_0 are specified by the initial conditions. Without loss of generality we can let $A \geq 0$. If $A = 0$, then $u = \frac{k}{h^2}$ giving a circular orbit. If $A > 0$, then u is maximised (at periapsis) when $\theta = \theta_0$, and u is minimised (at apoapsis) where $\theta = \theta_0 + \pi$. We will choose that $\theta_0 = 0$ for convenience; we will simply need to change the origin of our coordinate system if this does not hold. We will redefine other constants for convenience:

$$r = \frac{1}{u} = \frac{\ell}{1 + e\cos\theta}; \quad \ell = \frac{h^2}{k}; \quad e = A\frac{h^2}{k}$$

This is the equation of a conic section. Here, e is the eccentricity of the curve. We can rewrite this in Cartesian form:

$$\begin{aligned} r(1 + e\cos\theta) &= \ell \\ r &= \ell - ex \\ x^2 + y^2 &= (\ell - ex)^2 \\ (1 - e^2)x^2 + y^2 + 2e\ell x &= \ell^2 \end{aligned} \tag{†}$$

By inspection we can see that the value of $(1 - e^2)$ determines the shape of the conic section.

- $(0 \leq e < 1)$ This forms an ellipse; the orbit is bounded by $\frac{\ell}{1+e} \leq r \leq \frac{\ell}{1-e}$. We can rewrite (†) as
$$\frac{(x + ea)^2}{a^2} + \frac{y^2}{b^2} = 1$$
where $a = \frac{\ell}{1-e^2}$ and $b = \frac{\ell}{\sqrt{1-e^2}}$, and therefore clearly $b \leq a$. Note that $e = 0$ is the special case of a circle. The origin lies at one of the foci of the ellipse.
- $(e > 1)$ This forms a hyperbola. This is an unbounded orbit, since there exists a value α such that as $\theta \to \alpha$, we have $r \to \infty$. Note that $\alpha = \arccos\left(\frac{-1}{e}\right) \in \left(\frac{\pi}{2}, \pi\right)$ We can transform (†) as before:
$$\frac{(x - ea)^2}{a^2} - \frac{y^2}{b^2} = 1$$

where $a = \frac{\ell}{e^2-1}$ and $b = \frac{\ell}{\sqrt{e^2-1}}$. This hyperbolic orbit represents an incoming body with large velocity, which is deflected by the gravitational force. The asymptotes are

$$y = \pm\frac{b}{a}(x - ea)$$

In other words,

$$bx \mp ay = eab$$

The normal vectors to the asymptotes are

$$\hat{\mathbf{n}} = \frac{(b, \pm a)}{\sqrt{a^2 + b^2}}$$

The (asymptotic) perpendicular distance between the incoming particle and the central mass (the origin) is given by

$$\mathbf{r} \cdot \hat{\mathbf{n}} = (x, y) \cdot \frac{(b, \pm a)}{\sqrt{a^2 + b^2}} = \frac{bx \mp ay}{\sqrt{a^2 + b^2}} = \frac{eab}{\sqrt{a^2 + b^2}} = b$$

This is sometimes called the impact parameter, since it is the distance away from impacting the central mass.

- ($e = 1$) This is the form of a parabola. This can be seen as the 'transitional' case between the ellipse and the hyperbola.

$$r = \frac{\ell}{1 + \cos\theta}$$

Hence, $r \to \infty$ as $\theta \to \pm\pi$. In Cartesian coordinates,

$$y^2 = \ell(\ell - 2x)$$

The other variables have very useful geometric interpretations; review diagrams of conic sections for more information.

7.6. Energy and eccentricity

Recall that

$$E = \frac{1}{2}m\left(\dot{r}^2 + r^2\dot{\theta}^2\right) - \frac{mk}{r}$$

We can rewrite this in terms of u, using $\dot{r} = -h\frac{\mathrm{d}u}{\mathrm{d}\theta}$:

$$\begin{aligned} E &= \frac{1}{2}mh^2\left(\left(\frac{\mathrm{d}u}{\mathrm{d}\theta}\right)^2 + u^2\right) - mku \\ &= \frac{1}{2}mh^2\left(e^2\sin^2\theta + (1 + e\cos\theta)^2\right)\frac{1}{e^2} - \frac{mk}{\ell}(1 + \cos\theta) \\ &= \frac{mk}{2\ell}(e^2 - 1) \end{aligned}$$

So the energy is positive (unbounded orbits) if $|e| > 1$, and negative (bounded orbits) if $|e| < 1$. The marginal case is at $e = 1$, and $E = 0$.

7.7. Kepler's laws of planetary motion

Kepler's Laws state:

(i) The orbit of a planet is an ellipse, with the sun at one focus.

(ii) The line between a planet and the sun sweeps out equal area in equal times.

(iii) The period P and the semi-major axis a are related: $P^2 \propto a^3$.

Note that the first law is consistent with our solution of the orbit equation for bound orbits. The second law can be rewritten as approximating the sector area with $\frac{1}{2}r^2\delta\theta$, giving a rate of change with respect to time of $\frac{1}{2}r^2\dot{\theta}$, which is half of the angular momentum h. So this law can be seen as stating that the angular momentum is constant. Using dimensional analysis, we can get close to the third law, but we need to be a little more precise to verify it completely. The area of the ellipse is $\pi ab = \frac{h}{2}P$ since in one period the line sweeps out the entire area of the ellipse. We can then derive that $P^2 = \frac{4\pi^2}{k}a^3$. Note that two ellipses with equal semi-major but differing semi-minor axes have the same period.

7.8. Rutherford scattering

Consider a positive charge fired towards another, fixed, positive charge. The particle will be deflected by the electrostatic force between the two particles. What is the angle β by which the particle is deflected? This is motion under a repulsive inverse square law force.

$$V(r) = \frac{mk}{r}; \quad F(r) = \frac{mk}{r^2}$$

We have already solved this problem for an attractive inverse square law force; this was the orbit equation. We can replace k with $-k$ to model a repulsive force.

$$u = \frac{-k}{h^2} + A\cos(\theta - \theta_0); \quad \theta_0 = 0, A \geq 0$$

We can rewrite this as

$$r = \frac{\ell}{e\cos\theta - 1}; \quad \ell = \frac{h^2}{k}, e = \frac{Ah^2}{k}$$

Since we want $r > 0$, we need $e > 1$ such that for some θ, $r > 0$. Then, $r \to \infty$ as $\theta \to \pm\alpha$, with $\arccos(e^{-1}) \in \left(0, \frac{\pi}{2}\right)$. This gives a hyperbolic orbit. We find

$$\frac{(x - ea)^2}{a^2} - \frac{y^2}{b^2} = 1; \quad a = \frac{\ell}{e^2 - 1}, b = \frac{\ell}{\sqrt{e^2 - 1}}$$

h is given by $|\mathbf{r} \times \dot{\mathbf{r}}|$. b, the impact parameter, is the asymptotic distance of the moving particle from impacting the fixed particle. On the incoming asymptote, $\dot{\mathbf{r}} \approx v\mathbf{e}_\parallel$, and $\mathbf{r} \approx b\mathbf{e}_\perp + z\mathbf{e}_\parallel$ for some z. Hence, $h = bv$. Since $\tan\alpha = \sqrt{e^2 - 1}$, we have

$$b = \frac{\ell}{\tan\alpha} = \frac{\ell}{\tan\left(\frac{\pi}{2} - \beta\right)} = \frac{h^2}{k}\tan\left(\frac{\beta}{2}\right) = \frac{v^2b^2}{k}\tan\left(\frac{\beta}{2}\right)$$

Hence

$$\beta = 2 \arctan\left(\frac{k}{bv^2}\right)$$

8. Rotating frames

8.1. Introduction

Newton's laws of motion are only valid in inertial frames of reference. Hence, the laws of dynamics are different from the perspective of a rotating, or non-inertial, frame of reference. Let S be an inertial frame, and let S' be a non-inertial frame, rotating around the z-axis in S with angular velocity $\omega = \dot{\theta}$ where θ is the angle between the x or y axis in S or S'. We will denote the basis vectors $\mathbf{e}_i = \{\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}\}$ for S and $\mathbf{e}'_i = \{\hat{\mathbf{x}}', \hat{\mathbf{y}}', \hat{\mathbf{z}}'\}$ for S'. Consider a particle at rest in S', viewed in S, with position vector $\mathbf{r}$.

$$\left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right)_S = \boldsymbol{\omega} \times \mathbf{r}; \quad \boldsymbol{\omega} = \omega\hat{\mathbf{z}}$$

This angular velocity vector is aligned with the axis of rotation. The convention is that viewed from the direction of the vector, the rotation is anticlockwise. The same formula applies to any vector which is fixed in S', not just the position vector. In particular, this applies to the basis vectors:

$$\left(\frac{\mathrm{d}\mathbf{e}'_i}{\mathrm{d}t}\right)_S = \boldsymbol{\omega} \times \mathbf{e}'_i$$

Here, for instance, $\left(\frac{\mathrm{d}\mathbf{e}'_3}{\mathrm{d}t}\right)_S = 0$. Consider a general time-dependent vector $\mathbf{a}$, defined by the components of the basis vectors in S':

$$\mathbf{a}(t) = \sum_{i=1}^{3} a'_i(t)\mathbf{e}'_i(t)$$

Then we can deduce the key identity:

$$\left(\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{a}(t)\right)_{S'} = \sum_{i=1}^{3}\left(\frac{\mathrm{d}}{\mathrm{d}t}a'_i(t)\right)\mathbf{e}'_i(t)$$

$$\begin{aligned}\left(\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{a}(t)\right)_S &= \sum_{i=1}^{3}\left(\frac{\mathrm{d}}{\mathrm{d}t}a'_i(t)\right)\mathbf{e}'_i(t) + \sum_{i=1}^{3} a'_i(t)\left(\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}'_i(t)\right) \\ &= \sum_{i=1}^{3}\left(\frac{\mathrm{d}}{\mathrm{d}t}a'_i(t)\right)\mathbf{e}'_i(t) + \sum_{i=1}^{3} a'_i(t)\left(\boldsymbol{\omega} \times \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{e}'_i(t)\right) \\ &= \left(\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{a}(t)\right)_{S'} + \boldsymbol{\omega} \times \mathbf{a}\end{aligned}$$

We can apply this identity to the position vector $\mathbf{r}$, and the velocity $\dot{\mathbf{r}}$.

$$\left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right)_S = \left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right)_{S'} + \boldsymbol{\omega} \times \mathbf{r}$$

For the purposes of this derivation, we will allow $\boldsymbol{\omega}$ to depend on time.

$$\begin{aligned}\left(\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2}\right)_S &= \left\{\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)_{S'} + \boldsymbol{\omega}\times\right\}\left\{\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)_{S'} + \boldsymbol{\omega}\times\right\}\mathbf{r} \\ &= \left(\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2}\right)_{S'} + 2\boldsymbol{\omega}\times\left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right)_{S'} + \dot{\boldsymbol{\omega}}\times\mathbf{r} + \boldsymbol{\omega}\times(\boldsymbol{\omega}\times\mathbf{r})\end{aligned}$$

Now, let us write down Newton's equation of motion in a rotating frame.

$$m\left(\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2}\right)_S = \mathbf{F}$$

$$m\left(\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2}\right)_{S'} + 2m\boldsymbol{\omega}\times\left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right)_{S'} + m\dot{\boldsymbol{\omega}}\times\mathbf{r} + m\boldsymbol{\omega}\times(\boldsymbol{\omega}\times\mathbf{r}) = \mathbf{F}$$

Note that we do not need to distinguish between $\left(\frac{\mathrm{d}\boldsymbol{\omega}}{\mathrm{d}t}\right)_S$ and $\left(\frac{\mathrm{d}\boldsymbol{\omega}}{\mathrm{d}t}\right)_{S'}$, since any difference vanishes under the cross product with $\boldsymbol{\omega}$. These extra terms apart from $m\left(\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2}\right)_{S'}$ can be referred to as 'fictitious' forces, since they only appear to be there as perceived by an observer in a rotating (or more general non-inertial) frame. According to this rotating observer, these fictitious forces act in the negative direction:

$$m\left(\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2}\right)_{S'} = \mathbf{F} - 2m\boldsymbol{\omega}\times\left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right)_{S'} - m\dot{\boldsymbol{\omega}}\times\mathbf{r} - m\boldsymbol{\omega}\times(\boldsymbol{\omega}\times\mathbf{r})$$

We can give each fictitious force a name:

- $-2m\boldsymbol{\omega}\times\left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right)_{S'}$ is the Coriolis force;
- $-m\dot{\boldsymbol{\omega}}\times\mathbf{r}$ is the Euler force;
- $-m\boldsymbol{\omega}\times(\boldsymbol{\omega}\times\mathbf{r})$ is the centrifugal force.

In many applications, we take the Euler force to be zero, since this only is relevant when the angular velocity is changing.

8.2. The centrifugal force

$$\begin{aligned}-m\boldsymbol{\omega}\times(\boldsymbol{\omega}\times\mathbf{r}) &= -m((\boldsymbol{\omega}\cdot\mathbf{r})\boldsymbol{\omega} - \omega^2\mathbf{r}) \\ &= m\omega^2(\mathbf{r} - \hat{\boldsymbol{\omega}}(\hat{\boldsymbol{\omega}}\cdot\mathbf{r})) \\ &= m\omega^2\mathbf{r}_\perp\end{aligned}$$

where $\mathbf{r}_\perp$ is the component of $\mathbf{r}$ which is perpendicular to $\boldsymbol{\omega}$. Note that $|\mathbf{r}_\perp|$ is the perpendicular distance from the point $\mathbf{r}$ to the axis of rotation, and $\mathbf{r}_\perp$ is directed away from this axis.

Hence the centrifugal force is always directed away from the rotation axis, perpendicular to it, with its magnitude proportional to the particle's distance from the axis. Note that

$$\mathbf{r}_\perp^2 = \mathbf{r}^2 - (\mathbf{r}\cdot\hat{\boldsymbol{\omega}})^2 = |\hat{\boldsymbol{\omega}}\times\mathbf{r}|^2$$

And

$$\nabla\mathbf{r}_\perp^2 = 2\mathbf{r} - 2\hat{\boldsymbol{\omega}}(\hat{\boldsymbol{\omega}}\cdot\mathbf{r}) = 2\mathbf{r}_\perp$$

Hence,

$$m\omega^2\mathbf{r}_\perp = \nabla\left(\frac{1}{2}m\omega^2\mathbf{r}_\perp^2\right)$$

Therefore the centrifugal force is conservative. On a rotating planet such as the earth, it is often convenient to combine the centrifugal force and the gravitational force into an 'effective gravity'

$$\mathbf{g}_{\text{eff}} = \mathbf{g} + \omega^2\mathbf{r}_\perp$$

As an example, consider a spherical planet which rotates through an axis through a pole. Point P is at latitude λ, i.e. it is λ radians above the equator. On this point, we have $\hat{\mathbf{z}}$ normal to the Earth's surface, $\hat{\mathbf{y}}$ in the north direction parallel to the surface and $\hat{\mathbf{x}}$ in the east direction parallel to the surface. The earth has radius R. Now,

$$\mathbf{r} = R\hat{\mathbf{z}};\quad \boldsymbol{\omega} = \omega(\hat{\mathbf{y}}\cos\lambda + \hat{\mathbf{z}}\sin\lambda)$$

Hence,

$$\begin{aligned}\mathbf{g}_{\text{eff}} &= -g\hat{\mathbf{z}} + \omega^2\mathbf{r}_\perp\\ &= -g\hat{\mathbf{z}} + \omega^2 R\cos\lambda(\hat{\mathbf{z}}\cos\lambda - \hat{\mathbf{y}}\sin\lambda)\\ &= \hat{\mathbf{z}}(\omega^2 R\cos^2\lambda - g) - \hat{\mathbf{y}}(\omega^2 R\cos\lambda\sin\lambda)\end{aligned}$$

The angle α between $\mathbf{g}_{\text{eff}}$ and $\hat{\mathbf{z}}$ is

$$\alpha = \arctan\frac{\omega^2 R\cos\lambda\sin\lambda}{g - \omega^2 R\cos^2\lambda}$$

For earth, $\omega = \frac{2\pi}{86400} \approx 7.3\times10^{-5}\,\mathrm{s}^{-1}$ and $R \approx 6.4\times10^6\,\mathrm{m}$, hence $\frac{\omega^2 R}{g} \approx 3.5\times10^{-3}$. Neglecting the ω^2 term in the denominator, α is very small for the earth.

8.3. The Coriolis force

$$-2m\boldsymbol{\omega}\times\left(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\right)_{S'} = -2m\boldsymbol{\omega}\times\mathbf{v}$$

where $\mathbf{v}$ is as observed in the rotating frame. The force is proportional to, and perpendicular to, the velocity. Consequently, this force does not do any work. Considering the previous

example of the earth, let us consider a velocity tangential to the surface of the planet, specifically $\mathbf{v} = v_x\hat{\mathbf{x}} + v_y\hat{\mathbf{y}}$. The angular velocity has components $\boldsymbol{\omega} = \omega(\hat{\mathbf{y}}\cos\lambda + \hat{\mathbf{z}}\sin\lambda)$. Hence,

$$-2m\boldsymbol{\omega}\times\mathbf{v} = \underbrace{2m\omega\sin\lambda(v_y\hat{\mathbf{x}} - v_x\hat{\mathbf{y}})}_{\text{horizontal}} + \underbrace{2m\omega\cos\lambda(v_x\hat{\mathbf{z}})}_{\text{vertical}}$$

The horizontal component of the Coriolis force gives an acceleration to the right of the horizontal velocity in the Northern hemisphere, and the acceleration is to the left in the southern hemisphere. This appears due to the $\sin\lambda$ term, where the sign changes depending on the hemisphere.

This force can be balanced by another force, notably a pressure gradient, which can be useful for predicting weather patterns in meteorology. Hence, in the northern hemisphere, an area of low pressure implies an anticlockwise flow of fluid around it; in the southern hemisphere this would imply a clockwise flow of fluid. This is called a cyclone. A high-pressure environment (in either hemisphere), would have the opposite direction of flow, and can be called an anticyclone.

8.4. Dropping a particle in a rotating frame

As an example, let us consider dropping a ball from the top of a tower. Where does it land, if we are in a rotating frame?

$$\ddot{\mathbf{r}} = \mathbf{g} - 2\boldsymbol{\omega}\times\dot{\mathbf{r}} - \boldsymbol{\omega}\times(\boldsymbol{\omega}\times\mathbf{r})$$

We will assume the rotation is slow, i.e. $\omega^2 R/g$ is small (we can accurately say 'small' in this case since $\omega^2 R/g$ is a dimensionless constant).

$$\begin{aligned}\ddot{\mathbf{r}} &= \mathbf{g} - 2\omega\times\dot{\mathbf{r}} + o(\omega^2)\\ \dot{\mathbf{r}} &= \mathbf{g}t - 2\omega\times\mathbf{r} + o(\omega^2) + \underbrace{2\omega\times\mathbf{r}_0}_{\text{to match the initial condition}}\end{aligned}$$

Hence, neglecting $o(\omega^3)$,

$$\begin{aligned}\ddot{\mathbf{r}} &= \mathbf{g} - 2\omega\times\mathbf{g}t + o(\omega^2)\\ \mathbf{r} &= \frac{1}{2}\mathbf{g}t^2 - \frac{1}{3}\boldsymbol{\omega}\times\mathbf{g}t^3 + \mathbf{r}_0 + o(\omega^2)\end{aligned}$$

Now, consider $\mathbf{g} = (0,0,-g)$ and $\boldsymbol{\omega} = (0,\omega,0)$, corresponding to the equator. Let $\mathbf{r}_0 = (0,0,R+h)$. Hence,

$$\mathbf{r}(t) = \left(0,0,-\frac{1}{2}gt^2\right) + \left(\frac{1}{3}\omega g t^3, 0, 0\right) + (0,0,R+h)$$

The particle hits the ground when $h = \frac{1}{2}gt^2$, hence $t = \sqrt{2h/g}$, and the corresponding horizontal displacement is therefore

$$\Delta x = \frac{1}{3}\omega g\left(\frac{2h}{g}\right)^{\frac{3}{2}}$$

So the particle hits the ground to the east of the tower's base. This is consistent with conservation of angular momentum.

Example (the Foucault pendulum)**.** Consider a pendulum at the north pole. It will swing in the plane fixed in an inertial frame; the earth rotates relative to this frame. From the point of view of an observer on the earth, the plane in which the pendulum moves is rotating to the west.

If we're at the north pole, the plane of rotation is observed to rotate once per day. This can be explained using a fictitious force from the perspective of the rotating frame of reference of the pendulum. At a general latitude λ, the plane of rotation completes a circuit in $\csc \lambda$ days. We can derive this result by considering the dynamics of the pendulum under the Coriolis force.

9. Systems of particles

9.1. Basic setup

Consider a system of N particles of mass m_i with position vectors $\mathbf{r}_i(t)$ and momentum $\mathbf{p}_i(t) = m_i\dot{\mathbf{r}}_i$. Newton's second law applies to the ith particle individually, but not necessarily to the whole group without any further derivation.

$$m_i\ddot{\mathbf{r}}_i = \dot{\mathbf{p}}_i = \mathbf{F}_i$$

We will make a distinction between internal and external forces;

$$\mathbf{F}_i = \mathbf{F}_i^{\text{ext}} + \sum_{j=1}^{N} \mathbf{F}_{ij}$$

where the $\mathbf{F}_i^{\text{ext}}$ is the external force on the ith particle, and the $\mathbf{F}_{ij}$ is the force exerted on the ith particle by the jth particle. Note that $\mathbf{F}_{ii} = 0$ since particles do not affect themselves. Newton's third law gives a further constraint:

$$\mathbf{F}_{ij} = -\mathbf{F}_{ji}$$

9.2. Centre of mass

The total mass of the system, M, is given by

$$M = \sum_{i=1}^{N} m_i$$

The centre of mass, $\mathbf{R}$, is given by

$$\mathbf{R} = \frac{1}{M}\sum_{i=1}^{N} m_i\mathbf{r}_i$$

The total linear momentum, $\mathbf{P}$, is given by

$$\mathbf{P} = \sum_{i=1}^{N} m_i\dot{\mathbf{r}}_i = \sum_{i=1}^{N} \mathbf{p}_i = M\dot{\mathbf{R}}$$

which is the same momentum as a single particle of mass M and position vector $\mathbf{R}$ would have. Then, by Newton's second law, taking into account the fact that the $\mathbf{F}_{ij}$ are antisym-

metric,

$$\begin{aligned}\dot{\mathbf{P}} &= M\ddot{\mathbf{R}} \\ &= \sum_{i=1}^{N} \dot{\mathbf{p}}_i \\ &= \sum_{i=1}^{N} \mathbf{F}_i^{\text{ext}} + \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbf{F}_{ij} \\ &= \sum_{i=1}^{N} \mathbf{F}_i^{\text{ext}} \\ &= \mathbf{F}^{\text{ext}}\end{aligned}$$

So the centre of mass moves as if it were the position of a mass M under the influence of a force $\mathbf{F}^{\text{ext}}$. This extends Newton's second law to a system of particles. If $\mathbf{F}^{\text{ext}} = \mathbf{0}$ then we have conservation of the total momentum $\mathbf{P}$. In this case, there will be an inertial frame tracking the centre of mass at its origin.

9.3. Motion relative to the centre of mass

Let $\mathbf{r}_i = \mathbf{R} + \mathbf{s}_i$, then $\mathbf{s}_i$ is the position vector of the ith particle relative to the centre of mass. Then

$$\sum_{i=1}^{N} m_i \mathbf{s}_i = \sum_{i=1}^{N} m_i(\mathbf{r}_i - \mathbf{R}) = \sum_{i=1}^{N} m_i \mathbf{r}_i - \sum_{i=1}^{N} m_i \mathbf{R} = \mathbf{0}$$

Further,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\sum_{i=1}^{N} m_i \mathbf{s}_i\right) = \mathbf{0}$$

The total linear momentum is

$$\mathbf{P} = \sum_{i=1}^{N} m_i(\dot{\mathbf{R}} + \dot{\mathbf{s}}_i) = \sum_{i=1}^{N} m_i \dot{\mathbf{R}} = M\dot{\mathbf{R}}$$

as expected.

9.4. Angular momentum

The total angular momentum $\mathbf{L}$ is defined as

$$\mathbf{L} = \sum_{i=1}^{N} \mathbf{r}_i \times \mathbf{p}_i$$

Then

$$\begin{aligned}\dot{\mathbf{L}} &= \sum_{i=1}^{N} \dot{\mathbf{r}}_i \times \mathbf{p}_i + \sum_{i=1}^{N} \mathbf{r}_i \times \dot{\mathbf{p}}_i \\ &= \sum_{i=1}^{N} \mathbf{r}_i \times \dot{\mathbf{p}}_i \\ &= \sum_{i=1}^{N} \mathbf{r}_i \times \left(\mathbf{F}_i^{\text{ext}} + \sum_{j=1}^{N} \mathbf{F}_{ij} \right) \\ &= \sum_{i=1}^{N} \mathbf{r}_i \times \mathbf{F}_i^{\text{ext}} + \sum_{i=1}^{N} \mathbf{r}_i \times \sum_{j=1}^{N} \mathbf{F}_{ij}\end{aligned}$$

The latter term is not necessarily zero, but for example if $\mathbf{F}_{ij} \parallel (\mathbf{r}_i - \mathbf{r}_j)$ then it is zero. If $\mathbf{F}_{ij} \parallel (\mathbf{r}_i - \mathbf{r}_j)$ then

$$\dot{\mathbf{L}} = \sum_{i=1}^{N} \mathbf{r}_i \times \mathbf{F}_i^{\text{ext}} = \mathbf{G}^{\text{ext}}$$

where $\mathbf{G}^{\text{ext}}$ is the total external torque on the system. Relative to the centre of mass, we can write instead

$$\begin{aligned}\mathbf{L} &= \sum_{i=1}^{N} m_i(\mathbf{R} + \mathbf{s}_i) \times (\dot{\mathbf{R}} + \dot{\mathbf{s}}_i) \\ &= \sum_{i=1}^{N} m_i \mathbf{R} \times \dot{\mathbf{R}} + \underbrace{\sum_{i=1}^{N} m_i \mathbf{R} \times \dot{\mathbf{s}}_i}_{=0} + \underbrace{\sum_{i=1}^{N} m_i \mathbf{s}_i \times \dot{\mathbf{R}}}_{=0} + \sum_{i=1}^{N} m_i \mathbf{s}_i \times \dot{\mathbf{s}}_i \\ &= \sum_{i=1}^{N} m_i \mathbf{R} \times \dot{\mathbf{R}} + \sum_{i=1}^{N} m_i \mathbf{s}_i \times \dot{\mathbf{s}}_i\end{aligned}$$

So the total angular momentum is essentially the sum of the angular momentum of a particle of mass M at $\mathbf{R}$ moving with velocity $\dot{\mathbf{R}}$, and the angular momentum associated with the particles relative to the centre of mass.

9.5. Energy

The total kinetic energy T is given by

$$\begin{aligned} T &= \sum_{i=1}^{N} \frac{1}{2} m_i \dot{\mathbf{r}}_i^2 \\ &= \sum_{i=1}^{N} \frac{1}{2} m_i (\dot{\mathbf{R}} + \dot{\mathbf{s}}_i)^2 \\ &= \frac{1}{2} \dot{\mathbf{R}}^2 \sum_{i=1}^{N} m_i + \underbrace{\sum_{i=1}^{N} m_i \dot{\mathbf{R}} \cdot \dot{\mathbf{s}}_i}_{=0} + \sum_{i=1}^{N} \frac{1}{2} m_i \dot{\mathbf{s}}_i^2 \\ &= \frac{1}{2} M \dot{\mathbf{R}}^2 + \sum_{i=1}^{N} \frac{1}{2} m_i \dot{\mathbf{s}}_i^2 \end{aligned}$$

The total kinetic energy is the sum of the kinetic energy of a particle of mass M at $\mathbf{R}$ moving with velocity $\dot{\mathbf{R}}$, and the kinetic energy associated with the particles relative to the centre of mass. Let us consider the rate of change of kinetic energy:

$$\begin{aligned} \frac{\mathrm{d}T}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t} \sum_{i=1}^{N} \frac{1}{2} m_i \dot{\mathbf{r}}_i^2 \\ &= \sum_{i=1}^{N} \frac{1}{2} m_i \dot{\mathbf{r}}_i \cdot \ddot{\mathbf{r}}_i \\ &= \sum_{i=1}^{N} \dot{\mathbf{r}}_i \cdot \mathbf{F}^{\text{ext}} + \sum_{i=1}^{N} \dot{\mathbf{r}}_i \cdot \sum_{j=1}^{N} \mathbf{F}_{ij} \\ &= \sum_{i=1}^{N} \dot{\mathbf{r}}_i \cdot \mathbf{F}^{\text{ext}} + \sum_{i=1}^{N} \sum_{j=i+1}^{N} (\dot{\mathbf{r}}_i - \dot{\mathbf{r}}_j) \cdot \mathbf{F}_{ij} \end{aligned}$$

If the external forces are defined by a potential

$$\mathbf{F}_i^{\text{ext}} = -\nabla_{\mathbf{r}_i} V_i^{\text{ext}}$$

and the internal forces are defined by a potential

$$\mathbf{F}_{ij} = -\nabla_{\mathbf{r}_i} V(\mathbf{r}_i - \mathbf{r}_j)$$

then

$$\frac{\mathrm{d}T}{\mathrm{d}t} = -\frac{\mathrm{d}}{\mathrm{d}t} \sum_{i=1}^{N} V_i^{\text{ext}} - \frac{\mathrm{d}}{\mathrm{d}t} \sum_{i=1}^{N} \sum_{j=1+1}^{N} V(\mathbf{r}_i - \mathbf{r}_j)$$

Hence we have conservation of energy if the given properties are true.

10. Applications of orbits

10.1. Two body problem

Consider two bodies of mass m_1, m_2 experiencing gravitational attraction to the other, with no external forces. Let m_1 be at position $\mathbf{r}_1$, and m_2 at $\mathbf{r}_2$, with the centre of mass at $\mathbf{R}$, and total mass $M = m_1 + m_2$. Then certainly,

$$\mathbf{R} = \frac{1}{M}(m_1\mathbf{r}_1 + m_2\mathbf{r}_2)$$

We will define the separation vector $\mathbf{r} = \mathbf{r}_1 - \mathbf{r}_2$. We can then further say that

$$\mathbf{r}_1 = \mathbf{R} + \frac{m_2}{M}\mathbf{r}; \quad \mathbf{r}_2 = \mathbf{R} - \frac{m_1}{M}\mathbf{r}$$

Since $\mathbf{F}^{\text{ext}} = \mathbf{0}$, the centre of mass $\mathbf{R}$ does not accelerate; it moves with constant velocity. Now, let us consider $\mathbf{r}$.

$$\ddot{\mathbf{r}} = \ddot{\mathbf{r}}_1 + \ddot{\mathbf{r}}_2 = \frac{\mathbf{F}_{12}}{m_1} - \frac{\mathbf{F}_{21}}{m_2} = \mathbf{F}_{12}\left(\frac{1}{m_1} + \frac{1}{m_2}\right)$$

Equivalently, we can write

$$\mu\ddot{\mathbf{r}} = \mathbf{F}_{12}; \quad \mu = \frac{m_1 m_2}{m_1 + m_2}$$

Notice that μ has the dimension of mass; we call it the 'reduced' mass since it is less than m_1 and m_2. This can be seen as the equation of motion of a particle of mass μ under the effect of force $\mathbf{F}_{12}$. In the case of a gravitational force, we have

$$\mu\ddot{\mathbf{r}} = \frac{-Gm_1m_2}{|\mathbf{r}|^3}\mathbf{r}$$

Hence,

$$\ddot{\mathbf{r}} = \frac{-G(m_1 + m_2)}{|\mathbf{r}|^3}\mathbf{r}$$

This is the motion of a particle under the effect of a gravitational force due to a mass $m_1 + m_2$ fixed at the origin. The total kinetic energy T is

$$T = \frac{1}{2}M\dot{\mathbf{R}}^2 + \frac{1}{2}\mu\dot{\mathbf{r}}^2$$

The total angular momentum $\mathbf{L}$ is

$$\mathbf{L} = M\mathbf{R} \times \dot{\mathbf{R}} + \mu\mathbf{r} \times \dot{\mathbf{r}}$$

Example. Let us consider the orbit of the earth and the sun. Both particles move around the centre of mass, and both orbits have the same shape. However, the sizes of the orbits are very different. The ratio of masses is around 3×10^{-4}, and the radius of orbit is approximately 1.5×10^7 km. Hence the displacement of the sun is around 450 km.

10.2. Variable mass problems and the rocket problem

Consider a rocket which ejects mass (exhaust gases) at a high speed in order to propel itself forward. We cannot apply Newton's second law to the rocket alone, since in this system mass is not conserved. Consider the rocket moving in one dimension, with speed $v(t)$ and mass $m(t)$. The mass is being expelled at velocity u relative to the rocket. At time t, the rocket has momentum $v(t)m(t)$. At time $t + \delta t$, the momentum is $v(t + \delta t)m(t + \delta t)$. The exhaust gases emitted during δt have velocity $v(t) - u + O(\delta t)$ and mass $m(t) - m(t + \delta t)$. The total momentum at $t + \delta t$ is

$$v(t + \delta t)m(t + \delta t) + (v(t) - u + O(\delta t))(m(t) - m(t + \delta t))$$

So the change in momentum is

$$\begin{aligned}\delta p &= v(t + \delta t)m(t + \delta t) + (v(t) - u + O(\delta t))(m(t) - m(t + \delta t)) - v(t)m(t) \\ &= \left(\frac{\mathrm{d}m}{\mathrm{d}t}u + m\frac{\mathrm{d}v}{\mathrm{d}t}\right)\delta t + O(\delta t^2)\end{aligned}$$

But since momentum is conserved,

$$\frac{\mathrm{d}m}{\mathrm{d}t}u + m\frac{\mathrm{d}v}{\mathrm{d}t} = 0$$

This is called the rocket equation. We can generalise this to $\frac{\mathrm{d}m}{\mathrm{d}t}u + m\frac{\mathrm{d}v}{\mathrm{d}t} = \mathbf{F}^{\text{ext}}$ in the presence of external forces. In the absence of such external forces,

$$\begin{aligned}\frac{\mathrm{d}m}{\mathrm{d}t}u &= -m\frac{\mathrm{d}v}{\mathrm{d}t} \\ \implies v(t) &= v(0) + u\log\left(\frac{m(0)}{m(t)}\right)\end{aligned}$$

11. Rigid bodies

11.1. Definition

A rigid body is an extended object of finite size that can be considered as a multi-particle system such that the distance between any two particles in the body remains constant, i.e.

$$|\mathbf{r}_i - \mathbf{r}_j| = \text{constant}$$

The possible motion of a rigid body is therefore constrained to some combination of the two basic isometries of Euclidean space, rotations and translations. We exclude reflections from this, since this would alter the 'ordering' of the points in some sense.

11.2. Recap of angular velocity

Consider a particle rotating about an axis through the origin with angular velocity $\boldsymbol{\omega}$. Let $r_\perp$ be the perpendicular distance from $\mathbf{r}$ to the axis of rotation. Then

$$\dot{\mathbf{r}} = \boldsymbol{\omega} \times \mathbf{r}; \quad |\dot{\mathbf{r}}| = \omega r_\perp$$

If the particle has mass m, then the kinetic energy T is given by

$$T = \frac{1}{2} m \dot{\mathbf{r}}^2 = \frac{1}{2} m (\boldsymbol{\omega} \times \mathbf{r}) \cdot (\boldsymbol{\omega} \times \mathbf{r}) = \frac{1}{2} m \omega^2 r_\perp^2$$

Note that if $\boldsymbol{\omega} = \omega \mathbf{n}$, then $r_\perp = |\mathbf{n} \times \mathbf{r}|$. We will define the moment of inertia I to be

$$I = m r_\perp^2 \implies T = \frac{1}{2} I \omega^2$$

11.3. Moment of inertia for a rigid body

Consider a rigid body to be made up of N particles, rotating about an axis through the origin, with angular velocity $\boldsymbol{\omega}$. For each particle in the body,

$$\dot{\mathbf{r}}_i = \boldsymbol{\omega} \times \mathbf{r}_i$$

Note that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} |\mathbf{r}_i - \mathbf{r}_j|^2 &= 2(\mathbf{r}_i - \mathbf{r}_j) \cdot (\dot{\mathbf{r}}_i - \dot{\mathbf{r}}_j) \\ &= 2(\mathbf{r}_i - \mathbf{r}_j) \cdot (\omega \times (\mathbf{r}_i - \mathbf{r}_j)) \\ &= 0 \end{aligned}$$

which is consistent with the expected properties of the rigid body. Consider the kinetic energy of the entire body, which is the sum of the energies of the component particles.

$$\begin{aligned} T &= \sum_{i=1}^{N} \frac{1}{2} m_i \dot{\mathbf{r}}_i^2 \\ &= \sum_{i=1}^{N} \frac{1}{2} m_i |\boldsymbol{\omega} \times \mathbf{r}_i|^2 \\ &= \frac{1}{2}\omega^2 \sum_{i=1}^{N} m_i |\mathbf{n} \times \mathbf{r}_i|^2 \\ &= \frac{1}{2} I \omega^2 \end{aligned}$$

where $I = \sum_{i=1}^{N} m_i |\mathbf{n} \times \mathbf{r}_i|^2$ is the moment of inertia of the body for a rotation of axis $\mathbf{n}$ through the origin. Now, we can consider the angular momentum.

$$\mathbf{L} = \sum_{i=1}^{N} m_i \mathbf{r}_i \times (\boldsymbol{\omega} \times \mathbf{r}_i)$$

In the case that $\boldsymbol{\omega} = \omega \mathbf{n}$, we have

$$\mathbf{L} = \omega \sum_{i=1}^{N} m_i \mathbf{r}_i \times (\mathbf{n} \times \mathbf{r}_i)$$

Now, we will consider just the component of $\mathbf{L}$ that is parallel to the rotation axis.

$$\begin{aligned} \mathbf{L} \cdot \mathbf{n} &= \omega \sum_{i=1}^{N} m_i \mathbf{n} \cdot (\mathbf{r}_i \times (\mathbf{n} \times \mathbf{r}_i)) \\ &= \omega \sum_{i=1}^{N} m_i |\mathbf{n} \times \mathbf{r}_i|^2 \\ &= \omega \sum_{i=1}^{N} m_i \dot{r}_{i\perp}^2 \\ &= I\omega \end{aligned}$$

Therefore the component of the angular momentum in the direction of the rotation axis is $I\omega$. However, it is not the case that $\mathbf{L}$ *only* has a component in the direction of the rotation axis; indeed it is possible that it may have more components in other directions. We can derive that

$$\begin{aligned} \mathbf{L} &= \omega \sum_{i=1}^{N} m_i \mathbf{r}_i \times (\mathbf{n} \times \mathbf{r}_i) \\ &= \sum_{i=1}^{N} m_i (|\mathbf{r}_i|^2 \boldsymbol{\omega} - (\mathbf{r}_i \cdot \boldsymbol{\omega}) \mathbf{r}_i) \end{aligned}$$

which is a linear function of the vector $\boldsymbol{\omega}$. For instance, in terms of suffix notation (which is not examinable),

$$\mathbf{L}_\alpha = I_{\alpha\beta}\omega_\beta$$

for some symmetric tensor I (symmetric since $I_{\alpha\beta} = I_{\beta\alpha}$). In fact, we can deduce

$$I_{\alpha\beta} = \sum_{i=1}^{N} m_i \left\{ |\mathbf{r}_i|^2 \delta_{\alpha\beta} - (\mathbf{r}_i)_\alpha (\mathbf{r}_i)_\beta \right\}$$

In general therefore, there are three principal axes; three linearly independent directions $\boldsymbol{\omega}$ such that $I \cdot \boldsymbol{\omega}$ is parallel to $\boldsymbol{\omega}$. If a body is rotated about one of these principal axes, the angular momentum $\mathbf{L}$ will be parallel to $\boldsymbol{\omega}$. This holds for any shape of body, since it is simply a property of matrices. To recap, if we choose to rotate in a direction such that $\mathbf{L}$ is parallel to $\boldsymbol{\omega}$, then

$$\mathbf{L} = I(\mathbf{n})\boldsymbol{\omega}$$

where $I(\mathbf{n})$ is the moment of inertia about this axis $\mathbf{n}$. Note that since we often consider bodies which are symmetric about a particular axis, rotating about this axis guarantees this above property. Further note the similarities between the equations for angular and linear velocities and energies:

$$T = \frac{1}{2} I\omega^2, \mathbf{L} = I\boldsymbol{\omega}; \quad T = \frac{1}{2} m v^2, \mathbf{p} = m\mathbf{v}$$

11.4. Calculating moments of inertia

For a solid body, instead of considering finite sums of particles we instead consider integrals. Consider a body occupying a volume V, with mass density $\rho(\mathbf{r})$. Then we can compute the total mass m by

$$M = \int_V \rho \, \mathrm{d}V$$

The centre of mass is given by

$$\mathbf{R} = \frac{1}{M} \int_V \rho \mathbf{r} \, \mathrm{d}V$$

The moment of inertia about an axis $\mathbf{n}$ is

$$I = \int_V \rho |\mathbf{r}_\perp|^2 \, \mathrm{d}V = \int_V \rho |\mathbf{n} \times \mathbf{r}|^2 \, \mathrm{d}V$$

We can alternatively formulate these volume integrals as surface or line integrals in order to compute these quantities for mass distributed on a sheet or along a curve. We can explicitly calculate these values for simple shapes.

(i) Consider a uniform thin ring of total mass M and radius a. Let ρ be the mass per unit length, which is therefore $M/2\pi a$. The moment of inertia about an axis through the centre of the ring and perpendicular to the plane of the ring is given by

$$I = \int_0^{2\pi} \frac{M}{2\pi a} a^2 a \, \mathrm{d}\theta = a^2 M$$

This is easy to compute since every point in the body has $r_\perp = a$.

(ii) Consider a uniform thin rod of total mass M and length ℓ. The axis of rotation is at one end of the rod, and the rod is rotating about an axis perpendicular to its length. Here, $\rho = M/\ell$.

$$I = \int_0^{\ell} \frac{M}{\ell} x^2 \, \mathrm{d}x = \frac{1}{3} M \ell^2$$

(iii) Consider a uniform thin disc of mass M, radius a with the axis of rotation through the centre of the disc, perpendicular to the plane of the disc. We will use an area integral, and let $\rho = \frac{M}{\pi a^2}$ be the mass per unit area. In plane polar coordinates,

$$I = \int_{r=0}^{a} \mathrm{d}r \int_{\theta=0}^{2\pi} \mathrm{d}\theta \, \frac{M}{\pi a^2} r^2 r = \frac{1}{2} M a^2$$

(iv) Consider the same disc, but with the axis of rotation through the centre, in the plane of the disc. Again in plane polar coordinates, we can let θ be the angle between the axis of rotation and the line through the point and the centre of mass. Therefore $r_\perp = r \sin\theta$. Hence,

$$I = \int_{r=0}^{a} \mathrm{d}r \int_{\theta=0}^{2\pi} \mathrm{d}\theta \, \frac{M}{\pi a^2} r^2 \sin^2\theta r = \frac{1}{4} M a^2$$

(v) Consider a uniform sphere with mass M and radius a, with axis of rotation through the centre of the sphere. Then ρ, the density per unit volume, is $\frac{3M}{4\pi a^3}$. In spherical polar coordinates, we can let the $\theta = 0$ axis be the axis of rotation. Then

$$I = \int_{r=0}^{a} \mathrm{d}r \int_{\theta=0}^{\pi} \mathrm{d}\theta \int_{\phi=0}^{2\pi} \mathrm{d}\phi \, \frac{3M}{4\pi a^3} r^2 \sin^2\theta r^2 \sin\theta = \frac{2}{5} M a^2$$

11.5. Results on moments of inertia

Theorem (Perpendicular Axes Theorem)**.** For a two-dimensional body (a lamina),

$$I_z = I_x + I_y$$

where I_z is the moment of inertia about the axis perpendicular to the lamina, and the I_x and I_y are the moments of inertia in perpendicular directions in the plane of the lamina.

Proof. Let A be the lamina as shown. Then

$$I_x = \int_A \rho y^2 \, \mathrm{d}A \, ; \quad I_y = \int_A \rho x^2 \, \mathrm{d}A$$

where x, y are the plane Cartesian components of the position vector of a point. Then

$$I_z = \int_A \rho(x^2 + y^2) \, \mathrm{d}A = I_x + I_y$$

as required. □

This theorem is useful when there is a level of symmetry in the problem where $I_x = I_y$.

Theorem (Parallel Axes Theorem)**.** Consider a rigid body of mass M with a moment of inertia I_c about some axis through the centre of mass. Then the moment of inertia about a parallel axis a distance d from the centre of mass has moment of inertia

$$I = I_c + Md^2$$

Proof. Let us consider Cartesian coordinates, with the origin at the centre of mass. The moment of inertia about an axis in the z direction through the origin is I_c, and the moment about the axis passing through the point $(d, 0, 0)$ is I. Let us denote the volume of the body as V. Then

$$I_c = \int_V \rho(x^2 + y^2) \, \mathrm{d}V \, ; \quad I = \int_V \rho((x - d)^2 + y^2) \, \mathrm{d}V$$

Hence,

$$I = \int_V \rho(x^2 + y^2) \, \mathrm{d}V - 2 \underbrace{\int_V \rho x d \, \mathrm{d}V}_{=0} + \int_V \rho d^2 \, \mathrm{d}V$$

The middle term is zero since the origin is the centre of mass, and we are integrating over the x coordinate multiplied by a constant multiple of density.

$$I = I_c + Md^2$$

as required. □

11.6. General motion of a rigid body

In general, the motion of a rigid body can be described by a combination of

- the translation of the centre of mass, following a trajectory $\mathbf{R}(t)$, and
- the rotation about an axis through the centre of mass.

Like before, we define the position vector of a point i in the body as $\mathbf{r}_i = \mathbf{R} + \mathbf{s}_i$ where the $\mathbf{s}_i$ are relative to the centre of mass. Recall that $\sum_{i=1}^{N} \mathbf{s}_i = \mathbf{0}$. If a body is rotating about the centre of mass with angular velocity $\boldsymbol{\omega}$, then

$$\dot{\mathbf{s}}_i = \boldsymbol{\omega} \times \mathbf{s}_i; \quad \dot{\mathbf{r}}_i = \dot{\mathbf{R}} + \boldsymbol{\omega} \times \dot{\mathbf{s}}_i$$

Recall that the kinetic energy is

$$T = \frac{1}{2}M\dot{\mathbf{R}}^2 + \frac{1}{2}\sum_{i=1}^{N} m_i \dot{\mathbf{s}}_i^2 = \frac{1}{2}M\dot{\mathbf{R}}^2 + \frac{1}{2}I_c\omega^2$$

where I_c is the moment of inertia about the axis of rotation $\mathbf{n} = \boldsymbol{\omega}/\omega$ through the centre of mass. We can therefore consider T as the sum of a 'translational' kinetic energy and a 'rotational' kinetic energy. Recall that in a general multiparticle system, linear momentum and angular momentum satisfy

$$\dot{\mathbf{p}} = \mathbf{F}; \quad \dot{\mathbf{L}} = \mathbf{G}$$

where $\mathbf{F}$ is the total external force and $\mathbf{G}$ is the total external torque. For a rigid body, these two equations determine the translational and rotational components of motion entirely. Note that sometimes we can determine the motion in a simpler way by using energy conservation laws. Note further that $\mathbf{L}$ and $\mathbf{G}$ depend on the choice of origin, which could be defined as any point fixed in an inertial frame, or alternatively we could define them with respect to the centre of mass. In this case, the equation $\dot{\mathbf{L}} = \mathbf{G}$ still holds. Indeed,

$$\begin{aligned}
\underbrace{\mathbf{G}}_{\text{external torque about origin}} &= \frac{\mathrm{d}}{\mathrm{d}t}\left(M\mathbf{R} \times \dot{\mathbf{R}} + \sum_{i=1}^{N} m_i \mathbf{s}_i \times \dot{\mathbf{s}}_i\right) \\
&= M\dot{\mathbf{R}} \times \dot{\mathbf{R}} + M\mathbf{R} \times \ddot{\mathbf{R}} + \frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^{N} m_i \mathbf{s}_i \times \dot{\mathbf{s}}_i \\
&= \mathbf{R} \times \mathbf{F}^{\text{ext}} + \frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^{N} m_i \mathbf{s}_i \times \dot{\mathbf{s}}_i
\end{aligned}$$

Hence the rate of change of the angular momentum about the centre of mass $\frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^{N} m_i \mathbf{s}_i \times \dot{\mathbf{s}}_i$ is exactly $\mathbf{G} - \mathbf{R} \times \mathbf{F}^{\text{ext}}$. Therefore,

$$\begin{aligned}
\mathbf{G}_c &= \sum_{i=1}^{N} \mathbf{r}_i \times \mathbf{F}_i^{\text{ext}} - \mathbf{R} \times \mathbf{F}^{\text{ext}} \\
&= \sum_{i=1}^{N} (\mathbf{r}_i - \mathbf{R}) \times \mathbf{F}_i^{\text{ext}}
\end{aligned}$$

Hence the rate of change of the angular momentum about the centre of mass is exactly the external torque about the centre of mass.

Example. Consider the motion of a rigid body in a uniform gravitional field with constant acceleration $\mathbf{g}$. The total gravitational force and torque acting on the rigid body are the same as those that would act on a particle of the same mass located at the rigid body's centre of mass. In a gravitational field, the centre of mass is often referred to as the 'centre of gravity'. Indeed,

$$\begin{aligned}\mathbf{F} &= \sum_{i=1}^{N} \mathbf{F}_i^{\text{ext}} \\ &= \sum_{i=1}^{N} m_i \mathbf{g} \\ &= M\mathbf{g}\end{aligned}$$

Correspondingly, the total torque is given by

$$\begin{aligned}\mathbf{G} &= \sum_{i=1}^{N} \mathbf{G}_i^{\text{ext}} \\ &= \sum_{i=1}^{N} \mathbf{r}_i \times m_i \mathbf{g} \\ &= \sum_{i=1}^{N} m_i \mathbf{r}_i \times \mathbf{g} \\ &= M\mathbf{R} \times \mathbf{g}\end{aligned}$$

Note that the gravitational torque about the centre of mass is exactly zero, since

$$\begin{aligned}\mathbf{G}_c &= \sum_{i=1}^{N} \mathbf{s}_i \times m_i \mathbf{g} \\ &= \sum_{i=1}^{N} m_i \mathbf{s}_i \times \mathbf{g} \\ &= \mathbf{0}\end{aligned}$$

Note further that the external potential V^{ext}, which is exactly the gravitational potential, will be given by

$$\begin{aligned}V^{\text{ext}} &= -\sum_{i=1}^{N} m_i \mathbf{r}_i \cdot \mathbf{g} \\ &= -M\mathbf{R} \cdot \mathbf{g}\end{aligned}$$

Consider a stick thrown into the air. The centre of mass will follow a parabola, and the angular acceleration about the centre of mass is zero.

11.7. Simple pendulum

Consider a uniform rod of length ℓ and mass M, fixed at one end to a pivot point O. The centre of mass is the midpoint of the rod, at a distance of $\ell/2$ from the pivot. The angle between the rod and the rest position (when the rod is pointing downwards from the pivot) is θ. We can consider the angular momentum about the pivot point.

$$\omega = \dot{\theta}; \quad L = I\dot{\theta} = \frac{1}{3}M\ell^2\dot{\theta}$$

The torque produced by the gravitational force is

$$G = -Mg\frac{\ell}{2}\sin\theta$$

The torque associated with the force at the pivot will be zero, since it acts on the line of the rod. We have

$$\dot{L} = G \implies I\ddot{\theta} = -Mg\frac{\ell}{2}\sin\theta \implies \ddot{\theta} = \frac{-3g}{2\ell}\sin\theta$$

which is equivalent to a simple pendulum of length $2\ell/3$, and small oscillations will have period $2\pi\sqrt{2\ell/3g}$. We could alternatively solve this using conservation of energy.

$$T + V = \frac{1}{2}I\dot{\theta}^2 - \frac{Mg\ell}{2}\cos\theta = E$$

where E is constant. Then

$$I\dot{\theta}\ddot{\theta} + \frac{Mg\ell}{2}\dot{\theta}\sin\theta = 0$$

So either $\dot{\theta} = 0$ everywhere, or

$$I\ddot{\theta} + \frac{Mg\ell}{2}\sin\theta = 0$$

which gives the equation of motion we found earlier. In general, when solving a problem, there are three methods:

(i) use Newton's second law for the centre of mass, and use the rate of change of angular momentum about the centre of mass;

(ii) use the rate of change of angular momentum about a fixed point; and

(iii) use conservation of energy (less useful in general, since it removes dimensions).

11.8. Comparison of sliding and rolling

Consider a cylinder or a sphere with radius a, moving along a stationary horizontal surface. The general motion is some combination of the rotation of the centre of mass with angular velocity ω and the translation of the centre of mass with velocity v. The point P is the instantaneous point of contact between the body and the surface. The horizontal velocity of the point of contact is given by

$$v_{\text{slip}} = v - a\omega$$

In general, the point of contact P slips, and there may be some kinetic frictional force associated with this slip. We can categorise rolling and sliding as follows.

- A 'pure sliding' motion is given by $\omega = 0$, and $v = v_{\text{slip}} \neq 0$. In this case, the body slides across the surface without rotation.
- A 'pure rolling' motion is given by $v_{\text{slip}} = 0$, but $v \neq 0$ and $\omega \neq 0$. In this case, the point of contact P is stationary. A rolling body can be described instantaneously as rotating about the point of contact with angular velocity ω.

As an example, consider a body rolling downhill, where the hill has a constant incline α to the horizontal. Let x be the displacement of the centre of mass from its initial position, so $v = \dot{x}$. Let Mg be the gravitational force, N be the normal force, and F be the frictional force. Now, we know that the rolling condition is that $v - a\omega = 0$. We will analyse the motion of this body, under the assumption that it is rolling, by considering conservation of energy.

$$T = \frac{1}{2}Mv^2 + \frac{1}{2}I\omega^2 = \frac{1}{2}\left(M + \frac{I}{a^2}\right)v^2; \quad V = -Mgx\sin\alpha$$

The normal force does not do any work, since it is perpendicular to the direction of motion, and the frictional force does not do work because the point of contact is instantaneously stationary. Hence, energy is conserved, giving

$$\frac{1}{2}\left(M + \frac{I}{a^2}\right)v^2 - Mgx\sin\alpha = E$$

Hence,

$$\left(M + \frac{I}{a^2}\right)\dot{x}\ddot{x} - Mg\dot{x}\sin\alpha = 0$$

We have therefore deduced that

$$\left(M + \frac{I}{a^2}\right)\ddot{x} = Mg\sin\alpha$$

which is a second order differential equation with constant coefficients, which we can solve. Note that due to the $\frac{I}{a^2}$ term, the total acceleration is less than it would be for a frictionless particle (since such a particle would not rotate). For example, a cylinder would have $I = \frac{1}{2}Ma^2$ hence $\ddot{x} = \frac{2}{3}Mg\sin\alpha$. Alternatively, we could analyse the forces and torques. We can use Newton's second law to deduce

$$M\dot{v} = Mg\sin\alpha - F$$

Further, the rate of change of angular momentum about the centre of mass is

$$I\dot{\omega} = aF$$

The rolling condition implies that $\dot{v} = a\dot{\omega}$, hence

$$\frac{I\dot{v}}{a} = aF \implies M\dot{v} = Mg\sin\alpha - \frac{I\dot{v}}{a^2} \implies \left(M + \frac{I}{a^2}\right)\dot{v} = Mg\sin\alpha$$

We could also alternatively look at the torque about the point P. In this case, using the parallel axes theorem,

$$I_P = I + Ma^2$$

Then,

$$I_P\dot{\omega} = Mga\sin\alpha \implies (Ma^2 + I)\frac{\dot{v}}{a} = Mga\sin\alpha$$

and the substitution $v = a\omega$ gives the equation we found before.

11.9. Transition from sliding to rolling

Consider a sphere with radius a that begins by sliding across a horizontal surface, for instance a snooker ball being hit parallel to the table, through the centre of mass, by a cue. Eventually, the ball will transition from sliding to rolling across the table. Initially, $v = v_0$ and $\omega = 0$. The kinetic frictional force F is given by

$$F = \mu_k N = \mu_k Mg$$

Considering linear motion, we have

$$M\dot{v} = -F$$

Considering angular motion,

$$I\dot{\omega} = aF \implies \frac{2}{5}Ma\dot{\omega} = F$$

Hence,

$$M\dot{v} + \frac{2}{5}Ma\dot{\omega} = 0 \implies v = v_0 - \mu_k gt; \omega = \frac{5}{2a}\mu_k gt$$

We can now compute the slip velocity.

$$v_{\text{slip}} = v - a\omega = v_0 - \frac{7}{2}\mu_k gt$$

There is slipping when $v_{\text{slip}} > 0$, which occurs for

$$0 \le t < \frac{2v_0}{7\mu_k g}$$

Rolling begins when $t = \frac{2v_0}{7\mu_k g} = t_{\text{roll}}$. Note that at this time,

$$T = \frac{1}{2}Mv^2 + \frac{1}{2}I\omega^2 = \frac{1}{2}M\left(1 + \frac{2}{5}\right)v_{\text{roll}}^2 = \frac{5}{7}\left(\frac{1}{2}Mv_0^2\right)$$

So during the sliding phase, we have lost $\frac{2}{7}$ of the initial kinetic energy. We can check the loss of kinetic energy due to friction, giving

$$\int_0^{t_{\text{roll}}} Fv_{\text{slip}}\,\mathrm{d}t = \int_0^{t_{\text{roll}}} F\left(v_0 - \frac{7}{2}\mu_k gt\right)\mathrm{d}t = \frac{2}{7}\left(\frac{1}{2}Mv_0^2\right)$$

as expected.

12. Special relativity

12.1. Introduction and postulates

When velocities get comparable to the speed of light $c = 299\,792\,458\,\mathrm{m\,s^{-1}}$, the Newtonian theory of dynamics is no longer a good approximation to real-world dynamical systems. In this case, we need to consider the Special Theory of Relativity in order to get a better understanding of the real world. The theory of special relativity is based on two postulates:

(i) The laws of physics are the same in all inertial frames.

(ii) The speed of light in a vacuum is the same in all inertial frames.

The first postulate is consistent with the Newtonian theory of dynamics. The second postulate arises from the fact that we cannot detect any change in the speed of light in inertial frames moving at different velocities. This second postulate then has major consequences; in fact, we must rewrite our understanding of space and time, as well as the relationships between energy, momentum and mass.

12.2. Lorentz transformations

Consider two inertial frames S and S', which are related by a Galilean transformation given by

$$x' = x - vt; \quad y' = y; \quad z' = z; \quad t' = t$$

Consider the path of a ray of light travelling in the x direction in S. It has position $x = ct$. In S', we have

$$x' = x - vt = ct - vt = (c - v)t'$$

This contradicts the second postulate, since this would imply that the speed of light is not in fact the same in all inertial frames. Therefore, the assumption that there exists a Galilean transformation between the inertial frames was incorrect. In order to rectify this apparent contradiction, we must let space and time interact with each other under this transformation. The particular transformation that satisfies both postulates of special relativity is called the Lorentz transformation; we will now construct such a transformation.

Consider inertial frames S and S' that have the same origin when $t = t' = 0$. S' is moving at a speed v in the x direction relative to S, and we will assume that $y' = y$ and $z' = z$. Postulate 1 implies that a particle moving at a constant velocity in S must appear to be moving at a constant velocity in S'. So the transformation $(x, t) \mapsto (x', t')$ must preserve straight lines, hence it must be a linear transformation. We know that the origin in S' (called O') moves with speed v, hence

$$x' = \gamma(x - vt) \tag{1}$$

where γ is a function of $|v|$, since there is no directional preference in our system of physics. Further, O moves with speed $-v$ in S', so in the same way,

$$x = \gamma(x' + vt') \tag{2}$$

The γ value is the same since $\gamma(v) = \gamma(-v)$. Consider a light ray passing through O and O' at $t = t' = 0$, moving in the x direction.

$$x = ct; \quad x' = ct'$$

We can now substitute these equations into the results we found before.

$$x = ct = \gamma(x' + vt') = \gamma(c + v)t'; \quad x' = ct' = \gamma(x - vt) = \gamma(c - v)t$$

For consistency,

$$\gamma^2\left(1 - \frac{v}{c}\right)\left(1 + \frac{v}{c}\right) = 1$$

Hence,

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$

We call γ the Lorentz factor. Using (1) and (2), we can deduce that

$$\begin{aligned} vt' &= \frac{x}{\gamma} - x' \\ &= \frac{x}{\gamma} - \gamma(x - vt) \\ &= \gamma\left(\frac{1}{\gamma^2} - 1\right)x + \gamma vt \\ &= \gamma\left(vt - \frac{v^2}{c^2}x\right) \end{aligned}$$

Hence,

$$t' = \gamma\left(t - \frac{vx}{c^2}\right)$$

It is easy to deduce the inverse transformation

$$t = \gamma\left(t' + \frac{vx'}{c^2}\right)$$

Note also that y and z are unchanged. Note that $\gamma(v) \geq 1$, and $\gamma \to \infty$ as $|v| \to c$. In particular, the Galilean transformation is recovered when $\gamma \to 1$. Also, as $v \to c$, we have approximately

$$\gamma \approx \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{1 - \frac{v}{c}}}$$

Example. Consider a light ray travelling in the x direction. In S, $x = ct$, $y = 0$, $z = 0$. In S',

$$x' = \gamma(x - vt) = \gamma(c - v)t; \quad t' = \gamma\left(t - \frac{vx}{c^2}\right) = \gamma\left(1 - \frac{v}{c}\right)t$$

and additionally $y = 0, z = 0$. Combined, we find

$$\frac{x'}{t'} = \frac{\gamma(c - v)}{\gamma\left(1 - \frac{v}{c}\right)}$$

Now instead, consider a light ray travelling in the y direction in S. In S, $x = 0, y = ct, z = 0$. In S',

$$x' = \gamma(x - vt); \quad t' = \gamma\left(t - \frac{vx}{c^2}\right)$$

and $y' = y = ct, z' = z = 0$. Since $x = 0$ at all time, we have

$$x' = -\gamma vt; \quad t' = \gamma t$$

The square of the speed of the light ray in S' is given by the x component squared plus the y component squared.

$$c'^2 = v^2 + \frac{c^2}{\gamma^2} = c^2$$

as expected. Note that while the speed of light has remained fixed, the direction has changed.

12.3. General properties of Lorentz transformation

Note that the following always holds.

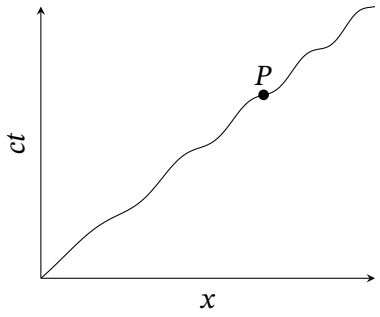
$$\begin{aligned} c^2t'^2 - r'^2 &= c^2t'^2(x'^2 + y'^2 + z'^2) \\ &= c^2\gamma^2\left(t - \frac{vx}{c^2}\right)^2 - (x - vt)^2\gamma^2 - y^2 - z^2 \\ &= c^2\gamma^2\left(t^2 - \frac{2vxt}{c^2} + \frac{v^2x^2}{c^2}\right) - \gamma(x^2 - 2vxt + v^2t^2) - y^2 - z^2 \\ &= c^2t^2 - x^2 - y^2 - z^2 \\ &= c^2t^2 - r^2 \end{aligned}$$

This quantity is invariant under Lorentz transformations. So, considering a radial emission of light rays, if $r' = ct'$, then $r = ct$.

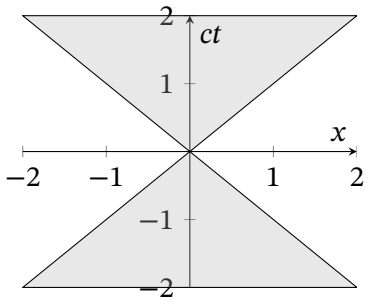
13. Space-time diagrams, simultaneity and causality

13.1. Space-time diagrams

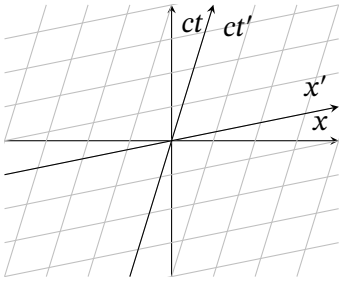
Consider one spatial dimension x and one temporal dimension t in an inertial frame S. We can plot x on the horizontal axis and ct on the vertical axis, in order to make the units match. This combination of space and time in one diagram is called Minkowski spacetime. Each point P in spacetime represents an event labelled by coordinates (x, ct). A moving particle traces out a curve in this diagram, called the world line.



The world line would be a straight line if the particle is moving at a constant velocity. In particular, light rays have gradient 1. Since particles cannot travel faster than the speed of light, world lines are restricted to certain regions (drawn in red) of the space time plane, given that the particle is at $x = 0$ when $t = 0$.



We can also draw the axes of a different frame S' on the same diagram, moving at speed v relative to S. The t' axis corresponds to the equation $x' = 0$ and therefore corresponds to $x = vt$, or equivalently $x = \frac{v}{c} \cdot ct$. The x' axis corresponds to $t' = 0$, which is $ct = \frac{v}{c} \cdot x$.



The angle between the x and x' axes matches the angle between the ct and ct' axes; they are symmetric about the diagonal (as are the original x and ct axes). Note that the diagonal is given by $x = ct$ and $x' = ct'$, which is the same light ray.

13.2. Comparing velocities

Consider a particle moving with constant velocity u' in S', where S' is travelling at velocity v with respect to S. The world line of the particle in S' is simply $x' = u't'$. Correspondingly in S, $x = ut$. Now, using the Lorentz transformation,

$$x = \gamma(x' + vt') = \gamma(u' + v)t'$$

$$t = \gamma\left(t' + \frac{vx'}{c^2}\right) = \gamma\left(1 + \frac{vu'}{c^2}\right)t'$$

Hence,

$$u = \frac{x}{t} = \frac{u' + v}{1 + \frac{u'v}{c^2}}$$

Note that

$$c - u = \frac{(c - u')(c - v)}{1 + \frac{u'v}{c^2}}$$

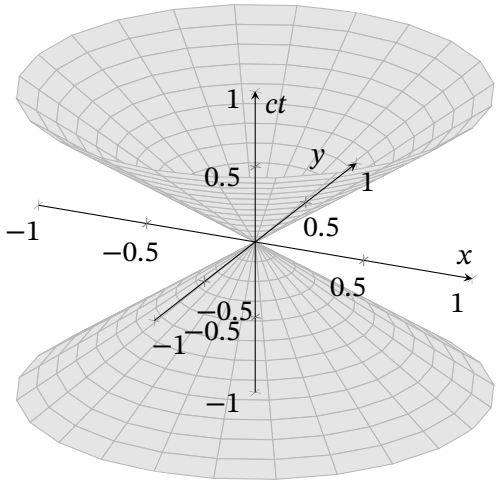
which is always positive if $u' < c$ and $v < c$. Therefore, a Lorentz transformation preserves the property that a speed is smaller than the speed of light.

13.3. Simultaneity

Two events P_1 and P_2 are simultaneous in S if they occur at the same time in S. This is a line parallel to the space axis in the spacetime diagram. In another reference frame, this line of constant time might be at a different angle. So events simultaneous in S' may not correspond to events simultaneous in S. We can use the above formulae to deduce the exact time that an event happens in a different frame of reference.

13.4. Causality

Different observers may disagree on the time ordering of events, but we can construct a viewpoint which gives a consistent description of 'cause' and 'effect', so special relativity does not break causality. Note that lines of simultaneity cannot have an angle greater than $\frac{\pi}{2}$ since the speed of the moving frame must be less than c. We can construct a 'light cone' from all lines or surfaces from an event P at an angle $\frac{\pi}{2}$ to the time axis, which represents the possible effects of an event.



The cone above the origin is the 'future light cone' and the cone below is called the 'past light cone'. Note that this cone is fixed under Lorentz transformations. If an event occurs in the future light cone, then all observers agree that this event occurs after that the event at the origin. Likewise, if an event occurs in the past light cone, all observers agree that this event occurs before the event at the origin. Note that if an event P is not in the light cone, then it cannot cause, or be caused by, the event at the origin, since nothing travels faster than c. Hence, an event at the origin can only be influenced by events inside the past light cone, and may only influence events inside the future light cone.

13.5. Time dilation

Consider first a clock which is stationary in S', which ticks at constant intervals $\Delta t'$. What is the time interval between ticks as perceived in S? We can use the Lorentz transformation, noting that $x' = 0$ since the clock is stationary in S', to get

$$t = \gamma\left(t' + \frac{vx'}{c^2}\right) = \gamma t'$$

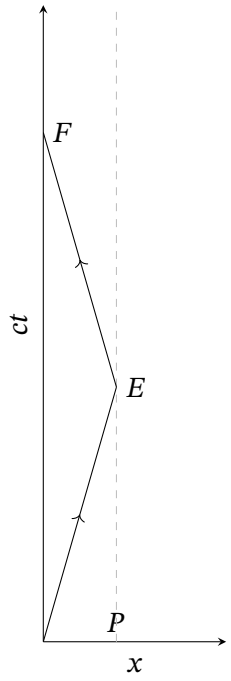
Hence,

$$\Delta t = \gamma \Delta t'$$

So moving clocks run slowly.

13.6. The twin paradox

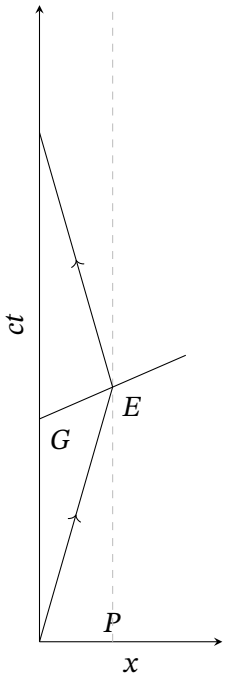
Consider two twins A and B. Twin A stays on Earth (considered to be an inertial frame), and B travels at a constant speed v to a distant planet P, then she turns around and returns to Earth. In the frame of reference of A,



E is the point where B reaches P. The event E occurs at time T as perceived by A, so E has coordinates $(x, ct) = (vT, cT)$. The time experienced by B on her outward journey is

$$T' = \gamma\Big(T - \frac{v}{c} \cdot vT\Big) = \frac{T}{\gamma}$$

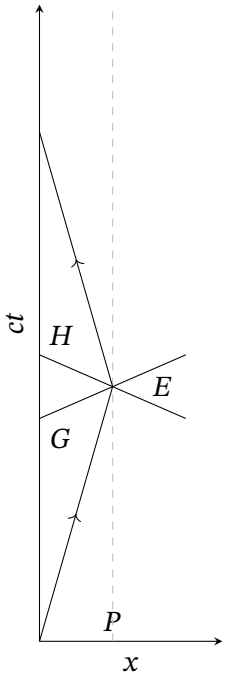
On her return to event F, twin A has aged by $2T$ but twin B has aged by $2T' < 2T$. However, from twin B's perspective, twin A has aged less than she has, since the problem is seemingly symmetric. This would be a paradox. To rectify this, consider the frame of reference of B's outward journey. At E, $x' = 0$ and $t' = T/\gamma$. Consider an event G simultaneous to E in the frame of reference of S'. The new line drawn in the following diagram is a line of constant t'.



At E,

$$t' = \gamma\left(t - \frac{vx}{c^2}\right) = t\gamma \implies t = \frac{t'}{\gamma} = \frac{T}{\gamma^2}$$

So each of them thinks that the other has aged less, when B is at E, by a factor of γ^{-1}. On the return,



The new line is a line of constant t' as measured by B on the return journey, at E. So for the return journey, A sees B age from the event E to the event F. However, B sees A age from the event H to the event F. So there is a time gap between G and H as observed by B, which is not considered by the naive model of this problem. B sees A age instantaneously at the point when she changes direction. In particular, the frame of B as she changes direction is not inertial.

13.7. Length contraction

The length of an object is dependent on the choice of frame. Consider a rod of length L' in S', which is stationary in S'. The world lines of the ends of the rod are vertical. The length of the rod at time t' is the distance in x' between the two world lines. In S,

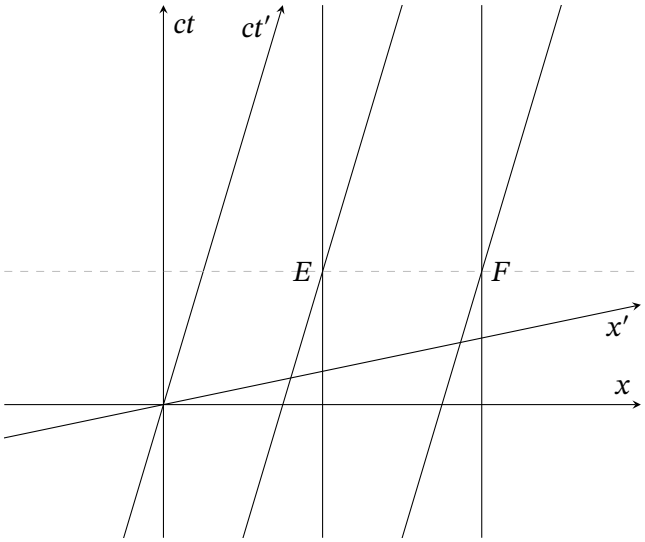
$$x' = 0 \implies \gamma(x - vt) = 0$$

Further,

$$x' = L' \implies \gamma(x - vt) = L'$$

Therefore, the distance between the two x points at the same t is $L = L'/\gamma < L'$. So the length of a moving object shrinks in the direction it is moving. Sometimes, analogously to 'proper time', we consider the 'proper length' of an object, which is the length as measured in the rest frame of the object.

For example, does a train of (proper) length $2L$ fit alongside a platform of length L if it is travelling along the tracks at a speed such that $\gamma = 2$? For observers on the platform, the train indeed contracts to length L, so indeed it fits. On the other hand, for observers on the train, the platform contracts to a length $\frac{1}{2}L$, so the train would not fit. To resolve the uncertainty, we will draw a spacetime diagram, from the frame of reference S where the platform is stationary. The vertical lines represent the end points of the platform. The world lines for the end points of the train are the diagonal lines intersecting E and F. E is the event when the rear of the train is at the rear of the platform, and F is the event where the front of the train is at the front of the platform.



Let E correspond to $t = 0$ and $t' = 0$. The front of the train is at $x' = 2L$, and the front of the platform is at $x = L$. In the S frame, events E and F occur at the same time $t = 0$.

$$x' = \gamma(x - vt) \implies 2L = \gamma(L - vt) = 2(L - vt) \implies t = 0$$

Further,

$$x = \gamma(x' + vt') \implies L = \gamma(2L + vt') = 2(2L + vt') \implies t' = \frac{L - 4L}{2v} = \frac{-3L}{2v} < 0$$

Hence, the time t' at which F occurs is before the event E. So from the perspective of the train, the front of the train has already passed the front of the platform by the time that the back of the train passes the back of the platform, so from this perspective the train does not fit.

14. Geometry of spacetime

14.1. Invariant interval

Consider two events P and Q with space-time coordinates (ct_1, x_1) and (ct_2, x_2), where the time coordinate is given first. The time separation is $\Delta t = t_1 - t_2$, and the space separation $\Delta x = x_1 - x_2$. These two separations are dependent on the choice of inertial frame. The invariant interval between P and Q is defined as

$$\Delta s^2 = c^2 \Delta t^2 - \Delta x^2$$

This is invariant under a Lorentz transformation. In three spatial dimensions, we simply replace this Δx^2 with $\Delta x^2 + \Delta y^2 + \Delta z^2$, so

$$\Delta s^2 = c^2 \Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2$$

If the separation between P and Q is very small, we can define the infinitesimal invariant interval as

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2$$

Note that spacetime with three spatial dimensions (Minkowski spacetime) is topologically equivalent to $\mathbb{R}^4$, where the distance measure is $\mathrm{d}s^2$ as defined above. Note that this distance quantity, although squared, can be either positive or negative. Sometimes this arrangement of one temporal and three spatial dimensions is denoted by the abbreviation '1 + 3 dimensions'.

14.2. Signs of the invariant interval

As noted before, Δs^2 can have either a positive or negative sign.

- Events with $\Delta s^2 > 0$ are 'time-like separated'. In this case, there exists a frame of reference in which the events occur in the same spatial position, but at different times. In particular, the two events appear in each other's light cones. The time ordering of the two events is unambiguous.
- Events with $\Delta s^2 < 0$ are 'space-like separated'. Here, there exists a frame of reference in which the events occur at the same time, but in different places. The two events are outside of each other's light cones, and the ordering of the two events can change depending on the choice of frame of reference.
- If $\Delta s^2 = 0$, the events are 'light-like separated'. The events lie exactly on each other's light cones, and this does *not* imply that the two events are the same (unlike in Euclidean space, where a distance measure of zero implies that two points are equal).

14.3. The Lorentz group

The coordinates of an event P in some frame S can be written as a 4-vector X.

$$X^\mu = \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}$$

where the ct coordinate is given by $\mu = 0$ and the spatial dimensions are given by $\mu = 1, 2, 3$ as usual. The invariant interval between P and the origin is written as the inner product of X with itself:

$$X \cdot X \coloneqq X^\intercal \eta X$$

or alternatively,

$$X \cdot X = \eta_{\mu\nu} X^\mu X^\nu$$

where η is the Minkowski metric given by

$$\eta = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

Then

$$X \cdot X = c^2 t^2 - x^2 - y^2 - z^2$$

We can classify 4-vectors as 'space-like', 'time-like' and 'light-like' as before, by considering the sign of $\eta_{\mu\nu} X^\mu X^\nu$. The Lorentz transformation is a linear transformation that converts the components of X into the components of X in S'. Therefore, any Lorentz transform can be represented as a 4×4 matrix Λ. We now define Lorentz transforms as such linear transformations that preserve the Minkowski metric. So considering a sets of coordinates X and X' in S and S', we have $X' = \Lambda X$, and $X' \cdot X' = X \cdot X$. This then implies that

$$\Lambda^\intercal \eta \Lambda = \eta \tag{$*$}$$

Two classes of possible Λ are

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & a & b & c \\ 0 & d & e & f \\ 0 & g & h & i \end{pmatrix}; \quad R = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$$

where $R^\intercal R = I$, giving that R may be a spatial rotation or a reflection. We could also have

$$\Lambda = \begin{pmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

where $\beta = \frac{v}{c}$. This expresses a Lorentz transformation where the two frames are moving at a constant velocity v relative to each other, as discussed before in $1+1$ spacetime. We denote the Lorentz group as $O(1,3)$, defined by the set of Λ that satisfy $(*)$. Note that this includes the group generated by the above two transformations (notably including spatial reflections), as well as time reflections. We define the *proper* Lorentz group as $SO(1,3)$, which is the kernel of the determinant homomorphism on the Lorentz group. Note that this includes the *composition* of both temporal and spatial reflection. The subgroup that forbids any kind of reflection is called the *restricted* Lorentz group, denoted $SO^+(1,3)$, generated by compositions of rotations and boosts, as shown in the above two examples (excluding the case when R is a reflection).

14.4. Rapidity

While a 4×4 matrix can be useful for computation, it is sometimes more convenient to label a Lorentz transformation using a concept of 'rapidity'. In $1+1$ spacetime, we write

$$\Lambda[\beta] = \begin{pmatrix} \gamma & -\gamma\beta \\ -\gamma\beta & \gamma \end{pmatrix}; \quad \gamma = (1-\beta^2)^{-\frac{1}{2}}$$

This represents a boost in the x direction. Combining two boosts, we get

$$\begin{aligned} \Lambda[\beta_1]\Lambda[\beta_2] &= \begin{pmatrix} \gamma_1 & -\gamma_1\beta_1 \\ -\gamma_1\beta_1 & \gamma_1 \end{pmatrix}\begin{pmatrix} \gamma_2 & -\gamma_2\beta_2 \\ -\gamma_2\beta_2 & \gamma_2 \end{pmatrix} \\ &= \begin{pmatrix} \gamma_1\gamma_2(1+\beta_1\beta_2) & -\gamma_1\gamma_2(\beta_1+\beta_2) \\ -\gamma_1\gamma_2(\beta_1+\beta_2) & \gamma_1\gamma_2(1+\beta_1\beta_2) \end{pmatrix} \\ &= \Lambda\left[\frac{\beta_1+\beta_2}{1+\beta_1\beta_2}\right] \end{aligned}$$

Note the relation to the velocity transformation law. Recall that with spatial rotations, we can characterise a rotation R by some parameter θ, where $R(\theta_1)R(\theta_2) = R(\theta_1+\theta_2)$. This is the same kind of composition law. For Lorentz boosts, we can define ϕ such that $\beta = \tanh\phi$, and then we can redefine Λ to be in terms of ϕ, giving this new composition law

$$\Lambda[\phi_1]\Lambda[\phi_2] = \Lambda[\phi_1+\phi_2]$$

Note that $\gamma = \cosh\phi$, and $\gamma\beta = \sinh\phi$. This suggests that Lorentz boosts can be thought of as hyperbolic rotations in spacetime.

15. Relativistic physics

15.1. Proper time

A particle moves along a trajectory $\mathbf{x}(t)$. The velocity of this particle is $\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{u}(t)$. The path in spacetime is parametrised by t. Both $\mathbf{x}$ and t vary under a Lorentz transformation. Now, consider a particle at rest in S' with $\mathbf{x}' = 0$. The invariant interval on the world line is

$$\Delta s^2 = c^2 \Delta t^2$$

We define the proper time τ as

$$\Delta\tau = \frac{1}{c}\Delta s$$

In particular, in S', $\Delta\tau = \Delta t$, so the proper time is the time experienced in the rest frame of the particle. However, the equation $\Delta\tau = \frac{1}{c}\Delta s$ holds in all frames, since Δs is Lorentz invariant. Note further that Δs is real since this always represents a timelike interval, as it represents a particle travelling through spacetime. We can therefore instead parametrise this particle's world line by its proper time, rather than by considering the time in any particular frame. So $\mathbf{x}$ and t are both functions of τ in any given reference frame. Further, infinitesimal changes are related by

$$\begin{aligned}
\mathrm{d}\tau &= \frac{\mathrm{d}s}{c} \\
&= \frac{1}{c}\sqrt{c^2\,\mathrm{d}t^2 - |\mathrm{d}\mathbf{x}|^2} \\
&= \frac{1}{c}\sqrt{c^2\,\mathrm{d}t^2 - |\mathbf{u}|^2\,\mathrm{d}t^2} \\
&= \left(1 - \frac{\mathbf{u}^2}{c^2}\right)^{\frac{1}{2}} \mathrm{d}t \\
\therefore \frac{\mathrm{d}t}{\mathrm{d}\tau} &= \gamma_{\mathbf{u}}
\end{aligned}$$

where $\gamma_{\mathbf{u}} = \left(1 - \frac{\mathbf{u}^2}{c^2}\right)^{\frac{1}{2}}$. Now, the total time observed by a particle moving along its world line is

$$T = \int \mathrm{d}\tau = \int \frac{\mathrm{d}t}{\gamma_{\mathbf{u}}}$$

15.2. 4-velocity

We can parametrise the position 4-vector of a particle using τ, written

$$X(\tau) = \begin{pmatrix} ct(\tau) \\ \mathbf{x}(\tau) \end{pmatrix}$$

We define the 4-velocity as

$$U = \frac{\mathrm{d}}{\mathrm{d}\tau}X = \begin{pmatrix} c\,\mathrm{d}t/\mathrm{d}\tau \\ \mathrm{d}\mathbf{x}/\mathrm{d}\tau \end{pmatrix} = \frac{\mathrm{d}t}{\mathrm{d}\tau}\begin{pmatrix} c \\ \mathbf{u} \end{pmatrix} = \gamma_{\mathbf{u}}\begin{pmatrix} c \\ \mathbf{u} \end{pmatrix}$$

Since $X' = \Lambda X$, we also have that

$$U' = \Lambda U$$

because τ is invariant. Note that any quantity whose components transform according to this rule is called a 4-vector, and in particular, the derivative of a 4-vector with respect to an invariant is also a 4-vector. Also, the scalar product $U \cdot U$ is invariant under Lorentz transforms. Indeed, in the rest frame of a particle moving with 4-velocity U, in this frame we have $U \cdot U = c^2$. In other frames,

$$U \cdot U = \gamma^2(c^2 - \mathbf{u}^2) = c^2$$

as expected.

15.3. Transformation of velocities

We have found that in special relativity, we cannot simply add velocities together. Consider a transformation Λ from S to S', where S' is moving (relative to S) at a speed v in the x direction. Consider a particle moving in S at speed u at an angle θ to the x axis (with no component in the z axis). In S', it moves with speed u' at an angle θ'. We can write the 4-velocities as

$$U = \begin{pmatrix} \gamma_{\mathbf{u}}c \\ \gamma_{\mathbf{u}}u\cos\theta \\ \gamma_{\mathbf{u}}u\sin\theta \\ 0 \end{pmatrix}; \quad U' = \begin{pmatrix} \gamma_{\mathbf{u}'}c \\ \gamma_{\mathbf{u}'}u'\cos\theta' \\ \gamma_{\mathbf{u}'}u'\sin\theta' \\ 0 \end{pmatrix}$$

and further,

$$U' = \Lambda U$$

where

$$\Lambda = \begin{pmatrix} \gamma_v & -\gamma_v\frac{v}{c} & 0 & 0 \\ -\gamma_v\frac{v}{c} & \gamma_v & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Carrying out the matrix multiplication, we find

$$\begin{pmatrix} \gamma_{\mathbf{u}'}c \\ \gamma_{\mathbf{u}'}u'\cos\theta' \\ \gamma_{\mathbf{u}'}u'\sin\theta' \\ 0 \end{pmatrix} = \begin{pmatrix} \gamma_v & -\gamma_v\frac{v}{c} & 0 & 0 \\ -\gamma_v\frac{v}{c} & \gamma_v & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} \gamma_{\mathbf{u}}c \\ \gamma_{\mathbf{u}}u\cos\theta \\ \gamma_{\mathbf{u}}u\sin\theta \\ 0 \end{pmatrix} \implies \begin{cases} u'\cos\theta' = \dfrac{u\cos\theta - v}{1 - uv\cos\theta/c^2} \\ \tan\theta' = \dfrac{u\sin\theta}{\gamma_{\mathbf{u}}(u\cos\theta - v)} \end{cases}$$

The first equation corresponds to the normal transformation law for Lorentz transforms. The second equation, corresponding to a change in angle due to the motion of the observer, is called aberration. In particular, when $u = c$, we can see that light rays appear to change direction due to the relative motion of the emitter and the observer.

15.4. Energy-momentum 4-vector

We define the 4-momentum of a particle of mass m and 4-velocity U to be

$$P = mU = m\gamma_{\mathbf{u}} \begin{pmatrix} c \\ \mathbf{u} \end{pmatrix}$$

Since U is a 4-vector, we must have that m is invariant under a Lorentz transformation. We will call this m the 'rest mass' of the object, defined as the mass as measured in the rest frame of the particle. The 4-momentum of a system of particles is defined as the sum of the 4-momenta of its individual particles. The spatial components of P, given by $\mu = 1, 2, 3$, can be referred to as the relativistic 3-momentum, given by $\mathbf{p} = \gamma_{\mathbf{u}} m\mathbf{u}$. This matches with the definition as seen in Newtonian physics, except that the mass m is replaced by $\gamma_{\mathbf{u}} m$. We call this quantity the 'apparent mass' of the particle or system of particles, as it represents the mass of the particle as observed by a different reference frame. Note that $|\mathbf{p}|$ and $\gamma_{\mathbf{u}} m$ both tend to infinity as the particle approaches the speed of light. Note that the first component of P, P^0, is

$$\gamma_{\mathbf{u}} mc = \frac{mc}{\sqrt{1 - \frac{\mathbf{u}^2}{c^2}}} = \frac{1}{c}\left(mc^2 + \frac{1}{2}m\mathbf{u}^2 + \ldots\right)$$

We recognise the $\frac{1}{2}m\mathbf{u}^2$ term as the kinetic energy of the particle. We interpret P^0 as an energy, divided by c (to conserve units).

$$P = \begin{pmatrix} \frac{1}{c}E \\ \mathbf{p} \end{pmatrix}$$

where

$$E = \gamma_{\mathbf{u}} mc^2 = mc^2 + \frac{1}{2}m\mathbf{u}^2 + \ldots$$

Note that as $|\mathbf{u}| \to c$, $E \to \infty$. Since P contains an energy term as well as a momentum term, we also call P the energy-momentum 4-vector. Note that for a stationary particle of rest mass m, we have

$$E = mc^2$$

This implies that mass is a form of energy. The energy of a moving particle is

$$E = mc^2 + \underbrace{(\gamma_{\mathbf{u}} - 1)mc^2}_{\text{relativistic kinetic energy}}$$

Since $P \cdot P = \frac{E^2}{c^2} - |\mathbf{p}|^2$ is Lorentz invariant, we have

$$P \cdot P = m^2c^2$$

Hence,

$$\frac{E^2}{c^2} = |\mathbf{p}|^2 + m^2c^2$$

In Newtonian physics, mass is conserved, and energy is also conserved. In relativistic physics, mass is not conserved by itself, since it is a form of energy. From this derivation, it is theoretically possible to convert between mass and kinetic energy.

15.5. Massless particles

A massless particle has zero rest mass. Such particles can have nonzero momentum and nonzero energy, because they are travelling at the speed of light, giving $\gamma_{\mathbf{u}} = \infty$. Since $P \cdot P = m^2c^2$, there are no factors of γ in this expression giving

$$P \cdot P = 0$$

So such a particle travels along a light-like trajectory. Therefore there is no Lorentz transformation that brings a given reference frame into the rest frame of the particle, so we cannot define proper time for such a particle. Since $m^2c^2 = 0$, we must have

$$\frac{E^2}{c^2} = |\mathbf{p}|^2 \implies E = |\mathbf{p}|c$$

Then,

$$P = \frac{E}{c}\begin{pmatrix} 1 \\ \hat{\mathbf{n}} \end{pmatrix}$$

where $\hat{\mathbf{n}}$ is a unit 3-vector in the direction of travel of the particle.

15.6. Newton's second law

Now that we have defined P for all particles, we can rewrite Newton's second law in special relativity as

$$\frac{\mathrm{d}P}{\mathrm{d}\tau} = F$$

where F is the 4-force. If the 3-force is $\mathbf{F}$, we have

$$F = \gamma_{\mathbf{u}}\begin{pmatrix} \mathbf{F} \cdot \mathbf{u}/c \\ \mathbf{F} \end{pmatrix}$$

Hence,

$$\frac{\mathrm{d}E}{\mathrm{d}\tau} = \gamma_{\mathbf{u}}\mathbf{F} \cdot \mathbf{u}; \quad \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}\tau} = \gamma_{\mathbf{u}}\mathbf{F}$$

giving

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \mathbf{F} \cdot \mathbf{u}; \quad \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = \mathbf{F}$$

which are the familiar Newtonian expressions for rate of work and rate of change of momentum. We can now define 4-acceleration:

$$F = mA$$

where m is the rest mass. Hence,

$$\frac{\mathrm{d}U}{\mathrm{d}\tau} = A$$

15.7. Special relativity with particle physics

In Newtonian physics, when two particles collide, we must consider the conservation of 3-momentum. In special relativity however, we must instead consider the conservation of 4-momentum:

$$P = \begin{pmatrix} \frac{E}{c} \\ \mathbf{p} \end{pmatrix}$$

It is often convenient, when dealing with systems of particles, to let the origin of our frame of reference be the centre of momentum. This is the frame such that the total 3-momentum of the system is zero. However, this cannot be done when dealing with massless particles since there does not exist such a rest frame.

15.8. Particle decay

Consider a particle of mass m_1 with 3-momentum $\mathbf{p}_1$ which decays into two particles of mass m_2 and m_3 with 3-momenta $\mathbf{p}_2, \mathbf{p}_3$. Since 4-momentum is conserved, we get $P_1 = P_2 + P_3$. First, consider the 0 component (the timelike component) of P.

$$E_1 = E_2 + E_3$$

Now, consider the $1, 2, 3$ components (the spacelike components) of the 4-momentum. We have

$$\mathbf{p}_1 = \mathbf{p}_2 + \mathbf{p}_3$$

Let us look at this in the centre of momentum frame, so $\mathbf{p}_1 = 0$. Hence

$$\mathbf{p}_2 = -\mathbf{p}_3$$

Because we are in the centre of momentum frame, we have $E_1 = m_1 c^2$ hence

$$\frac{E_1}{c} = m_1 c = \frac{E_2}{c} + \frac{E_3}{c}$$

Further,

$$\frac{E_2}{c} = \sqrt{\mathbf{p}_2^2 + m_2^2 c^2}; \quad \frac{E_3}{c} = \sqrt{\mathbf{p}_3^2 + m_3^2 c^2}$$

Hence,

$$m_1 c = \sqrt{\mathbf{p}_2^2 + m_2^2 c^2} + \sqrt{\mathbf{p}_3^2 + m_3^2 c^2} \geq m_2 c + m_3 c$$

Hence the rest mass of the initial particle must be *at least* the sum of the rest masses of the particles that result from the decay.

15.9. Higgs to photon decay

Consider the decay of the Higgs particle h into two photons γ_1, γ_2. By conservation of 4-momentum,

$$P_h = P_{\gamma_1} + P_{\gamma_2}$$

In the Higgs rest frame,

$$P_h = \begin{pmatrix} m_h c \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \frac{E_{\gamma_1}}{c} \\ \mathbf{p}_{\gamma_1} \end{pmatrix} + \begin{pmatrix} \frac{E_{\gamma_2}}{c} \\ \mathbf{p}_{\gamma_2} \end{pmatrix}$$

Looking at the 1, 2, 3 components we find

$$\mathbf{p}_{\gamma_1} = -\mathbf{p}_{\gamma_2}$$

Looking at the 0 component we find

$$m_h c = \frac{E_{\gamma_1}}{c} + \frac{E_{\gamma_2}}{c}$$

Since $\frac{E^2}{c^2} = \mathbf{p}^2 + m^2 c^2$, because the photons have zero rest mass we have

$$\frac{E_{\gamma_1}}{c} = |\mathbf{p}_{\gamma_1}| = |\mathbf{p}_{\gamma_2}| = \frac{E_{\gamma_2}}{c}$$

Hence,

$$E_{\gamma_1} = E_{\gamma_2} = \frac{1}{2} m_h c^2$$

Note that mass has been lost, but kinetic energy has been gained.

15.10. Particle scattering

Consider two identical particles colliding, without decaying into new particles. In frame S, particle 1 is moving horizontally with 3-velocity $\mathbf{u}$, and particle 2 starts at rest. After the collision, particle 1 has 3-velocity $\mathbf{q}$ and particle 2 has 3-velocity $\mathbf{r}$, where $\mathbf{q}$ has angle θ to the horizontal and $\mathbf{r}$ has angle ϕ to the horizontal. In the centre of momentum frame S', particles 1 and 2 move towards each other horizontally with 3-momenta $\mathbf{p}_1$ and $\mathbf{p}_2 = -\mathbf{p}_1$. After the collision, particle 1 moves with 3-momentum $\mathbf{p}_3$ and particle 2 moves with 3-momentum $\mathbf{p}_4 = -\mathbf{p}_3$. The angle of deflection is θ'. By conservation of 4-momentum,

$$P_1 + P_2 = P_3 + P_4$$

Since particles 1 and 2 have the same mass, their speeds (in S') are equal both before and after the collision. Let the speed before the collision be v and the speed after the collision be w.

$$P_1' = \begin{pmatrix} m\gamma_v c \\ m\gamma_v v \\ 0 \\ 0 \end{pmatrix}; \quad P_2' = \begin{pmatrix} m\gamma_v c \\ -m\gamma_v v \\ 0 \\ 0 \end{pmatrix}; \quad P_3' = \begin{pmatrix} m\gamma_w c \\ m\gamma_w w \cos\theta' \\ m\gamma_w w \sin\theta' \\ 0 \end{pmatrix}; \quad P_4' = \begin{pmatrix} m\gamma_w c \\ -m\gamma_w w \cos\theta' \\ -m\gamma_w w \sin\theta' \\ 0 \end{pmatrix}$$

Looking at the 0 component,

$$2m\gamma_v c = 2m\gamma_w c$$

Since m is the same on both sides,

$$v = w$$

Now we will apply a Lorentz transformation to return to S.

$$\Lambda = \begin{pmatrix} \gamma_v & \gamma_v \frac{v}{c} & 0 & 0 \\ \gamma_v \frac{v}{c} & \gamma_v & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Now, since u is the initial velocity of particle 1 in S,

$$P_1 = \Lambda P_1' = \begin{pmatrix} m\gamma_v^2\left(c + \frac{v^2}{c}\right) \\ m\gamma_v^2(v + v) \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} m\gamma_u c \\ m\gamma_u u \\ 0 \\ 0 \end{pmatrix}$$

After the collision, as seen in S, particle 1's 4-momentum is

$$P_3 = \Lambda P_3' = \begin{pmatrix} m\gamma_v^2\left(c + \frac{v^2}{c}\cos\theta'\right) \\ m\gamma_v^2(v + v\cos\theta') \\ m\gamma_v v\sin\theta' \\ 0 \end{pmatrix} = \begin{pmatrix} m\gamma_q c \\ m\gamma_q q\cos\theta \\ m\gamma_q q\sin\theta \\ 0 \end{pmatrix}$$

By dividing the 1 and 2 components on both sides, we deduce

$$\tan\theta = \frac{m\gamma_v v\sin\theta'}{m\gamma_v^2 v(1+\cos\theta')} = \frac{1}{\gamma_v}\tan\frac{1}{2}\theta'$$

For the second particle, we can do the same calculation to get

$$\tan\phi = \frac{m\gamma_v v\sin\theta'}{m\gamma_v^2 v(1-\cos\theta')} = \frac{1}{\gamma_v}\cot\frac{1}{2}\theta'$$

So given the knowledge of the exact setup of the particles, we can find the angles between the particles as viewed in a different reference frame. In particular,

$$\tan\theta\cdot\tan\phi = \frac{1}{\gamma_v^2} = \frac{2}{1+\gamma_u} \leq 1$$

This is a generalisation of the Newtonian result, where $\gamma_u = 1$ giving

$$\tan\theta\cdot\tan\phi = 1$$

So the angle between the trajectories in the Newtonian case is $\frac{\pi}{2}$.

15.11. Particle creation

Consider equal particles 1 and 2 of mass m moving towards each other horizontally with speed v in S, with 4-momenta P_1 and P_2. After the collision, particles 1 and 2 have 4-momenta P_3 and P_4, and a new particle 3 with 4-momentum P_5 is created with mass M. Note that S is the centre of momentum frame. By conservation of 4-momentum, we have

$$P_1 + P_2 = P_3 + P_4 + P_5$$

We have

$$P_2 + P_2 = \begin{pmatrix} 2m\gamma_v c \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \frac{E_3}{c} + \frac{E_4}{c} + \frac{E_5}{c} \\ \mathbf{0} \end{pmatrix}$$

Certainly we have

$$2m\gamma_v c^2 = E_3 + E_4 + E_5 \geq (m + m + M)c^2 = (2m + M)c^2$$

Hence, for the particle's creation to be possible, we must have

$$\gamma_v \geq 1 + \frac{M}{2m}$$

So the initial kinetic energy in S must satisfy

$$2m(\gamma_v - 1)c^2 \geq Mc^2$$

Consider some other reference frame S' where one particle moves with speed u and the other is at rest. Then

$$u = \frac{2v}{1 + \frac{v^2}{c^2}}$$

Hence, by the result above in the particle scattering experiment,

$$\gamma_u = 2(\gamma_v^2 - 1) \geq 2\left(1 + \frac{M}{2m}\right)^2 - 1 = 1 + \frac{2M}{m} + \frac{M^2}{2m^2}$$

Hence, in this frame, the kinetic energy $mc^2(\gamma_u - 1)$ must satisfy

$$mc^2(\gamma_u - 1) \geq mc^2\left(\frac{2M}{m} + \frac{M^2}{2m^2}\right) \geq 2Mc^2 + \frac{M^2c^2}{2m}$$

This extra $\frac{M^2c^2}{2m}$ term (compared to the Mc^2 expression in S) is produced by the transformation between frames. So in a frame where one particle is at rest, we require significantly more kinetic energy. So a particle accelerator is most efficiently utilised by accelerating two particles into each other, rather than by accelerating one particle into a fixed target.

VI. Probability

Lectured in Lent 2021 by Dr. P. Sousi

In this course, we establish the rules of probability spaces, which are the mathematical framework for dealing with randomness. Potential states of a mathematical system are called outcomes, and we look at particular sets of outcomes called events. For instance, rolling two six-sided dice produces 36 outcomes, and we might be interested in the event 'rolled a double'. Each event can be assigned a probability of occurring; in this case, one sixth. By carefully reasoning about probabilities of events using the rules of probability spaces, we can avoid many apparent paradoxes of probability, such as Simpson's paradox.

When there are many different possible outcomes (or even infinitely many), it becomes helpful to think of certain events as tied to random variables. For example, the amount of coin flips needed before getting a head is a random variable, and its value could be any integer at least 1. The statement 'at least three coin flips were needed' is an example of an event linked to this random variable. The values that a random variable can be, as well as the probabilities that they occur, form the distribution of the random variable. We study many different examples of distributions and their properties to gain a better understanding of random variables.

Contents

1. Probability spaces 348
1.1. Probability spaces and σ-algebras 348
1.2. Properties of the probability measure 348
1.3. Combinatorial analysis 349
1.4. Stirling's formula 350
1.5. Countable subadditivity 353
1.6. Continuity of probability measures 354
2. Inclusion-exclusion 355
2.1. Inclusion-exclusion formula 355
2.2. Bonferroni inequalities 356
2.3. Counting using inclusion-exclusion 357
2.4. Counting derangements 357
3. Independence and dependence of events 359
3.1. Independence of events 359
3.2. Conditional probability 360
3.3. Law of total probability 360
3.4. Bayes' formula 361
3.5. Bayes' formula for medical tests 362
3.6. Probability changes under extra knowledge 362
3.7. Simpson's paradox 363
4. Discrete distributions 365
4.1. Discrete distributions 365
4.2. Bernoulli distribution 365
4.3. Binomial distribution 365
4.4. Multinomial distribution 365
4.5. Geometric distribution 366
4.6. Poisson distribution 366
5. Discrete random variables 367
5.1. Random variables 367
5.2. Expectation 368
5.3. Expectation of binomial distribution 369
5.4. Expectation of Poisson distribution 370
5.5. Expectation of a general random variable 370
5.6. Properties of the expectation 370
5.7. Countable additivity for the expectation 371
5.8. Expectation of indicator function 371
5.9. Expectation under function application 371

5.10. Calculating expectation with cumulative probabilities 372
5.11. Inclusion-exclusion formula with indicators 372
6. Variance and covariance 374
6.1. Variance 374
6.2. Covariance 375
6.3. Expectation of functions of a random variable 376
6.4. Covariance of independent variables 376
7. Inequalities for random variables 378
7.1. Markov's inequality 378
7.2. Chebyshev's inequality 378
7.3. Cauchy–Schwarz inequality 378
7.4. Equality in Cauchy–Schwarz 379
7.5. Jensen's inequality 380
7.6. Arithmetic mean and geometric mean inequality 381
8. Combinations of random variables 382
8.1. Conditional expectation and law of total expectation 382
8.2. Joint distribution 382
8.3. Convolution 383
8.4. Conditional expectation 384
8.5. Properties of conditional expectation 386
9. Random walks 388
9.1. Definition 388
9.2. Gambler's ruin estimate 388
9.3. Expected time to absorption 389
10. Probability generating functions 390
10.1. Definition 390
10.2. Finding moments and probabilities 390
10.3. Sums of random variables 392
10.4. Common probability generating functions 392
10.5. Random sums of random variables 393
11. Branching processes 395
11.1. Introduction 395
11.2. Expectation of generation size 395
11.3. Probability generating functions 395
11.4. Probability of extinction 396
12. Continuous random variables 399
12.1. Probability distribution function 399
12.2. Defining a continuous random variable 400
12.3. Expectation 400
12.4. Computing the expectation 401

12.5. Variance . . . 402
12.6. Uniform distribution . . . 402
12.7. Exponential distribution . . . 402
12.8. Functions of continuous random variables . . . 403
12.9. Normal distribution . . . 404
13. Multivariate density functions . . . 406
13.1. Standardising normal distributions . . . 406
13.2. Multivariate density functions . . . 406
13.3. Independence of events . . . 407
13.4. Marginal density . . . 408
13.5. Sum of random variables . . . 408
13.6. Conditional density . . . 410
13.7. Conditional expectation . . . 410
13.8. Transformations of multidimensional random variables . . . 410
13.9. Order statistics of a random sample . . . 411
13.10. Order statistics on exponential distribution . . . 412
14. Moment generating functions . . . 414
14.1. Moment generating functions . . . 414
14.2. Gamma distribution . . . 414
14.3. Moment generating function of the normal distribution . . . 416
14.4. Cauchy distribution . . . 417
14.5. Multivariate moment generating functions . . . 417
15. Limit theorems . . . 418
15.1. Convergence in distribution . . . 418
15.2. Weak law of large numbers . . . 418
15.3. Types of convergence . . . 418
15.4. Strong law of large numbers . . . 419
15.5. Central limit theorem . . . 420
15.6. Applications of central limit theorem . . . 423
15.7. Sampling error via central limit theorem . . . 423
15.8. Buffon's needle . . . 424
15.9. Bertrand's paradox . . . 425
16. Gaussian vectors . . . 427
16.1. Multidimensional Gaussian random variables . . . 427
16.2. Expectation and variance . . . 427
16.3. Moment generating function . . . 428
16.4. Constructing Gaussian vectors . . . 428
16.5. Density . . . 430
16.6. Diagonal variance . . . 431
16.7. Bivariate Gaussian vectors . . . 431

16.8. Density of bivariate Gaussian 432
16.9. Conditional expectation 432
16.10. Multivariate central limit theorem 433
17. Simulation of random variables 434
17.1. Sampling from uniform distribution 434
17.2. Rejection sampling 434

1. Probability spaces

1.1. Probability spaces and σ-algebras

Definition. Suppose Ω is a set, and $\mathcal{F}$ is a collection of subsets of Ω. We call $\mathcal{F}$ a σ-algebra if

(i) $\Omega \in \mathcal{F}$

(ii) if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$

(iii) for any countable collection $(A_n)_{n\geq 1}$ with $A_n \in \mathcal{F}$ for all n, we must also have that $\bigcup_n A_n \in \mathcal{F}$

Definition. Suppose that $\mathcal{F}$ is a σ-algebra on Ω. A function $\mathbb{P} : \mathcal{F} \to [0,1]$ is called a probability measure if

(i) $\mathbb{P}(\Omega) = 1$

(ii) for any countable disjoint collection of sets $(A_n)_{n\geq 1}$ in $\mathcal{F}$ ($A_n \in \mathcal{F}$ for all n), then $\mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) = \sum_{n\geq 1} \mathbb{P}(A_n)$ (this is called 'countable additivity')

We say that $\mathbb{P}(A)$ is 'the probability of A'. We call $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space, where Ω is the sample space, $\mathcal{F}$ is the σ-algebra, and $\mathbb{P}$ is the probability measure.

Remark. When Ω is countable, we take $\mathcal{F}$ to be all subsets of Ω, i.e. $\mathcal{F} = \mathcal{P}(\Omega)$, its power set.

Definition. The elements of Ω are called outcomes, and the elements of $\mathcal{F}$ are called events.

Note that $\mathbb{P}$ is dependent on $\mathcal{F}$ but not on Ω. We talk about probabilities of *events*, not probabilities of *outcomes*. For example, if you pick a uniform number from the interval $[0,1]$; then the probability of getting any specific outcome is zero—but we can define useful events that have nonzero probabilities.

1.2. Properties of the probability measure

- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, since A and A^c are disjoint sets, whose union is Ω
- $\mathbb{P}(\varnothing) = 0$, since it is the complement of Ω
- if $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ using the inclusion-exclusion theorem

Example. Consider the following examples of probability spaces and probability measures.

- Rolling a fair 6-sided die:
  - $\Omega = \{1,2,3,4,5,6\}$
  - $\mathcal{F} = \mathcal{P}(\Omega)$

– $\forall \omega \in \Omega, \mathbb{P}(\{\omega\}) = \frac{1}{6}$, and if $A \subseteq \Omega$ then $\mathbb{P}(A) = \frac{|A|}{6}$

- Equally likely outcomes (more generally). Suppose Ω is some finite set, e.g. $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$. Then we define $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$. In classical probability, this models picking a random element of Ω.
- Picking balls from a bag. Suppose we have n balls with n labels from the set $\{1, \dots, n\}$, indistinguishable by touching. Let us pick $k \leq n$ balls at random from the bag, *without replacement*. Here, 'at random' just means that all possible outcomes are equally likely, and their probability measures should be equal.

  We will take $\Omega = \{A \subseteq \{1, \dots, n\} : |A| = k\}$. Then $|\Omega| = \binom{n}{k}$. Then of course $\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|}$, since all outcomes (combinations, in this case) are equally likely.
- Consider a well-shuffled deck of 52 cards, i.e. it is equally likely that each possible permutation of the 52 cards will appear. $\Omega = \{\text{all permutations of 52 cards}\}$, and $|\Omega| = 52!$

  The probability that the top two cards are aces is therefore $\frac{4\times 3\times 50!}{52!} = \frac{1}{221}$, since there are $4 \times 3 \times 50!$ outcomes that produce such an event.
- Consider a string of n random digits from $\{0, \dots, 9\}$. Then $\Omega = \{0, \dots, 9\}^n$, and $|\Omega| = 10^n$. We define $A_k = \{\text{no digit exceeds } k\}$, and $B_k = \{\text{largest digit is } k\}$. Then $\mathbb{P}(B_k) = \frac{|B_k|}{|\Omega|}$. Notice that $B_k = A_k \setminus A_{k-1}$. $|A_k| = (k+1)^n$, so $|B_k| = (k+1)^n - k^n$, so $\mathbb{P}(B_k) = \frac{(k+1)^n - k^n}{10^n}$.
- The birthday problem. There are n people; what is the probability that at least two of them share a birthday? We assume that each year has exactly 365 days, i.e. nobody is born on 29$^{\text{th}}$ of February, and that the probabilities of being born on any given day are equal.

  Let $\Omega = \{1, \dots, 365\}^n$, and $\mathcal{F} = \mathcal{P}(\Omega)$. Since all outcomes are equally likely, we take $\mathbb{P}(\{\omega\}) = \frac{1}{365^n}$. Let $A = \{\text{at least two people share the same birthday}\}$. $A^c = \{\text{all } n \text{ birthdays are different}\}$. Since $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$, it suffices to calculate $\mathbb{P}(A^c)$, which is $\frac{|A^c|}{|\Omega|} = \frac{365!}{(365-n)!365^n}$. So the answer is $\mathbb{P}(A) = 1 - \frac{365!}{(365-n)!365^n}$.

  Note that at $n = 22$, $\mathbb{P}(A) \approx 0.476$ and at $n = 23$, $\mathbb{P}(A) \approx 0.507$. So when there are at least 23 people in a room, the probability that two of them share a birthday is around 50%.

1.3. Combinatorial analysis

Let Ω be a finite set, and suppose $|\Omega| = n$. We want to partition Ω into k disjoint subsets $\Omega_1, \dots, \Omega_k$ with $|\Omega_i| = n_i$ and $\sum_{i=1}^k n_i = n$. How many ways of doing such a partition are

there? The result is

$$\underbrace{\binom{n}{n_1}}_{\text{choose first set}} \underbrace{\binom{n-n_1}{n_2}}_{\text{choose second set}} \cdots \underbrace{\binom{n-(n_1+n_2+\cdots+n_{k-1})}{n_k}}_{\text{choose last set}} = \frac{n!}{n_1!n_2!\dots n_k!}$$

So we will write

$$\binom{n}{n_1,\dots,n_k} = \frac{n!}{n_1!n_2!\dots n_k!}$$

Now, let $f : \{1,\dots,k\} \to \{1,\dots,n\}$. f is strictly increasing if $x < y \implies f(x) < f(y)$. f is increasing if $x < y \implies f(x) \le f(y)$. How many strictly increasing functions f exist? Note that if we know the range of f, the function is completely determined. The range is a subset of $\{1,\dots,n\}$ of size k, i.e. a k-subset of an n-set, which yields $\binom{n}{k}$ choices, and thus there are $\binom{n}{k}$ strictly increasing functions.

How many increasing functions f exist? Let us define a bijection from the set of increasing functions $\{f : \{1,\dots,k\} \to \{1,\dots,n\}\}$ to the set of *strictly* increasing functions $\{g : \{1,\dots,k\} \to \{1,\dots,n+k-1\}\}$. For any increasing function f, we define $g(i) = f(i) + i - 1$. Then g is clearly strictly increasing, and takes values in the range $\{1,\dots,n+k-1\}$. By extension, we can define an increasing function f from any strictly increasing function g. So the total number of increasing functions $f : \{1,\dots,k\} \to \{1,\dots,n\}$ is $\binom{n+k-1}{k}$.

1.4. Stirling's formula

Let (a_n) and (b_n) be sequences. We will write $a_n \sim b_n$ if $\frac{a_n}{b_n} \to 1$ as $n \to \infty$. This is asymptotic equality.

Theorem (Stirling's Formula)**.** $n! \sim n^n \sqrt{2\pi n}\, e^{-n}$ as $n \to \infty$.

Let us first prove the weaker statement $\log(n!) \sim n \log n$.

Proof. Let us define $l_n = \log(n!) = \log 2 + \log 3 + \cdots + \log n$. For $x \in \mathbb{R}$, we write $\lfloor x \rfloor$ for the integer part of x. Note that

$$\log\lfloor x \rfloor \le \log x \le \log\lfloor x+1 \rfloor$$

Let us integrate this from 1 to n.

$$\sum_{k=1}^{n-1} \log k \le \int_1^n \log x \,\mathrm{d}x \le \sum_{k=2}^{n} \log k$$

$$l_{n-1} \le n \log n - n + 1 \le l_n$$

For all n, therefore:

$$n \log n - n + 1 \le l_n \le (n+1)\log(n+1) - (n+1) + 1$$

Dividing through by $n \log n$, we get

$$\frac{l_n}{n \log n} \to 1$$

as $n \to \infty$. □

The following complete proof is non-examinable.

Proof. For any twice-differentiable function f, for any $a < b$ we have

$$\int_a^b f(x)\,\mathrm{d}x = \frac{f(a)+f(b)}{2}(b-a) - \frac{1}{2}\int_a^b (x-a)(b-x)f''(x)\,\mathrm{d}x$$

Now let $f(x) = \log x$, $a = k$ and $b = k+1$. Then

$$\begin{aligned}\int_k^{k+1} \log x\,\mathrm{d}x &= \frac{\log k + \log(k+1)}{2} + \frac{1}{2}\int_k^{k+1} \frac{(x-k)(k+1-x)}{x^2}\,\mathrm{d}x \\ &= \frac{\log k + \log(k+1)}{2} + \frac{1}{2}\int_0^1 \frac{x(1-x)}{(x+k)^2}\,\mathrm{d}x\end{aligned}$$

Let us take the sum for $k = 1, \dots, n-1$ of the equality.

$$\int_1^n \log x\,\mathrm{d}x = \frac{\log((n-1)!) + \log(n!)}{2} + \frac{1}{2}\sum_{k=1}^{n-1}\int_0^1 \frac{x(1-x)}{(x+k)^2}\,\mathrm{d}x$$

$$n\log n - n + 1 = \log(n!) - \frac{\log n}{2} + \sum_{k=1}^{n-1} a_k; \quad a_k = \frac{1}{2}\int_0^1 \frac{x(1-x)}{(x+k)^2}\,\mathrm{d}x$$

$$\log(n!) = n\log n - n + \frac{\log n}{2} + 1 - \sum_{k=1}^{n-1} a_k$$

$$n! = n^n e^{-n}\sqrt{n}\,\exp\left(1 - \sum_{k=1}^{n-1} a_k\right)$$

Now, note that

$$a_k \le \frac{1}{2}\int_0^1 \frac{x(1-x)}{k^2}\,\mathrm{d}x = \frac{1}{12k^2}$$

So the sum of all a_k converges. We set

$$A = \exp\left(1 - \sum_{k=1}^{\infty} a_k\right)$$

and then

$$n! = n^n e^{-n}\sqrt{n}A\,\exp\left(\underbrace{\sum_{k=n}^{\infty} a_k}_{\text{converges to zero}}\right)$$

Therefore,

$$n! \sim n^n \sqrt{n}\, e^{-n} A$$

To finish the proof, we must show that $A = \sqrt{2\pi}$. We can utilise the fact that $n! \sim n^n \sqrt{n}\, e^{-n} A$.

$$\begin{aligned} 2^{-2n}\binom{2n}{n} &= 2^{-2n} \cdot \frac{2n!}{(n!)^2} \\ &\sim 2^{-2n} \frac{(2n)^{2n} \cdot \sqrt{2n} \cdot A \cdot e^{-2n}}{n^n\, e^{-n} \sqrt{n} A\, n^n\, e^{-n} \sqrt{n} A} \\ &= \frac{\sqrt{2}}{A\sqrt{n}} \end{aligned}$$

Using a different method, we will prove that $2^{-2n}\binom{2n}{n} \sim \frac{1}{\sqrt{\pi n}}$, which then forces $A = \sqrt{2\pi}$. Consider

$$I_n = \int_0^{\frac{\pi}{2}} (\cos\theta)^n \, \mathrm{d}\theta\,; \quad n \geq 0$$

So $I_0 = \frac{\pi}{2}$ and $I_1 = 1$. By integrating by parts,

$$I_n = \frac{n-1}{n} I_{n-2}$$

Therefore,

$$I_{2n} = \frac{2n-1}{2n} I_{2n-2} = \frac{(2n-1)(2n-3)\ldots(3)(1)}{(2n)(2n-2)\ldots(2)} I_0$$

Multiplying the numerator and denominator by the denominator, we have

$$I_{2n} = \frac{(2n)!}{(n! \cdot 2^n)^2} \cdot \frac{\pi}{2} = 2^{-2n} \frac{2n}{n} \cdot \frac{\pi}{2}$$

In the same way, we can deduce that

$$I_{2n+1} = \frac{(2n)(2n-2)\ldots(2)}{(2n+1)(2n-1)\ldots(3)(1)} I_1 = \frac{1}{2n+1} \left(2^{-2n}\binom{2n}{n}\right)^{-1}$$

From $I_n = \frac{n-1}{n} I_{n-2}$, we get that

$$\frac{I_n}{I_{n-2}} \to 1$$

as $n \to \infty$. We now want to show that $\frac{I_{2n}}{I_{2n+1}} \to 1$. We see from the definition of I_n that I is a decreasing function of n. Therefore,

$$\frac{I_{2n}}{I_{2n+1}} \leq \frac{I_{2n-1}}{I_{2n+1}} \to 1$$

and also

$$\frac{I_{2n}}{I_{2n+1}} \geq \frac{I_{2n}}{I_{2n-2}} \to 1$$

So

$$\frac{I_{2n}}{I_{2n+1}} \to 1$$

which means that

$$\frac{2^{-2n}\binom{2n}{n}\frac{\pi}{2}}{\left(2^{-2n}\binom{2n}{n}\right)^{-1}\frac{1}{2n+1}} \to 1 \implies \left(2^{-2n}\binom{2n}{n}\right)^2 \frac{\pi}{2}(2n+1) \to 1$$

Therefore,

$$\left(2^{-2n}\binom{2n}{n}\right)^2 \sim \frac{2}{\pi(2n+1)} \sim \frac{1}{\pi n}$$

Finally,

$$A = \sqrt{2\pi}$$

completes the proof. □

1.5. Countable subadditivity

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $(A_n)_{n\geq 1}$ be a (not necessarily disjoint) sequence of events in $\mathcal{F}$. Then

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

Proof. Let us define a new sequence $B_1 = A_1$ and $B_n = A_n \setminus (A_1 \cup A_2 \cup \cdots \cup A_{n-1})$. So by construction, the sequence B_n is a disjoint sequence of events in $\mathcal{F}$. Note further that the union of all B_n is equal to the union of all A_n. So

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n\right)$$

By the countable additivity axiom,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(B_n)$$

But $B_n \subseteq A_n$. So $\mathbb{P}(B_n) \leq \mathbb{P}(A_n)$. Therefore,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

□

1.6. Continuity of probability measures

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $(A_n)_{n\geq 1}$ be an increasing sequence in $\mathcal{F}$, i.e. $A_n \in \mathcal{F}$, and $A_n \subseteq A_{n+1}$. Then $\mathbb{P}(A_n) \leq \mathbb{P}(A_{n+1})$. We want to show that

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_n A_n\right)$$

Proof. Let $B_1 = A_1$, and for all $n \geq 2$, let $B_n = A_n \setminus (A_1 \cup A_2 \cup \cdots \cup A_{n-1})$. Then the union over B_i up to n is equal to the union over A_i up to n. So

$$\mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{k=1}^{n} B_k\right) = \sum_{k=1}^{n} \mathbb{P}(B_k) \to \sum_{k=1}^{\infty} \mathbb{P}(B_k) = \mathbb{P}\left(\bigcup_n B_n\right) = \mathbb{P}\left(\bigcup_n A_n\right)$$

□

We can say that probability measures are continuous; an increasing sequence of events has a probability which tends to some limit. Similarly, if (A_n) is decreasing, then the limit probability is the probability of the intersection of all A_n.

2. Inclusion-exclusion

2.1. Inclusion-exclusion formula

Suppose that $A, B \in \mathcal{F}$. Then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$. Now let also $C \in \mathcal{F}$. Then

$$\begin{aligned}\mathbb{P}(A \cup B \cup C) &= \mathbb{P}(A \cup B) + \mathbb{P}(C) - \mathbb{P}((A \cup B) \cap C) \\ &= \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) \\ &- \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) \\ &+ \mathbb{P}(A \cap B \cap C)\end{aligned}$$

Let $A_1, \ldots, A_n$ be events in $\mathcal{F}$. Then

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{i=1}^{n} \mathbb{P}(A_i) \\ &- \sum_{1 \le i_1 < i_2 \le n} \mathbb{P}\left(A_{i_1} \cap A_{i_2}\right) \\ &+ \sum_{1 \le i_1 < i_2 < i_3 \le n} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap A_{i_3}\right) \\ &- \cdots \\ &+ (-1)^{n+1} \mathbb{P}(A_1 \cap \cdots \cap A_n)\end{aligned}$$

Or more concisely,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \mathbb{P}\left(A_{i_1} \cap \cdots \cap A_{i_k}\right)$$

Proof. The case for $n = 2$ has been verified, so we can use induction on n. Now, let us assume this holds for $n - 1$ events.

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \mathbb{P}\left(\left(\bigcup_{i=1}^{n-1} A_i\right) \cup A_n\right) \\ &= \mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) + \mathbb{P}(A_n) - \mathbb{P}\left(\left(\bigcup_{i=1}^{n-1} A_i\right) \cap A_n\right) \\ &= \mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) + \mathbb{P}(A_n) - \mathbb{P}\left(\bigcup_{i=1}^{n-1} (A_i \cap A_n)\right)\end{aligned}$$

Let $B_i = A_i \cap A_n$ for all i. By the inductive hypothesis, we have

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) + \mathbb{P}(A_n) - \mathbb{P}\left(\bigcup_{i=1}^{n-1} B_n\right) \\ &= \sum_{k=1}^{n-1}(-1)^{k+1} \sum_{1\le i_1<\cdots<i_k\le n-1} \mathbb{P}\left(A_{i_1} \cap \cdots \cap A_{i_k}\right) \\ &\quad - \sum_{k=1}^{n-1}(-1)^{k+1} \sum_{1\le i_1<\cdots<i_k\le n-1} \mathbb{P}\left(B_{i_1} \cap \cdots \cap B_{i_k}\right) \\ &\quad + \mathbb{P}(A_n)\end{aligned}$$

which gives the claim as required. □

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with $|\Omega| < \infty$ and $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$. Let $A_1, \ldots, A_n \in \mathcal{F}$. Then

$$|A_1 \cup \cdots \cup A_n| = \sum_{k=1}^{n}(-1)^{k+1} \sum_{1\le i_1<\cdots<i_k\le n} \left|A_{i_1} \cap \cdots \cap A_{i_k}\right|$$

2.2. Bonferroni inequalities

Truncating the sum in the inclusion-exclusion formula at the rth term yields an estimate for the probability. The Bonferroni inequalities state that if r is odd, it is an overestimate, and if r is even, it is an underestimate.

$$\begin{aligned} r \text{ odd} &\implies \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \le \sum_{k=1}^{r}(-1)^{k+1} \sum_{1\le i_1<\cdots<i_k\le n} \mathbb{P}\left(A_{i_1} \cap \cdots \cap A_{i_k}\right) \\ r \text{ even} &\implies \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \ge \sum_{k=1}^{r}(-1)^{k+1} \sum_{1\le i_1<\cdots<i_k\le n} \mathbb{P}\left(A_{i_1} \cap \cdots \cap A_{i_k}\right)\end{aligned}$$

Proof. Again, we will use induction. The $n = 2$ case is trivial. Suppose that r is odd. Then

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) + \mathbb{P}(A_n) - \mathbb{P}\left(\bigcup_{i=1}^{n} B_i\right) \tag{$*$}$$

where $B_i = A_i \cap A_n$. Since r is odd,

$$\mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) \le \sum_{k=1}^{r}(-1)^{k+1} \sum_{1\le i_1<\cdots<i_k\le n-1} \mathbb{P}\left(A_{i_1} \cap \cdots \cap A_{i_k}\right)$$

Since $r-1$ is even, we can apply the inductive hypothesis to $\mathbb{P}\left(\bigcup_{i=1}^{n-1} B_i\right)$.

$$\mathbb{P}\left(\bigcup_{i=1}^{n-1} B_i\right) \geq \sum_{k=1}^{r-1}(-1)^{k+1} \sum_{1\leq i_1<\cdots<i_k\leq n-1} \mathbb{P}\left(B_{i_1}\cap\cdots\cap B_{i_k}\right)$$

We can substitute both bounds into $(*)$ to get an overestimate. □

2.3. Counting using inclusion-exclusion

We can apply the inclusion-exclusion formula to count various things. How many functions $f : \{1,\dots,n\} \to \{1,\dots,m\}$ are surjective? Let Ω be the set of such functions, and $A = \{f \in \Omega : f \text{ is a surjection}\}$. For all $i \in \{1,\dots,m\}$, we define $A_i = \{f \in \Omega : i \notin \{f(1), f(2),\dots,f(n)\}\}$. Then $A = {A_1}^c \cap {A_2}^c \cap \cdots \cap {A_m}^c = (A_1 \cup A_2 \cup \cdots \cup A_m)^c$. Then

$$|A| = |\Omega| - |A_1 \cup \cdots \cup A_m| = m^n - |A_1 \cup \cdots \cup A_m|$$

Now, let us use the inclusion-exclusion formula.

$$|A_1 \cup \cdots \cup A_m| = \sum_{k=1}^{n}(-1)^{k+1} \sum_{1\leq i_1<\cdots<i_k\leq n} \left|A_{i_1}\cap\cdots\cap A_{i_k}\right|$$

Note that $A_{i_1}\cap\cdots\cap A_{i_k}$ is the set of functions where k distinct numbers are not included in the function's range. There are $(m-k)^n$ such functions.

$$\begin{aligned}
|A_1 \cup \cdots \cup A_m| &= \sum_{k=1}^{n}(-1)^{k+1} \sum_{1\leq i_1<\cdots<i_k\leq n} (m-k)^n \\
&= \sum_{k=1}^{n}(-1)^{k+1}\binom{m}{k}(m-k)^n \\
|A| &= m^n - \sum_{k=1}^{n}(-1)^{k+1}\binom{m}{k}(m-k)^n \\
|A| &= \sum_{k=0}^{n}(-1)^{k}\binom{m}{k}(m-k)^n
\end{aligned}$$

2.4. Counting derangements

A derangement is a permutation which has no fixed point, i.e. $\forall i, \sigma(i) \neq i$. We will let Ω be the set of permutations of $\{1,\dots,n\}$, i.e. S_n. Let A be the set of derangements in Ω. Let us pick a permutation σ at random from Ω. What is the probability that it is a derangement? We define $A_i = \{f \in \Omega : f(i) = i\}$, then $A = A_1^c \cap \cdots \cap A_n^c = \left(\bigcup_{i=1}^{n} A_i\right)^c$, so $\mathbb{P}(A) =$

$1 - \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right)$. By the inclusion-exclusion formula,

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \mathbb{P}\left(A_{i_1} \cap \cdots \cap A_{i_k}\right) \\
&= \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \frac{(n-k)!}{|\Omega|} \\
&= \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \frac{(n-k)!}{n!} \\
&= \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} \frac{(n-k)!}{n!} \\
&= \sum_{k=1}^{n} (-1)^{k+1} \frac{n!}{k!(n-k)!} \cdot \frac{(n-k)!}{n!} \\
&= \sum_{k=1}^{n} (-1)^{k+1} \frac{1}{k!}
\end{aligned}$$

So

$$\mathbb{P}(A) = 1 - \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = 1 - \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k!} = \sum_{k=0}^{n} \frac{(-1)^k}{k!}$$

As $n \to \infty$, this value tends to $e^{-1} \approx 0.3678$.

3. Independence and dependence of events

3.1. Independence of events

Definition. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $A, B \in \mathcal{F}$. A and B are called independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

We write $A \perp B$, or $A \perp\!\!\!\perp B$. A countable collection of events (A_n) is said to be independent if for all distinct $i_1, \dots, i_k$, we have

$$\mathbb{P}\left(A_{i_1} \cap \cdots \cap A_{i_k}\right) = \prod_{j=1}^{k} \mathbb{P}\left(A_{i_j}\right)$$

Remark. To show that a collection of events is independent, it is insufficient to show that events are pairwise independent. For example, consider tossing a fair coin twice, so $\Omega = \{(0,0),(0,1),(1,0),(1,1)\}$. $\mathbb{P}(\{\omega\}) = \frac{1}{4}$. Consider the events A, B, C where

$$A = \{(0,0),(0,1)\}; \quad B = \{(0,0),(1,0)\}; \quad C = \{(1,0),(0,1)\}$$

$$\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2}$$

$$\begin{aligned}
\mathbb{P}(A \cap B) &= \mathbb{P}(\{(0,0)\}) = \frac{1}{4} = \mathbb{P}(A) \cdot \mathbb{P}(B) \\
\mathbb{P}(A \cap C) &= \mathbb{P}(\{(0,1)\}) = \frac{1}{4} = \mathbb{P}(A) \cdot \mathbb{P}(C) \\
\mathbb{P}(B \cap C) &= \mathbb{P}(\{(1,0)\}) = \frac{1}{4} = \mathbb{P}(B) \cdot \mathbb{P}(C)
\end{aligned}$$

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(\varnothing) = 0$$

Claim. If $A \perp B$, then $A \perp B^c$.

Proof.

$$\begin{aligned}
\mathbb{P}(A \cap B^c) &= \mathbb{P}(A) - \mathbb{P}(A \cap B) \\
&= \mathbb{P}(A) - \mathbb{P}(A) \cdot \mathbb{P}(B) \\
&= \mathbb{P}(A)(1 - \mathbb{P}(B)) \\
&= \mathbb{P}(A)\,\mathbb{P}(B^c)
\end{aligned}$$

as required. □

3.2. Conditional probability

Definition. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. We define the conditional probability of A given B, written $\mathbb{P}(A \mid B)$, as

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Note that if A and B are independent, then

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A) \cdot \mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A)$$

Claim. Suppose that (A_n) is a disjoint sequence in $\mathcal{F}$. Then

$$\mathbb{P}\left(\bigcup A_n \,\middle|\, B\right) = \sum_n \mathbb{P}(A_n \mid B)$$

This is the countable additivity property for conditional probability.

Proof.

$$\begin{aligned}\mathbb{P}\left(\bigcup A_n \,\middle|\, B\right) &= \frac{\mathbb{P}\left(\left(\bigcup A_n\right) \cap B\right)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}\left(\bigcup (A_n \cap B)\right)}{\mathbb{P}(B)}\end{aligned}$$

By countable additivity, since the $(A_n \cap B)$ are disjoint,

$$\begin{aligned}&= \sum_n \frac{\mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} \\ &= \sum_n \mathbb{P}(A_n \mid B)\end{aligned}$$

□

We can think of $\mathbb{P}(\,\cdot \mid B)$ as a new probability measure for the same Ω.

3.3. Law of total probability

Claim. Suppose (B_n) is a disjoint collection of events in $\mathcal{F}$, such that $\bigcup B = \Omega$, and for all n, we have $\mathbb{P}(B_n) > 0$. If $A \in \mathcal{F}$ then

$$\mathbb{P}(A) = \sum_n \mathbb{P}(A \mid B_n) \cdot \mathbb{P}(B_n)$$

Proof.

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \cap \Omega) \\ &= \mathbb{P}\left(A \cap \left(\bigcup B_n\right)\right) \\ &= \mathbb{P}\left(\bigcup (A \cap B_n)\right)\end{aligned}$$

By countable additivity,

$$\begin{aligned}&= \sum_n \mathbb{P}(A \cap B_n) \\ &= \sum_n \mathbb{P}(A \mid B_n)\,\mathbb{P}(B_n)\end{aligned}$$

□

3.4. Bayes' formula

Claim. Suppose (B_n) is a disjoint sequence of events with $\bigcup B_n = \Omega$ and $\mathbb{P}(B_n) > 0$ for all n. Then

$$\mathbb{P}(B_n \mid A) = \frac{\mathbb{P}(A \mid B_n)\,\mathbb{P}(B_n)}{\sum_k \mathbb{P}(A \mid B_k)\,\mathbb{P}(B_k)}$$

Proof.

$$\begin{aligned}\mathbb{P}(B_n \mid A) &= \frac{\mathbb{P}(B_n \cap A)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(A \mid B_n)\,\mathbb{P}(B_n)}{\mathbb{P}(A)}\end{aligned}$$

By the law of total probability,

$$= \frac{\mathbb{P}(A \mid B_n)\,\mathbb{P}(B_n)}{\sum_k \mathbb{P}(A \mid B_k)\,\mathbb{P}(B_k)}$$

□

Note that on the right hand side, the numerator appears somewhere in the denominator. This formula is the basis of Bayesian statistics. It allows us to reverse the direction of a conditional probability—knowing the probabilities of the events (B_n), and given a model of $\mathbb{P}(A \mid B_n)$, we can calculate the posterior probabilities of B_n given that A occurs.

3.5. Bayes' formula for medical tests

Consider the probability of getting a false positive on a test for a rare condition. Suppose 0.1% of the population have condition A, and we have a test which is positive for 98% of the affected population, and 1% of those unaffected by the disease. Picking an individual at random, what is the probability that they suffer from A, given that they have a positive test?

We define A to be the set of individuals suffering from the condition, and P is the set of individuals testing positive. Then by Bayes' formula,

$$\mathbb{P}(A \mid P) = \frac{\mathbb{P}(P \mid A)\,\mathbb{P}(A)}{\mathbb{P}(P \mid A)\,\mathbb{P}(A) + \mathbb{P}(P \mid A^c)\,\mathbb{P}(A^c)} = \frac{0.98 \cdot 0.001}{0.98 \cdot 0.001 + 0.01 \cdot 0.999} \approx 0.09 = 9\%$$

Why is this so low? We can rewrite this instance of Bayes' formula as

$$\mathbb{P}(A \mid P) = \frac{1}{1 + \frac{\mathbb{P}(P|A^c)\mathbb{P}(A^c)}{\mathbb{P}(P|A)\mathbb{P}(A)}}$$

Here, $\mathbb{P}(A^c) \approx 1, \mathbb{P}(P \mid A) \approx 1$. So

$$\mathbb{P}(A \mid P) \approx \frac{1}{1 + \frac{\mathbb{P}(P|A^c)}{\mathbb{P}(A)}}$$

So this is low because $\mathbb{P}(P \mid A^c) \gg \mathbb{P}(A)$. Suppose that there is a population of 1000 people and about 1 suffers from the disease. Among the 999 not suffering from A, about 10 will test positive. So there will be about 11 people who test positive, and only 1 out of 11 (9%) of those actually has the disease.

3.6. Probability changes under extra knowledge

Consider these three statements:

(a) I have two children, (at least) one of whom is a boy.

(b) I have two children, and the eldest one is a boy.

(c) I have two children, one of whom is a boy born on a Thursday.

What is the probability that I have two boys, given a, b or c? Since no further information is given, we will assume that all outcomes are equally likely. We define:

- BG is the event that the elder sibling is a boy, and the younger is a girl;
- GB is the event that the elder sibling is a girl, and the younger is a boy;
- BB is the event that both children are boys; and
- GG is the event that both children are girls.

Now, we have

(a) $\mathbb{P}(BB \mid BB \cup BG \cup GB) = \frac{1}{3}$

(b) $\mathbb{P}(BB \mid BB \cup BG) = \frac{1}{2}$

(c) Let us define GT to be the event that the elder sibling is a girl, and the younger is a boy born on a Thursday, and define TN to be the event that the elder sibling is a boy born on a Thursday and the younger is a boy not born on a Thursday, and other events are defined similarly. So

$$\begin{aligned}\mathbb{P}(TT \cup TN \cup NT \mid GT \cup TG \cup TT \cup TN \cup NT) &= \frac{\mathbb{P}(TT \cup TN \cup NT)}{\mathbb{P}(GT \cup TG \cup TT \cup TN \cup NT)} \\ &= \frac{\frac{1}{2}\frac{1}{7}\frac{1}{2}\frac{1}{7} + 2 \cdot \frac{1}{2}\frac{1}{7}\frac{1}{2}\frac{6}{7}}{2 \cdot \frac{1}{2}\frac{1}{2}\frac{1}{7} + \frac{1}{2}\frac{1}{7}\frac{1}{2}\frac{1}{7} + 2 \cdot \frac{1}{2}\frac{1}{7}\frac{1}{2}\frac{6}{7}} \\ &= \frac{13}{27} \approx 48\%\end{aligned}$$

3.7. Simpson's paradox

Consider admissions by men and women from state and independent schools to a university given by the tables

All applicants	Admitted	Rejected	% Admitted
State	25	25	50%
Independent	28	22	56%

Men only	Admitted	Rejected	% Admitted
State	15	22	41%
Independent	5	8	38%

Women only	Admitted	Rejected	% Admitted
State	10	3	77%
Independent	23	14	62%

This is seemingly a paradox; both women and men are more likely to be admitted if they come from a state school, but when looking at all applicants, they are more likely to be admitted if they come from an independent school. This is called Simpson's paradox; it arises when we aggregate data from disparate populations. Let A be the event that an individual is admitted, B be the event that an individual is a man, and C be the event that an individual comes from a state school. We see that

$$\begin{gathered}\mathbb{P}(A \mid B \cap C) > \mathbb{P}(A \mid B \cap C^c) \\ \mathbb{P}(A \mid B^c \cap C) > \mathbb{P}(A \mid B^c \cap C^c) \\ \mathbb{P}(A \mid C) < \mathbb{P}(A \mid C^c)\end{gathered}$$

First, note that

$$\begin{aligned}
\mathbb{P}(A \mid C) &= \mathbb{P}(A \cap B \mid C) + \mathbb{P}(A \cap B^c \mid C) \\
&= \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(C)} + \frac{\mathbb{P}(A \cap B^c \cap C)}{\mathbb{P}(C)} \\
&= \frac{\mathbb{P}(A \mid B \cap C)\,\mathbb{P}(B \cap C)}{\mathbb{P}(C)} + \frac{\mathbb{P}(A \mid B^c \cap C)\,\mathbb{P}(B^c \cap C)}{\mathbb{P}(C)} \\
&= \mathbb{P}(A \mid B \cap C)\,\mathbb{P}(B \mid C) + \mathbb{P}(A \mid B^c \cap C)\,\mathbb{P}(B^c \mid C) \\
&> \mathbb{P}(A \mid B \cap C^c)\,\mathbb{P}(B \mid C) + \mathbb{P}(A \mid B^c \cap C^c)\,\mathbb{P}(B^c \mid C)
\end{aligned}$$

Let us also assume that $\mathbb{P}(B \mid C) = \mathbb{P}(B \mid C^c)$. Then

$$\begin{aligned}
\mathbb{P}(A \mid C) &> \mathbb{P}(A \mid B \cap C^c)\,\mathbb{P}(B \mid C^c) + \mathbb{P}(A \mid B^c \cap C^c)\,\mathbb{P}(B^c \mid C^c) \\
&= \mathbb{P}(A \mid C^c)
\end{aligned}$$

So we needed to further assume that $\mathbb{P}(B \mid C) = \mathbb{P}(B \mid C^c)$ in order for the 'intuitive' result to hold. The assumption was not valid in the example, so the result did not hold.

4. Discrete distributions

4.1. Discrete distributions

In a discrete probability distribution on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, Ω is either finite or countable, i.e. $\Omega = \{\omega_1, \omega_2, \dots\}$, and as stated before, $\mathcal{F}$ is the power set of Ω. If we know $\mathbb{P}(\{\omega_i\})$, then this completely determines $\mathbb{P}$. Indeed, let $A \subseteq \Omega$, then

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{i\,:\,\omega_i \in A} \{\omega_i\}\right) = \sum_{i\,:\,\omega_i \in A} \mathbb{P}(\{\omega_i\})$$

by countable additivity. We will see later that this is not true if Ω is uncountable. We write $p_i = \mathbb{P}(\{\omega_i\})$, and we then call this a discrete probability distribution. It has the following key properties:

- $p_i \geq 0$
- $\sum_i p_i = 1$

4.2. Bernoulli distribution

We model the outcome of a test with two outcomes (e.g. the toss of a coin) with the Bernoulli distribution. Let $\Omega = \{0, 1\}$. We will denote $p = p_1$, then clearly $p_0 = 1 - p$.

4.3. Binomial distribution

The binomial distribution B has parameters $N \in \mathbb{Z}^+$, $p \in [0, 1]$. This distribution models a sequence of N independent Bernoulli distributions of parameter p. We then count the amount of 'successes', i.e. trials in which the result was 1. $\Omega = \{0, 1, \dots, N\}$.

$$\mathbb{P}(\{k\}) = p_k = \binom{N}{k} p^k (1-p)^{N-k}$$

4.4. Multinomial distribution

The multinomial distribution is a generalisation of the binomial distribution. M has parameters $N \in \mathbb{Z}^+$ and $p_1, p_2, \dots \in [0, 1]$ where $\sum_{i=1}^{k} p_i = 1$. This models a sequence of N independent trials in which a number from 1 to N is selected, where the probability of selecting i is p_i. $\Omega = \left\{(n_1, \dots, n_k) \in \mathbb{N}^k : \sum_{i=1}^{k} n_i = N\right\}$, in other words, ordered partitions of N. Therefore

$$\begin{aligned}\mathbb{P}(n_1 \text{ outcomes had value } 1, \dots, n_k \text{ outcomes had value } k) &= \mathbb{P}((n_1, \dots, n_k)) \\ &= \binom{N}{n_1, \dots, n_k} p_1^{n_1} \dots p_k^{n_k}\end{aligned}$$

4.5. Geometric distribution

Consider a Bernoulli distribution of parameter p. The geometric distribution models running this trial many times independently until the first 'success' (i.e. the first result of value 1). Then $\Omega = \{1, 2, \dots\} = \mathbb{Z}^+$. Then

$$p_k = (1-p)^{k-1}p$$

We can compute the infinite geometric series $\sum p_k$ which gives 1. We could alternatively model the distribution using $\Omega' = \{0, 1, \dots\} = \mathbb{N}$ which records the amount of failures before the first success. Then

$$p'_k = (1-p)^k p$$

Again, the sum converges to 1.

4.6. Poisson distribution

This is used to model the number of occurrences of an event in a given interval of time. $\Omega = \{0, 1, 2, \dots\} = \mathbb{N}$. This distribution has one parameter $\lambda \in \mathbb{R}$. We have

$$p_k = e^{-\lambda}\frac{\lambda^k}{k!}$$

Then

$$\sum_{k=0}^{\infty} p_k = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{-\lambda}\cdot e^{\lambda} = 1$$

Suppose customers arrive into a shop during the time interval $[0, 1]$. We will subdivide $[0, 1]$ into N intervals $\left[\frac{i-1}{N}, \frac{i}{N}\right]$. In each interval, a single customer arrives with probability p, independent of other time intervals. In this example,

$$\mathbb{P}(k \text{ customers arrive}) = \binom{N}{k}p^k(1-p)^{N-k}$$

Let $p = \frac{\lambda}{N}$ for $\lambda > 0$. We will show that as $N \to \infty$, this binomial distribution converges to the Poisson distribution.

$$\begin{aligned}\binom{N}{k}p^k(1-p)^{N-k} &= \frac{N!}{k!(N-k)!}\left(\frac{\lambda}{n}\right)^k\cdot\left(1-\frac{\lambda}{n}\right)^{N-k}\\ &= \frac{\lambda_k}{k!}\cdot\frac{N!}{N^k(N-k)!}\cdot\left(1-\frac{\lambda}{N}\right)^{N-k}\\ &\to \frac{\lambda_k}{k!}\cdot 1\cdot e^{-\lambda}\end{aligned}$$

which matches the Poisson distribution.

5. Discrete random variables

5.1. Random variables

Definition. Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A random variable X is a function $X : \Omega \to \mathbb{R}$ satisfying

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$$

for any given x.

Suppose $A \subseteq \mathbb{R}$. Then typically we write

$$\{X \in A\} = \{\omega : X(\omega) \in A\}$$

as shorthand. Given $A \in \mathcal{F}$, we define the indicator of A to be

$$1_A(\omega) = 1(\omega \in A) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

Because $A \in \mathcal{F}$, 1_A is a random variable. Suppose X is a random variable. We define the probability distribution function of X to be

$$F_X : \mathbb{R} \to [0,1]; \quad F_X(x) = \mathbb{P}(X \leq x)$$

Definition. $(X_1, \dots, X_n)$ is called a random variable in $\mathbb{R}^n$ if $(X_1, \dots, X_n) : \Omega \to \mathbb{R}^n$, and for all $x_1, \dots, x_n \in \mathbb{R}$ we have

$$\{X_1 \leq x_1, \dots, X_n \leq x_n\} = \{\omega : X_1(\omega) \leq x_1, \dots, X_n(\omega) \leq x_n\} \in \mathcal{F}$$

This definition is equivalent to saying that $X_1, \dots, X_n$ are all random variables in $\mathbb{R}$. Indeed,

$$\{X_1 \leq x_1, \dots, X_n \leq x_n\} = \{X_1 \leq x_1\} \cap \cdots \cap \{X_n \leq x_n\}$$

which, since $\mathcal{F}$ is a σ-algebra, is an element of $\mathcal{F}$.

Definition. A random variable X is called discrete if it takes values in a countable set. Suppose X takes values in the countable set S. For every $x \in S$, we write

$$p_x = \mathbb{P}(X = x) = \mathbb{P}(\{\omega : X(\omega) = x\})$$

We call $(p_x)_{x \in S}$ the probability mass function of X, or the distribution of X. If (p_x) is Bernoulli for example, then we say that X is a Bernoulli (or such) random variable, or that X has the Bernoulli distribution.

Definition. Suppose $X_1, \dots, X_n$ are discrete random variables taking variables in $S_1, \dots, S_n$. We say that the random variables $X_1, \dots, X_n$ are independent if

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n) \quad \forall x_1 \in S_1, \dots, x_n \in S_n$$

As an example, suppose we toss a p-biased coin n times independently. Let $\Omega = \{0,1\}^n$. For every $\omega \in \Omega$,

$$p_\omega = \prod_{k=1}^{n} p^{\omega_k}(1-p)^{1-\omega_k}; \quad \text{where we write } \omega = (\omega_1, \dots, \omega_n)$$

We define a set of discrete random variables $X_k(\omega) = \omega_k$. Then X_k gives the output of the kth toss. We have

$$\mathbb{P}(X_k = 1) = \mathbb{P}(\omega_k = 1) = p; \quad \mathbb{P}(X_k = 0) = \mathbb{P}(\omega_k = 0) = 1 - p$$

So X_k has the Bernoulli distribution with parameter p. We can also show that the X_i are independent. Let $x_1, \dots, x_n \in \{0,1\}$. Then

$$\begin{aligned}
\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) &= \mathbb{P}(\omega = (x_1, \dots, x_n)) \\
&= p_{(x_1,\dots,x_n)} \\
&= \prod_{k=1}^{N} p^{x_k}(1-p)^{1-x_k} \\
&= \prod_{k=1}^{N} \mathbb{P}(X_k = x_k)
\end{aligned}$$

as required. Now, we define $S_n(\omega) = X_1(\omega) + \cdots + X_n(\omega)$. This is the number of heads in N tosses. So $S_n : \Omega \to \{0, \dots, N\}$, and

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

So S_n has the binomial distribution with parameters n and p.

5.2. Expectation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space such that Ω is countable. Let $X : \Omega \to \mathbb{R}$ be a random variable, which is necessarily discrete. We say that X is non-negative if $X \geq 0$. We define the expectation of X to be

$$\mathbb{E}[X] = \sum_{\omega} X(\omega) \cdot \mathbb{P}(\{\omega\})$$

We will write

$$\Omega_X = \{X(\omega) : \omega \in \Omega\}$$

So

$$\Omega = \bigcup_{x \in \Omega_X} \{X = x\}$$

So we have partitioned Ω using X. Note that

$$\begin{aligned}\mathbb{E}[X] &= \sum_{\omega} X(\omega)\mathbb{P}(\{\omega\}) \\ &= \sum_{x\in\Omega_X}\sum_{\omega\in\{X=x\}} X(\omega)\mathbb{P}(\{\omega\}) \\ &= \sum_{x\in\Omega_X}\sum_{\omega\in\{X=x\}} x\mathbb{P}(\{\omega\}) \\ &= \sum_{x\in\Omega_X} x\mathbb{P}(\{X=x\})\end{aligned}$$

which matches the more familiar definition of the expectation; the average of the values taken by X, weighted by the probability of the event occurring. So

$$\mathbb{E}[X] = \sum_{x\in\Omega_X} x p_x$$

5.3. Expectation of binomial distribution

Let $X \sim \text{Bin}(N, p)$. Then

$$\forall k = 0, \dots, N, \quad \mathbb{P}(X = k) = \binom{N}{k} p^k (1-p)^{N-k}$$

So using the second definition,

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=0}^{N} k\mathbb{P}(X=k) \\ &= \sum_{k=0}^{N} k\binom{n}{k} p^k(1-p)^{N-k} \\ &= \sum_{k=0}^{N} \frac{k\cdot N!}{k!\cdot(N-k)!} p^k(1-p)^{N-k} \\ &= \sum_{k=1}^{N} \frac{(N-1)!\cdot N\cdot p}{(k-1)!\cdot(N-k)!} p^{k-1}(1-p)^{N-k} \\ &= Np\sum_{k=1}^{N}\binom{N-1}{k-1} p^{k-1}(1-p)^{N-k} \\ &= Np\sum_{k=0}^{N-1}\binom{N-1}{k} p^{k}(1-p)^{N-1-k} \\ &= Np(p+1-p)^{N-1} \\ &= Np\end{aligned}$$

5.4. Expectation of Poisson distribution

Let $X \sim \text{Poi}(\lambda)$, so

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Hence

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=0}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} \\ &= \sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^{k-1}\lambda}{(k-1)!} \\ &= e^{-\lambda}\cdot e^{\lambda}\cdot\lambda \\ &= \lambda\end{aligned}$$

5.5. Expectation of a general random variable

Let X be a general (not necessarily non-negative) discrete random variable. Then we define

$$X^+ = \max(X, 0); \quad X^- = \max(-X, 0)$$

Then $X = X^+ - X^-$. Note that X^+ and X^- are non-negative random variables, which has a well-defined expectation. So if at least one of $\mathbb{E}[X^+], \mathbb{E}[X^-]$ is finite, we define

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$$

If both are infinite, then we say that the expectation of X is not defined. Whenever we write $\mathbb{E}[X]$, it is assumed to be well-defined. If $\mathbb{E}[|X|] < \infty$, we say that X is integrable. When $\mathbb{E}[X]$ is well-defined, we have again that

$$\mathbb{E}[X] = \sum_{x\in\Omega_x} x\cdot\mathbb{P}(X = x)$$

5.6. Properties of the expectation

The following properties follow immediately from the definition.

(i) If $X \geq 0$, then $\mathbb{E}[X] \geq 0$.

(ii) If $X \geq 0$ and $\mathbb{E}[X] = 0$, then $\mathbb{P}(X = 0) = 1$.

(iii) If $c \in \mathbb{R}$, then $\mathbb{E}[cX] = c\mathbb{E}[X]$, and $\mathbb{E}[c + X] = c + \mathbb{E}[X]$.

(iv) If X, Y are two integrable random variables, then $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

(v) More generally, let $c_1, \dots, c_n \in \mathbb{R}$ and $X_1, \dots, X_n$ integrable random variables. Then

$$\mathbb{E}[c_1X_1 + \cdots + c_nX_n] = c_1\mathbb{E}[X_1] + \cdots + c_n\mathbb{E}[X_n]$$

So the expectation is a linear operator over finitely many inputs.

5.7. Countable additivity for the expectation

Suppose $X_1, X_2, \ldots$ are non-negative random variables. Then

$$\mathbb{E}\left[\sum_n X_n\right] = \sum_n \mathbb{E}[X_n]$$

The non-negativity constraint allows us to guarantee that the sums are well-defined; they could be infinite, but at least their values are well-defined. We will construct a proof assuming that Ω is countable, however the result holds regardless of the choice of Ω.

Proof.

$$\begin{aligned}\mathbb{E}\left[\sum_n X_n\right] &= \sum_\omega \sum_n X_n(\omega)\mathbb{P}(\{\omega\}) \\ &= \sum_n \sum_\omega X_n(\omega)\mathbb{P}(\{\omega\}) \\ &= \sum_n \mathbb{E}[X_n]\end{aligned}$$

□

We are allowed to rearrange the sums since all relevant terms are non-negative.

5.8. Expectation of indicator function

If $X = 1(A)$ where $A \in \mathcal{F}$, then $\mathbb{E}[X] = \mathbb{P}(A)$. This is obvious from the second definition of the expectation.

5.9. Expectation under function application

If $g : \mathbb{R} \to \mathbb{R}$, we can define $g(X)$ to be the random variable given by

$$g(X)(\omega) = g(X(\omega))$$

Then

$$\mathbb{E}[g(X)] = \sum_{x \in \Omega_X} g(x) \cdot \mathbb{P}(X = x)$$

Proof. Let $Y = g(X)$. Then

$$\mathbb{E}[Y] = \sum_{y \in \Omega_Y} y \cdot \mathbb{P}(Y = y)$$

Note that

$$\begin{aligned}\{Y = y\} &= \{\omega : Y(\omega) = y\} \\ &= \{\omega : g(X(\omega)) = y\} \\ &= \{\omega : X(\omega) \in g^{-1}(\{y\})\} \\ &= \{X \in g^{-1}(\{y\})\}\end{aligned}$$

where $g^{-1}(\{y\})$ is the set of all x such that $g(x) \in \{y\}$. So

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{y \in \Omega_y} y \cdot \mathbb{P}\left(X \in g^{-1}(\{y\})\right) \\ &= \sum_{y \in \Omega_Y} y \cdot \sum_{x \in g^{-1}(\{y\})} \mathbb{P}(X = x) \\ &= \sum_{y \in \Omega_Y} \sum_{x \in g^{-1}(\{y\})} g(x)\mathbb{P}(X = x) \\ &= \sum_{x \in \Omega_X} g(x)\mathbb{P}(X = x)\end{aligned}$$

□

5.10. Calculating expectation with cumulative probabilities

If $X \geq 0$ and takes integer values, then

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k) = \sum_{k=0}^{\infty} \mathbb{P}(X > k)$$

Proof. Since X takes non-negative integer values,

$$X = \sum_{k=1}^{\infty} 1(X \geq k) = \sum_{k=0}^{\infty} 1(X > k)$$

This represents the fact that any integer is the sum of that many ones, e.g. $4 = 1+1+1+1+0+0+\ldots$ to infinity. Taking the expectation of the above formula, using that $\mathbb{E}[1(A)] = \mathbb{P}(A)$ and countable additivity, we have the result as claimed. □

5.11. Inclusion-exclusion formula with indicators

We can provide another proof of the inclusion-exclusion formula, using some basic properties of indicator functions.

- $1(A^c) = 1 - 1(A)$
- $1(A \cap B) = 1(A) \cdot 1(B)$

- Following from the above, $1(A \cup B) = 1 - (1 - 1(A))(1 - 1(B))$.

More generally,

$$1(A_1 \cup \cdots \cup A_n) = 1 - \prod_{i=1}^{n}(1 - 1(A_i))$$

which gives the inclusion-exclusion formula. Taking the expectation of both sides, we can see that

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{i_1<i_2} \mathbb{P}\left(A_{i_1} \cap A_{i_2}\right) + \cdots + (-1)^{n+1}\mathbb{P}(A_1 \cap \cdots \cap A_n)$$

which is the result as previously found.

6. Variance and covariance

6.1. Variance

Let X be a random variable, and $r \in \mathbb{N}$. If it is well-defined, we call $\mathbb{E}[X^r]$ the rth moment of X. We define the variance of X by

$$\operatorname{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

If the variance is small, X is highly concentrated around $\mathbb{E}[X]$. If the variance is large, X has a wide distribution including values not necessarily near $\mathbb{E}[X]$. We call $\sqrt{\operatorname{Var}(X)}$ the standard deviation of X, denoted with σ. The variance has the following basic properties:

- $\operatorname{Var}(X) \geq 0$, and if $\operatorname{Var}(X) = 0$, $\mathbb{P}(X = \mathbb{E}[X]) = 1$.
- If $c \in \mathbb{R}$, then $\operatorname{Var}(cX) = c^2 \operatorname{Var}(X)$, and $\operatorname{Var}(X + c) = \operatorname{Var}(X)$.
- $\operatorname{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$. This follows since

$$\begin{aligned}\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] &= \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right] \\ &= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\,\mathbb{E}[X] + \mathbb{E}[X]^2 \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2\end{aligned}$$

- $\operatorname{Var}(X) = \min_{c \in \mathbb{R}} \mathbb{E}\left[(X - c)^2\right]$, and this minimum is achieved at $c = \mathbb{E}[X]$. Indeed, if we let $f(c) = \mathbb{E}\left[(X - c)^2\right]$, then $f(c) = \mathbb{E}\left[X^2\right] - 2c\mathbb{E}[X] + c^2$. Minimising f, we get $f(\mathbb{E}[X]) = \operatorname{Var}(X)$ as required.

As an example, consider $X \sim \operatorname{Bin}(n, p)$. Then $\mathbb{E}[X] = np$, as we found before. Note that we can also represent this binomial distribution as the sum of n Bernoulli distributions of parameter p to get the same result. The variance of X is

$$\operatorname{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$

In fact, in order to compute $\mathbb{E}\left[X^2\right]$ it is easier to find $\mathbb{E}[X(X-1)]$.

$$\begin{aligned}\mathbb{E}[X(X-1)] &= \sum_{k=2}^{n} k \cdot (k-1) \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \\ &= \sum_{k=2}^{n} \frac{k(k-1)n!p^k(1-p)^{n-k}}{(n-k)!k!} \\ &= \sum_{k=2}^{n} \frac{n!p^k(1-p)^{n-k}}{((n-2)-(k-2))!(k-2)!} \\ &= n(n-1)p^2 \sum_{k=2}^{n} \binom{n-2}{k-2} p^{k-2}(1-p)^{n-k} \\ &= n(n-1)p^2\end{aligned}$$

Hence,

$$\operatorname{Var}(X) = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - \mathbb{E}[X]^2 = n(n-1)p^2 + np - (np)^2 = np(1-p)$$

As a second example, if $X \sim \operatorname{Poi}(\lambda)$, we have $\mathbb{E}[X] = \lambda$. Because of the factorial term, it is easier to use $X(X-1)$ than X^2.

$$\begin{aligned}\mathbb{E}[X(X-1)] &= \sum_{k=2}^{\infty} k(k-1)e^{-\lambda}\frac{\lambda^k}{k!} \\ &= e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda_{k-2}}{(k-2)!}\cdot\lambda^2 \\ &= \lambda^2\end{aligned}$$

Hence,

$$\operatorname{Var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$$

6.2. Covariance

Definition. Let X and Y be random variables. Their covariance is defined

$$\operatorname{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

It is a measure of how dependent X and Y are.

Immediately we can deduce the following properties.

- $\operatorname{Cov}(X,Y) = \operatorname{Cov}(Y,X)$
- $\operatorname{Cov}(X,X) = \operatorname{Var}(X)$
- $\operatorname{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\cdot\mathbb{E}[Y]$. Indeed, $(X - \mathbb{E}[X])(Y - \mathbb{E}[Y]) = XY - X\mathbb{E}[Y] - Y\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]$ and the result follows.
- Let $c \in \mathbb{R}$. Then $\operatorname{Cov}(cX,Y) = c\operatorname{Cov}(X,Y)$, and $\operatorname{Cov}(c+X,Y) = \operatorname{Cov}(X,Y)$.
- $\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y)$. Indeed, we have

  $\operatorname{Var}(X+Y) = \mathbb{E}\left[(X - \mathbb{E}[X] + Y - \mathbb{E}[Y])^2\right]$ which gives

  $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] + \mathbb{E}\left[(Y - \mathbb{E}[Y])^2\right] + 2\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$ as required.
- For all $c \in \mathbb{R}$, $\operatorname{Cov}(c,X) = 0$
- If X, Y, Z are random variables, then $\operatorname{Cov}(X+Y,Z) = \operatorname{Cov}(X,Z) + \operatorname{Cov}(Y,Z)$. More generally, for $c_1,\dots,c_n,d_1,\dots,d_m$ real numbers, and for $X_1,\dots,X_n,Y_1,\dots,Y_m$ random variables, we have

$$\operatorname{Cov}\left(\sum_{i=1}^{n} c_i X_i, \sum_{j=1}^{m} d_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} c_i d_j \operatorname{Cov}\left(X_i, Y_j\right)$$

In particular, if we apply this to $X_i = Y_i$, and $c_i = d_i = 1$, then we have

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}(X_i) + \sum_{i \neq j} \operatorname{Cov}(X_i, X_j)$$

6.3. Expectation of functions of a random variable

Recall that X and Y are independent if for all x and y,

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$$

We would like to prove that given positive functions $f, g : \mathbb{R} \to \mathbb{R}_+$, if X and Y are independent we have

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)] \cdot \mathbb{E}[g(Y)]$$

Proof.

$$\begin{aligned}
\mathbb{E}[f(X)g(Y)] &= \sum_{(x,y)} f(x)g(y)\mathbb{P}(X = x, Y = y) \\
&= \sum_{(x,y)} f(x)g(y)\mathbb{P}(X = x)\,\mathbb{P}(Y = y) \\
&= \sum_{x} f(x)\mathbb{P}(X = x) \cdot \sum_{y} g(y)\mathbb{P}(Y = y) \\
&= \mathbb{E}[f(X)] \cdot \mathbb{E}[g(Y)]
\end{aligned}$$

□

The same result holds for general functions, provided the required expectations exist.

6.4. Covariance of independent variables

Suppose X and Y are independent. Then

$$\operatorname{Cov}(X, Y) = 0$$

This is because

$$\begin{aligned}
\operatorname{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\
&= \mathbb{E}[X - \mathbb{E}[X]] \cdot \mathbb{E}[Y - \mathbb{E}[Y]] \\
&= 0 \cdot 0 \\
&= 0
\end{aligned}$$

In particular, we can deduce that

$$\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$$

Note, however, that the covariance being equal to zero does not imply independence. For instance, let X_1, X_2, X_3 be independent Bernoulli random variables with parameter $\frac{1}{2}$. Let us now define $Y_1 = 2X_1 - 1$, $Y_2 = 2X_2 - 1$, and $Z_1 = X_3 Y_1$, $Z_2 = X_3 Y_2$. Now, we have

$$\mathbb{E}[Y_1] = \mathbb{E}[Y_2] = \mathbb{E}[Z_1] = \mathbb{E}[Z_2] = 0$$

We can find that

$$\operatorname{Cov}(Z_1, Z_2) = \mathbb{E}[Z_1 \cdot Z_2] = \mathbb{E}\left[X_3^2 Y_1 Y_2\right] = \mathbb{E}\left[X_3^2\right] \cdot 0 \cdot 0 = 0$$

However, Z_1 and Z_2 are in fact not independent. Since Y_1, Y_2 are never zero,

$$\mathbb{P}(Z_1 = 0, Z_2 = 0) = \mathbb{P}(X_3 = 0) = \frac{1}{2}$$

But also

$$\mathbb{P}(Z_1 = 0) = \mathbb{P}(Z_2 = 0) = \mathbb{P}(X_3 = 0) = \frac{1}{2} \implies \mathbb{P}(Z_1 = 0) \cdot \mathbb{P}(Z_2 = 0) = 0$$

So the events are not independent.

7. Inequalities for random variables

7.1. Markov's inequality

The following useful inequality, and the others derived from it, hold in the discrete and the continuous case.

Theorem. Let $X \geq 0$ be a non-negative random variable. Then for all $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Proof. Observe that $X \geq a \cdot 1(X \geq a)$. This can be seen to be true simply by checking both cases, $X < a$ and $X \geq a$. Taking expectations, we get

$$\mathbb{E}[X] \geq \mathbb{E}[a \cdot 1(X \geq a)] = \mathbb{E}[a \cdot \mathbb{P}(X \geq a)] = a \cdot \mathbb{P}(X \geq a)$$

and the result follows. □

7.2. Chebyshev's inequality

Theorem. Let X be a random variable with finite expectation. Then for all $a > 0$,

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\operatorname{Var}(X)}{a^2}$$

Proof. Note that $\mathbb{P}(|X - \mathbb{E}[X]| \geq a) = \mathbb{P}\left(|X - \mathbb{E}[X]|^2 \geq a^2\right)$. Then we can apply Markov's inequality to this non-negative random variable to get

$$\mathbb{P}\left(|X - \mathbb{E}[X]|^2 \geq a^2\right) \leq \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}{a^2} = \frac{\operatorname{Var}(X)}{a^2}$$

□

7.3. Cauchy–Schwarz inequality

Theorem. If X and Y are random variables, then

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]}$$

Proof. It suffices to prove this statement for X and Y which have finite second moments, i.e. $\mathbb{E}\left[X^2\right]$ and $\mathbb{E}\left[X^2\right]$ are finite. Clearly if they are infinite, then the upper bound is infinite which is trivially true. We need to show that $|\mathbb{E}[XY]|$ is finite. Here we can apply the additional assumption that X and Y are non-negative, since we are taking the absolute value:

$$XY \leq \frac{1}{2}\left(X^2 + Y^2\right) \implies \mathbb{E}[XY] \leq \frac{1}{2}\left(\mathbb{E}\left[X^2\right] + \mathbb{E}\left[Y^2\right]\right)$$

Now, we can assume $\mathbb{E}\left[X^2\right] > 0$ and $\mathbb{E}\left[Y^2\right] > 0$. If this were not the case, the result is trivial since if at least one of them were equal to zero, the corresponding random variable would be identically zero. Let $t \in \mathbb{R}$ and consider

$$0 \leq (X - tY)^2 = X^2 - 2tXY + t^2Y^2$$

Hence

$$\mathbb{E}\left[X^2\right] - 2t\mathbb{E}\left[XY\right] + t^2\mathbb{E}\left[Y^2\right] \geq 0$$

We can view this left hand side as a function $f(t)$. The minimum value of this function is achieved at $t_* = \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}$. Then

$$f(t_*) \geq 0 \implies \mathbb{E}\left[X^2\right] - \frac{2\mathbb{E}\left[XY\right]}{\mathbb{E}\left[Y^2\right]} + \frac{\mathbb{E}\left[XY\right]^2}{\mathbb{E}\left[Y^2\right]} \geq 0$$

Hence,

$$\mathbb{E}\left[XY\right]^2 \leq \mathbb{E}\left[X^2\right] \cdot \mathbb{E}\left[Y^2\right]$$

and the result follows. □

Note that we also have

$$\mathbb{E}\left[|XY|\right] \leq \sqrt{\mathbb{E}\left[X^2\right] \cdot \mathbb{E}\left[Y^2\right]}$$

This is because we can redefine $X \mapsto |X|$ and $Y \mapsto |Y|$, giving

$$\begin{aligned} \left|\mathbb{E}\left[|X| \cdot |Y|\right]\right| &\leq \sqrt{\mathbb{E}\left[X^2\right] \cdot \mathbb{E}\left[Y^2\right]} \\ \mathbb{E}\left[|XY|\right] &\leq \sqrt{\mathbb{E}\left[X^2\right] \cdot \mathbb{E}\left[Y^2\right]} \end{aligned}$$

7.4. Equality in Cauchy–Schwarz

In what cases do we get equality in the Cauchy–Schwarz inequality? Recall that the inequality states

$$\left|\mathbb{E}\left[XY\right]\right| \leq \sqrt{\mathbb{E}\left[X^2\right] \cdot \mathbb{E}\left[Y^2\right]}$$

Recall that in the proof, we considered the random variable $(X - tY)^2$ where X and Y were non-negative, and had finite second moments. The expectation of this random variable was called $f(t)$, and we found that $f(t)$ was minimised when $t = \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}$. We have equality exactly when $f(t) = 0$ for this value of t. But $(X - tY)^2$ is a non-negative random variable, with expectation zero, so it must be zero with probability 1. So we have equality if and only if X is exactly tY.

7.5. Jensen's inequality

Definition. A function $f : \mathbb{R} \to \mathbb{R}$ is called convex if $\forall x, y \in \mathbb{R}$ and for all $t \in [0, 1]$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

This can be visualised as linearly interpolating the values of the function at two points, x and y. The linear interpolation of those points is always greater than the function applied to the linear interpolation of the input points.

Theorem. Let X be a random variable, and let f be a convex function. Then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

We can remember the direction of this inequality by considering the variance: $\operatorname{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$ which is non-negative. Further, $\operatorname{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$ hence $\mathbb{E}\left[X^2\right] \geq \mathbb{E}[X]^2$. Squaring is an example of a convex function, so Jensen's inequality holds in this case. We will first prove a basic lemma about convex functions.

Lemma. Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function. Then f is the supremum of all the lines lying below it. More formally, $\forall m \in \mathbb{R}$, $\exists a, b \in \mathbb{R}$ such that $f(m) = am + b$ and $f(x) \geq ax + b$ for all x.

Proof. Let $m \in \mathbb{R}$. Let $x < m < y$. Then we can express m as $tx + (1-t)y$ for some t in the interval $[0, 1]$. By convexity,

$$f(m) \leq tf(x) + (1-t)f(y)$$

And hence,

$$\begin{aligned} tf(m) + (1-t)f(m) &\leq tf(x) + (1-t)f(y) \\ t(f(m) - f(x)) &\leq (1-t)(f(y) - f(m)) \\ \frac{f(m) - f(x)}{m - x} &\leq \frac{f(y) - f(m)}{y - m} \end{aligned}$$

So the slope of the line joining m to a point on its left is smaller than the slope of the line joining m to a point on its right. So we can produce a value $a \in \mathbb{R}$ given by

$$a = \sup_{x<m} \frac{f(m) - f(x)}{m - x}$$

such that

$$\frac{f(m) - f(x)}{m - x} \leq a \leq \frac{f(y) - f(m)}{y - m}$$

for all $x < m < y$. We can rearrange this to give

$$f(x) \geq a(x - m) + f(m) = ax + (f(m) - am)$$

for all x. □

We may now prove Jensen's inequality.

Proof. Set $m = \mathbb{E}[X]$. Then from the lemma above, there exists $a, b \in \mathbb{R}$ such that

$$f(m) = am + b \implies f(\mathbb{E}[X]) = a\mathbb{E}[X] + b \tag{$*$}$$

and for all x, we have

$$f(x) \geq ax + b$$

We can now apply this inequality to X to get

$$f(X) \geq aX + b$$

Taking the expectation, by $(*)$ we get

$$\mathbb{E}[f(X)] \geq a\mathbb{E}[X] + b = f(\mathbb{E}[X])$$

as required. □

Like the Cauchy–Schwarz inequality, we would like to consider the cases of equality. Let X be a random variable, and f be a convex function such that if $m = \mathbb{E}[X]$, then $\exists a, b \in \mathbb{R}$ such that

$$f(m) = am + b; \quad \forall x \neq m,\ f(x) > ax + b$$

We know that $f(X) \geq aX + b$, since f is convex. Then $f(X) - (aX + b) \geq 0$ is a non-negative random variable. Taking expectations,

$$\mathbb{E}[f(X) - (aX + b)] \geq 0$$

But $\mathbb{E}[aX + b] = am + b = f(m) = f(\mathbb{E}[X])$. We assumed that $\mathbb{E}[f(X)] = f(\mathbb{E}[X])$, hence $\mathbb{E}[aX + b] = \mathbb{E}[f(X)]$ and $\mathbb{E}[f(X) - (aX + b)] = 0$. But since $f(X) \geq aX + b$, this forces $f(X) = aX + b$ everywhere. By our assumption, for all $x \neq m$, $f(x) > ax + b$. This forces $X = m$ with probability 1.

7.6. Arithmetic mean and geometric mean inequality

Let f be a convex function. Suppose $x_1, \dots, x_n \in \mathbb{R}$. Then, from Jensen's inequality,

$$\frac{1}{n}\sum_{k=1}^{n} f(x_k) \geq f\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)$$

Indeed, we can define a random variable X to take values $x_1, \dots, x_n$ all with equal probability. Then, $\mathbb{E}[f(X)]$ gives the left hand side, and $f(\mathbb{E}[X])$ gives the right hand side. Now, let $f(x) = -\log x$. This is a convex function as required. Hence

$$-\frac{1}{n}\sum_{k=1}^{n} \log x_k \geq -\log\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)$$

$$\left(\prod_{k=1}^{n} x_k\right)^{\frac{1}{n}} \leq \frac{1}{n}\sum_{k=1}^{n} x_k$$

Hence the geometric mean is less than or equal to the arithmetic mean.

8. Combinations of random variables

8.1. Conditional expectation and law of total expectation

Recall that if $B \in \mathcal{F}$ with $\mathbb{P}(B) \geq 0$, we defined

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Now, let X be a random variable, and let B be an event as above with nonzero probability. We can then define

$$\mathbb{E}[X \mid B] = \frac{\mathbb{E}[X \cdot 1(B)]}{\mathbb{P}(B)}$$

The numerator is notably zero when $1(B) = 0$, so in essence we are excluding the case where X is not B.

Theorem (law of total expectation)**.** Suppose $X \geq 0$. Let (Ω_n) be a partition of Ω into disjoint events, so $\Omega = \bigcup_n \Omega_n$. Then

$$\mathbb{E}[X] = \sum_n \mathbb{E}[X \mid \Omega_n] \cdot \mathbb{P}(\Omega_n)$$

Proof. We can write $X = X \cdot 1(\Omega)$, where

$$1(\Omega) = \sum_n 1(\Omega_n)$$

Taking expectations, we get

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_n X \cdot 1(\Omega_n)\right]$$

By countable additivity of expectation, we have

$$\mathbb{E}[X] = \sum_n \mathbb{E}[X \cdot 1(\Omega_n)] = \sum_n \mathbb{E}[X \mid \Omega_n] \cdot \mathbb{P}(\Omega_n)$$

as required. □

8.2. Joint distribution

Definition. Let $X_1, \ldots, X_n$ be discrete random variables. Their joint distribution is defined as

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$$

for all $x_i \in \Omega_i$.

Now, we have

$$\mathbb{P}(X_1 = x_1) = \mathbb{P}\left(\{X_1 = x_1\} \cap \bigcup_{i=2}^{n} \bigcup_{x_i} \{X_i = x_i\}\right) = \sum_{x_2,\dots,x_n} \mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

In general,

$$\mathbb{P}(X_i = x_i) = \sum_{x_1,x_2,\dots,x_{i-1},x_{i+1},\dots,x_n} \mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

We call $(\mathbb{P}(X_i = x_i))_i$ the marginal distribution of X_i. Let X, Y be random variables. The conditional distribution of X given $Y = y$ where $y \in \Omega_y$ is defined to be

$$\mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$$

We can find

$$\mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y \mathbb{P}(X = x \mid Y = y)\,\mathbb{P}(Y = y)$$

which is the law of total probability.

8.3. Convolution

Let X and Y be independent, discrete random variables. We would like to find $\mathbb{P}(X + Y = z)$. Clearly this is

$$\begin{aligned}\mathbb{P}(X + Y = z) &= \sum_y \mathbb{P}(X + Y = z, Y = y) \\ &= \sum_y \mathbb{P}(X = z - y, Y = y) \\ &= \sum_y \mathbb{P}(X = z - y) \cdot \mathbb{P}(Y = y)\end{aligned}$$

This last sum is called the convolution of the distributions of X and Y. Similarly,

$$\mathbb{P}(X + Y = z) = \sum_x \mathbb{P}(X = x) \cdot \mathbb{P}(Y = z - x)$$

As an example, let $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\mu)$ be independent. Then

$$\begin{aligned}
\mathbb{P}(X+Y=n) &= \sum_{r=0}^{n} \mathbb{P}(X=r)\,\mathbb{P}(Y=n-r) \\
&= \sum_{r=0}^{n} e^{-\lambda}\frac{\lambda^r}{r!} \cdot e^{-\mu}\frac{\mu^{n-r}}{(n-r)!} \\
&= e^{-(\lambda+\mu)} \sum_{r=0}^{n} \frac{\lambda^r \mu^{n-r}}{r!(n-r)!} \\
&= \frac{e^{-(\lambda+\mu)}}{n!} \sum_{r=0}^{n} \frac{\lambda^r \mu^{n-r} \cdot n!}{r!(n-r)!} \\
&= \frac{e^{-(\lambda+\mu)}}{n!} \sum_{r=0}^{n} \binom{n}{r} \lambda^r \mu^{n-r} \\
&= \frac{e^{-(\lambda+\mu)}}{n!} (\lambda+\mu)^n
\end{aligned}$$

which is the probability mass function of a Poisson random variable with parameter $\lambda + \mu$. In other words, $X + Y \sim \text{Poi}(\lambda + \mu)$.

8.4. Conditional expectation

Let X and Y be discrete random variables. Then the conditional expectation of X given that $Y = y$ is

$$\begin{aligned}
\mathbb{E}[X \mid Y=y] &= \frac{\mathbb{E}[X \cdot 1(Y=y)]}{\mathbb{P}(Y=y)} \\
&= \frac{1}{\mathbb{P}(Y=y)} \sum_{x} x \cdot \mathbb{P}(X=x, Y=y) \\
&= \sum_{x} x \cdot \mathbb{P}(X=x \mid Y=y)
\end{aligned}$$

Observe that for every $y \in \Omega_y$, this expectation is purely a function of y. Let $g(y) = \mathbb{E}[X \mid Y=y]$. Now, we define the conditional expectation of X given Y as $\mathbb{E}[X \mid Y] = g(Y)$.

Note that $\mathbb{E}[X \mid Y]$ is a random variable, dependent only on Y. We have

$$\begin{aligned}\mathbb{E}[X \mid Y] &= g(Y) \cdot 1 \\ &= g(Y) \sum_y 1(Y = y) \\ &= \sum_y g(Y) \cdot 1(Y = y) \\ &= \sum_y g(y) \cdot 1(Y = y) \\ &= \sum_y \mathbb{E}[X \mid Y = y] \cdot 1(Y = y)\end{aligned}$$

This is perhaps a clearer way to see that it depends only on Y. As an example, let us consider tossing a p-biased coin n times independently. We write X_i for the indicator function that the ith toss was a head. Let $Y_n = X_1 + \cdots + X_n$. What is $\mathbb{E}[X_1 \mid Y_n]$? Let $g(y) = \mathbb{E}[X_1 \mid Y_n = y]$. Then $\mathbb{E}[X_1 \mid Y_n] = g(Y_n)$. We therefore need to find g. Let $y \in \{0, \dots, n\}$, then

$$\begin{aligned}g(y) &= \mathbb{E}[X_1 \mid Y_n = y] \\ &= \mathbb{P}(X_1 = 1 \mid Y_n = y)\end{aligned}$$

Clearly if $y = 0$, then $\mathbb{P}(X_1 = 1 \mid Y_n = 0) = 0$. Now, suppose $y \neq 0$. We have

$$\begin{aligned}\mathbb{P}(X_1 = 1 \mid Y_n = y) &= \frac{\mathbb{P}(X_1 = 1, Y_n = y)}{\mathbb{P}(Y_n = y)} \\ &= \frac{\mathbb{P}(X_1 = 1, X_2 + \cdots + X_n = y - 1)}{\mathbb{P}(Y_n = y)} \\ &= \frac{\mathbb{P}(X_1 = 1) \cdot \mathbb{P}(X_2 + \cdots + X_n = y - 1)}{\mathbb{P}(Y_n = y)} \\ &= \frac{p \cdot \binom{n-1}{y-1} \cdot p^{y-1}(1-p)^{n-y}}{\mathbb{P}(Y_n = y)} \\ &= \frac{\binom{n-1}{y-1} \cdot p^y (1-p)^{n-y}}{\binom{n}{y} p^y (1-p)^{n-y}} \\ &= \frac{\binom{n-1}{y-1}}{\binom{n}{y}} \\ &= \frac{y}{n}\end{aligned}$$

Hence

$$g(y) = \frac{y}{n}$$

We can then find that

$$\mathbb{E}[X_1 \mid Y_n] = g(Y_n) = \frac{Y_n}{n}$$

which is indeed a random variable dependent only on Y_n.

8.5. Properties of conditional expectation

The following properties hold.

- For all $c \in \mathbb{R}$, $\mathbb{E}[cX \mid Y] = c\mathbb{E}[X \mid Y]$, and $\mathbb{E}[c \mid Y] = c$.
- Let $X_1, \dots, X_n$ be random variables. Then $\mathbb{E}\left[\sum_{i=1}^n X_i \mid Y\right] = \sum_{i=1}^n \mathbb{E}[X_i \mid Y]$.
- $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$.

The last property is not obvious from the definition, so it warrants its own proof. We can see by the standard properties of the expectation that

$$\begin{aligned}
\mathbb{E}[X \mid Y] &= \sum_y 1(Y = y)\mathbb{E}[X \mid Y = y] \\
\therefore \mathbb{E}[\mathbb{E}[X \mid Y]] &= \sum_y \mathbb{E}[1(Y = y)]\,\mathbb{E}[X \mid Y = y] \\
&= \sum_y \mathbb{P}(Y = y)\,\mathbb{E}[X \mid Y = y] \\
&= \sum_y \mathbb{P}(Y = y)\,\frac{\mathbb{E}[X \cdot 1(Y = y)]}{\mathbb{P}(Y = y)} \\
&= \sum_y \mathbb{E}[X \cdot 1(Y = y)] \\
&= \mathbb{E}\left[\sum_y X \cdot 1(Y = y)\right] \\
&= \mathbb{E}\left[X \sum_y 1(Y = y)\right] \\
&= \mathbb{E}[X]
\end{aligned}$$

Alternatively, we could expand the inner expectation as a sum:

$$\sum_y \mathbb{E}[X \mid Y = y] \cdot \mathbb{P}(Y = y) = \sum_x \sum_y x \cdot \mathbb{P}(X = x \mid Y = y) \cdot \mathbb{P}(Y = y)$$

and the result follows as required. The final property relates conditional probability to independence. Let X and Y be independent. Then $\mathbb{E}[X \mid Y] = \mathbb{E}[X]$. Indeed,

$$\begin{aligned}
\mathbb{E}[X \mid Y] &= \sum_y 1(Y = y)\mathbb{E}[X \mid Y = y] \\
&= \sum_y 1(Y = y)\mathbb{E}[X] \\
&= \mathbb{E}[X]
\end{aligned}$$

Proposition. Suppose Y and Z are independent random variables. Then

$$\mathbb{E}[\mathbb{E}[X \mid Y] \mid Z] = \mathbb{E}[X]$$

Proof. Let $\mathbb{E}[X \mid Y] = g(Y)$ be a random variable that is a function only of Y. Since Y and Z are independent, $f(Y)$ is also independent of Z for any function f. Then $\mathbb{E}[g(Y) \mid Z] = \mathbb{E}[g(Y)] = \mathbb{E}[X]$. □

Proposition. Suppose $h : \mathbb{R} \to \mathbb{R}$ is a function. Then

$$\mathbb{E}[h(Y) \cdot X \mid Y] = h(Y) \cdot \mathbb{E}[X \mid Y]$$

We can 'take out what is known', since we know what Y is.

Proof. Note that

$$\mathbb{E}[h(Y) \cdot X \mid Y = y] = \mathbb{E}[h(y) \cdot X \mid Y = y] = h(y) \cdot \mathbb{E}[X \mid Y = y]$$

Then

$$\mathbb{E}[h(Y) \cdot X \mid Y] = h(y) \cdot \mathbb{E}[X \mid Y]$$

as required. □

Corollary. $\mathbb{E}[\mathbb{E}[X \mid Y] \mid Y] = \mathbb{E}[X \mid Y]$, and $\mathbb{E}[X \mid X] = X$.

Let $X_i = 1(i\text{th toss is a head})$, and $Y_n = X_1 + \cdots + X_n$. We found before that $\mathbb{E}[X_1 \mid Y_n] = \frac{Y_n}{n}$. By symmetry, for all i we have $\mathbb{E}[X_i \mid Y_n] = \mathbb{E}[X_1 \mid Y_n]$. Hence

$$\mathbb{E}[Y_n \mid Y_n] = \mathbb{E}\left[\sum_{i=1}^{n} X_i \mid Y_n\right] = \sum_{i=1}^{n} \mathbb{E}[X_i \mid Y_n] = n \cdot \mathbb{E}[X_1 \mid Y_n]$$

which yields the same result.

9. Random walks

9.1. Definition

A random process, also known as a stochastic process, is a sequence of random variables X_n for $n \in \mathbb{N}$. A random walk is a random process that can be expressed as

$$X_n = x + Y_1 + \cdots + Y_n$$

where the Y_i are independent and identically distributed, and x is a deterministic number. We will focus on the simple random walk on $\mathbb{Z}$, which is defined by taking

$$\mathbb{P}(Y_i = 1) = p; \quad \mathbb{P}(Y_i = -1) = 1 - p = q$$

This can be thought of as a specific case of a Markov chain; it has the property that the path to X_i does not matter, all that matters is the value that we are at, at any point in time.

9.2. Gambler's ruin estimate

What is the probability that X_n reaches some value a before it falls to 0? We will write $\mathbb{P}_x$ for the probability measure $\mathbb{P}$ with the condition that $X_0 = x$, i.e.

$$\mathbb{P}_x(A) = \mathbb{P}(A \mid X_0 = x)$$

We define $h(x) = \mathbb{P}_x((X_n) \text{ hits } a \text{ before hitting } 0)$. We can define a recurrence relation. By the law of total probability, we have, for $0 < x < a$,

$$\begin{aligned} h(x) &= \mathbb{P}_x((X_n) \text{ hits } a \text{ before hitting } 0 \mid Y_1 = 1) \cdot \mathbb{P}_x(Y_1 = 1) \\ &\quad + \mathbb{P}_x((X_n) \text{ hits } a \text{ before hitting } 0 \mid Y_1 = -1) \cdot \mathbb{P}_x(Y_1 = -1) \\ &= p \cdot h(x+1) + q \cdot h(x-1) \end{aligned}$$

Note that

$$h(0) = 0; \quad h(a) = 1$$

There are two cases; $p = q = \frac{1}{2}$ and $p \neq q$. If $p = q = \frac{1}{2}$, then

$$h(x) - h(x+1) = h(x-1) - h(x)$$

We can then solve this to find

$$h(x) = \frac{x}{a}$$

If $p \neq q$, then

$$h(x) = ph(x+1) + qh(x-1)$$

We can try a solution of the form λ^x. Substituting gives

$$p\lambda^2 - \lambda + q = 0 \implies \lambda = 1, \frac{q}{p}$$

The general solution can be found by using the boundary conditions.

$$h(x) = A + B\left(\frac{q}{p}\right)^x \implies h(x) = \frac{\left(\frac{q}{p}\right)^x - 1}{\left(\frac{q}{p}\right)^a - 1}$$

This is known as the 'gambler's ruin' estimate, since it determines whether a gambler will reach a target before going bankrupt.

9.3. Expected time to absorption

Let us define T to be the first time that $x = 0$ or $x = a$. Then $T = \min\{n \geq 0 : X_n \in \{0, a\}\}$. We want to find $\mathbb{E}_x[T] = \tau_x$. We can apply a condition on the first step, and use the law of total expectation to give

$$\tau_x = p\mathbb{E}_x[T \mid Y_1 = 1] + q\mathbb{E}_x[T \mid Y_1 = -1]$$

Hence

$$\tau_x = p(\tau_{x+1} + 1) + q(\tau_{x-1} + 1)$$

We can deduce that, for $0 < x < a$,

$$\tau_x = 1 + p\tau_{x+1} + q\tau_{x-1}$$

and $\tau_0 = \tau_a = 0$. If $p = q = \frac{1}{2}$, then we can try a solution of the form Ax^2.

$$Ax^2 = 1 + \frac{1}{2}A(x+1)^2 + \frac{1}{2}A(x-1)^2$$

This gives a general solution of the form

$$A = -1 \implies \tau_x = -x^2 + Bx + C \implies \tau_x = x(a - x)$$

If $p \neq q$, then we will try a solution of the form Cx, giving

$$C = \frac{1}{q - p}$$

The general solution has the form

$$\tau_x = \frac{x}{q-p} + A + B\left(\frac{q}{p}\right)^x \implies \tau_x = \frac{x}{q-p} - \frac{q}{q-p} \cdot \frac{\left(\frac{q}{p}\right)^x - 1}{\left(\frac{q}{p}\right)^a - 1}$$

10. Probability generating functions

10.1. Definition

Let X be a random variable with values in the positive integers, $\mathbb{N}$. Let $p_r = \mathbb{P}(X = r)$ be the probability mass function. Then the probability generating function is defined to be

$$p(z) = \sum_{r=0}^{\infty} p_r z^r = \mathbb{E}\left[z^X\right] \text{ for } |z| \leq 1$$

When $|z| \leq 1$, the probability generating function converges absolutely, since $\left|\sum_{r=0}^{\infty} p_r z^r\right| \leq \sum_{r=0}^{\infty} p_r = 1$. So $p(z)$ is well-defined and has a radius of convergence of at least 1.

Theorem. The probability generating function of X uniquely determines the distribution of X.

Proof. Suppose (p_r) and (q_r) are two probability mass functions with

$$\sum_{r=0}^{\infty} p_r z^r = \sum_{r=0}^{\infty} q_r z^r, \forall |z| \leq 1$$

We will show that $p_r = q_r$ for all r. First, set $z = 0$, then clearly $p_0 = q_0$. Then by induction, suppose that $p_r = q_r$ for all $r \leq n$. Then we would like to show that $p_{n+1} = q_{n+1}$. We know that

$$\sum_{r=n+1}^{\infty} p_r z^r = \sum_{r=n+1}^{\infty} q_r z^r$$

Hence, dividing by z^{n+1}, and taking the limit as $z \to 0$, we have $p_{n+1} = q_{n+1}$ as required. □

10.2. Finding moments and probabilities

Theorem.

$$\lim_{z \to 1^-} p'(z) = p'(1^-) = \mathbb{E}[X]$$

Proof. We will first assume that $\mathbb{E}[X]$ is finite; we will then extend the proof to the infinite case. Let $0 < z < 1$, then since the series $p(z)$ is absolutely convergent, we can interchange the sum and the derivative operators, giving

$$p'(z) = \sum_{r=0}^{\infty} r p_r z^{r-1}$$

We can make an upper bound for this sum:

$$\sum_{r=0}^{\infty} r p_r z^{r-1} \leq \sum_{r=0}^{\infty} r p_r = \mathbb{E}[X]$$

Since $0 < z < 1$, we see that $p'(z)$ is an increasing function of z. This implies that there exists a limit of $p'(z)$ as $z \to 1^-$, which is upper bounded by $\mathbb{E}[X]$. Now, let $\varepsilon > 0$ and let N be an integer large enough such that

$$\sum_{r=0}^{N} r p_r \geq \mathbb{E}[X] - \varepsilon$$

We have further that, since $0 < z < 1$,

$$p'(z) \geq \sum_{r=1}^{N} r p_r z^{r-1}$$

So

$$\lim_{z \to 1^-} p'(z) \geq \sum_{r=1}^{N} r p_r \geq \mathbb{E}[X] - \varepsilon$$

which is true for any ε. Therefore $\lim_{z \to 1^-} p'(z) = \mathbb{E}[X]$. Now, in the case that $\mathbb{E}[X]$ is infinite, for any M we can find a sufficiently large N such that

$$\sum_{r=0}^{N} r p_r \geq M$$

From above, we know that

$$\lim_{z \to 1^-} p'(z) \geq \sum_{r=1}^{N} r p_r \geq M$$

Since this is true for any M, this limit is equal to ∞. □

In exactly the same way, we can prove that

$$p''(1^-) = \mathbb{E}[X(X-1)]$$

and in general,

$$p^{(k)}(1^-) = \mathbb{E}[X(X-1)\cdots(X-k+1)]$$

In particular,

$$\operatorname{Var}(X) = p''(1^-) + p'(1^-) - p'(1^-)^2$$

Further,

$$\mathbb{P}(X = n) = \frac{1}{n!} \frac{\mathrm{d}^n}{\mathrm{d}z^n} p(z) \Big|_{z=0}$$

10.3. Sums of random variables

Suppose that $X_1, \dots, X_n$ are independent random variables with probability generating functions $q_1, \dots, q_n$ respectively. Then

$$p(z) = \mathbb{E}\left[z^{X_1+\cdots+X_n}\right]$$

Recall that if X and Y are independent, then for all functions f and g, we have $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)]$. Therefore,

$$p(z) = \mathbb{E}\left[z^{X_1}z^{X_2}\cdots z^{X_n}\right] = \mathbb{E}\left[z^{X_1}\right]\cdots\mathbb{E}\left[z^{X_n}\right] = q_1(z)\cdots q_n(z)$$

So the probability generating function factorises into its independent parts. In particular, if all the X_i are independent and identically distributed, then

$$p(z) = q(z)^n$$

10.4. Common probability generating functions

Suppose that $X \sim \mathrm{Bin}(n, p)$. Then

$$\begin{aligned} p(z) &= \mathbb{E}\left[z^X\right] \\ &= \sum_{r=0}^{n} z^r \binom{n}{r} p^r (1-p)^{n-r} \\ &= \sum_{r=0}^{n} \binom{n}{r} (pz)^r (1-p)^{n-r} \\ &= (pz + 1 - p)^n \end{aligned}$$

Now, let $X \sim \mathrm{Bin}(n, p)$, $Y \sim \mathrm{Bin}(m, p)$ be independent random variables. Then the probability generating function of $X + Y$ is

$$(pz + 1 - p)^n \cdot (pz + 1 - p)^m = (pz + 1 - p)^{n+m}$$

which is the probability generating function of a binomial distribution where the number of trials is $n + m$. Now, suppose that $X \sim \mathrm{Geo}(p)$. Then

$$\begin{aligned} p(z) &= \mathbb{E}\left[z^X\right] \\ &= \sum_{r=0}^{n} z^r (1-p)^r p \\ &= \frac{p}{1 - z(1-p)} \end{aligned}$$

Now, suppose that $X \sim \text{Poi}(\lambda)$. Then

$$\begin{aligned} p(z) &= \mathbb{E}\left[z^X\right] \\ &= \sum_{r=0}^{n} z^r e^{-\lambda} \frac{\lambda^r}{r!} \\ &= e^{\lambda(z-1)} \end{aligned}$$

10.5. Random sums of random variables

Consider the sum of a random number of random variables. Let $X_1, \ldots$ be independent and identically distributed, and let N be an independent random variable with values in $\mathbb{N}$. Now, we can define the random variables S_n to be

$$S_n = X_1 + \cdots + X_n$$

Then

$$S_N = X_1 + \cdots + X_N$$

is a random variable dependent on N. For all $\omega \in \Omega$,

$$\begin{aligned} S_N(\omega) &= X_1(\omega) + \cdots + X_{N(\omega)}(\omega) \\ &= \sum_{i=1}^{N(\omega)} X_i(\omega) \end{aligned}$$

Now, let q be the probability generating function of N, and p be the probability generating function of X_1 (or equivalently, any X_i). Then let

$$\begin{aligned} r(z) &= \mathbb{E}\left[z^{S_N}\right] \\ &= \sum_n \mathbb{E}\left[z^{X_1+\cdots+X_N} \cdot 1(N=n)\right] \\ &= \sum_n \mathbb{E}\left[z^{X_1+\cdots+X_n} \cdot 1(N=n)\right] \\ &= \sum_n \mathbb{E}\left[z^{X_1+\cdots+X_n}\right] \mathbb{E}\left[1(N=n)\right] \\ &= \sum_n \mathbb{E}\left[z^{X_1+\cdots+X_n}\right] \mathbb{P}\left(N=n\right) \\ &= \sum_n \mathbb{E}\left[z^{X_1}\right]^n \mathbb{P}\left(N=n\right) \\ &= \sum_n p(z)^n \mathbb{P}\left(N=n\right) \\ &= q(p(z)) \end{aligned}$$

Here is an alternative proof using conditional expectation.

$$\begin{aligned} r(z) &= \mathbb{E}\left[z^{S_N}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[z^{S_N} \mid N\right]\right] \end{aligned}$$

We can see that

$$\begin{aligned} \mathbb{E}\left[z^{S_N} \mid N = n\right] &= \mathbb{E}\left[z^{S_n} \mid N = n\right] \\ &= \mathbb{E}\left[z^{X_1}\right]^n \\ &= p(z)^n \end{aligned}$$

Therefore,

$$\begin{aligned} r(z) &= \mathbb{E}\left[p(z)^N\right] \\ &= q(p(z)) \end{aligned}$$

Using this expression for r, we can find that

$$\mathbb{E}[S_N] = r'(1^-) = q'(p(1^-)) \cdot p'(1^-) = q'(1^-) \cdot p'(1^-) = \mathbb{E}[N]\,\mathbb{E}[X_1]$$

Similarly,

$$\operatorname{Var}(S_N) = \mathbb{E}[N] \operatorname{Var}(X_1) + \operatorname{Var}(N)\left(\mathbb{E}[X_1]\right)^2$$

11. Branching processes

11.1. Introduction

Let $(X_n : n \geq 0)$ be a random process, where X_n is the number of individuals in generation n, and $X_0 = 1$. The individual in generation 0 produces a random number of offspring with distribution

$$g_k = \mathbb{P}(X_1 = k)$$

Then every individual in generation 1 produces an independent number of offspring with the same distribution. This is called a branching process. We can write a recursive formula for X_n. First, let $(Y_{k,n} : k \geq 1, n \geq 0)$ be an independent and identically distributed sequence with distribution $(g_k)_k$. So $Y_{k,n}$ is the number of offspring of the kth individual in generation n.

$$X_{n+1} = \begin{cases} Y_{1,n} + \cdots + Y_{X_n,n} & \text{when } X_n \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

11.2. Expectation of generation size

Theorem.

$$\mathbb{E}[X_n] = \mathbb{E}[X_1]^n$$

Proof. Inductively,

$$\begin{aligned} \mathbb{E}[X_{n+1}] &= \mathbb{E}[\mathbb{E}[X_{n+1} \mid X_n]] \\ \mathbb{E}[X_{n+1} \mid X_n = m] &= \mathbb{E}\left[Y_{1,n} + \cdots + Y_{X_n,n} \mid X_n = m\right] \\ &= \mathbb{E}\left[Y_{1,n} + \cdots + Y_{m,n} \mid X_n = m\right] \\ &= m\mathbb{E}\left[Y_{1,n}\right] \\ &= m\mathbb{E}[X_1] \\ \therefore \mathbb{E}[X_{n+1} \mid X_n] &= X_n \cdot \mathbb{E}[X_1] \\ \therefore \mathbb{E}[X_{n+1}] &= \mathbb{E}[X_n \cdot \mathbb{E}[X_1]] \\ &= \mathbb{E}[X_n] \cdot \mathbb{E}[X_1] \end{aligned}$$

□

11.3. Probability generating functions

Theorem. Let $G(z) = \mathbb{E}\left[z^{X_1}\right]$ be the probability generating function of X_1, and $G_n(z) = \mathbb{E}\left[z^{X_n}\right]$ be the probability generating function of X_n. Then

$$G_{n+1}(z) = G(G_n(z)) = G(G(\cdots G(z) \cdots)) = G_n(G(z))$$

Proof.

$$\begin{aligned} G_{n+1}(z) &= \mathbb{E}\left[z^{X_{n+1}}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[z^{X_{n+1}} \mid X_n\right]\right] \\ \mathbb{E}\left[z^{X_{n+1}} \mid X_n = m\right] &= \mathbb{E}\left[z^{Y_{1,n}+\cdots+Y_{m,n}} \mid X_n = m\right] \\ &= \mathbb{E}\left[z^{X_1}\right]^m \\ &= G(z)^m \\ \therefore \mathbb{E}\left[\mathbb{E}\left[z^{X_{n+1}} \mid X_n\right]\right] &= \mathbb{E}\left[G(z)^{X_n}\right] \\ &= G_n(G(z)) \end{aligned}$$

□

11.4. Probability of extinction

We define the extinction probability q as the probability that $X_n = 0$ for some $n \geq 1$, and $q_n = \mathbb{P}(X_n = 0)$. It is clear that $X_n = 0$ implies that $X_{n+1} = 0$. So the sequence of events $(A_n) = (\{X_n = 0\})$ is an increasing sequence of events. So by the continuity of the probability measure, $\mathbb{P}(A_n)$ converges to $\mathbb{P}\left(\bigcup A_n\right)$ as $n \to \infty$. Note that the event $\bigcup A_n$ is the event that there will be extinction. Therefore, $q_n \to q$ as $n \to \infty$.

Claim. $q_{n+1} = G(q_n)$ and $q = G(q)$.

Proof. Using the above theorem on q,

$$\begin{aligned} q_{n+1} &= \mathbb{P}(X_{n+1} = 0) \\ &= G_{n+1}(0) \\ &= G(G_n(0)) \\ &= G(q_n) \end{aligned}$$

Since G is continuous, taking the limit as $n \to \infty$ and using that $q_n \to q$ gives $G(q) = q$. □

We can form another proof for the first part of the above claim.

Proof. Instead of conditioning on the previous generation, let us condition on the first generation, i.e. $X_1 = m$. Note that after the first generation, we will have m independent subtrees on the family tree. Each tree is identically distributed to the entire tree as a whole. Hence,

$$X_{n+1} = X_n^{(1)} + \cdots + X_n^{(m)}$$

where the $X_i^{(j)}$ are independent and identically distributed random processes each with the

same offspring distribution. By the law of total probability,

$$\begin{aligned} q_{n+1} &= \mathbb{P}(X_{n+1} = 0) \\ &= \sum_m \mathbb{P}(X_{n+1} = 0 \mid X_1 = m) \cdot \mathbb{P}(X_1 = m) \\ &= \sum_m \mathbb{P}\left(X_n^{(1)} = 0, \dots, X_n^{(m)} = 0\right) \cdot \mathbb{P}(X_1 = m) \\ &= \sum_m \mathbb{P}\left(X_n^{(1)} = 0\right)^m \cdot \mathbb{P}(X_1 = m) \\ &= \sum_m q_n^m \cdot \mathbb{P}(X_1 = m) \\ &= G(q_n) \end{aligned}$$

□

Theorem. The extinction probability q is the minimal non-negative solution to $G(t) = t$. Further, supposing that $\mathbb{P}(X_1 = 1) < 1$, we have that $q < 1$ if and only if $\mathbb{E}[X_1] > 1$.

Proof. First, we will prove the minimality of q. Let t be the smallest non-negative solution to $G(t) = t$. We will prove inductively that $q_n \leq t$ for all n, and then by taking limits we have that $q \leq t$. Since q is a solution, this will imply that $q = t$. Now, as a base case, $q_0 = 0 = \mathbb{P}(X_0 = 0) \leq t$. Inductively let us suppose that $q_n \leq t$. We know that $q_{n+1} = G(q_n)$. G is an increasing function on $[0, 1]$, and since $q_n \leq t$ we have $q_{n+1} = G(q_n) \leq G(t) = t$.

Now, we can take $\mathbb{P}(X_1 = 1) < 1$. Let us use the notation $g_r = \mathbb{P}(X_1 = r)$ for simplicity. Consider the function $H(z) = G(z) - z$. Let us assume further that $g_0 + g_1 < 1$, since otherwise we cannot possibly ever increase the amount of individuals in future generations, as $\mathbb{E}[X_1] = \mathbb{P}(X_1 = 1) < 1$. In this case, $G(z) = g_0 + g_1 z = 1 - \mathbb{E}[X_1] + \mathbb{E}[X_1] \cdot z$, and solving $G(z) = z$ we would get only $z = 1$ since $\mathbb{E}[X_1] < 1$. Now,

$$H''(z) = \sum_{r=2}^{\infty} r(r-1) g_r z^{r-2} > 0 \quad \forall z \in (0, 1)$$

This implies that $H'(z)$ is a strictly increasing function in $(0, 1)$. Hence, $H(z)$ has at most one root different from 1 in $(0, 1)$, which follows from Rolle's theorem; indeed, if it had two roots different from 1, then H' would be zero once in (z_1, z_2) and once in $(z_2, 1)$, which contradicts the fact that H' is strictly increasing.

Let us first consider the case where H has no other root apart from 1. In this case, $H(1) = 0$ and $H(0) = g_0 \geq 0 \implies H(z) \geq 0$ for all $z \in [0, 1]$. We can find that

$$H'(1^-) = \lim_{z \to 1^-} \frac{H(z) - H(1)}{z - 1} = \frac{H(z)}{z - 1} < 0$$

since the numerator is positive, and the denominator is negative. We know that $H'(1^-) = G'(1^-) - 1$, and $H'(1^-) \leq 0 \implies G'(1^-) \leq 1$, and $G'(1^-) = \mathbb{E}[X_1]$. So when $q = 1$, then $\mathbb{E}[X_1] \leq 1$.

In the other case, H has one other root $r < 1$ as well as 1. We have that $H(r) = 0$ and $H(1) = 0$. By Rolle's theorem, there exists $z \in (r, 1)$ such that $H'(z) = 0$. Further, $H'(x) = G'(x) - 1$ therefore $G'(z) = 1$. Now,

$$G'(x) = \sum_{r=1}^{\infty} r g_r x^{r-1} \implies H''(x) = G''(x) = \sum_{r=2}^{\infty} r(r-1) g_r x^{r-2}$$

Under the assumption that $g_0 + g_1 < 1$, we have that $G''(x) > 0$ for all $x \in (0, 1)$, hence G' is strictly increasing for all $x \in (0, 1)$. Therefore, $G'(1^-) > G'(z) = 1$ giving $\mathbb{E}[X_1] > 1$. So if $q < 1$, then $\mathbb{E}[X_1] > 1$. □

12. Continuous random variables

12.1. Probability distribution function

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Then, as defined before, $X : \Omega \to \mathbb{R}$ is a random variable if

$$\forall x \in \mathbb{R}, \{X \leq x\} = \{\omega : X(\omega) \leq x\} \in \mathcal{F}$$

We define the probability distribution function $F : \mathbb{R} \to [0, 1]$ as

$$F(x) = \mathbb{P}(X \leq x)$$

Theorem. The following properties hold.

(i) If $x \leq y$, then $F(x) \leq F(y)$.

(ii) For all $a < b$, $\mathbb{P}(a < x \leq b) = F(b) - F(a)$.

(iii) F is a right continuous function, and left limits always exist. In other words,

$$F(x^+) = \lim_{y \to x^+} F(y) = F(x); \quad F(x^-) = \lim_{y \to x^-} F(y) \leq F(x)$$

(iv) For all $x \in \mathbb{R}$, $F(x^-) = \mathbb{P}(X < x)$.

(v) We have $\lim_{x \to \infty} F(x) = 1$ and $\lim_{x \to -\infty} F(x) = 0$.

Proof. (i) The first statement is immediate from the definition of the probability measure.

(ii) We can deduce

$$\begin{aligned} \mathbb{P}(a < X \leq b) &= \mathbb{P}(\{a < X\} \cap \{X \leq b\}) \\ &= \mathbb{P}(X \leq b) - \mathbb{P}(\{X \leq b\} \cap \{X \leq a\}) \\ &= \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) \\ &= F(b) - F(a) \end{aligned}$$

(iii) For right continuity, we want to prove $\lim_{n \to \infty} F\left(x + \frac{1}{n}\right) = F(x)$. We will define $A_n = \left\{x < X \leq x + \frac{1}{n}\right\}$. Then the A_n are decreasing events, and the intersection of all A_n is the empty set $\varnothing$. Hence, by continuity of the probability measure, $\mathbb{P}(A_n) \to 0$ as $n \to \infty$. But $\mathbb{P}(A_n) = \mathbb{P}\left(x < X \leq x + \frac{1}{n}\right) = F\left(x + \frac{1}{n}\right) - F(x)$, hence $F(x + \frac{1}{n}) \to F(x)$ as required. Now, we want to show that left limits always exist. This is clear since F is an increasing function, and is always bounded above by 1.

(iv) We know $F(x^-) = \lim_{n \to \infty} F\left(x - \frac{1}{n}\right)$. Consider $B_n = \left\{X \leq x - \frac{1}{n}\right\}$. Then the B_n is an increasing sequence of events, and their union is $\{X < x\}$. Hence $\mathbb{P}(B_n)$ converges to $\mathbb{P}(X < x)$, so $F(x^-) = \mathbb{P}(X < x)$.

(v) This is evident from the properties of the probability measure.

□

12.2. Defining a continuous random variable

For a discrete random variable, F is a step function, which of course is right continuous with left limits.

Definition. A random variable X is called *continuous* if F is a continuous function. In this case, clearly left limits and right limits give the same value, and $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$.

In this course, we will consider only *absolutely* continuous random variables. A continuous random variable is absolutely continuous if F is differentiable. We will make the convention that $F'(x) = f(x)$, where $f(x)$ is called the probability density function of X. The following immediate properties hold.

(i) $f \geq 0$

(ii) $\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1$

(iii) $F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$

(iv) For $S \subseteq \mathbb{R}$, $\mathbb{P}(X \in S) = \int_S f(x)\,\mathrm{d}x$

Here is an intuitive explanation of the probability density function. Suppose Δx is a small quantity. Then

$$\mathbb{P}(x < X \leq x + \Delta x) = \int_x^{x+\Delta x} f(y)\,\mathrm{d}y \approx f(x) \cdot \Delta x$$

So we can think of $f(x)$ as the continuous analogy to $\mathbb{P}(X = x)$.

12.3. Expectation

Consider a continuous random variable $X : \Omega \to \mathbb{R}$, with probability distribution function $F(x)$ and probability density function $f(x) = F'(x)$. We define the expectation of such a *non-negative* random variable as

$$\mathbb{E}[X] = \int_0^\infty x f(x)\,\mathrm{d}x$$

In this case, the expectation is either non-negative and finite, or positive infinity. Now, let X be a general continuous random variable, that is not necessarily non-negative. Suppose $g \geq 0$. Then,

$$\mathbb{E}[g(X)] = \int_{-\infty}^\infty g(x) f(x)\,\mathrm{d}x$$

We can define $X_+ = \max(X, 0)$ and $X_- = \max(-X, 0)$. If at least one of $\mathbb{E}[X_+]$ or $\mathbb{E}[X_-]$ is finite, then clearly

$$\mathbb{E}[X] \coloneqq \mathbb{E}[X_+] - \mathbb{E}[X_-] = \int_{-\infty}^\infty x f(x)\,\mathrm{d}x$$

It is easy to verify that the expectation is a linear function, due to the linearity property of the integral.

12.4. Computing the expectation

Claim. Let $X \geq 0$. Then

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq x)\,\mathrm{d}x$$

Proof. Using the definition of the expectation,

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x f(x)\,\mathrm{d}x \\
&= \int_0^\infty \left(\int_0^x \mathrm{d}y\right) f(x)\,\mathrm{d}x \\
&= \int_0^x \mathrm{d}y \int_y^\infty f(x)\,\mathrm{d}x \\
&= \int_0^\infty \mathrm{d}y \left(1 - \int_{-\infty}^y f(x)\,\mathrm{d}x\right) \\
&= \int_0^\infty \mathrm{d}y\, \mathbb{P}(X \geq y)
\end{aligned}$$

□

Here is an alternative proof.

Proof. For every $\omega \in \Omega$, we can write

$$X(\omega) = \int_0^\infty 1(X(\omega) \geq x)\,\mathrm{d}x$$

Taking expectations, we get

$$\mathbb{E}[X] = \mathbb{E}\left[\int_0^\infty 1(X(\omega) \geq x)\,\mathrm{d}x\right]$$

We will interchange the integral and the expectation, although this step is not justified or rigorous.

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty \mathbb{E}[1(X(\omega) \geq x)]\,\mathrm{d}x \\
&= \int_0^\infty \mathbb{P}(X \geq x)\,\mathrm{d}x
\end{aligned}$$

□

12.5. Variance

We define the variance of a continuous random variable as

$$\operatorname{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$

12.6. Uniform distribution

Consider the uniform distribution defined by $a, b \in \mathbb{R}$.

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

We write $X \sim U[a,b]$. For some $x \in [a,b]$, we can write

$$\mathbb{P}(X \leq x) = \int_a^x f(y)\,\mathrm{d}y = \frac{x-a}{b-a}$$

Hence, for $x \in [a,b]$,

$$F(x) = \begin{cases} 1 & x > b \\ \frac{x-a}{b-a} & x \in [a,b] \\ 0 & x < a \end{cases}$$

Then,

$$\mathbb{E}[X] = \int_a^b \frac{x}{b-a}\,\mathrm{d}x = \frac{a+b}{2}$$

12.7. Exponential distribution

The exponential distribution is defined by $f(x) = \lambda e^{-\lambda x}$ for $\lambda > 0$, $x > 0$. We write $X \sim \operatorname{Exp}(\lambda)$.

$$F(x) = \mathbb{P}(X \leq x) = \int_0^x \lambda e^{-\lambda y}\,\mathrm{d}y = 1 - e^{-\lambda x}$$

Further,

$$\mathbb{E}[X] = \int_0^\infty \lambda x e^{-\lambda x}\,\mathrm{d}x = \frac{1}{\lambda}$$

We can view the exponential distribution as a limit of geometric distributions. Suppose that $T \sim \operatorname{Exp}(\lambda)$, and let $T_n = \lfloor nT \rfloor$ for all $n \in \mathbb{N}$. We have

$$\mathbb{P}(T_n \geq k) = \mathbb{P}\left(T \geq \frac{k}{n}\right) = e^{-\lambda k/n} = \left(e^{-\lambda/n}\right)^k$$

Hence T_n is a geometric distribution with parameter $p_n = e^{-\lambda/n}$. As $n \to \infty$, $p_n \sim \frac{\lambda}{n}$, and $\frac{T_n}{n} \sim T$. Hence the exponential distribution is the limit of a scaled version of the geometric

distribution. A key property of the exponential distribution is that it has no memory. If $T \sim \text{Exp}(\lambda)$, $\mathbb{P}(T > t + s \mid T > s) = \mathbb{P}(T > t)$. In fact, the distribution is uniquely characterised by this property.

Proposition. Let T be a positive continuous random variable not identically zero or infinity. Then T has the memoryless property $\mathbb{P}(T > t + s \mid T > s) = \mathbb{P}(T > t)$ if and only if $T \sim \text{Exp}(\lambda)$ for some $\lambda > 0$.

Proof. Clearly if $T \sim \text{Exp}(\lambda)$, then $\mathbb{P}(T > t + s \mid T > s) = e^{-\lambda t} = \mathbb{P}(T > t)$ as required. Now, given that T has this memoryless property, for all s and t, we have $\mathbb{P}(T > t + s) = \mathbb{P}(T > t)\mathbb{P}(T > s)$. Let $g(t) = \mathbb{P}(T > t)$; we would like to show that $g(t) = e^{-\lambda t}$. Then g satisfies $g(t+s) = g(t)g(s)$. Then for all $m \in \mathbb{N}$, $g(mt) = (g(t))^m$. Setting $t = 1$, $g(m) = g(1)^m$. Now, $g(m/n)^n = g(mn/n) = g(m)$ hence $g(m/n) = g(1)^{m/n}$. So for all rational numbers $q \in \mathbb{Q}$, $g(q) = g(1)^q$.

Now, $g(1) = \mathbb{P}(T > 1) \in (0, 1)$. Indeed, $g(1) \neq 0$ since in this case, for any rational number q we would have $g(q) = 0$ contradicting the assumption that T was not identically zero, and $g(1) \neq \infty$ because in this case T would be identically infinity. Now, let $\lambda = -\log \mathbb{P}(T > 1) > 0$. We have now proven that $g(t) = e^{-\lambda t}$ for all $t \in \mathbb{Q}$.

Let $t \in \mathbb{R}_+$. Then for all $\varepsilon > 0$, there exist $r, s \in \mathbb{Q}$ such that $r \leq t \leq s$ and $|r - s| \leq \varepsilon$. In this case, $e^{-\lambda s} = \mathbb{P}(T > s) \leq \mathbb{P}(T > t) \leq \mathbb{P}(T > r) = e^{-\lambda r}$. Sending $\varepsilon \to 0$ finishes the proof, showing that $g(t) = e^{-\lambda t}$ for all positive reals. □

12.8. Functions of continuous random variables

Theorem. Suppose that X is a continuous random variable with density f. Let g be a monotonic continuous function (either strictly increasing or strictly decreasing), such that g^{-1} is differentiable. Then $g(X)$ is a continuous random variable with density $fg^{-1}(x)\left|\frac{\mathrm{d}}{\mathrm{d}x}g^{-1}(x)\right|$.

Proof. Suppose that g is strictly increasing. We have

$$\mathbb{P}(g(X) \leq x) = \mathbb{P}\left(X \leq g^{-1}(x)\right) = F(g^{-1}(x))$$

Hence,

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(g(X) \leq x) = F'(g^{-1}(x)) \cdot \frac{\mathrm{d}}{\mathrm{d}x}g^{-1}(x) = f(g^{-1}(x))\frac{\mathrm{d}}{\mathrm{d}x}g^{-1}(x)$$

Note that since g is strictly increasing, so is g^{-1}. Now, suppose the g is strictly decreasing. Since the random variable is continuous,

$$\mathbb{P}(g(X) \leq x) = \mathbb{P}\left(X \geq g^{-1}(x)\right) = 1 - F(g^{-1}(x))$$

Hence,

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(g(X) \leq x) = -F'(g^{-1}(x)) \cdot \frac{\mathrm{d}}{\mathrm{d}x}g^{-1}(x) = f(g^{-1}(x))\left|\frac{\mathrm{d}}{\mathrm{d}x}g^{-1}(x)\right|$$

Likewise, in this case, g is strictly decreasing. □

12.9. Normal distribution

The normal distribution is characterised by $\mu \in \mathbb{R}$ and $\sigma > 0$. We define

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$f(x)$ is indeed a probability density function:

$$I = \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\mathrm{d}x$$

Applying the substitution $x \mapsto \frac{x-\mu}{\sigma}$, we have

$$I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{x^2}{2}\right\}\mathrm{d}x$$

We can evaluate this integral by considering I^2.

$$I^2 = \frac{2}{\pi} \int_0^{\infty} \int_0^{\infty} e^{\frac{-(u^2+v^2)}{2}}\,\mathrm{d}u\,\mathrm{d}v$$

Using polar coordinates $u = r\cos\theta$ and $v = r\sin\theta$, we have

$$I^2 = \frac{2}{\pi} \int_0^{\infty} \mathrm{d}r \int_0^{\frac{\pi}{2}} \mathrm{d}\theta\, r e^{-\frac{r^2}{2}} = 1 \implies I = \pm 1$$

But clearly $I > 0$, so $I = 1$. Hence f really is a probability density function. Now, if $X \sim \mathrm{N}(\mu, \sigma^2)$,

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\mathrm{d}x \\
&= \underbrace{\int_{-\infty}^{\infty} \frac{x-\mu}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\mathrm{d}x}_{\text{odd function around } \mu \text{ hence } 0} + \mu \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\mathrm{d}x}_{I=1 \text{ by above}} \\
&= \mu
\end{aligned}$$

We can also compute the variance, using the substitution $u = \frac{x-\mu}{\sigma}$, giving

$$\begin{aligned}
\mathrm{Var}(X) &= \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\mathrm{d}x \\
&= \sigma^2 \int_{-\infty}^{\infty} \frac{u^2}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2}\right\}\mathrm{d}u \\
&= \sigma^2
\end{aligned}$$

In particular, when $\mu = 0$ and $\sigma^2 = 1$, we call the distribution $\mathrm{N}(\mu, \sigma^2) = \mathrm{N}(0, 1)$ the standard normal distribution. We define

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \, \mathrm{d}u\,; \quad \phi(x) = \Phi'(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Hence $\Phi(x) = \mathbb{P}(X \leq x)$ if X has the standard normal distribution. Since $\phi(x) = \phi(-x)$, we have $\Phi(x) + \Phi(-x) = 1$, hence $\mathbb{P}(X \leq x) = 1 - \mathbb{P}(X \leq -x)$.

13. Multivariate density functions

13.1. Standardising normal distributions

Suppose $X \sim \mathrm{N}(\mu, \sigma^2)$. Let $a \neq 0, b \in \mathbb{R}$, and let $g(x) = ax+b$. We define $Y = g(X) = aX+b$. We can find the density f_Y of Y, by noting that g is a monotonic function and the inverse has a derivative. We can then use the theorem in the last lecture to show that

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \cdot \left|\frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)\right| \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\frac{y-b}{a} - \mu)^2}{2\sigma^2}\right) \cdot \frac{1}{2a} \\ &= \frac{1}{\sqrt{2\pi a^2\sigma^2}} \exp\left(-\frac{(y - a\mu + b)^2}{2a^2\sigma^2}\right) \end{aligned}$$

Hence $Y \sim \mathrm{N}(a\mu + b, a^2\sigma^2)$. In particular, $\frac{X-\mu}{\sigma}$ is exactly the standard normal distribution.

Definition. Suppose X is a continuous random variable. Then the median of X, denoted by m, is the number satisfying

$$\mathbb{P}(X \leq m) = \mathbb{P}(X \geq m) = \frac{1}{2}$$

If $X \sim \mathrm{N}(\mu, \sigma^2)$, then $\mathbb{P}(X \leq \mu) = \Phi(0) = \frac{1}{2}$ hence μ is the median of the normal distribution.

13.2. Multivariate density functions

Suppose $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ is a random variable. We say that X has density f if

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} f(y_1, \dots, y_n)\, \mathrm{d}y_1 \dots \mathrm{d}y_n$$

Then,

$$f(x_1, \dots, x_n) = \frac{\partial^n}{\partial x_1 \dots \partial x_n} F(x_1, \dots, x_n)$$

This generalises the fact that for all (reasonable) $B \subseteq \mathbb{R}^n$,

$$\mathbb{P}((X_1, \dots, X_n) \in B) = \int_B f(y_1, \dots, y_n)\, \mathrm{d}y_1 \dots \mathrm{d}y_n$$

13.3. Independence of events

In the continuous case, we can no longer use the definition

$$\mathbb{P}(X = a, Y = b) = \mathbb{P}(X = a)\,\mathbb{P}(Y = b)$$

since the probability of a random variable being a specific value is always zero. Instead, we define that $X_1, \dots, X_n$ are independent if for all $x_1, \dots, x_n \in \mathbb{R}$,

$$\mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n) = \mathbb{P}(X_1 \le x_1) \cdots \mathbb{P}(X_n \le x_n)$$

Theorem. Suppose $X = (X_1, \dots, X_n)$ has density f.

(a) Suppose $X_1, \dots, X_n$ are independent with densities $f_1, \dots, f_n$. Then $f(x_1, \dots, x_n) = f_1(x_1) \cdots f_n(x_n)$.

(b) Suppose that f factorises as $f(x_1, \dots, x_n) = f_1(x_1) \cdots f_n(x_n)$ for some non-negative functions $f_1, \dots, f_n$. Then $X_1, \dots, X_n$ are independent with densities proportional to $f_1, \dots, f_n$. (In order to have a density function, we require that it integrates to 1, so we choose a scaling factor such that this requirement holds.)

In other words, f factorises if and only if it is comprised of independent events.

Proof. (a) We know that

$$\begin{aligned}\mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n) &= \mathbb{P}(X_1 \le x_1) \cdots \mathbb{P}(X_n \le x_n) \\ &= \int_{-\infty}^{x_1} f_1(y_1)\,\mathrm{d}y_1 \cdots \int_{-\infty}^{x_n} f_n(y_n)\,\mathrm{d}y_n \\ &= \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} \prod_{i=1}^{n} f_i(y_i)\,\mathrm{d}y_i\end{aligned}$$

So the density of $(X_1, \dots, X_n)$ is the product of the (f_i).

(b) Suppose f factorises. Let $B_1, \dots, B_n \subseteq \mathbb{R}$. Then

$$\mathbb{P}(X_1 \in B_1, \dots, X_n \in B_n) = \int_{B_1} \cdots \int_{B_n} f_1(x_1) \cdots f_n(x_n)\,\mathrm{d}y_1 \cdots \mathrm{d}y_n$$

Now, let $B_j = \mathbb{R}$ for all $j \ne i$. Then

$$\mathbb{P}(X_i \in B_i) = \mathbb{P}\left(X_i \in B_i, X_j \in B_j\ \forall j \ne i\right) = \int_{B_i} f_i(y_i)\,\mathrm{d}y_i \cdot \prod_{j \ne 1} \int_{B_j} f_j(x_j)\,\mathrm{d}y_j$$

Since f is a density function,

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \dots, x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n = 1$$

But f is the product of the f_i, so

$$\prod_j \int_{-\infty}^{\infty} f_j(y)\,\mathrm{d}y = 1 \implies \prod_{j\neq i} \int_{-\infty}^{\infty} f_j(y)\,\mathrm{d}y = \frac{1}{\int_{-\infty}^{\infty} f_i(y)\,\mathrm{d}y}$$

Hence,

$$\mathbb{P}(X_i \in B_i) = \frac{\int_{B_i} f_i(y)\,\mathrm{d}y}{\int_{-\infty}^{\infty} f_i(y)\,\mathrm{d}y}$$

This shows that the density of X_i is

$$\frac{f_i}{\int_{-\infty}^{\infty} f_i(y)\,\mathrm{d}y}$$

The X_i are independent, since

$$\begin{aligned}\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) &= \frac{\int_{-\infty}^{x_1} f_1(y_1)\,\mathrm{d}y_1 \cdots \int_{-\infty}^{x_n} f_n(y_n)\,\mathrm{d}y_n}{\int_{-\infty}^{\infty} f_1(y_1)\,\mathrm{d}y_1 \cdots \int_{-\infty}^{\infty} f_n(y_n)\,\mathrm{d}y_n} \\ &= \mathbb{P}(X_1 \leq x_1) \cdots \mathbb{P}(X_n \leq x_n)\end{aligned}$$

□

13.4. Marginal density

Suppose that $(X_1, \dots, X_n)$ has density f. Now we can compute the marginal density as follows.

$$\begin{aligned}\mathbb{P}(X_1 \leq x) &= \mathbb{P}(X_1 \leq x, X_2 \in \mathbb{R}, \dots, X_n \in \mathbb{R}) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \dots, x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= \int_{-\infty}^{x} \mathrm{d}x_1 \underbrace{\left(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \dots, x_n)\,\mathrm{d}x_2 \cdots \mathrm{d}x_n \right)}_{\text{marginal density of } X_1}\end{aligned}$$

13.5. Sum of random variables

Recall that in the discrete case, for independent random variables X and Y we have

$$\begin{aligned}\mathbb{P}(X + Y = z) &= \sum_y \mathbb{P}(X + Y = z, Y = y) \\ &= \sum_y \mathbb{P}(X = z - y)\,\mathbb{P}(Y = y) \\ &= \sum_y p_x(z - y) p_y(y)\end{aligned}$$

which was called the convolution. In the continuous case,

$$\begin{aligned}
\mathbb{P}(X+Y\le z) &= \iint_{\{x+y\le z\}} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{z-x} f_X(x)f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{z} f_X(x)f_Y(y-x)\,\mathrm{d}y\right)\mathrm{d}x \ \text{ (using } y\mapsto y+x) \\
&= \int_{-\infty}^{z}\mathrm{d}y\,\underbrace{\left(\int_{-\infty}^{\infty} f_Y(y-x)f_X(x)\,\mathrm{d}x\right)}_{g(y)}
\end{aligned}$$

Hence the density of $X+Y$ is $g(y)$, where

$$g(y)=\int_{-\infty}^{\infty} f_Y(y-x)f_X(x)\,\mathrm{d}x$$

Definition. Let f, g be density functions. Then the convolution of f and g is

$$(f\star g)(y)=\int_{-\infty}^{\infty} f_Y(y-x)f_X(x)\,\mathrm{d}x$$

Here is a non-rigorous argument, which can be used as a heuristic.

$$\begin{aligned}
\mathbb{P}(X+Y\le z) &= \int_{-\infty}^{\infty}\mathbb{P}(X+Y\le z, Y\in\mathrm{d}y) \\
&= \int_{-\infty}^{\infty}\mathbb{P}(X+Y\le z, Y\in\mathrm{d}y) \\
&= \int_{-\infty}^{\infty}\mathbb{P}(X\le z-y)\,\mathbb{P}(Y\in\mathrm{d}y) \\
&= \int_{-\infty}^{\infty}\mathbb{P}(X\le z-y)\,f_Y(y)\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}F_X(z-y)f_Y(y)\,\mathrm{d}y \\
\frac{\mathrm{d}}{\mathrm{d}z}\mathbb{P}(X+Y\le z) &= \int_{-\infty}^{\infty}\frac{\mathrm{d}}{\mathrm{d}z}F_X(z-y)f_Y(y)\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}f_X(z-y)f_Y(y)\,\mathrm{d}y
\end{aligned}$$

13.6. Conditional density

We will now define the conditional density of a continuous random variable, given the value of another continuous random variable. Let X and Y be continuous random variables with joint density $f_{X,Y}$ and marginal densities f_X and f_Y. Then we define the conditional density of X given that $Y = y$ is defined as

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Then we can find the law of total probability in the continuous case.

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x, y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} f_{X|Y}(x \mid y) f_Y(y)\,\mathrm{d}y \end{aligned}$$

13.7. Conditional expectation

We want to define $\mathbb{E}[X \mid Y]$ to be some function $g(Y)$ for some function g. We will define

$$g(y) = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\,\mathrm{d}x$$

which is the analogous expression to $\mathbb{E}[X \mid Y = y]$ from the discrete case. Then we just set $\mathbb{E}[X \mid Y] = g(Y)$ to be the conditional expectation.

13.8. Transformations of multidimensional random variables

Theorem. Let X be a continuous random variable with values in $D \subseteq \mathbb{R}^d$, with density f_X. Now, let g be a bijection D to $g(D)$ which has a continuous derivative, and $\det g'(x) \neq 0$ for all $x \in D$. Then the random variable $Y = g(X)$ has density

$$f_Y(y) = f_X(x) \cdot |J| \text{ where } x = g^{-1}(y)$$

where J is the Jacobian

$$J = \det\left(\left(\frac{\partial x_i}{\partial y_j}\right)_{i,j=1}^{d}\right)$$

No proof will be given for this theorem. As an example, let X and Y be independent continuous random variables with the standard normal distribution. The point (X, Y) in $\mathbb{R}^2$ has polar coordinates (R, Θ). What are the densities of R and Θ? We have $X = R\cos\Theta$ and $Y = R\sin\Theta$. The Jacobian is

$$J = \det\begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} = r$$

Hence,

$$\begin{aligned} f_{R,\Theta}(r,\theta) &= f_{X,Y}(r\cos\theta, r\sin\theta)|J| \\ &= f_{X,Y}(r\cos\theta, r\sin\theta)r \\ &= f_X(r\cos\theta)f_Y(r\sin\theta)r \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{r^2\cos^2\theta}{2}} \cdot \frac{1}{\sqrt{2\pi}}e^{-\frac{r^2\sin^2\theta}{2}} \cdot r \\ &= \frac{1}{2\pi}e^{-\frac{r^2}{2}} \cdot r \end{aligned}$$

for all $r > 0$ and $\theta \in [0, 2\pi]$. Note that the joint density factorises into marginal densities:

$$f_{R,\Theta}(r,\theta) = \underbrace{\frac{1}{2\pi}}_{f_\Theta} \underbrace{re^{-\frac{r^2}{2}}}_{f_R}$$

so the random variables R and Θ are independent, where $\Theta \sim U[0, 2\pi]$ and R has density $re^{\frac{-r^2}{2}}$ on $(0, \infty)$.

13.9. Order statistics of a random sample

Let $X_1, \dots, X_n$ be independent and identically distributed random variables with distribution function F and density function f. We can put them in increasing order:

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

and let $Y_i = X_{(i)}$. The (Y_i) are the order statistics.

$$\begin{aligned} \mathbb{P}(Y_1 \leq x) &= \mathbb{P}(\min(X_1, \dots, X_n) \leq x) \\ &= 1 - \mathbb{P}(\min(X_1, \dots, X_n) > x) \\ &= 1 - \mathbb{P}(X_1 > x) \cdots \mathbb{P}(X_n > x) \\ &= 1 - (1 - F(x))^n \end{aligned}$$

Further,

$$\begin{aligned} f_{Y_1}(x) &= \frac{\mathrm{d}}{\mathrm{d}x}(1 - (1 - F(x))^n) \\ &= n(1 - F(x))^{n-1} f(x) \end{aligned}$$

We can compute an analogous result for the maximum.

$$\begin{aligned} \mathbb{P}(Y_n \leq x) &= (F(x))^n \\ f_{Y_n}(x) &= n(F(x))^{n-1} f(x) \end{aligned}$$

What are the densities of the other random variables? First, let $x_1 < x_2 < \cdots < x_n$. Then, we can first find the joint distribution $\mathbb{P}(Y_1 \leq x_1, \dots, Y_n \leq x_n)$. Note that this is simply the sum

over all possible permutations of the (X_i) of $\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n)$. But since the variables are independent and identically distributed, these probabilities are the same. Hence,

$$\begin{aligned}\mathbb{P}(Y_1 \leq x_1, \dots, Y_n \leq x_n) &= n! \cdot \mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n, X_1 < \cdots < X_n) \\ &= n! \int_{-\infty}^{x_1} \int_{u_1}^{x_2} \cdots \int_{u_{n-1}}^{x_n} f(u_1) \cdots f(u_n) \, \mathrm{d}u_1 \cdots \mathrm{d}u_n \\ \therefore f_{Y_1,\dots,Y_n}(x_1, \dots, x_n) &= n! f(x_1) \cdots f(x_n)\end{aligned}$$

when $x_1 < x_2 < \cdots < x_n$, and the joint density is zero otherwise. Note that this joint density does not factorise as a product of densities, since we must always consider the indicator function that $x_1 < x_2 < \cdots < x_n$.

13.10. Order statistics on exponential distribution

Let $X \sim \mathrm{Exp}(\lambda)$, $Y \sim \mathrm{Exp}(\mu)$ be independent continuous random variables. Let $Z = \min(X, Y)$.

$$\mathbb{P}(Z \geq z) = \mathbb{P}(X \geq z, Y \geq z) = \mathbb{P}(X \geq z)\,\mathbb{P}(Y \geq z) = e^{-\lambda z} \cdot e^{-\mu z} = e^{-(\lambda+\mu)z}$$

Hence Z has the exponential distribution with parameter $\lambda + \mu$. More generally, if $X_1, \dots, X_n$ are independent continuous random variables with $X_i \sim \mathrm{Exp}(\lambda_i)$, then $Z = \min(X_1, \dots, X_n)$ has distribution $\mathrm{Exp}\left(\sum_{i=1}^n \lambda_i\right)$. Now, let $X_1, \dots, X_n$ be independent identically distributed random variables with distribution $\mathrm{Exp}(\lambda)$, and let Y_i be their order statistics. Then

$$Z_1 = Y_1; \quad Z_2 = Y_2 - Y_1; \quad Z_i = Y_i - Y_{i-1}$$

So the Z_i are the 'durations between consecutive results' from the X_i. What is the density of these Z_i? First, note that

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} = A \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}; \quad A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

Note that $\det A = 1$, and $Z = AY$, and note further that

$$y_j = \sum_{i=1}^{j} z_i$$

Now,

$$\begin{aligned} f_{(Z_1,\dots,Z_n)}(z_1,\dots,z_n) &= f_{(Y_1,\dots,Y_n)}(y_1,\dots,y_n)\underbrace{|A|}_{=1} \\ &= n!f(y_1)\cdots f(y_n) \\ &= n!(\lambda e^{-\lambda y_1})\cdots(\lambda e^{-\lambda y_n}) \\ &= n!\lambda^n e^{-\lambda(nz_1+(n-1)z_2+\cdots+z_n)} \\ &= \prod_{i=1}^{n}(n-i+1)\lambda e^{-\lambda(n-i+1)z_i} \end{aligned}$$

The density function of the vector Z factorises into functions of the z_i, so $Z_1,\dots,Z_n$ are independent and $Z_i \sim \text{Exp}(\lambda(n-i+1))$.

14. Moment generating functions

14.1. Moment generating functions

Consider a continuous random variable X with density f. Then the moment generating function of X is defined as

$$m(\theta) = \mathbb{E}\left[e^{\theta X}\right] = \int_{-\infty}^{\infty} e^{\theta x} f(x)\,\mathrm{d}x$$

whenever this integral is finite. Note that $m(0) = 1$.

Theorem. The moment generating function uniquely determines the distribution of a continuous random variable, provided that it is defined on some open interval (a, b) of values of θ.

No proof will be given.

Theorem. Suppose the moment generating function is defined on an open interval of values of θ. Then

$$\frac{\mathrm{d}^r}{\mathrm{d}\theta^r} m(\theta)\bigg|_{\theta=0} = \mathbb{E}\left[X^r\right]$$

Theorem. Suppose $X_1, \dots, X_n$ are independent random variables. Then

$$m(\theta) = \mathbb{E}\left[e^{\theta(X_1+\cdots+X_n)}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{\theta X_i}\right]$$

Proof. Since the X_i are independent, we can move the product outside of the expectation. □

14.2. Gamma distribution

Let X be a random variable with density

$$f(x) = e^{-\lambda x} \frac{\lambda^n x^{n-1}}{(n-1)!}$$

where $\lambda > 0$, $n \in \mathbb{N}$, $x \geq 0$. We can say that $X \sim \Gamma(n, \lambda)$. First, we check that f is indeed a density.

$$\begin{aligned} I_n &= \int_0^\infty f(x)\,\mathrm{d}x \\ &= \int_0^\infty \lambda e^{-\lambda x} \frac{\lambda^n x^{n-1}}{(n-1)!}\,\mathrm{d}x \\ &= \int_0^\infty \frac{e^{-\lambda x}\lambda^{n-1}(n-1)x^{n-2}}{(n-1)!}\,\mathrm{d}x \\ &= \int_0^\infty \frac{e^{-\lambda x}\lambda^{n-1}x^{n-2}}{(n-2)!}\,\mathrm{d}x \\ &= I_{n-1} = \cdots = I_1 \end{aligned}$$

Note that for $n = 1$, $f(x) = \lambda e^{-\lambda x}$ which is the density of the exponential distribution. Therefore, $I_n = 1$ as required, so f really is a density. Now,

$$m(\theta) = \int_0^\infty \frac{e^{\theta x} e^{-\lambda x}\lambda^n x^{n-1}}{(n-1)!}\,\mathrm{d}x$$

If $\lambda > \theta$, then we have a finite integral. If $\lambda \leq \theta$, then the exponential term $e^{\theta x}$ will dominate and we will have an infinite integral. So, let $\lambda > \theta$.

$$\begin{aligned} m(\theta) &= \int_0^\infty \frac{e^{\theta x} e^{-\lambda x}\lambda^n x^{n-1}}{(n-1)!}\,\mathrm{d}x \\ &= \left(\frac{\lambda}{\lambda-\theta}\right)^n \int_0^\infty \frac{e^{-(\lambda-\theta)x}(\lambda-\theta)^n x^{n-1}}{(n-1)!}\,\mathrm{d}x \end{aligned}$$

The integral on the right hand side is the probability distribution function of a random variable $Y \sim \Gamma(n, \lambda - \theta)$, which gives 1 since the integral is taken over the entire domain. Hence,

$$m(\theta) = \left(\frac{\lambda}{\lambda-\theta}\right)^n$$

Now, let $X \sim \Gamma(n, \lambda)$ and $Y \sim \Gamma(m, \lambda)$ be independent continuous random variables. Then

$$m(\theta) = \mathbb{E}\left[e^{\theta(X+Y)}\right] = \mathbb{E}\left[e^{\theta X}\right]\mathbb{E}\left[e^{\theta Y}\right] = \left(\frac{\lambda}{\lambda-\theta}\right)^{n+m}$$

So by the uniqueness property we saw earlier, we get that $X + Y \sim \Gamma(n + m, \lambda)$. In particular, this implies that if $X_1, \ldots, X_n$ are independent and identically distributed with the distribution $\mathrm{Exp}(\lambda) = \Gamma(1, \lambda)$, then

$$X_1 + \cdots + X_n \sim \Gamma(n, \lambda)$$

We could alternatively consider $\Gamma(\alpha, \lambda)$ for $\alpha > 0$ by replacing $(n-1)!$ with

$$\Gamma(\alpha) = \int_0^\infty e^{-x} x^{\alpha-1}\,\mathrm{d}x$$

which agrees with this factorial function for integer values of α.

14.3. Moment generating function of the normal distribution

Recall that

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Now,

$$m(\theta) = \int_0^\infty e^{\theta x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x = \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\theta x - \frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x$$

Note that

$$\theta x - \frac{(x-\mu)^2}{2\sigma^2} = \theta\mu + \frac{\theta^2\sigma^2}{2} - \frac{\left(x - (\mu + \theta\sigma^2)\right)^2}{2\sigma^2}$$

Hence,

$$\begin{aligned} m(\theta) &= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\theta\mu + \frac{\theta^2\sigma^2}{2} - \frac{\left(x - (\mu + \theta\sigma^2)\right)^2}{2\sigma^2}\right) \mathrm{d}x \\ &= \exp\left(\theta\mu + \frac{\theta^2\sigma^2}{2}\right) \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x - (\mu + \theta\sigma^2)\right)^2}{2\sigma^2}\right) \mathrm{d}x \end{aligned}$$

Note that the integral on the right hand side has the form of the probability distribution function of a variable $Y \sim \mathrm{N}(\mu + \theta\sigma^2, \sigma^2)$, hence it integrates to 1.

$$m(\theta) = \exp\left(\theta\mu + \frac{\theta^2\sigma^2}{2}\right)$$

Recall that if $X \sim \mathrm{N}(\mu, \sigma^2)$, then $aX + b \sim \mathrm{N}(a\mu + b, a^2\sigma^2)$. We can then deduce that

$$\mathbb{E}\left[e^{\theta(aX+b)}\right] = \exp\left(\theta(a\mu + b) + \frac{\theta^2 a^2\sigma^2}{2}\right)$$

Now, suppose that $X \sim \mathrm{N}(\mu, \sigma^2)$ and $Y \sim \mathrm{N}(\nu, \tau^2)$ are independent. Then

$$\begin{aligned} \mathbb{E}\left[e^{\theta(X+Y)}\right] &= \mathbb{E}\left[e^{\theta X}\right] \mathbb{E}\left[e^{\theta Y}\right] \\ &= \exp\left(\theta\mu + \frac{\theta^2\sigma^2}{2}\right) \exp\left(\theta\nu + \frac{\theta^2\tau^2}{2}\right) \\ &= \exp\left(\theta(\mu + \nu) + \frac{\theta^2(\sigma^2 + \tau^2)}{2}\right) \end{aligned}$$

Hence $X + Y \sim \mathrm{N}(\mu + \nu, \sigma^2 + \tau^2)$.

14.4. Cauchy distribution

Suppose that a continuous random variable X has density

$$f(x) = \frac{1}{\pi(1+x^2)}$$

where $x \in \mathbb{R}$. Now,

$$m(\theta) = \mathbb{E}\left[e^{\theta X}\right] = \int_{-\infty}^{\infty} \frac{e^{\theta x}}{\pi(1+x^2)} = \begin{cases} \infty & \theta \neq 0 \\ 1 & \theta = 0 \end{cases}$$

Suppose $X \sim f$. Then $X, 2X, 3X, \ldots$ have the same moment generating function, but they do not have the same distribution. This is because $m(\theta)$ is not finite on an open interval.

14.5. Multivariate moment generating functions

Let $X = (X_1, \ldots, X_n)$ be a random variable with values in $\mathbb{R}^n$. Then the moment generating function of X is defined as

$$m(\theta) = \mathbb{E}\left[e^{\theta^\intercal X}\right] = \mathbb{E}\left[e^{\theta_1 X_1 + \cdots + \theta_n X_n}\right]; \quad \theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_n \end{pmatrix}$$

Theorem. If the moment generating function is finite for a range of values of θ, it uniquely determines the distribution of X. Also,

$$\left.\frac{\partial^r m}{\partial \theta_i^r}\right|_{\theta=0} = \mathbb{E}\left[X_i^r\right]$$

and

$$\left.\frac{\partial^{r+s} m}{\partial \theta_i^r \partial \theta_j^s}\right|_{\theta=0} = \mathbb{E}\left[X_i^r X_j^s\right]$$

Further,

$$m(\theta) = \prod_{i=1}^{n} \mathbb{E}\left[e^{\theta_i X_i}\right]$$

if and only if $X_1, \ldots, X_n$ are independent.

No proof is provided.

15. Limit theorems

15.1. Convergence in distribution

Definition. Let $(X_n : n \in \mathbb{N})$ be a sequence of random variables and let X be another random variable. We say that X_n converges to X in distribution, written $X_n \xrightarrow{d} X$, if

$$F_{X_n}(x) \to F_X(x)$$

for all $x \in \mathbb{R}$ that are continuity points of F_X.

Theorem (Continuity property for moment generating functions)**.** Let X be a continuous random variable with $m(\theta) < \infty$ for some $\theta \neq 0$. Suppose that $m_n(\theta) \to m(\theta)$ for all $\theta \in \mathbb{R}$, where $m_n(\theta) = \mathbb{E}\left[e^{\theta X_n}\right]$, and $m(\theta) = \mathbb{E}\left[e^{\theta X}\right]$. Then $X_n \xrightarrow{d} X$.

15.2. Weak law of large numbers

Theorem. Let $(X_n : n \in \mathbb{N})$ be a sequence of independent and identically distributed random variables, with $\mu = \mathbb{E}[X_1] < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then for all $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \to 0$$

as $n \to \infty$.

We will give a proof assuming that the variance of X_1 is finite.

Proof. By Chebyshev's inequality,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) &= \mathbb{P}\left(|S_n - n\mu| > \varepsilon n\right)\\ &\leq \frac{\operatorname{Var}(S_n)}{\varepsilon^2 n^2}\\ &= \frac{n\sigma^2}{\varepsilon^2 n^2}\\ &\to 0\end{aligned}$$

□

15.3. Types of convergence

Definition. A sequence (X_n) converges to X *in probability*, written $X_n \xrightarrow{\mathbb{P}} X$ as $n \to \infty$ if for all $\varepsilon > 0$,

$$\mathbb{P}\left(|X_n - X| > \varepsilon\right) \to 0; \quad n \to \infty$$

Definition. A sequence (X_n) converges to X *almost surely* (with probability 1), if

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1$$

This second definition is a stronger form of convergence. If a sequence (X_n) converges to zero almost surely, then $X_n \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$.

Proof. We want to show that given any $\varepsilon > 0$, $\mathbb{P}(|X_n| > \varepsilon) \to 0$ as $n \to \infty$, or equivalently, $\mathbb{P}(|X_n| \leq \varepsilon) \to 1$.

$$\mathbb{P}(|X_n| \leq \varepsilon) \geq \mathbb{P}\left(\underbrace{\bigcap_{m=n}^{\infty} \{|X_m| \leq \varepsilon\}}_{A_n}\right)$$

Note that A_n is an increasing sequence of events, and

$$\bigcup_n A_n = \{|X_m| \leq \varepsilon \text{ for all } m \text{ sufficiently large}\}$$

Hence, as $n \to \infty$,

$$\mathbb{P}(A_n) \to \mathbb{P}\left(\bigcup A_n\right)$$

Therefore,

$$\lim_{n\to\infty} \mathbb{P}(|X_n| \leq \varepsilon) \geq \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup A_n\right) \geq \mathbb{P}\left(\lim_{n\to\infty} X_n = 0\right)$$

Since X_n converges to zero almost surely, this event on the right hand side has probability 1, so in particular the limit on the left has probability 1, as required. □

15.4. Strong law of large numbers

Theorem. Let $(X_n)_{n\in\mathbb{N}}$ be an independent and identically distributed sequence of random variables, with $\mu = \mathbb{E}[X_1]$ finite. Let $S_n = X_1 + \cdots + X_n$. Then

$$\frac{S_n}{n} \to \mu \text{ as } n \to \infty \text{ almost surely}$$

In other words,

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} \to \mu\right) = 1$$

The following proof, made under the assumption of a finite fourth moment, is non-examinable. A proof can be formulated without this assumption, but it is more complicated.

Proof. Let $Y_i = X_i - \mu$. Then $\mathbb{E}[Y_i] = 0$, and $\mathbb{E}\left[Y_i^4\right] \leq 2^4(\mathbb{E}\left[X_i^4\right] + \mu^4) < \infty$. It then suffices to show that

$$\frac{S_n}{n} \to 0 \text{ a.s.}$$

where $S_n = \sum_1^n X_i$ and $\mathbb{E}[X_i] = 0$, $\mathbb{E}\left[X_i^4\right] < \infty$. First,

$$S_n^4 = \left(\sum_{i=1}^n X_i\right)^4 = \sum_{i=1}^n X_i^4 + \binom{4}{2}\sum_{i=1}^n X_i^2 X_j^2 + R$$

where R is a sum of terms of the form $X_i^2 X_j X_k$ or $X_i^3 X_j$ or $X_i X_j X_k X_\ell$ for i, j, k, l distinct. Once we take expectations, each term in R will have no contribution to the result, since they all contain an $\mathbb{E}[X_i] = 0$ term.

$$\begin{aligned}
\mathbb{E}\left[S_n^4\right] &= n\mathbb{E}\left[X_i^4\right] + \binom{4}{2}\frac{n(n-1)}{2}\mathbb{E}\left[X_i^2 X_j^2\right] + \mathbb{E}[R] \\
&= n\mathbb{E}\left[X_1^4\right] + 3n(n-1)\mathbb{E}\left[X_1^2\right]\mathbb{E}\left[X_1^2\right] \\
&\leq n\mathbb{E}\left[X_1^4\right] + 3n(n-1)\mathbb{E}\left[X_1^4\right] \\
&= 3n^2\mathbb{E}\left[X_1^4\right]
\end{aligned}$$

by Jensen's inequality. Now,

$$\mathbb{E}\left[\sum_{n=1}^\infty \left(\frac{S_n}{n}\right)^4\right] \leq \sum_{n=1}^\infty \frac{3}{n^2}\mathbb{E}\left[X_1^4\right] < \infty$$

Hence,

$$\sum_{n=1}^\infty \left(\frac{S_n}{n}\right)^4 < \infty \text{ with probability 1}$$

Then since the sum of infinitely many positive terms is finite, the terms must converge to zero.

$$\lim_{n\to\infty} \frac{S_n}{n} \to 0 \text{ a.s.}$$

□

15.5. Central limit theorem

Suppose, like before, that we have a sequence of independent and identically distributed random variables X_n, and suppose further that $\mathbb{E}[X_1] = \mu$, and $\operatorname{Var}(X_1) = \sigma^2 < \infty$.

$$\operatorname{Var}\left(\frac{S_n}{n} - \mu\right) = \frac{\sigma^2}{n}$$

We can normalise this new random variable $\frac{S_n}{n} - \mu$ by dividing by its standard deviation.

$$\frac{\frac{S_n}{n} - \mu}{\sqrt{\operatorname{Var}\left(\frac{S_n}{n} - \mu\right)}} = \frac{\frac{S_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{Sn - n\mu}{\sigma\sqrt{n}}$$

Theorem. For all $x \in \mathbb{R}$,

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) \to \Phi(x) = \int_{-\infty}^{x} \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}\,\mathrm{d}y$$

In other words,

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Z$$

where Z is the standard normal distribution.

Less formally, we might say that the central limit theorem shows that, for a large n,

$$S_n \approx n\mu + \sigma\sqrt{n}Z \sim N(n\mu, n\sigma^2)$$

Proof. Consider $Y_i = \frac{X_i - \mu}{\sigma}$. Then the Y_i have zero expectation and unit variance. It then suffices to prove the central limit theorem when the X_i have zero expectation and unit variance. We assume further that there exists $\delta > 0$ such that

$$\mathbb{E}\left[e^{\delta X_1}\right] < \infty; \quad \mathbb{E}\left[e^{-\delta X_1}\right] < \infty$$

We will show that

$$\frac{S_n}{n} \xrightarrow{d} \mathrm{N}(0, 1)$$

By the continuity property of moment generating functions, it is sufficient to show that for all $\theta \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{E}\left[e^{\theta\frac{S_n}{n}}\right] = \mathbb{E}\left[e^{\theta Z}\right] = e^{\frac{\theta^2}{2}}$$

Let $m(\theta) = \mathbb{E}\left[e^{\theta X_1}\right]$. Then

$$\mathbb{E}\left[e^{\theta\frac{S_n}{n}}\right] = \mathbb{E}\left[e^{\frac{\theta}{\sqrt{n}}X_1}\right]^n = \left(m\left(\frac{\theta}{\sqrt{n}}\right)\right)^n$$

We now need to show that

$$\lim_{n\to\infty}\left(m\left(\frac{\theta}{\sqrt{n}}\right)\right)^n = e^{\frac{\theta^2}{2}}$$

Now, let $|\theta| < \frac{\delta}{2}$. In this case,

$$\begin{aligned}
m(\theta) &= \mathbb{E}\left[e^{\theta X_1}\right] \\
&= \mathbb{E}\left[1 + \theta X_1 + \frac{\theta^2}{2}X_1^2 + \sum_{k=3}^{\infty}\frac{\theta^k}{k!}X_1^k\right] \\
&= \mathbb{E}[1] + \mathbb{E}[\theta X_1] + \mathbb{E}\left[\frac{\theta^2}{2}X_1^2\right] + \mathbb{E}\left[\sum_{k=3}^{\infty}\frac{\theta^k}{k!}X_1^k\right] \\
&= 1 + \frac{\theta^2}{2} + \mathbb{E}\left[\sum_{k=3}^{\infty}\frac{\theta^k}{k!}X_1^k\right]
\end{aligned}$$

Now, it suffices to prove that $\left|\mathbb{E}\left[\sum_{k=3}^{\infty}\frac{\theta^k}{k!}X_1^k\right]\right| = o(\theta^2)$ as $\theta \to 0$. Indeed, if we have this bound, then $m\left(\frac{\theta}{\sqrt{n}}\right) = 1 + \frac{\theta^2}{2n} + o\left(\frac{\theta^2}{n}\right)$, and hence $\lim_{n\to\infty}\left(m\left(\frac{\theta}{\sqrt{n}}\right)\right)^n = e^{\frac{\theta^2}{2}}$. To find this bound, we know that

$$\begin{aligned}\left|\mathbb{E}\left[\sum_{k=3}^{\infty}\frac{\theta^k}{k!}X_1^k\right]\right| &\leq \mathbb{E}\left[\sum_{k=3}^{\infty}\frac{|\theta|^k|X_1|^k}{k!}\right]\\ &= \mathbb{E}\left[|\theta X_1|^3\sum_{k=0}^{\infty}\frac{|\theta X_1|^k}{(k+3)!}\right]\\ &\leq \mathbb{E}\left[|\theta X_1|^3\sum_{k=0}^{\infty}\frac{|\theta X_1|^k}{k!}\right]\end{aligned}$$

Since $|\theta| \leq \frac{\delta}{2}$,

$$\mathbb{E}\left[|\theta X_1|^3\sum_{k=0}^{\infty}\frac{|\theta X_1|^k}{k!}\right] \leq \mathbb{E}\left[|\theta X_1|^3 e^{\frac{\delta}{2}|X_1|}\right]$$

Now,

$$|\theta X_1|^3 e^{\frac{\delta}{2}|X_1|} = |\theta|^3\frac{\left(\frac{\delta}{2}|X_1|\right)^3}{3!}\cdot\frac{3!}{\left(\frac{\delta}{2}\right)^3}\cdot e^{\frac{\delta}{2}|X_1|}$$

Note that

$$\frac{\left(\frac{\delta}{2}|X_1|\right)^3}{3!} \leq \sum_{k=0}^{\infty}\frac{\left(\frac{\delta}{2}|X_1|\right)^k}{k!} = e^{\frac{\delta}{2}|X_1|}$$

Hence,

$$|\theta X_1|^3 e^{\frac{\delta}{2}|X_1|} \leq |\theta|^3 e^{\frac{\delta}{2}|X_1|}\cdot\frac{3!}{\left(\frac{\delta}{2}\right)^3}\cdot e^{\frac{\delta}{2}|X_1|} = \frac{3!|\theta|^3}{\left(\frac{\delta}{2}\right)^3}e^{\delta|X_1|} = 3!\left(\frac{2|\theta|}{\delta}\right)^3 e^{\delta|X_1|}$$

Therefore,

$$e^{\delta|X_1|} \leq e^{\delta X_1} + e^{-\delta X_1}$$

So finally,

$$\mathbb{E}\left[|\theta X_1|^3\sum_{k=0}^{\infty}\frac{|\theta X_1|^k}{k!}\right] \leq 3!\left(\frac{2|\theta|}{\delta}\right)^3\mathbb{E}\left[e^{\delta X_1} + e^{-\delta X_1}\right] = o(|\theta|^2)$$

as $\theta \to 0$. □

15.6. Applications of central limit theorem

We can use the central limit theorem to approximate the binomial distribution using the normal distribution. Suppose that $S_n \sim \text{Bin}(n, p)$. Then $S_n = \sum_{i=1}^{n} X_i$, where the X_i have the Bernoulli distribution with parameter p. We know that $\mathbb{E}[S_n] = np$, and $\text{Var}(S_n) = np(1-p)$. Therefore, in particular,

$$S_n \approx \text{N}(np, np(1-p))$$

for n large. Note that we showed before that

$$\text{Bin}\left(n, \frac{\lambda}{n}\right) \to \text{Poi}(\lambda)$$

Note that with this approximation to the binomial, we let the parameter p depend on n. Since this is the case, we can no longer apply the central limit theorem, and we get a Poisson distributed approximation.

We can, however, use the central limit theorem to find a normal approximation for a Poisson random variable $S_n \sim \text{Poi}(n)$, since S_n can be written as $\sum_{i=1}^{n} X_i$ where the $X_i \sim \text{Poi}(1)$. Then

$$S_n \approx \text{N}(n, n)$$

15.7. Sampling error via central limit theorem

Suppose individuals independently vote 'yes' (with probability p) or 'no' (with probability $1-p$). We can sample the population to find an approximation for p. Pick N individuals at random, and let $\hat{p}_N = \frac{S_N}{N}$, where S_n is the number of individuals who voted 'yes'. We would like to find the minimum N such that $|\hat{p}_N - p| \leq 4\%$ with probability at least 99%. We have

$$S_N \sim \text{Bin}(N, p) \approx Np + \sqrt{Np(1-p)}Z; \quad Z \sim \text{N}(0, 1)$$

Hence,

$$\frac{S_N}{N} \approx p + \sqrt{\frac{p(1-p)}{N}}Z \implies |\hat{p}_N - p| \approx \sqrt{\frac{p(1-p)}{N}}|Z|$$

We then want to find N such that

$$\mathbb{P}\left(\sqrt{\frac{p(1-p)}{N}}|Z| \leq 0.04\right) \geq 0.99$$

We can compute this from the tables of the standard normal distribution. If $z = 2.58$, then $\mathbb{P}(|Z| \geq 2.58) = 0.01$, hence we need an N such that

$$0.04\sqrt{\frac{N}{p(1-p)}} \geq 2.58$$

In the worst case scenario, $p = \frac{1}{2}$ would give the largest N. So we need $N \geq 1040$ to get a good result for all p.

15.8. Buffon's needle

Consider a set of parallel lines on a plane, all a distance L apart. Imagine dropping a needle of length $\ell \leq L$ onto this plane at random. What is the probability that it intersects at least one line?

We will interpret a random drop to be represented by independent values x and θ, where x is the perpendicular distance from the lower end of the needle to the nearest line above it, and θ is the angle between the horizontal and the needle, where a value of $\theta = 0$ means that the needle is horizontal, and higher values of θ mean that the needle has been rotated θ radians anticlockwise. We assume that $\Theta \sim \mathrm{U}[0, \pi]$, and $X \sim \mathrm{U}[0, L]$, and that they are independent. The needle intersects a line if and only if $\ell \sin\theta \geq x$. We have

$$\begin{aligned}\mathbb{P}(\text{intersection}) &= \mathbb{P}(X \leq \ell \sin\Theta) \\ &= \int_0^L \int_0^\pi \frac{1}{\pi L} 1(x \leq \ell \sin\theta)\,\mathrm{d}x\,\mathrm{d}\theta \\ &= \frac{2\ell}{\pi L}\end{aligned}$$

Let this probability be denoted by p. So we can compute an approximation to π by finding

$$\pi = \frac{2\ell}{pL}$$

We can use the sampling error calculation above to find the amount of needles required to get a good approximation to π (within 0.1%) with probability ast least 99%, so we want

$$\mathbb{P}(|\hat{\pi}_n - \pi| \leq 0.001) \geq 0.99$$

Let S_n be the number of needles intersecting a line. Then $S_n \sim \mathrm{Bin}(n, p)$. So by the central limit theorem,

$$S_n \approx np + \sqrt{np(1-p)}Z \implies \hat{p}_n = \frac{S_n}{n} = p + \sqrt{\frac{p(1-p)}{n}}Z$$

Hence,

$$\hat{p}_n - p \approx \sqrt{\frac{p(1-p)}{n}}Z$$

Now, let $f(x) = 2\ell/xL$. Then $f(p) = \pi$, $f'(p) = -\frac{\pi}{p}$, and $\hat{\pi}_n = f(\hat{p}_n)$. We can then use a Taylor expansion to find

$$\hat{\pi}_n = f(\hat{p}_n) \approx f(p) + (\hat{p}_n - p)f'(p) \implies \hat{\pi}_n \approx \pi - (\hat{p}_n - p)\frac{\pi}{p}$$

Hence,

$$\hat{\pi}_n - \pi \approx -\frac{\pi}{p}\sqrt{\frac{p(1-p)}{n}} = -\pi\sqrt{\frac{1-p}{pn}}Z$$

We want

$$\mathbb{P}\left(\pi\sqrt{\frac{1-p}{pn}}|Z| \leq 0.001\right) \geq 0.99$$

So using tables, we find in the worst case scenario that $n \approx 3.75 \times 10^7$. So this approximation becomes good very slowly.

15.9. Bertrand's paradox

Consider a circle of radius r, and draw a random chord on the circle. What is the probability that its length C is less than r? There are two interpretations of the words 'random chord', that give different results. This is Bertrand's paradox.

(i) First, let us interpret 'random chord' as follows. Let $X \sim \mathrm{U}[0, r]$, and then we draw a chord perpendicular to a radius, such that it intersects the radius at a distance of X from the origin. Then we have formed a triangle between this intersection point, one end of the chord, and the circle's centre. By Pythagoras' theorem, the length of the chord is then twice the height of this triangle, so $C = 2\sqrt{r^2 - X^2}$. Hence,

$$\begin{aligned}\mathbb{P}(C \leq r) &= \mathbb{P}\left(2\sqrt{r^2 - X^2} \leq r\right) \\ &= \mathbb{P}\left(4(r^2 - X^2) \leq r^2\right) \\ &= \mathbb{P}\left(X \geq \frac{\sqrt{3}}{2}r\right) \\ &= 1 - \frac{\sqrt{3}}{2} \approx 0.134\end{aligned}$$

(ii) Instead, let us fix one end point of the chord A, and let $\Theta \sim \mathrm{U}[0, 2\pi]$. Let the other end point B be such that the angle between the radii OA and OB is Θ. Then if $\Theta \in [0, \pi]$, the length of the chord can be found by splitting this triangle in two by dropping a perpendicular from the centre, giving

$$C = 2r \sin\frac{\Theta}{2}$$

If $\Theta \in [\pi, 2\pi]$, then

$$C = 2r \sin\frac{2\pi - \Theta}{2} = 2r \sin\frac{\Theta}{2}$$

as before. Now,

$$\begin{aligned}
\mathbb{P}(C \leq r) &= \mathbb{P}\left(2r \sin\frac{\Theta}{2} \leq r\right) \\
&= \mathbb{P}\left(\sin\frac{\Theta}{2} \leq \frac{1}{2}\right) \\
&= \mathbb{P}\left(\Theta \leq \frac{\pi}{3}\right) + \mathbb{P}\left(\Theta \geq \frac{5\pi}{3}\right) \\
&= \frac{1}{6} + \frac{1}{6} \\
&= \frac{1}{3} \approx 0.333
\end{aligned}$$

Clearly, the two probabilities do not match.

16. Gaussian vectors

16.1. Multidimensional Gaussian random variables

Recall that a random variable X with values in $\mathbb{R}$ is called Gaussian (or normal) if

$$X = \mu + \sigma Z; \quad \mu \in \mathbb{R}, \sigma \geq 0, Z \sim \mathrm{N}(0,1)$$

The density function of X is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Now, let $X = (X_1, \dots, X_n)^\intercal$ with values in $\mathbb{R}^n$. Then we define that X is a Gaussian vector (also called Gaussian) if

$$\forall u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \in \mathbb{R}^n,\ u^\intercal X = \sum_{i=1}^{n} u_i X_i = \mu + \sigma Z$$

so any linear combination of the X_i is Gaussian. This does not require that the X_i are independent, just that their sum is always Gaussian.

Let X be Gaussian in $\mathbb{R}^n$. Suppose that A is an $m \times n$ matrix, and $b \in \mathbb{R}^m$. Then $AX + b$ is also Gaussian. Indeed, let $u \in \mathbb{R}^m$, and let $v = A^\intercal u$. Then

$$u^\intercal(AX + b) = u^\intercal AX + u^\intercal b = v^\intercal X + u^\intercal b$$

Since X is Gaussian, $v^\intercal X$ is also Gaussian. An additive constant preserves this property, so the entire expression is Gaussian.

16.2. Expectation and variance

We define the mean of a Gaussian vector X as

$$\mu = \mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}; \quad \mu_i = \mathbb{E}[X_i]$$

We further define

$$\begin{aligned} V &= \mathrm{Var}(X) = \mathbb{E}[(X-\mu)(X-\mu)^\intercal] \\ &= \begin{pmatrix} \mathbb{E}\left[(X_1-\mu_1)^2\right] & \mathbb{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \mathbb{E}[(X_1-\mu_1)(X_n-\mu_n)] \\ \mathbb{E}[(X_2-\mu_2)(X_1-\mu_1)] & \mathbb{E}\left[(X_2-\mu_2)^2\right] & \cdots & \mathbb{E}[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[(X_n-\mu_n)(X_1-\mu_1)] & \mathbb{E}[(X_n-\mu_1)(X_n-\mu_2)] & \cdots & \mathbb{E}\left[(X_n-\mu_n)^2\right] \end{pmatrix} \end{aligned}$$

Hence the components of V are

$$V_{ij} = \operatorname{Cov}(X_i, X_j)$$

In particular, V is a symmetric matrix, and

$$\mathbb{E}[u^\intercal X] = \mathbb{E}\left[\sum_{i=1}^n u_i X_i\right] = \sum_{i=1}^n u_i \mu_i = u^\intercal \mu$$

and

$$\operatorname{Var}(u^\intercal X) = \operatorname{Var}\left(\sum_{i=1}^n u_i X_i\right) = \sum_{i,j=1}^n u_i \operatorname{Cov}(X_i, X_j) u_j = u^\intercal V u$$

Hence $u^\intercal X \sim \mathrm{N}(u^\intercal \mu, u^\intercal V u)$. Further, V is a non-negative definite matrix. Indeed, let $u \in \mathbb{R}^n$. Then $\operatorname{Var}(u^\intercal X) = u^\intercal V u$. Since $\operatorname{Var}(u^\intercal X) \geq 0$, we have $u^\intercal V u \geq 0$.

16.3. Moment generating function

We define the moment generating function of X by

$$m(\lambda) = \mathbb{E}\left[e^{\lambda^\intercal X}\right]$$

where $\lambda \in \mathbb{R}^n$. Then, we know that $\lambda^\intercal X \sim \mathrm{N}(\lambda^\intercal \mu, \lambda^\intercal V \lambda)$. Hence $m(\lambda)$ is the moment generating function of a normal random variable with the above mean and variance, applied to the parameter $\theta = 1$.

$$m(\lambda) = \exp\left(\lambda^\intercal \mu + \frac{\lambda^\intercal V \lambda}{2}\right)$$

Since the moment generating function uniquely characterises the distribution, it is clear that a Gaussian vector is uniquely characterised by its mean vector μ and variance matrix V. In this case, we write $X \sim \mathrm{N}(\mu, V)$.

16.4. Constructing Gaussian vectors

Given a μ and a V matrix, we might like to create a Gaussian vector that has this mean and variance. Let $Z_1, \dots, Z_n$ be a list of independent and identically distributed standard normal random variables. Let $Z = (Z_1, \dots, Z_n)^\intercal$. Then Z is a Gaussian vector.

Proof. For any vector $u \in \mathbb{R}^n$, we have

$$u^\intercal Z = \sum_{i=1}^n u_i Z_i$$

Because the Z_i are independent, it is easy to take the moment generating function to get

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}u_iz_i\right)\right] &= \mathbb{E}\left[\prod_{i=1}^{n}\exp(\lambda u_iZ_i)\right]\\ &= \prod_{i=1}^{n}\mathbb{E}\left[\exp(\lambda u_iZ_i)\right]\\ &= \prod_{i=1}^{n}\exp\left(\frac{(\lambda u_i)^2}{2}\right)\\ &= \exp\left(\frac{\lambda^2|u|^2}{2}\right)\end{aligned}$$

So $u^\intercal Z \sim \mathrm{N}(0, |u|^2)$, which is normal as required. □

Now, $\mathbb{E}[Z] = 0$, and $\mathrm{Var}(Z) = I$, the identity matrix. We then write $Z \sim \mathrm{N}(0, I)$. Now, let $\mu \in \mathbb{R}^n$, and V be a non-negative definite matrix. We want to construct a Gaussian vector X such that its mean is μ and its expectation is V, by using Z. In the one-dimensional case, this is easy, since μ is a single value, and V contains only one element, σ^2. In this case therefore, $Z \sim \mathrm{N}(0,1)$ so $\mu + \sigma Z \sim \mathrm{N}(\mu, \sigma^2)$. In the general case, since V is non-negative definite, we can write

$$V = U^\intercal DU$$

where $U^{-1} = U^\intercal$, and D is a diagonal matrix with diagonal entries $\lambda_i \geq 0$. We define the square root of the matrix V to be

$$\sigma = U^\intercal\sqrt{D}U$$

where $\sqrt{D}$ is the diagonal matrix with diagonal entries $\sqrt{\lambda_i}$. Then clearly,

$$\sigma^2 = U^\intercal\sqrt{D}UU^\intercal\sqrt{D}U = U^\intercal\sqrt{D}\sqrt{D}U = U^\intercal DU = V$$

Now, let $X = \mu + \sigma Z$. We now want to show that $X \sim \mathrm{N}(\mu, V)$.

Proof. X is certainly Gaussian, since it is generated by a linear multiple of the Gaussian vector Z, with an added constant. By linearity,

$$\mathbb{E}[X] = \mu$$

and

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}\left[(X-\mu)(X-\mu)^\intercal\right]\\ &= \mathbb{E}\left[(\sigma Z)(\sigma Z)^\intercal\right]\\ &= \mathbb{E}\left[\sigma ZZ^\intercal\sigma^\intercal\right]\\ &= \sigma\mathbb{E}\left[ZZ^\intercal\right]\sigma^\intercal\\ &= \sigma\sigma^\intercal\\ &= \sigma\sigma\\ &= V\end{aligned}$$

□

16.5. Density

We can calculate the density of such a Gaussian vector $X \sim \mathrm{N}(\mu, V)$. First, consider the case where V is positive definite. Recall that in the one-dimensional case,

$$f_X(x) = f_Z(z)|J|; \quad x = \mu + \sigma z$$

In general, since V is positive definite, σ is invertible. So $x = \mu + \sigma z$ gives $z = \sigma^{-1}(x - \mu)$. Hence,

$$\begin{aligned} f_X(x) &= f_Z(z)|J| \\ &= \prod_{i=1}^{n} \frac{\exp\left(-\frac{z_i^2}{2}\right)}{\sqrt{2\pi}} |\det \sigma^{-1}| \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{|z|^2}{2}\right) \cdot \frac{1}{\sqrt{\det V}} \\ &= \frac{1}{\sqrt{(2\pi)^n \det V}} \exp\left(-\frac{z^\intercal z}{2}\right) \end{aligned}$$

Now,

$$\begin{aligned} z^\intercal z &= (\sigma^{-1}(x-\mu))^\intercal (\sigma^{-1}(x-\mu)) \\ &= (x-\mu)^\intercal (\sigma^{-1})^\intercal \sigma^{-1}(x-\mu) \\ &= (x-\mu)^\intercal \sigma^{-2}(x-\mu) \\ &= (x-\mu)^\intercal V^{-1}(x-\mu) \end{aligned}$$

Hence,

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \det V}} \exp\left(-\frac{(x-\mu)^\intercal V^{-1}(x-\mu)}{2}\right)$$

In the case where V is just non-negative definite (so it could have some zero eigenvalues), we can make an orthogonal change of basis, and assume that

$$V = \begin{pmatrix} U & 0 \\ 0 & 0 \end{pmatrix}; \quad \mu = \begin{pmatrix} \lambda \\ \nu \end{pmatrix}$$

where U is an $m \times m$ positive definite matrix, where $m < n$, and where $\lambda \in \mathbb{R}^m$, $\nu \in \mathbb{R}^{n-m}$. For U, we can then apply the result above. We can write

$$X = \begin{pmatrix} Y \\ \nu \end{pmatrix}$$

where Y has density

$$f_Y(y) = \frac{1}{\sqrt{(2\pi)^n \det U}} \exp\left(-\frac{(y-\lambda)^\intercal U^{-1}(y-\lambda)}{2}\right)$$

16.6. Diagonal variance

Note that if a Gaussian vector $X = (X_1, \dots, X_n)$ is comprised of independent normal random variables, then V is a diagonal matrix. Indeed, since the X_i are independent then $\operatorname{Cov}(X_i, X_j) = 0$ for all $i \neq j$, so V is diagonal.

Lemma. If V is diagonal, then the X_i are independent.

Note that zero covariance does not in general imply independence, as we saw earlier in the course, but in this specific case with Gaussian variables, this is true.

Proof. Since V is diagonal with diagonal entries λ_i, we have

$$(x-\mu)^\top V^{-1}(x-\mu) = \sum_{i=1}^{n} \frac{(x_i-\mu_i)^2}{\lambda_i}$$

Hence,

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \det V}} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu_i)^2}{2\lambda_i}\right)$$

So f_X factorises into a product. Hence the X_i are independent. □

We can construct an alternative proof using moment generating functions.

Proof.

$$\begin{aligned} m(\theta) &= \mathbb{E}\left[e^{\theta^\top X}\right] \\ &= \exp\left(\theta^\top \mu + \frac{\theta^\top V \theta}{2}\right) \\ &= \exp\left(\sum_{i=1}^{n} \theta_i \mu_i + \frac{1}{2}\sum_{i=1}^{n} \theta_i^2 \lambda_i\right) \end{aligned}$$

Hence $m(\theta)$ factorises into the moment generating functions of Gaussian random variables in $\mathbb{R}$. □

In summary, for Gaussian vectors, we have $(X_1, \dots, X_n)$ independent if and only if V is diagonal.

16.7. Bivariate Gaussian vectors

A bivariate Gaussian is a Gaussian vector of two variables ($n = 2$). Let $X = (X_1, X_2)$. Let $\mu_k = \mathbb{E}[X_k]$ and $\sigma_k^2 = \operatorname{Var}(X_k)$. We further define the *correlation*

$$\rho = \operatorname{Corr}(X_1, X_2) = \frac{\operatorname{Cov}(X_1, X_2)}{\sqrt{\operatorname{Var}(X_1)\operatorname{Var}(X_2)}}$$

Note that due to the Cauchy–Schwarz inequality, we have $\rho \in [-1, 1]$. We can write the variance matrix as

$$V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

This matrix V is non-negative definite. Indeed, let $u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$, then

$$\begin{aligned} u^\intercal V u &= (1-\rho)(\sigma_1^2 u_1^2 + \sigma_2^2 u_2^2) + \rho(\sigma_1 u_1 + \sigma_2 u_2)^2 \\ &= (1+\rho)(\sigma_1^2 u_1^2 + \sigma_2^2 u_2^2) - \rho(\sigma_1 u_1 - \sigma_2 u_2)^2 \end{aligned}$$

Since $\rho \in [-1, 1]$, this is non-negative for all choices of ρ.

16.8. Density of bivariate Gaussian

When $\rho = 0$ and $\sigma_1, \sigma_2 > 0$, we have

$$f_{X_1,X_2}(x_1, x_2) = \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x_k - \mu_k)^2}{2\sigma_k^2}\right)$$

So X_1 and X_2 are independent in this case.

16.9. Conditional expectation

Let (X_1, X_2) be a bivariate Gaussian vector. Then let $a \in \mathbb{R}$, and consider $X_2 - aX_1$. We have

$$\operatorname{Cov}(X_2 - aX_1, X_1) = \operatorname{Cov}(X_2, X_1) - a\operatorname{Cov}(X_1, X_1) = \operatorname{Cov}(X_2, X_1) - a\operatorname{Var}(X_1) = \rho\sigma_1\sigma_2 - a\sigma_1^2$$

Now, let $a = \frac{\rho\sigma_2}{\sigma_1}$, so $\operatorname{Cov}(X_2 - aX_1, X_1) = 0$. Since $Y = X_2 - aX_1$ is Gaussian, (X_1, Y) is a Gaussian vector, and so Y and X_1 are independent. Now, we can find

$$\begin{aligned} \mathbb{E}[X_2 \mid X_1] &= \mathbb{E}[Y + aX_1 \mid X_1] \\ &= \mathbb{E}[Y] + a\mathbb{E}[X_1 \mid X_1] \\ &= \mathbb{E}[X_2 - aX_1] + aX_1 \end{aligned}$$

In particular, since $X_2 = (X_2 - aX_1) + aX_1$, we can say that given X_1,

$$X_2 \sim \mathrm{N}(\mu_2 - a\mu_1 + aX_1, \operatorname{Var}(X_2 - aX_1))$$

and

$$\operatorname{Var}(X_2 - aX_1) = \operatorname{Var}(X_2) + a^2\operatorname{Var}(X_1) - 2a\operatorname{Cov}(X_1, X_2)$$

16.10. Multivariate central limit theorem

This subsection is non-examinable, but included for completeness. Let X be a random vector in $\mathbb{R}^k$ with $\mu = \mathbb{E}[X]$ and covariance matrix Σ. Let $X_1, X_2, \ldots$ be independent and identically distributed with the same distribution as X. Then

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i - \mathbb{E}[X_i] \xrightarrow{d} \mathrm{N}(\mu, \Sigma)$$

Convergence in distribution here means that for all reasonable $B \subseteq \mathbb{R}^k$, we have

$$\mathbb{P}(S_n \in B) \to \mathbb{P}(\mathrm{N}(\mu, \Sigma) \in B)$$

17. Simulation of random variables

17.1. Sampling from uniform distribution

It is easy for a computer to generate a random number in the interval $[0, 1)$.

We can use this as a source of randomness to simulate a random variable with an arbitrary density. Let $U \sim \mathrm{U}[0, 1]$, then let $X = -\log U$. Then

$$\mathbb{P}(X \leq x) = \mathbb{P}(\log U \leq x) = \mathbb{P}(U \geq e^{-x}) = 1 - e^{-x}$$

So X is exponentially distributed with parameter 1. More generally, we have the following.

Theorem. Let X be a continuous random variable with distribution function F. Then, if $U \sim \mathrm{U}[0, 1]$, then $F^{-1}(U) \sim F$.

Proof. Set $Y = F^{-1}(U)$. Then

$$\begin{aligned}\mathbb{P}(Y \leq x) &= \mathbb{P}\left(F^{-1}(U) \leq x\right) \\ &= \mathbb{P}(U \leq F(x)) \\ &= F(x)\end{aligned}$$

□

One way of thinking of this function F^{-1} function is that it takes an input probability p, and outputs the x value such that $\mathbb{P}(X \leq x) = p$. Then, if U is uniformly distributed, we are essentially sampling a random p.

17.2. Rejection sampling

In certain cases, finding such an F^{-1} function is difficult, if not impossible, especially where this function has jumps or has a higher dimension. Here is an alternative sampling method. Suppose $A \subset [0, 1]^d$. We then define

$$f(x) = \frac{1(x \in A)}{|A|}$$

where $|A|$ is the size or volume of this set A. Let X have density function f. How can we simulate X? Let (U_n) be an independent and identically distributed sequence of d-dimensional uniform random variables, i.e.

$$U_n = (U_{k,n} : k \in \{1, \dots, d\}); \quad (U_{k,n}) \sim \mathrm{U}[0, 1] \text{ i.i.d.}$$

Now, let

$$N = \min\{n \geq 1 : U_n \in A\}$$

So we keep generating random numbers until a U_n lies in A, and reject all other possibilities. We now show that $U_N \sim f$. In particular, we want to show that for all $B \subseteq [0,1]^d$,

$$\mathbb{P}(U_n \in B) = \int_B f(x)\,\mathrm{d}x$$

We have

$$\begin{aligned}
\mathbb{P}(U_n \in B) &= \sum_{n=1}^{\infty} \mathbb{P}(U_N \in B, N = n) \\
&= \sum_{n=1}^{\infty} \mathbb{P}(U_n \in A \cap B, U_{n-1} \notin A, \dots, U_1 \notin A) \\
&= \sum_{n=1}^{\infty} \mathbb{P}(U_n \in A \cap B)\,\mathbb{P}(U_{n-1} \notin A) \cdots \mathbb{P}(U_1 \notin A) \\
&= \sum_{n=1}^{\infty} |A \cap B|(1 - |A|)^{n-1} \\
&= \frac{|A \cap B|}{|A|} \\
&= \int_A \frac{1(x \in B)}{|A|}\,\mathrm{d}x \\
&= \int_B f(x)\,\mathrm{d}x
\end{aligned}$$

Now suppose that f is a density on $[0,1]^{d-1}$ which is bounded by $\lambda > 0$. We can use rejection sampling to sample a random variable X with this density. Consider the set

$$A = \left\{(x_1, \dots, x_d) \in [0,1]^d : \; x_d \leq \frac{f(x_1, \dots, x_{d-1})}{\lambda}\right\}$$

From the above, we can generate a uniform random variable $Y = (X_1, \dots, X_d)$ on A. Let $X = (X_1, \dots, X_{d-1})$, then we will show that $X \sim f$. In particular, we want to show that for all $B \subseteq [0,1]^{d-1}$,

$$\mathbb{P}(X \in B) = \int_B f(x)\,\mathrm{d}x$$

We find that

$$\begin{aligned}
\mathbb{P}(X \in B) &= \mathbb{P}((X_1, \dots, X_{d-1}) \in B) \\
&= \mathbb{P}((X_1, \dots, X_d) \in (B \times [0,1]) \cap A) \\
&= \frac{|(B \times [0,1]) \cap A|}{|A|} \\
|(B \times [0,1]) \cap A| &= \int \cdots \int 1((X_1, \dots, X_d) \in (B \times [0,1]) \cap A) \, \mathrm{d}x_1 \dots \mathrm{d}x_d \\
&= \int \cdots \int 1((X_1, \dots, X_{d-1}) \in B) \cdot 1\left(x_d \leq \frac{f(x_1, \dots, x_{d-1})}{\lambda}\right) \mathrm{d}x_1 \dots \mathrm{d}x_d \\
&= \int \cdots \int 1((X_1, \dots, X_{d-1}) \in B) \cdot \frac{f(x_1, \dots, x_{d-1})}{\lambda} \, \mathrm{d}x_1 \dots \mathrm{d}x_{d-1} \\
&= \frac{1}{\lambda} \int \cdots \int 1((X_1, \dots, X_{d-1}) \in B) \cdot f(x_1, \dots, x_{d-1}) \, \mathrm{d}x_1 \dots \mathrm{d}x_{d-1} \\
&= \frac{1}{\lambda} \int_B f(x) \, \mathrm{d}x \\
|A| &= \frac{1}{\lambda} \int_{[0,1]^{d-1}} f(x) \, \mathrm{d}x \\
&= \frac{1}{\lambda} \\
\therefore \mathbb{P}(X \in B) &= \int_B f(x) \, \mathrm{d}x
\end{aligned}$$

VII. Vector Calculus

Lectured in Lent 2021 by Dr. A. Ashton

This course brings the tools of calculus to higher dimensions. We move away from one-dimensional graphs and towards curves and surfaces, building the foundation for a subject called differential geometry. These new kinds of objects have different ways of calculating derivatives, giving rise to various differential operators. One such operator, the gradient operator, shows how the value of a function changes when the input point is moved slightly in all possible directions in space. These differential operators show up in many formulas in mathematics and physics, and we explore various tools to solve equations involving them.

We also study tensors, which can be thought of as a step up from vectors, matrices, or bilinear maps. We can also apply differential operators to tensors. Tensors can be seen in physics, such as the linear strain tensor which explains some features of how an elastic body deforms, or the inertia tensor which shows how mass is concentrated in a rigid body.

Contents

1. Differential geometry of curves 441
1.1. Notation 441
1.2. Parametrised curves and smoothness 441
1.3. Arc length 442
1.4. Choice of parametrisation of curves 444
1.5. Parametrisation according to arc length 445
1.6. Curvature 446
1.7. Torsion 446
1.8. Radius of curvature 447
1.9. Gaussian curvature (non-examinable) 447
2. Coordinates, differentials and gradients 448
2.1. Differentials and first order changes 448
2.2. Coordinates and line elements in $\mathbb{R}^2$ 449
2.3. Orthogonal curvilinear coordinates 450
2.4. Cylindrical polar coordinates 450
2.5. Spherical polar coordinates 451
2.6. Gradient operator 451
2.7. Gradient on curves 452
2.8. Gradient on surfaces 452
2.9. Coordinate-independent representation 453
2.10. Computing the gradient vector 454
3. Integration over lines 456
3.1. Line integrals 456
3.2. Closed curves 457
3.3. Conservative forces and exact differentials 457
4. Integration in Euclidean space 460
4.1. Definition of integral in two dimensions 460
4.2. Change of variables 461
4.3. Definition of integral in three dimensions 463
4.4. Calculating volumes 465
5. Integration over surfaces 466
5.1. Two-dimensional surfaces 466
5.2. Areas and integrals over surfaces 467
5.3. Choice of parametrisation of surfaces 468
6. Differential operators 469
6.1. Divergence, curl, and Laplacian 469
6.2. Explanation of divergence and curl 470

6.3. Identities . . . 470
6.4. Definitions in orthogonal curvilinear coordinate systems . . . 471
6.5. Laplacian of a vector field . . . 472
6.6. Relations between differential operators . . . 473
6.7. Irrotational and solenoidal forces . . . 473
7. Integral theorems . . . 474
7.1. Green's theorem . . . 474
7.2. Stokes' theorem . . . 475
7.3. Stokes' theorem on closed surfaces . . . 476
7.4. Zero circulation and irrotationality . . . 476
7.5. Intuition for curl as infinitesimal circulation . . . 477
7.6. Gauss' divergence theorem . . . 477
7.7. Intuition for divergence as infinitesimal flux . . . 479
7.8. Conservation laws . . . 479
7.9. Proof of divergence theorem . . . 480
7.10. Proof of Green's theorem . . . 481
7.11. Proof of Stokes' theorem . . . 482
8. Maxwell's equations . . . 484
8.1. Introduction and the equations . . . 484
8.2. Integral formulations of Maxwell's equations . . . 484
8.3. Electromagnetic waves . . . 485
8.4. Electrostatics and magnetostatics . . . 486
9. Poisson's and Laplace's equations . . . 488
9.1. The boundary value problem . . . 488
9.2. Uniqueness of solutions . . . 489
9.3. Gauss' flux method for spherically symmetric sources . . . 491
9.4. Cylindrical symmetry . . . 492
9.5. Superposition principle . . . 494
9.6. Integral solutions . . . 495
9.7. Harmonic functions . . . 496
9.8. Intuitive explanation of Laplacian . . . 497
9.9. Non-existence of maximum points . . . 498
10. Cartesian tensors . . . 500
10.1. Intuitive description of vectors and changes of basis . . . 500
10.2. Intuitive description of scalars and scalar products . . . 501
10.3. Intuitive description of linear maps . . . 501
10.4. Definition . . . 502
10.5. Kronecker δ and Levi-Civita ε . . . 503
10.6. Electrical conductivity tensor . . . 503
10.7. Indexed objects without tensor transformation properties . . . 504

10.8. Operations on tensors 504
10.9. Symmetric and antisymmetric tensors 504
11. Tensor calculus 506
11.1. Introduction 506
11.2. Differential operators producing tensor fields 507
11.3. Divergence theorem with tensor fields 507
12. Properties of tensors 509
12.1. Symmetry and antisymmetry 509
12.2. Isotropic tensors 512
12.3. Classifying isotropic tensors in three dimensions 513
12.4. Integrals with isotropic tensors 514
12.5. Bilinear and multilinear maps as tensors 516
12.6. Quotient theorem 516

1. Differential geometry of curves

1.1. Notation

Throughout this course, a column vector e.g.

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

should be interpreted as the vector

$$\mathbf{x} = a\mathbf{e}_x + b\mathbf{e}_y + x\mathbf{e}_z$$

where $\{\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z\}$ are the basis vectors aligned with the fixed Cartesian x, y, z axes in $\mathbb{R}^3$. We will be dealing with various kinds of basis vectors through the course, so it is useful to define now that column vectors written as above always represent the standard basis.

1.2. Parametrised curves and smoothness

A parametrised curve C in R^3 is the image of a continuous map $\mathbf{x} : [a, b] \to \mathbb{R}^3$, in which $t \mapsto \mathbf{x}(t)$. In Cartesian coordinates,

$$\mathbf{x}(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix} = \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}$$

The resultant curve has a direction, from $\mathbf{x}(a)$ to $\mathbf{x}(b)$.

Definition. We say that C is a differentiable curve if each of the components $\{x_i(t)\}$ are differentiable functions. C is regular if it is differentiable and $|\mathbf{x}'(t)| \neq 0$. If C is differentiable and regular, we say that C is smooth.

Note. We need this regularity condition because it is quite easy to create 'bad curves' with cusps and spikes using only differentiable functions, for example

$$\mathbf{x}(t) = (t^2, t^3)$$

The components are clearly differentiable, but $\mathbf{x}(t)$ has a cusp at $t = 0$. At this point, $|\mathbf{x}'(0)| = 0$.

Definition. Recall that $x_i(t)$ is called 'differentiable' at t if

$$x_i(t + h) = x_i(t) + x_i'(t)h + o(h)$$

where $o(h)$ represents a function that obeys

$$\lim_{h \to 0} \frac{o(h)}{h} = 0$$

In terms of vectors,

$$\mathbf{x}(t+h) = \mathbf{x}(t) + \mathbf{x}'(t)h + o(h)$$

where here $o(h)$ is a vector for which

$$\lim_{h\to 0} \frac{|o(h)|}{h} = 0$$

1.3. Arc length

We can approximate the length of a curve C by splitting it into small straight lines and summing the lengths of such lines. We will introduce a partition P of $[a,b]$ with $t_0 = a, t_N = b$ and

$$t_0 < t_1 < t_2 < \cdots < t_N$$

Let us now set $\Delta t_i = t_{i+1} - t_i$ and $\Delta t = \max_i \Delta t_i$. The length of the curve relative to P is defined as

$$\ell(C,P) = \sum_{i=0}^{N-1} |\mathbf{x}(t_{i+1}) - \mathbf{x}(t_i)|$$

As Δt gets smaller, we would expect $\ell(C,P)$ to give a better approximation to the true length of C, which we will call $\ell(C)$. Therefore we can define the length of C by

$$\ell(C) = \lim_{\Delta t\to 0} \sum_{i=0}^{N-1} |\mathbf{x}(t_{i+1}) - \mathbf{x}(t_i)| = \lim_{\Delta t\to 0} \ell(C,P)$$

If this limit doesn't exist, we say that the curve is *non-rectifiable*. Suppose C is differentiable. Then

$$\begin{aligned}\mathbf{x}(t_{i+1}) &= \mathbf{x}(t_i + t_{i+1} - t_i)\\ &= \mathbf{x}(t_i + \Delta t_i)\\ &= \mathbf{x}(t_i) + \mathbf{x}'(t_i)\Delta t_i + o(\Delta t_i)\end{aligned}$$

It follows then that

$$|\mathbf{x}(t_{i+1}) - \mathbf{x}(t_i)| = |\mathbf{x}'(t_i)|\Delta t_i + o(\Delta t_i)$$

So if C is differentiable,

$$\ell(C,P) = \lim_{\Delta t\to 0} \sum_{i=0}^{N-1} (|\mathbf{x}'(t_i)|\Delta t_i + o(\Delta t_i))$$

Recall that this $o(\Delta t_i)$ term represents a function for which $o(\Delta t_i)/\Delta t_i \to 0$. So for any $\varepsilon > 0$, if $\Delta t = \max_i \Delta t_i$ is sufficiently small, we have $|o(\Delta t_i)| < \frac{\varepsilon}{b-a}\Delta t_i$, for $i = 0, \ldots, N-1$. So by the Triangle Inequality, choosing Δt sufficiently small,

$$\left|\ell(C,P) - \sum_{i=0}^{N-1} |\mathbf{x}'(t_i)|\Delta t_i\right| = \left|\sum_{i=0}^{N-1} o(\Delta t_i)\right| < \frac{\varepsilon}{b-a} \underbrace{\sum_{i=0}^{N-1} \Delta t_i}_{b-a} = \varepsilon$$

So the left hand side tends to zero as $\Delta t \to 0$. We then get

$$\begin{aligned}\ell(C) &= \lim_{\Delta t \to 0} \ell(C, P) \\ &= \lim_{\Delta t \to 0} \sum_{i=0}^{N-1} |\mathbf{x}'(t_i)| \Delta t_i \\ &= \int_a^b |\mathbf{x}'(t)| \, \mathrm{d}t\end{aligned}$$

according to Analysis I, and the definition of the Riemann Integral. So in summary, if $C : [a, b] \ni t \mapsto \mathbf{x}(t)$, then

$$\begin{aligned}\ell(C) &= \int_a^b |\mathbf{x}'(t)| \, \mathrm{d}t \\ &= \int_C \mathrm{d}s\end{aligned}$$

where $\mathrm{d}s$ is the 'arc length element', i.e. $\mathrm{d}s = |\mathbf{x}'(t)| \, \mathrm{d}t$. Similarly, we define

$$\int_C f(\mathbf{x}) \, \mathrm{d}s = \int_a^b f(\mathbf{x}(t)) \, |\mathbf{x}'(t)| \, \mathrm{d}t$$

If C is made up of M smooth curves $C_1, \ldots, C_M$, we say that C is 'piecewise smooth'. We write $C = C_1 + \cdots + C_M$ and define

$$\int_C f(\mathbf{x}) \, \mathrm{d}s = \sum_{i=1}^{M} \int_{C_i} f(\mathbf{x}) \, \mathrm{d}s$$

Now note (informally) that

$$\mathrm{d}s = |\mathbf{x}'(t)| \, \mathrm{d}t = \sqrt{\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2 + \left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2 + \left(\frac{\mathrm{d}z}{\mathrm{d}t}\right)^2} \, \mathrm{d}t$$

i.e. (now very informally)

$$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$$

which is Pythagoras' Theorem.

Example. Let C be the circle of radius $r > 0$ in $\mathbb{R}^3$

$$\mathbf{x}(t) = \begin{pmatrix} r \cos t \\ r \sin t \\ 0 \end{pmatrix}; \quad t \in [0, 2\pi]$$

So

$$\mathbf{x}'(t) = \begin{pmatrix} -r \sin t \\ r \cos t \\ 0 \end{pmatrix}$$

Therefore

$$\int_C \mathrm{d}s = \int_0^{2\pi} |\mathbf{x}'(t)|\,\mathrm{d}t = \int_0^{2\pi} \sqrt{r^2\sin^2 t + r^2\cos^2 t}\,\mathrm{d}t = \int_0^{2\pi} r\,\mathrm{d}t = 2\pi r$$

Also, for example,

$$\int_C x^2 y\,\mathrm{d}s = \int_0^{2\pi} (r\cos t)^2 (r\sin t)\sqrt{r^2\sin^2 t + r^2\cos^2 t}\,\mathrm{d}t = \int_0^{2\pi} r^3\cos^2 t\sin t\,\mathrm{d}t = 0$$

1.4. Choice of parametrisation of curves

Does $\ell(C)$ depend on the choice of parametrisation of $\mathbf{x}(t)$? For example,

$$\mathbf{x}(t) = \begin{pmatrix} r\cos t \\ r\sin t \\ 0 \end{pmatrix}; \quad t \in [0, 2\pi]$$

and

$$\tilde{\mathbf{x}}(t) = \begin{pmatrix} r\cos 2t \\ r\sin 2t \\ 0 \end{pmatrix}; \quad t \in [0, \pi]$$

both give rise to a circle, but have different forms. Suppose that C has two different parametrisations,

$$\mathbf{x} = \mathbf{x}_1(t); \quad a \le t \le b$$

$$\mathbf{x} = \mathbf{x}_2(\tau); \quad \alpha \le \tau \le \beta$$

There must be some relationship $\mathbf{x}_2(\tau) = \mathbf{x}_1(t(\tau))$ for some function $t(\tau)$, since they represent the same curve. We can assume $\frac{\mathrm{d}t}{\mathrm{d}\tau} \neq 0$, so the map between t and τ is invertible and differentiable (see IB Analysis and Topology). Note that

$$\begin{aligned} \mathbf{x}_2'(\tau) &= \frac{\mathrm{d}}{\mathrm{d}\tau}\mathbf{x}_2(\tau) \\ &= \frac{\mathrm{d}}{\mathrm{d}\tau}\mathbf{x}_1(t(\tau)) \end{aligned}$$

By the Chain Rule,

$$= \frac{\mathrm{d}t}{\mathrm{d}\tau}\mathbf{x}_1'(t(\tau))$$

And now from the above definitions,

$$\int_C f(\mathbf{x})\,\mathrm{d}s = \int_a^b f(\mathbf{x}_1(t))\,|\mathbf{x}_1'(t)|\,\mathrm{d}t$$

Making the substitution $t = t(\tau)$, and assuming $\frac{\mathrm{d}t}{\mathrm{d}\tau} > 0$, the latter integral becomes

$$\int_\alpha^\beta f(\mathbf{x}_2(\tau)) \underbrace{\left|\mathbf{x}_1'(t(\tau))\right| \frac{\mathrm{d}t}{\mathrm{d}\tau}\mathrm{d}\tau}_{\left|\mathbf{x}_2'(\tau)\right|\,\mathrm{d}\tau} = \int_\alpha^\beta f(\mathbf{x}_2(\tau)) \left|\mathbf{x}_2'(\tau)\right| \mathrm{d}\tau$$

which is precisely the same as $\int_C f(\mathbf{x})\,\mathrm{d}s$ using the $\mathbf{x}_2(\tau)$ parametrisation. When $\frac{\mathrm{d}t}{\mathrm{d}\tau} < 0$, you get the same result. So the definition of $\int_C f(\mathbf{x})\,\mathrm{d}s$ does *not* depend on the choice of parametrisation of C.

1.5. Parametrisation according to arc length

We know that for any curve C there exist multiple unique parametrisations. We will define the arc-length function for a curve $[a, b] \ni t \mapsto \mathbf{x}(t)$ by

$$s(t) = \int_a^t |\mathbf{x}'(\tau)|\,\mathrm{d}\tau$$

So $s(a) = 0, s(b) = \ell(C)$. Using the Fundamental Theorem of Calculus, we have

$$s'(t) = |\mathbf{x}'(t)| \geq 0$$

For regular curves, we have that

$$s'(t) > 0$$

So we can invert the relationship between s and t; i.e. we can find t as a function of s. Hence, we can parametrise curves with respect to arc length. If we write

$$\mathbf{r}(s) = \mathbf{x}(t(s))$$

where $0 \leq s \leq \ell(C)$, then by the chain rule we have

$$\frac{\mathrm{d}t}{\mathrm{d}s} = \frac{1}{\frac{\mathrm{d}s}{\mathrm{d}t}} = \frac{1}{|\mathbf{x}'(t(s))|}$$

So

$$\mathbf{r}'(s) = \frac{\mathrm{d}}{\mathrm{d}s}\mathbf{x}(t(s)) = \frac{\mathrm{d}t}{\mathrm{d}s}\mathbf{x}'(t(s)) = \frac{\mathbf{x}'(t(s))}{|\mathbf{x}'(t(s))|}$$

In other words, $\mathbf{r}'(s)$ is a unit vector tangential to the curve. This (consistently) gives

$$\ell(C) = \int_0^{\ell(C)} |\mathbf{r}'(s)|\,\mathrm{d}s = \int_0^{\ell(C)} \mathrm{d}s$$

as previously found above.

1.6. Curvature

Throughout this section, we will be talking about a generic regular curve C, parametrised with respect to arc length, where a position vector on C is given by $\mathbf{r}(s)$. We will define the tangent vector

$$\mathbf{t}(s) = \mathbf{r}'(s)$$

We already know that $|\mathbf{t}(s)| = 1$. Therefore the only part of $\mathbf{t}$ that changes with respect to s is its direction. So $\mathbf{t}'(s) = \mathbf{r}''(s)$ only measures the change in the direction of the tangent as we move along the curve. So intuitively, if $|\mathbf{r}''(s)|$ is large then the curve is rapidly changing direction. If $|\mathbf{r}''(s)|$ is small, the curve is approximately flat; there is little change in direction. Using this intuition, we will define curvature as

$$\kappa(s) = |\mathbf{r}''(s)| = |\mathbf{t}'(s)|$$

In other words κ is the magnitude of the acceleration a particle experiences while moving along the curve at unit speed.

1.7. Torsion

Since $\mathbf{t} = \mathbf{r}'(s)$ is a unit vector, differentiating $\mathbf{t} \cdot \mathbf{t} = 1$ gives $\mathbf{t} \cdot \mathbf{t}' = 0$. We will define the principal normal $\mathbf{n}$ by the formula

$$\mathbf{t}' = \kappa \mathbf{n}$$

Note that $\mathbf{n}$ is everywhere normal to the curve C, since it is always perpendicular to the tangent vector $\mathbf{t}$, since $\mathbf{t} \cdot \mathbf{n} = 0$. We can extend the vectors $\{\mathbf{t}, \mathbf{n}\}$ into an orthonormal basis by computing the cross product:

$$\mathbf{b} = \mathbf{t} \times \mathbf{n}$$

We call $\mathbf{b}$ the binormal. It is a unit vector, since it is the cross product of two orthogonal unit vectors in $\mathbb{R}^3$. We also have that $\mathbf{b} \cdot \mathbf{b}' = 0$; also since $\mathbf{t} \cdot \mathbf{b} = 0$ and $\mathbf{n} \cdot \mathbf{b} = 0$, we must have

$$0 = (\mathbf{t} \cdot \mathbf{b})' = \mathbf{t}' \cdot \mathbf{b} + \mathbf{t} \cdot \mathbf{b}' = \kappa \mathbf{n} \cdot \mathbf{b} + \mathbf{t} \cdot \mathbf{b}' = \mathbf{t} \cdot \mathbf{b}'$$

So $\mathbf{b}'$ is orthogonal to both $\mathbf{t}$ and $\mathbf{b}$, i.e. it is parallel to $\mathbf{n}$. We will define the torsion τ of a curve by

$$\mathbf{b}' = -\tau \mathbf{n}$$

A physical interpretation of torsion is a kind of 'corkscrew' rotation in three dimensions.

Proposition (Fundamental Theorem of Differential Geometry of Curves)**.** The curvature $\kappa(s)$ and torsion $\tau(s)$ uniquely define a curve in $\mathbb{R}^3$, up to translation and orientation.

Proof. Since $\mathbf{n} = \mathbf{b} \times \mathbf{t}$, we have $\mathbf{t}' = \kappa(\mathbf{b} \times \mathbf{t})$ and $\mathbf{b}' = -\tau(\mathbf{b} \times \mathbf{t})$. This gives six equations (written in component form) for six unknowns. Given $\kappa(s)$ and $\tau(s)$, and given $\mathbf{t}(0)$ and $\mathbf{b}(0)$, we can construct the functions $\mathbf{t}(s), \mathbf{b}(s), \mathbf{n}(s) = \mathbf{b}(s) \times \mathbf{t}(s)$. □

1.8. Radius of curvature

A generic curve $s \mapsto \mathbf{r}(s)$ can be Taylor expanded around $s = 0$. Writing $\mathbf{t} = \mathbf{t}(0), \mathbf{n} = \mathbf{n}(0)$ and so on, we have

$$\begin{aligned}\mathbf{r}(s) &= \mathbf{r} + s\mathbf{r}' + \frac{1}{2}s^2\mathbf{r}'' + o(s^2)\\ &= \mathbf{r} + s\mathbf{t} + \frac{1}{2}s^2\kappa\mathbf{n} + o(s^2)\end{aligned}$$

What circle that touches the curve at $s = 0$ would be the best approximation for the curve at this point? Since the circle touches the curve, we know the position vectors (of the curve and the circle) match, and their first derivatives match. So we want to unify the second derivatives. The equation of such a circle of radius R is

$$\mathbf{x}(\theta) = \mathbf{r} + R(1 - \cos\theta)\mathbf{n} + R(\sin\theta)\mathbf{t}$$

Expanding this for small θ gives

$$\mathbf{x}(\theta) = \mathbf{r} + R\theta\mathbf{t} + \frac{1}{2}R\theta^2\mathbf{n} + o(\theta^2)$$

But the arc length on a circle is simply $R\theta$. So in terms of arc length,

$$\mathbf{x}(\theta) = \mathbf{r} + s\mathbf{t} + \frac{1}{2}s^2\frac{1}{R}\mathbf{n} + o(s^2)$$

Hence by comparing coefficients,

$$R = \frac{1}{\kappa}$$

We name this $R(s)$ the radius of curvature.

1.9. Gaussian curvature (non-examinable)

This subsection is non-examinable. How can we find the curvature of a surface? At any point $\mathbf{r}$ on a surface, we have a normal vector $\mathbf{n}$. We can construct a plane containing this normal; such a plane will then intersect the surface near $\mathbf{r}$. This intersection is a curve C, which has a curvature κ. The choice of plane is arbitrary, however. To unify all of these different possible results for κ, we can compute the Gaussian curvature κ_G by

$$\kappa_G = \kappa_{\min}\kappa_{\max}$$

- The Gaussian curvature of a flat plane is zero, since the minimum and maximum curvatures are both zero.
- On any point on a sphere of radius R, the Gaussian curvature is $\frac{1}{R^2}$, since any plane containing the normal produces a great circle of radius R, i.e. of curvature $\frac{1}{\kappa}$.

Theorem (Gauss's Remarkable Theorem)**.** The Gaussian curvature of a surface S is invariant under local isometries; i.e. if you bend the surface without stretching it.

2. Coordinates, differentials and gradients

2.1. Differentials and first order changes

Recall that for a function $f(u_1, \dots, u_n)$, we define the differential of f, written $\mathrm{d}f$, by

$$\mathrm{d}f = \frac{\partial f}{\partial u_i}\,\mathrm{d}u_i$$

noting that the summation convention applies. The $\mathrm{d}u_i$ are called differential forms, which can be thought of as linearly independent objects (if the coordinates $u_1, \dots, u_n$ are independent), i.e. $\alpha_i\,\mathrm{d}u_i = 0 \implies \alpha_i = 0$ for all i. Similarly, if we have a vector $\mathbf{x}(u_1, \dots, u_n)$, we define

$$\mathrm{d}\mathbf{x} = \frac{\partial \mathbf{x}}{\partial u_i}\,\mathrm{d}u_i$$

As an example, let $f(u, v, w) = u^2 + w\sin(v)$. Then

$$\mathrm{d}f = 2u\,\mathrm{d}u + w\cos(v)\,\mathrm{d}v + \sin(v)\,\mathrm{d}w$$

Similarly, given

$$\mathbf{x}(u, v, w) = \begin{pmatrix} u^2 - v^2 \\ w \\ e^v \end{pmatrix}$$

we can compute

$$\mathrm{d}\mathbf{x} = \begin{pmatrix} 2u \\ 0 \\ 0 \end{pmatrix}\mathrm{d}u + \begin{pmatrix} -2v \\ 0 \\ e^v \end{pmatrix}\mathrm{d}v + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}\mathrm{d}w$$

Differentials encode information about how a function (or vector field) changes when we change the coordinates by a small amount. By calculus,

$$f(u + \delta u_1, \dots, u_n + \delta u_n) - f(u_1, \dots, u_n) = \frac{\partial f}{\partial u_i}\delta u_i + o(\delta \mathbf{u})$$

So if δf denotes the change in $f(u_1, \dots, u_n)$ under this small change in coordinates, we have, to first order,

$$\delta f \approx \frac{\partial f}{\partial u_i}\delta u_i$$

The analogous result holds for vector fields:

$$\delta \mathbf{x} \approx \frac{\partial \mathbf{x}}{\partial u_i}\delta u_i$$

2.2. Coordinates and line elements in $\mathbb{R}^2$

We can create multiple different consistent coordinate systems by defining a relationship between them. For example, polar coordinates (r, θ) and Cartesian coordinates (x, y) can be related by

$$x = r\cos\theta; \quad y = r\sin\theta$$

Even though this relationship is not bijective (there are multiple polar coordinates mapping to the origin), it's still a useful coordinate system because the vast majority of points work well. Even coordinate systems with a countable amount of badly-behaved points are still useful.

A general set of coordinates (u, v) on $\mathbb{R}^2$ can be specified by their relationship to the standard Cartesian coordinates (x, y). We must specify smooth, invertible functions $x(u, v)$, $y(u, v)$. We would also like to have a small change in one coordinate system to be equivalent to a small change in the other coordinate system (i.e. the inverse is also smooth). The same principle applies in $\mathbb{R}^3$ for three coordinates, for example.

Consider the standard Cartesian coordinates in $\mathbb{R}^2$.

$$\mathbf{x}(x, y) = \begin{pmatrix} x \\ y \end{pmatrix} = x\mathbf{e}_x + y\mathbf{e}_y$$

Note that $\{\mathbf{e}_x, \mathbf{e}_y\}$ are orthonormal, and point in the same direction regardless of the value of $\mathbf{x}$: $\mathbf{e}_x$ points in the direction of changing x with y held constant, for example. Equivalently,

$$\mathbf{e}_x = \frac{\frac{\partial}{\partial x}\mathbf{x}(x, y)}{\left|\frac{\partial}{\partial x}\mathbf{x}(x, y)\right|}; \quad \mathbf{e}_y = \frac{\frac{\partial}{\partial y}\mathbf{x}(x, y)}{\left|\frac{\partial}{\partial y}\mathbf{x}(x, y)\right|}$$

Note that

$$\mathrm{d}\mathbf{x} = \frac{\partial \mathbf{x}}{\partial x}\,\mathrm{d}x + \frac{\partial \mathbf{x}}{\partial y}\,\mathrm{d}y = \mathrm{d}x\,\mathbf{e}_x + \mathrm{d}y\,\mathbf{e}_y$$

In other words, when applying the change in coordinate $x \mapsto x + \delta x$, the vector changes (to first order) to $\mathbf{x} \mapsto \mathbf{x} + \delta x\mathbf{e}_x$. In fact, in the case of Cartesian coordinates, this change is precisely correct for any size of δ, since the coordinate basis vectors are the same everywhere. We call $\mathrm{d}\mathbf{x}$ the line element; it tells us how small changes in coordinates produce changes in position vectors.

Now, let us consider polar coordinates in two-dimensional space. We can use the same idea as before, giving

$$\mathbf{e}_r = \frac{\frac{\partial}{\partial r}\mathbf{x}(r, \theta)}{\left|\frac{\partial}{\partial r}\mathbf{x}(r, \theta)\right|} = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}; \quad \mathbf{e}_\theta = \frac{\frac{\partial}{\partial \theta}\mathbf{x}(r, \theta)}{\left|\frac{\partial}{\partial \theta}\mathbf{x}(r, \theta)\right|} = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}$$

Therefore, we have

$$\mathbf{x}(r, \theta) = \begin{pmatrix} r\cos\theta \\ r\sin\theta \end{pmatrix} = r\mathbf{e}_r$$

Note that $\{\mathbf{e}_r, \mathbf{e}_\theta\}$ are also orthonormal at each (r, θ), but their exact values are not the same everywhere. Since the basis vectors are orthogonal, we can call r and θ orthogonal curvilinear coordinates. Also, we can compute the line element $\mathrm{d}\mathbf{x}$ as

$$\mathrm{d}\mathbf{x} = \frac{\partial \mathbf{x}}{\partial r}\,\mathrm{d}r + \frac{\partial \mathbf{x}}{\partial \theta}\,\mathrm{d}\theta = \begin{pmatrix}\cos\theta \\ \sin\theta\end{pmatrix}\mathrm{d}r + \begin{pmatrix}-r\sin\theta \\ r\cos\theta\end{pmatrix}\mathrm{d}\theta = \mathrm{d}r\ \mathbf{e}_r + r\,\mathrm{d}\theta\ \mathbf{e}_\theta$$

We see that a change in θ produces (up to first order) a change $\mathbf{x} \mapsto \mathbf{x} + r\,\delta\theta\,\mathbf{e}_\theta$, a change proportional to r. So a small change in θ could cause quite a large change in Cartesian coordinates.

2.3. Orthogonal curvilinear coordinates

We say that (u, v, w) are a set of orthogonal curvilinear coordinates if the vectors

$$\mathbf{e}_u = \frac{\frac{\partial \mathbf{x}}{\partial u}}{\left|\frac{\partial \mathbf{x}}{\partial u}\right|}; \quad \mathbf{e}_v = \frac{\frac{\partial \mathbf{x}}{\partial v}}{\left|\frac{\partial \mathbf{x}}{\partial v}\right|}; \quad \mathbf{e}_w = \frac{\frac{\partial \mathbf{x}}{\partial w}}{\left|\frac{\partial \mathbf{x}}{\partial w}\right|}$$

form a right-handed, orthonormal basis for each (u, v, w); but not necessarily the same basis over the entire vector field. It is standard to write

$$h_u = \left|\frac{\partial \mathbf{x}}{\partial u}\right|; \quad h_v = \left|\frac{\partial \mathbf{x}}{\partial v}\right|; \quad h_w = \left|\frac{\partial \mathbf{x}}{\partial w}\right|$$

We call h_u, h_v, h_w the scale factors. Note that the line element is

$$\begin{aligned}\mathrm{d}\mathbf{x} &= \frac{\partial \mathbf{x}}{\partial u}\,\mathrm{d}u + \frac{\partial \mathbf{x}}{\partial v}\,\mathrm{d}v + \frac{\partial \mathbf{x}}{\partial w}\,\mathrm{d}w \\ &= h_u\mathbf{e}_u\,\mathrm{d}u + h_v\mathbf{e}_v\,\mathrm{d}v + h_w\mathbf{e}_w\,\mathrm{d}w\end{aligned}$$

So the scale factors show how first-order changes in the coordinates are scaled into changes in $\mathbf{x}$.

2.4. Cylindrical polar coordinates

We define (ρ, ϕ, z) by

$$\mathbf{x}(\rho, \phi, z) = \begin{pmatrix}\rho\cos\phi \\ \rho\sin\phi \\ z\end{pmatrix}$$

where $0 \le \rho; 0 \le \phi < 2\pi; z \in \mathbb{R}$. So we can find

$$\mathbf{e}_\rho = \begin{pmatrix}\cos\phi \\ \sin\phi \\ 0\end{pmatrix}; \quad \mathbf{e}_\phi = \begin{pmatrix}-\sin\phi \\ \cos\phi \\ 0\end{pmatrix}; \quad \mathbf{e}_z = \begin{pmatrix}0 \\ 0 \\ 1\end{pmatrix}$$

The scale factors are

$$h_\rho = 1; \quad h_\phi = \rho; \quad h_z = 1$$

The line element is

$$\mathrm{d}\mathbf{x} = \mathrm{d}\rho\ \mathbf{e}_\rho + \rho\,\mathrm{d}\phi\ \mathbf{e}_\phi + \mathrm{d}z\ \mathbf{e}_z$$

Note that

$$\mathbf{x} = \rho\begin{pmatrix}\cos\phi\\ \sin\phi\\ 0\end{pmatrix} + z\begin{pmatrix}0\\ 0\\ 1\end{pmatrix} = \rho\mathbf{e}_\rho + z\mathbf{e}_z$$

2.5. Spherical polar coordinates

We define (r, θ, ϕ) by

$$\mathbf{x}(r,\theta,\phi) = \begin{pmatrix}r\cos\phi\sin\theta\\ r\sin\phi\sin\theta\\ r\cos\theta\end{pmatrix}$$

where $0 \le r; 0 \le \theta < \pi; 0 \le \phi < 2\pi$. So we can find

$$\mathbf{e}_r = \begin{pmatrix}\cos\phi\sin\theta\\ \sin\phi\sin\theta\\ \cos\theta\end{pmatrix}; \quad \mathbf{e}_\theta = \begin{pmatrix}\cos\phi\cos\theta\\ \sin\phi\cos\theta\\ -\sin\theta\end{pmatrix}; \quad \mathbf{e}_\phi = \begin{pmatrix}-\sin\phi\\ \cos\phi\\ 0\end{pmatrix}$$

The scale factors are

$$h_r = 1; \quad h_\theta = r; \quad h_\phi = r\sin\theta$$

The line element is

$$\mathrm{d}\mathbf{x} = \mathrm{d}r\ \mathbf{e}_r + r\,\mathrm{d}\theta\ \mathbf{e}_\theta + r\sin\theta\,\mathrm{d}\phi\ \mathbf{e}_\phi$$

Note that

$$\mathbf{x} = r\begin{pmatrix}\cos\phi\sin\theta\\ \sin\phi\sin\theta\\ \cos\theta\end{pmatrix} = r\mathbf{e}_r$$

2.6. Gradient operator

For $f: \mathbb{R}^3 \to \mathbb{R}$, we define the gradient of f, written ∇f, by

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x})\cdot\mathbf{h} + o(\mathbf{h}) \qquad (*)$$

as $|\mathbf{h}| \to 0$. The directional derivative of f in the direction $\mathbf{v}$, denoted by $D_\mathbf{v} f$ or $\frac{\partial f}{\partial \mathbf{v}}$, is defined by

$$D_\mathbf{v} f(\mathbf{x}) = \lim_{t\to 0}\frac{f(\mathbf{x}+t\mathbf{v}) - f(\mathbf{x})}{t}$$

Alternatively,

$$f(\mathbf{x}+t\mathbf{v}) = f(\mathbf{x}) + tD_\mathbf{v} f(\mathbf{x}) + o(t) \qquad (\dagger)$$

as $t \to 0$. Setting $\mathbf{h} = t\mathbf{v}$ in $(*)$, we have

$$f(\mathbf{x} + t\mathbf{v}) = f(\mathbf{x}) + t\nabla f(\mathbf{x}) \cdot \mathbf{v} + o(t)$$

This gives another way to interpret the gradient of f. Comparing this result to $(\dagger)$, we see that

$$D_{\mathbf{v}} f = \mathbf{v} \cdot \nabla f$$

By the Cauchy–Schwarz inequality, the dot product is maximised when the two vectors are parallel. Hence, the directional derivative is maximised when $\mathbf{v}$ points in the direction of ∇f. So ∇f points in the direction of greatest increase of f. Similarly, $-\nabla f$ points in the direction of greatest decrease of f. For example, suppose $f(x) = \frac{1}{2}|\mathbf{x}|^2$. Then

$$f(\mathbf{x} + \mathbf{h}) = \frac{1}{2}(\mathbf{x} + \mathbf{h}) \cdot (\mathbf{x} + \mathbf{h}) = \frac{1}{2}|\mathbf{x}|^2 + \frac{1}{2}(2\mathbf{x} \cdot \mathbf{h}) + \frac{1}{2}|\mathbf{h}|^2 = f(\mathbf{x}) + \mathbf{x} \cdot \mathbf{h} + o(\mathbf{h})$$

Hence $\nabla f(\mathbf{x}) = \mathbf{x}$.

2.7. Gradient on curves

Suppose we have a curve $t \mapsto \mathbf{x}(t)$. How does some function f change when moving along the curve? We will write $F(t) = f(\mathbf{x}(t))$, $\delta\mathbf{x} = \mathbf{x}(t + \delta t) - \mathbf{x}(t)$.

$$\begin{aligned} F(t + \delta t) &= f(\mathbf{x}(t + \delta t)) \\ &= f(\mathbf{x}(t) + \delta\mathbf{x}) \\ &= f(\mathbf{x}(t)) + \nabla f(\mathbf{x}(t)) \cdot \delta\mathbf{x} + o(\delta\mathbf{x}) \end{aligned}$$

Since $\delta\mathbf{x} = \mathbf{x}'(t)\,\delta t + o(\delta t)$, we have

$$F(t + \delta t) = F(t) + \mathbf{x}'(t) \cdot \nabla f(\mathbf{x}(t))\,\delta t + o(\delta t)$$

In other words,

$$\frac{\mathrm{d}F}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{x}(t)) = \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} \cdot \nabla f(\mathbf{x}(t))$$

2.8. Gradient on surfaces

Suppose we have a surface S in $\mathbb{R}^3$ defined implicitly by

$$S = \{\mathbf{x} \in \mathbb{R}^3 \,:\, f(\mathbf{x}) = 0\}$$

If $t \mapsto \mathbf{x}(t)$ is any curve in S, then $f(\mathbf{x}(t)) = 0$ everywhere. So

$$0 = \frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{x}(t)) = \nabla f(\mathbf{x}(t)) \cdot \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}$$

So $\nabla f(\mathbf{x}(t))$, the gradient, is orthogonal to $\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}$, the tangent vector of any chosen curve in S. So $\nabla f(\mathbf{x}(t))$ is normal to the surface.

2.9. Coordinate-independent representation

If we are working in an orthogonal curvilinear coordinate system (u, v, w), it is not immediately clear how to compute ∇f, since we need to represent this arbitrary perturbation $\mathbf{h}$ using (u, v, w). In Cartesian coordinates it is simple; to represent the change $\mathbf{x} \mapsto \mathbf{x} + \mathbf{h}$ we simply add the components of $\mathbf{x}$ and $\mathbf{h}$.

$$\begin{aligned} f(\mathbf{x}+\mathbf{h}) &= f((x+h_1, y+h_2, z+h_3)) \\ &= f(\mathbf{x}) + \frac{\partial f}{\partial x}h_1 + \frac{\partial f}{\partial y}h_2 + \frac{\partial f}{\partial z}h_3 + o(\mathbf{h}) \\ &= f(\mathbf{x}) + \begin{pmatrix} \partial f/\partial x \\ \partial f/\partial y \\ \partial f/\partial z \end{pmatrix} \cdot h + o(\mathbf{h}) \end{aligned}$$

So we have

$$\implies \nabla f = \begin{pmatrix} \partial f/\partial x \\ \partial f/\partial y \\ \partial f/\partial z \end{pmatrix}$$

Or, using suffix notation,

$$\nabla f = \mathbf{e}_i \frac{\partial f}{\partial x_i}; \quad [\nabla f]_i = \frac{\partial f}{\partial x_i}$$

We see that this ∇ is a kind of vector differential operator. In Cartesian coordinates,

$$\nabla = \mathbf{e}_x \frac{\partial}{\partial x} + \mathbf{e}_y \frac{\partial}{\partial y} + \mathbf{e}_z \frac{\partial}{\partial z} \equiv \mathbf{e}_i \frac{\partial}{\partial x_i}$$

From our previous example,

$$f(\mathbf{x}) = \frac{1}{2}(x^2 + y^2 + z^2) = \frac{1}{2}|\mathbf{x}|^2$$

$$\begin{aligned} [\nabla f]_i &= \frac{\partial}{\partial x_i}\left[\frac{1}{2}x_j x_j\right] \\ &= \frac{1}{2}\left[\delta_{ij}x_j + x_j\delta_{ij}\right] \\ &= x_i \\ \nabla f &= \mathbf{e}_i x_i \end{aligned}$$

Let us return back to computing the gradient in the general case. Recall that in Cartesian coordinates, the line element is simple:

$$\mathrm{d}\mathbf{x} = \mathrm{d}x_i\, \mathbf{e}_i$$

And also, if we have a function on $\mathbb{R}^3$ such as $f(x, y, z)$, it has the differential

$$\mathrm{d}f = \frac{\partial f}{\partial x_i}\, \mathrm{d}x_i$$

Then,

$$\begin{aligned}\nabla f \cdot \mathrm{d}\mathbf{x} &= \left(\mathbf{e}_i \frac{\partial f}{\partial x_i}\right) \cdot \left(\mathbf{e}_j \, \mathrm{d}x_j\right) \\ &= \frac{\partial f}{\partial x_i} \left(\mathbf{e}_i \cdot \mathbf{e}_j\right) \mathrm{d}x_j \\ &= \frac{\partial f}{\partial x_i} \delta_{ij} \, \mathrm{d}x_j \\ &= \frac{\partial f}{\partial x_i} \, \mathrm{d}x_i \\ &= \mathrm{d}f\end{aligned}$$

In other words, in *any* set of coordinates,

$$\nabla f \cdot \mathrm{d}\mathbf{x} = \mathrm{d}f$$

2.10. Computing the gradient vector

Proposition. If (u, v, w) are orthogonal curvilinear coordinates, and f is a function of the position vector (u, v, w), then

$$\nabla f = \frac{1}{h_u} \frac{\partial f}{\partial u} \mathbf{e}_u + \frac{1}{h_v} \frac{\partial f}{\partial v} \mathbf{e}_v + \frac{1}{h_w} \frac{\partial f}{\partial w} \mathbf{e}_w$$

Proof. If $f = f(u, v, w)$ and $\mathbf{x} = \mathbf{x}(u, v, w)$, then

$$\mathrm{d}f = \frac{\partial f}{\partial u} \, \mathrm{d}u + \frac{\partial f}{\partial v} \, \mathrm{d}v + \frac{\partial f}{\partial w} \, \mathrm{d}w$$

$$\mathrm{d}x = h_u \, \mathrm{d}u \, \mathbf{e}_u + h_v \, \mathrm{d}v \, \mathbf{e}_v + h_w \, \mathrm{d}w \, \mathbf{e}_w$$

Using the above result, we have

$$\nabla f \cdot \mathrm{d}\mathbf{x} = \mathrm{d}f$$

$$((\nabla f)_u \mathbf{e}_u + (\nabla f)_v \mathbf{e}_v + (\nabla f)_w \mathbf{e}_w) \cdot (h_u \, \mathrm{d}u \, \mathbf{e}_u + h_v \, \mathrm{d}v \, \mathbf{e}_v + h_w \, \mathrm{d}w \, \mathbf{e}_w) = \frac{\partial f}{\partial u} \, \mathrm{d}u + \frac{\partial f}{\partial v} \, \mathrm{d}v + \frac{\partial f}{\partial w} \, \mathrm{d}w$$

$$(\nabla f)_u h_u \, \mathrm{d}u + (\nabla f)_v h_v \, \mathrm{d}v + (\nabla f)_w h_w \, \mathrm{d}w = \frac{\partial f}{\partial u} \, \mathrm{d}u + \frac{\partial f}{\partial v} \, \mathrm{d}v + \frac{\partial f}{\partial w} \, \mathrm{d}w$$

Since u, v, w are independent coordinates, $\mathrm{d}u \, , \mathrm{d}v \, , \mathrm{d}w$ are linearly independent. So we can simply compare coefficients, getting

$$\nabla f = \frac{1}{h_u} \frac{\partial f}{\partial u} \mathbf{e}_u + \frac{1}{h_v} \frac{\partial f}{\partial v} \mathbf{e}_v + \frac{1}{h_w} \frac{\partial f}{\partial w} \mathbf{e}_w$$

as required. □

In cylindrical polar coordinates, we have

$$\nabla f = \frac{\partial f}{\partial \rho}\mathbf{e}_\rho + \frac{1}{\rho}\frac{\partial f}{\partial \phi}\mathbf{e}_\phi + \frac{\partial f}{\partial z}\mathbf{e}_z$$

In spherical polar coordinates, we have

$$\nabla f = \frac{\partial f}{\partial r}\mathbf{e}_r + \frac{1}{r}\frac{\partial f}{\partial \theta}\mathbf{e}_\theta + \frac{1}{r\sin\theta}\frac{\partial f}{\partial \phi}\mathbf{e}_\phi$$

Then using the familiar example $f(\mathbf{x}) = \frac{1}{2}|\mathbf{x}|^2$, we have

$$f = \begin{cases} \frac{1}{2}(x^2 + y^2 + z^2) & \text{in Cartesian coordinates} \\ \frac{1}{2}(\rho^2 + z^2) & \text{in cylindrical polar coordinates} \\ \frac{1}{2}r^2 & \text{in spherical polar coordinates} \end{cases}$$

Then we can check the value of ∇f in these different coordinate systems.

$$\begin{aligned} \nabla f &= \begin{cases} x\mathbf{e}_x + y\mathbf{e}_y + z\mathbf{e}_z & \text{in Cartesian coordinates} \\ \rho\mathbf{e}_\rho + z\mathbf{e}_z & \text{in cylindrical polar coordinates} \\ r\mathbf{e}_r & \text{in spherical polar coordinates} \end{cases} \\ &= \mathbf{x} \end{aligned}$$

3. Integration over lines

3.1. Line integrals

For a vector field $\mathbf{F}(\mathbf{x})$ and a piecewise smooth parametrised curve C defined by $[a,b] \ni t \mapsto \mathbf{x}(t)$, we define the line integral of F along C

$$\int_C \mathbf{F} \cdot \mathrm{d}\mathbf{x} = \int_a^b \mathbf{F}(\mathbf{x}(t)) \cdot \underbrace{\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}}_{\text{tangent vector}} \mathrm{d}t$$

Note that this tangent vector is not necessarily normalised, and note further that the curve direction matters. If we want to integrate in the other direction, it is common to write $\int_{-C}$ instead. We can think of this line integral as the work done by a particle moving along C in the presence of a force F. As an example, consider the vector field given by

$$\mathbf{F} = \begin{pmatrix} x^2 y \\ yz \\ 2xz \end{pmatrix}$$

Consider two curves connecting the origin to the position vector $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$.

$$C_1 : [0,1] \ni t \mapsto \begin{pmatrix} t \\ t \\ t \end{pmatrix}; \quad C_2 : [0,1] \ni t \mapsto \begin{pmatrix} t \\ t \\ t^2 \end{pmatrix}$$

$$\int_{C_1} \mathbf{F} \cdot \mathrm{d}\mathbf{x} = \int_0^1 \begin{pmatrix} t^3 \\ t^2 \\ 2t^2 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \mathrm{d}t = \frac{5}{4}$$

$$\int_{C_2} \mathbf{F} \cdot \mathrm{d}\mathbf{x} = \int_0^1 \begin{pmatrix} t^3 \\ t^3 \\ 2t^3 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 2t \end{pmatrix} \mathrm{d}t = \frac{13}{10}$$

In general, the result of the line integral depends on the path taken between the two points. In the force analogy, there might be a path between A and B that is very easy to traverse, and another path that is very difficult (i.e. uses a lot of energy).

Now, consider a particle at $\mathbf{x}$ experiencing a force $\mathbf{F}$, represented in cylindrical polar coordinates as

$$\mathbf{F}(\mathbf{x}) = z\rho \mathbf{e}_\phi$$

Consider the path C given by

$$C : [0, 2\pi] \ni t \mapsto \begin{pmatrix} a\cos t \\ a \sin t \\ t \end{pmatrix}$$

What is the work done by the particle travelling along C? Using the definition of the line element $\mathrm{d}\mathbf{x}$ in cylindrical polar coordinates, we can compute that $\mathbf{F}\mathrm{d}\mathbf{x} = z\rho^2\,\mathrm{d}\phi$. Note that in cylindrical polar coordinates, the path can be represented simply as $(\rho, \phi, z) = (a, t, t)$. Hence,

$$(\mathrm{d}\rho\,, \mathrm{d}\phi\,, \mathrm{d}z) = (0, \mathrm{d}t\,, \mathrm{d}t)$$

Therefore, $\mathbf{F} \cdot \mathrm{d}\mathbf{x} = z\rho^2\,\mathrm{d}t$. We can now compute the integral:

$$\int_C \mathbf{F} \cdot \mathrm{d}\mathbf{x} = a^2 \int_0^{2\pi} t\,\mathrm{d}t = 2\pi^2 a^2$$

3.2. Closed curves

A curve $[a, b] \ni t \mapsto \mathbf{x}(t)$ might be such that $\mathbf{x}(a) = \mathbf{x}(b)$. This is called a closed curve. The line integral around a closed loop is written

$$\oint_C \mathbf{F} \cdot \mathrm{d}\mathbf{x}$$

Sometimes, this is called the 'circulation' of $\mathbf{F}$ about C. Consider the first example from this lecture, with curves C_1 and C_2. Let $C = C_1 - C_2$. Then

$$\oint_C \mathbf{F} \cdot \mathrm{d}\mathbf{x} = \int_{C_1} \mathbf{F} \cdot \mathrm{d}\mathbf{x} - \int_{C_2} \mathbf{F} \cdot \mathrm{d}\mathbf{x} = \frac{-2}{15}$$

3.3. Conservative forces and exact differentials

We have seen how to interpret things like $\mathbf{F} \cdot \mathrm{d}\mathbf{x}$ when inside an integral. This is an example of a differential form; in orthogonal curvilinear coordinates (u, v, w) we have

$$\mathbf{F} \cdot \mathrm{d}\mathbf{x} = a\,\mathrm{d}u + b\,\mathrm{d}v + c\,\mathrm{d}w$$

for some a, b, c dependent on u, v, w. We say that $\mathbf{F} \cdot \mathrm{d}\mathbf{x}$ is exact if

$$\mathbf{F} \cdot \mathrm{d}\mathbf{x} = \mathrm{d}f$$

for some scalar function f. Recall that $\mathrm{d}f = \nabla f \cdot \mathrm{d}\mathbf{x}$. So equivalently, $\mathbf{F} \cdot \mathrm{d}\mathbf{x}$ is exact if and only if

$$\mathbf{F} = \nabla f$$

Such a vector field is called conservative. $\mathbf{F} \cdot \mathrm{d}\mathbf{x}$ is exact if and only if $\mathbf{F}$ is conservative. Using the properties that $\mathrm{d}(\alpha f + \beta g) = \alpha\,\mathrm{d}f + \beta\,\mathrm{d}g$, $\mathrm{d}(fg) = g\,\mathrm{d}f + f\,\mathrm{d}g$ and so on, it is usually easy to see if a differential form is exact.

Proposition. If θ is an exact differential form, then

$$\oint_C \theta = 0$$

for any closed curve C.

Proof. If θ is exact, then $\theta = \nabla f \cdot \mathrm{d}\mathbf{x}$ for some scalar function f. Given a curve $C : [a,b] \ni t \mapsto \mathbf{x}(t)$,

$$\begin{aligned}\oint_C \theta &= \oint_C \nabla f \cdot \mathrm{d}\mathbf{x} \\ &= \int_a^b \nabla f(\mathbf{x}(t)) \cdot \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\,\mathrm{d}t\end{aligned}$$

By the previous lecture,

$$\begin{aligned}&= \int_a^b \frac{\mathrm{d}}{\mathrm{d}t}\left[f(\mathbf{x}(t))\right]\mathrm{d}t \\ &= f(\mathbf{x}(a)) - f(\mathbf{x}(b)) \\ &= 0\end{aligned}$$

since $\mathbf{x}(a) = \mathbf{x}(b)$. □

Note, for example in cylindrical polar coordinates, that $f(\rho,\phi,z) = \phi$ is not a function on $\mathbb{R}^3$, since there are many possible values of ϕ for any given position vector. These are called multi-valued functions; for example the contour integral of this function over a circle where $\phi \in [0, 2\pi]$ is not well-defined, since $f(\rho,0,z) \neq f(\rho,2\pi,z)$.

Note that if $\mathbf{F}$ is conservative, then the circulation of $\mathbf{F}$ around any closed curve C vanishes. This means that the line integral between A and B is not dependent on the path chosen between the two points; simply choose the most convenient curve for the problem.

Let $(u,v,w) = (u_1,u_2,u_3)$ be a set of orthogonal curvilinear coordinates. Let

$$\mathbf{F}\cdot\mathrm{d}\mathbf{x} = \theta = \underbrace{A(u,v,w)\,\mathrm{d}u}_{\theta_1} + \underbrace{B(u,v,w)\,\mathrm{d}v}_{\theta_2} + \underbrace{C(u,v,w)\,\mathrm{d}w}_{\theta_3} = \theta_i\,\mathrm{d}u_i$$

A necessary condition for θ to be exact is

$$\frac{\partial\theta_i}{\partial u_j} = \frac{\partial\theta_j}{\partial u_i} \tag{†}$$

Indeed, if θ is exact, then $\theta = \mathrm{d}f$, so

$$\theta = \frac{\partial f}{\partial u_i}\,\mathrm{d}u_i \iff \theta_i = \frac{\partial f}{\partial u_i}$$

and therefore,

$$\frac{\partial\theta_i}{\partial u_j} = \frac{\partial^2 f}{\partial u_j \partial u_i} = \frac{\partial\theta_j}{\partial u_i}$$

A differential form $\theta = \theta_i\,\mathrm{d}u_i$ that obeys (†) is called a *closed* differential form. Certainly any exact differential form is closed. A differential form is exact if it is closed *and* the domain $\Omega \subset$

$\mathbb{R}^3$ on which θ is defined is simply connected, i.e. all closed loops in Ω can be continuously 'shrunk' to any point inside Ω without leaving it. This is notable, since one direction of implication is related to calculus, but the other direction is related to topology.

Now, let us consider an example. Let

$$\theta = y\,\mathrm{d}x - x\,\mathrm{d}y$$

Is this differential form exact? First, we will check if it is closed.

$$\frac{\partial}{\partial y}y = 1; \quad \frac{\partial}{\partial x}(-x) = -1$$

It is not closed, so it is not exact. As another example, let us compute the line integral

$$\int_C 3x^2y\,\mathrm{d}x + x^3\,\mathrm{d}y$$

where

$$C : [\alpha_1, \alpha_{100}] \ni t \mapsto \begin{pmatrix} \cos\left(\mathrm{Im}\left(\zeta\left(\frac{1}{2} + it\right)\right)\right) \\ \sin\left(\mathrm{Re}\left(\zeta\left(\frac{1}{2} + it\right)\right)\right) \\ 0 \end{pmatrix}$$

where α_1 and α_{100} are the 1st and 100th zeroes of $\zeta\left(\frac{1}{2} + it\right)$. The loop is closed and exact; $\mathrm{d}\left(x^3y\right) = 3x^2y\,\mathrm{d}x + x^3\,\mathrm{d}y$. So the result is zero. As a final example, consider a particle travelling along a curve $C : [a, b] \ni t \mapsto \mathbf{x}(t)$. Then the work done is

$$\begin{aligned} W &= \int_C \mathbf{F} \cdot \mathrm{d}\mathbf{x} \\ &= m\int_a^b \ddot{\mathbf{x}} + \dot{\mathbf{x}}\,\mathrm{d}t \\ &= \frac{1}{2}m\,|\dot{\mathbf{x}}|^2\Big|_a^b \end{aligned}$$

which is the change in kinetic energy. If $\mathbf{F} = -\nabla V$, i.e. $\mathbf{F}$ is conservative,

$$\int_C \mathbf{F} \cdot \mathrm{d}\mathbf{x} = -\int_C \nabla V \cdot \mathrm{d}\mathbf{x} = V(\mathbf{x}(a)) - V(\mathbf{x}(b))$$

So the change in kinetic energy is equal to the change in potential energy; energy is conserved.

4. Integration in Euclidean space

4.1. Definition of integral in two dimensions

We can integrate over a bounded region $D \subset \mathbb{R}^2$. To do this, we can cover D with small, disjoint sets A_{ij} each with area δA_{ij}. Each of these sets A_{ij} are contained in a disc of radius $\varepsilon > 0$. Let (x_i, y_j) be points contained in each A_{ij}. We now define

$$\int_D f(\mathbf{x}) \, \mathrm{d}A = \lim_{\varepsilon \to 0} \sum_{i,j} f(x_i, y_j) \, \delta A_{ij}$$

The integral exists if it is independent of the choice of partitions A_{ij} and the points (x_i, y_j). The obvious choice of partitioning D is to use rectangles where the area of each rectangle is $\delta A_{ij} = \delta x_i \delta y_j$. We can create horizontal 'strips' of height δy which we can integrate over. The possible x coordinates for this strip are $x_y = \{x : (x, y) \in D\}$. We can take the limit as $\delta x \to 0$, giving

$$\delta y \int_{x_y} f(x, y) \, \mathrm{d}x$$

Summing over each such strip, taking the limit as $\delta y \to 0$, we have

$$\int_D f(x, y) \, \mathrm{d}A = \int_Y \left(\int_{x_y} f(x, y) \, \mathrm{d}x \right) \mathrm{d}y$$

where Y is the set of all possible y coordinates, i.e. $Y = \{y : \exists x, (x, y) \in D\}$. We can equivalently sum over all vertical strips, and get

$$\int_D f(x, y) \, \mathrm{d}A = \int_X \left(\int_{y_x} f(x, y) \, \mathrm{d}y \right) \mathrm{d}x$$

More concisely, we can write the following (Fubini's Theorem):

$$\mathrm{d}A = \mathrm{d}x \, \mathrm{d}y = \mathrm{d}y \, \mathrm{d}x$$

Let us consider an example; let D be the triangle with vertices $(0, 0), (1, 0), (0, 1)$. If $f(x, y) = xy^2$, then by integrating over horizontal strips, we have

$$\begin{aligned} \int_D f(x, y) \, \mathrm{d}A &= \int_0^1 \left(\int_0^{1-y} xy^2 \, \mathrm{d}x \right) \mathrm{d}y \\ &= \int_0^1 \left[\frac{1}{2} x^2 y^2 \right]_0^{1-y} \mathrm{d}y \\ &= \int_0^1 \frac{1}{2} (1 - y)^2 y^2 \, \mathrm{d}y \\ &= \frac{1}{60} \end{aligned}$$

Instead, integrating over vertical strips, we have

$$\begin{aligned}\int_D f(x,y)\,\mathrm{d}A &= \int_0^1 \left(\int_0^{1-x} xy^2\,\mathrm{d}y\right)\mathrm{d}x \\ &= \int_0^1 \left[\frac{1}{3}xy^3\right]_0^{1-x}\mathrm{d}x \\ &= \int_0^1 \frac{1}{3}x(1-x)^3\,\mathrm{d}x \\ &= \frac{1}{60}\end{aligned}$$

Note that if $f(x,y) = g(x)\cdot h(y)$, and D is a rectangle $\{(x,y) : a \le x \le b, c \le y \le d\}$, then

$$\int_A f(x,y)\,\mathrm{d}A = \left(\int_a^b g(x)\,\mathrm{d}x\right)\left(\int_c^d h(y)\,\mathrm{d}y\right)$$

4.2. Change of variables

It can be useful to introduce a change of variables in order to compute the one-dimensional integral. For example, if x is represented as a function of u,

$$\int_a^b f(x)\,\mathrm{d}x = \int_{x^{-1}(a)}^{x^{-1}(b)} f(x(u))\frac{\mathrm{d}x}{\mathrm{d}u}\,\mathrm{d}u$$

Note that if $\frac{\mathrm{d}x}{\mathrm{d}u} > 0$, then the right hand side integral is taken over a limit from a smaller value to a larger one, but if $\frac{\mathrm{d}x}{\mathrm{d}u} < 0$, then the integral is the 'wrong way round'. If $I = [a,b]$ and $I' = x^{-1}I$, we have

$$\int_I f(x)\,\mathrm{d}x = \int_{I'} f(x(u))\left|\frac{\mathrm{d}x}{\mathrm{d}u}\right|\mathrm{d}u$$

where the absolute value is used since I' is defined as going from the lower limit to the upper limit. There is a similar formula in 2D.

Proposition. Let $\mathbf{x}(u,v) = (x(u,v), y(u,v))$ be a smooth, invertible transformation with a smooth inverse that maps the region D' in the (u,v) plane to the region D in the (x,y) plane. (This map must be a bijection; every point must have a unique inverse.) Then

$$\iint_D f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \iint_{D'} f(x(u,v),y(u,v))\left|\frac{\partial(x,y)}{\partial(u,v)}\right|\mathrm{d}u\,\mathrm{d}v$$

where

$$\frac{\partial(x,y)}{\partial(u,v)} = J = \det\begin{pmatrix}\partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v\end{pmatrix} = \det\left(\frac{\partial\mathbf{x}}{\partial u}\,\middle|\,\frac{\partial\mathbf{x}}{\partial v}\right)$$

is the Jacobian determinant. More concisely,

$$\mathrm{d}x\,\mathrm{d}y = |J|\,\mathrm{d}u\,\mathrm{d}v$$

It doesn't matter if the Jacobian vanishes at a single point, since the area of a single point is zero and hence will have no contribution to the result. The Jacobian being zero means that something non-smooth is happening at this point, so it is important to consider why this point is special.

Proof. We can form a partition of D by using the image of a rectangular partition of D'. Let the rectangular partition be characterised by a horizontal step δx and a vertical step of δy. Then each small rectangle in D' is mapped to some small (not necessarily rectangular) region in D', with vertices

$$\mathbf{x}(u_i, v_j), \mathbf{x}(u_{i+1}, v_j), \mathbf{x}(u_{i+1}, v_{j+1}), \mathbf{x}(u_i, v_{j+1})$$

To first order, the area of this region is the area of the parallelogram with the same vertices. Two of the sides of the parallelogram are

$$\mathbf{x}(u_{i+1}, v_j) - \mathbf{x}(u_i, v_j) \approx \frac{\partial \mathbf{x}}{\partial u}(u_i, v_j)\delta u$$

$$\mathbf{x}(u_i, v_{j+1}) - \mathbf{x}(u_i, v_j) \approx \frac{\partial \mathbf{x}}{\partial v}(u_i, v_j)\delta v$$

So the area of the parallelogram is approximately

$$\begin{aligned}\left|\frac{\partial \mathbf{x}}{\partial u}(u_i, v_j)\delta u \cdot \frac{\partial \mathbf{x}}{\partial v}(u_i, v_j)\delta v\right| &= \left|\det\left(\frac{\partial \mathbf{x}}{\partial u}(u_i, v_j) \,\middle|\, \frac{\partial \mathbf{x}}{\partial v}(u_i, v_j)\right)\right| \\ &= \left|J(u_i, v_j)\right| \delta u\, \delta v \\ &= \delta A_{ij}\end{aligned}$$

Hence,

$$\begin{aligned}\int_D f\,\mathrm{d}A &= \lim_{\varepsilon \to 0} \sum_{ij} f(x_i, y_j)\, \delta A_{ij} \\ &= \lim_{\varepsilon \to 0} \sum_{ij} f(x(u_i, v_j), y(u_i, v_j))\, \left|J(u_i, v_j)\right| \delta u\, \delta v \\ &= \iint_{D'} f(x(u, v), y(u, v))\, \left|J(u_i, v_j)\right| \mathrm{d}u\, \mathrm{d}v\end{aligned}$$

□

As an example, let us consider polar coordinates (ρ, ϕ), where

$$x(\rho, \phi) = \rho \cos \phi; \quad y(\rho, \phi) = \rho \sin \phi$$

Hence,

$$|J| = \left|\det \begin{pmatrix} \cos\phi & -\rho\sin\phi \\ \sin\phi & \rho\cos\phi \end{pmatrix}\right| = |\rho| = \rho$$

If $D = \{(x,y) : x > 0,\ y > 0,\ x^2 + y^2 < r^2\}$, which is a quarter-circle of radius r in the first quadrant, then $D' = \{(\rho,\phi) : 0 < \rho < r,\ 0 < \phi < \frac{\pi}{2}\}$. This is notably a rectangle in polar coordinates.

$$\iint_D f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \iint_{D'} f(\rho\cos\phi, \rho\sin\phi)\,\rho\,\mathrm{d}\rho\,\mathrm{d}\phi$$

So, for example, if we let $r \to \infty$, then

$$\int_{x=0}^{\infty}\int_{y=0}^{\infty} f(x,y)\,\mathrm{d}y\,\mathrm{d}x = \int_{\phi=0}^{\frac{\pi}{2}}\int_{\rho=0}^{\infty} f(\rho\cos\phi, \rho\sin\phi)\,\rho\,\mathrm{d}\rho\,\mathrm{d}\phi$$

Consider

$$I = \int_0^{\infty} e^{-x^2}\,\mathrm{d}x$$

Then,

$$\begin{aligned} I^2 &= \int_0^{\infty} e^{-x^2}\,\mathrm{d}x \cdot \int_0^{\infty} e^{-y^2}\,\mathrm{d}y \\ &= \int_{x=0}^{\infty}\int_{y=0}^{\infty} e^{-x^2-y^2}\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{\phi=0}^{\frac{\pi}{2}}\int_{\rho=0}^{\infty} e^{-\rho^2}\,\rho\,\mathrm{d}\rho\,\mathrm{d}\phi \\ &= \frac{\pi}{2}\int_0^{\infty} \frac{\mathrm{d}}{\mathrm{d}\rho}\left(-\frac{1}{2}e^{-\rho^2}\right)\mathrm{d}\rho \\ &= \frac{\pi}{4} \\ \implies I &= \frac{\sqrt{\pi}}{2} \end{aligned}$$

4.3. Definition of integral in three dimensions

To integrate over regions V in $\mathbb{R}^3$, we can use similar ideas to those discussed in the previous lecture.

$$\int_V f(\mathbf{x})\,\mathrm{d}V = \lim_{\varepsilon\to 0}\sum_{i,j,k} f(x_i, y_j, z_k)\,\delta V_{ijk}$$

where the δV_{ijk} partition V, and each contain the point (x_i, y_j, z_k). In this case, the volume element satisfies

$$\mathrm{d}V = \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$$

The integrals may be computed in any order. As an example, consider the simplex defined by

$$V = \{x > 0,\, y > 0,\, z > 0,\, x + y + z < 1\}$$

We can compute the volume using the integral

$$\begin{aligned} I &= \int_{z=0}^{1} \int_{y=0}^{1-z} \int_{x=0}^{1-y-z} 1 \,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \\ &= \int_{z=0}^{1} \int_{y=0}^{1-z} (1-y-z)\,\mathrm{d}y\,\mathrm{d}z \\ &= \int_{z=0}^{1} \left((1-z) - \frac{1}{2}(1-z)^2 - (1-z)z\right)\mathrm{d}z \\ &= \left[z - \frac{1}{2}z^2 - \frac{1}{2}z + \frac{1}{2}z^2 - \frac{1}{6}z^3 - \frac{1}{2}z^2 + \frac{1}{3}z^3\right]_{z=0}^{1} \\ &= \frac{1}{6} \end{aligned}$$

We can compute things like the centre of mass, assuming it has constant density $\rho = 1$. Then

$$\mathbf{X} = \frac{1}{m}\int_V \rho \mathbf{x}\,\mathrm{d}V = \frac{1}{4}\begin{pmatrix}1\\1\\1\end{pmatrix}$$

Proposition. Let $x(u,v,w), y(u,v,w), z(u,v,w)$ be a continuously differentiable bijection with a continuously differentiable inverse, that maps the volume V' to V. The integral

$$\iiint_V f(x,y,z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z = \iiint_{V'} f(x(u,v,w), y(u,v,w), z(u,v,w))\,|J|\,\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}w$$

where

$$J = \det\left(\frac{\partial \mathbf{x}}{\partial u}\,\middle|\,\frac{\partial \mathbf{x}}{\partial v}\,\middle|\,\frac{\partial \mathbf{x}}{\partial w}\right)$$

More concisely,

$$\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z = |J|\,\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}w$$

The Jacobian comes from the fact that the volume of a parallepiped generated by the vectors

$$\frac{\partial \mathbf{x}}{\partial u}\delta u,\ \frac{\partial \mathbf{x}}{\partial v}\delta v,\ \frac{\partial \mathbf{x}}{\partial w}\delta w$$

is precisely the determinant of the Jacobian matrix multiplied by $\delta u\,\delta v\,\delta w$. The rest of this proof follows from the two-dimensional case. As an example, let us consider cylindrical polar coordinates $(u,v,w) = (\rho, \phi, z)$.

$$\mathrm{d}V = \rho\,\mathrm{d}\rho\,\mathrm{d}\phi\,\mathrm{d}z\,; \quad |J| = \rho$$

In spherical polar coordinates $(u,v,w) = (r,\theta,\phi)$,

$$\mathrm{d}V = r^2 \sin\theta\,\mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}\phi\,; \quad |J| = r^2\sin\theta$$

4.4. Calculating volumes

We can use the volume element to calculate, for example, the volume of a ball of radius R. To begin, let us use Cartesian coordinates.

$$\begin{aligned}
\int_V \mathrm{d}V &= \int_{z=-R}^{R} \mathrm{d}z \int_{y=-\sqrt{R^2-z^2}}^{\sqrt{R^2-z^2}} \mathrm{d}y \int_{x=-\sqrt{R^2-z^2-y^2}}^{\sqrt{R^2-z^2-y^2}} \mathrm{d}x \\
&= \int_{z=-R}^{R} \mathrm{d}z \int_{y=-\sqrt{R^2-z^2}}^{\sqrt{R^2-z^2}} \mathrm{d}y \left[2\sqrt{R^2-z^2-y^2}\right] \\
&= \int_{z=-R}^{R} \mathrm{d}z \left[y\sqrt{R^2-z^2-y^2} + (R^2-z^2)\arctan\left(\frac{y}{\sqrt{R^2-z^2-y^2}}\right)_{y=-\sqrt{R^2-z^2}}^{\sqrt{R^2-z^2}} \right] \\
&= \int_{z=-R}^{R} \mathrm{d}z \left[\pi(R^2-z^2)\right] \\
&= \frac{4}{3}\pi R^3
\end{aligned}$$

We can alternatively use spherical polar coordinates.

$$\begin{aligned}
\int_V \mathrm{d}V &= \int_{r=0}^{R} \mathrm{d}r \int_{\theta=0}^{\pi} \mathrm{d}\theta \int_{\phi=0}^{2\pi} \mathrm{d}\phi \cdot r^2 \sin\theta \\
&= \int_{r=0}^{R} r^2\,\mathrm{d}r \int_{\theta=0}^{\pi} \sin\theta\,\mathrm{d}\theta \int_{\phi=0}^{2\pi} \mathrm{d}\phi \\
&= \int_{r=0}^{R} r^2\,\mathrm{d}r \cdot \int_{\theta=0}^{\pi} \sin\theta\,\mathrm{d}\theta \cdot \int_{\phi=0}^{2\pi} \mathrm{d}\phi \\
&= \frac{1}{3}R^3 \cdot 2 \cdot 2\pi \\
&= \frac{4}{3}\pi R^3
\end{aligned}$$

This is clearly a much cleaner computation. Now, consider the a ball of radius a with cylinder of radius $b < a$ removed from the centre aligned with the z axis. To calculate this volume, the symmetry of the problem suggests we might want to use cylindrical polar coordinates.

$$V = \{(\rho, \phi, z) : \ 0 < \rho^2 + z^2 < a^2,\ b < \rho < a\}$$

$$\begin{aligned}
\int_V \mathrm{d}V &= \int_{\rho=b}^{a} \rho\,\mathrm{d}\rho \int_{\phi=0}^{2\pi} \mathrm{d}\phi \int_{z=-\sqrt{a^2-\rho^2}}^{\sqrt{a^2-\rho^2}} \mathrm{d}z \\
&= 2\pi \int_b^a 2\rho\sqrt{a^2-\rho^2}\,\mathrm{d}\rho \\
&= \frac{4}{3}\pi(a^2-b^2)^{\frac{3}{2}}
\end{aligned}$$

5. Integration over surfaces

5.1. Two-dimensional surfaces

A two-dimensional surface in $\mathbb{R}^3$ can be defined implicitly using a function $f : \mathbb{R}^3 \to \mathbb{R}$, with

$$S = \{\mathbf{x} \in \mathbb{R}^3 : \ f(\mathbf{x}) = 0\}$$

The normal to S at $\mathbf{x}$ is parallel to $\nabla f(\mathbf{x})$. We call the surface regular if $\nabla f(\mathbf{x}) \neq \mathbf{0}$ everywhere on the surface. For example, consider

$$S = \{(x, y, z) : \ x^2 + y^2 + z^2 - 1 = 0\}$$

Then

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 2x \\ 2y \\ 2z \end{pmatrix} = 2\mathbf{x}$$

which is clearly normal to S at $\mathbf{x}$. Some surfaces have a boundary, for instance a hemisphere.

$$S = \{(x, y, z) : \ x^2 + y^2 + z^2 - 1 = 0, \ z \geq 0\}$$

We label the boundary ∂S, so

$$\partial S = \{(x, y, z) : \ x^2 + y^2 = 1, \ z = 0\}$$

In this course, a surface will either have no boundary or its boundary will be made of piecewise smooth curves. If S has no boundary, we say that S is a closed surface. It is often useful to parametrise a surface using some coordinates (u, v).

$$S = \{\mathbf{x} = \mathbf{x}(u, v) : \ (u, v) \in D\}$$

where D is some region in the u-v plane. For a hemisphere, we can use spherical polar coordinates:

$$S = \left\{ \mathbf{x} = \mathbf{x}(\theta, \phi) = \begin{pmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{pmatrix} : \ 0 \leq \theta \leq \frac{\pi}{2}, \ 0 \leq \phi \leq 2\pi \right\}$$

We call a parametrisation of S regular if

$$\frac{\partial \mathbf{x}}{\partial u} \times \frac{\partial \mathbf{x}}{\partial v} \neq \mathbf{0}$$

everywhere on the surface. Note that $\frac{\partial \mathbf{x}}{\partial u}$ is the tangent in one direction, and $\frac{\partial \mathbf{x}}{\partial v}$ is the tangent in another direction, so their cross product should be normal to the surface.

$$\hat{\mathbf{n}} = \frac{\frac{\partial \mathbf{x}}{\partial u} \times \frac{\partial \mathbf{x}}{\partial v}}{\left| \frac{\partial \mathbf{x}}{\partial u} \times \frac{\partial \mathbf{x}}{\partial v} \right|}$$

This normal will vary smoothly with respect to u and v, if we are moving across a smooth part of the curve. Choosing a consistent normal over S gives a way to give an orientation to the boundary ∂S. We make the convention that normal vectors near you should be on your left as you traverse ∂S.

5.2. Areas and integrals over surfaces

Consider a parametrised surface

$$S = \{\mathbf{x} = \mathbf{x}(u, v) : \ (u, v) \in D\}$$

The integral over S cannot be of the form

$$\iint_D \mathrm{d}u \, \mathrm{d}v$$

since a patch of area $\delta u \, \delta v$ in D will not in general correspond to a patch of area $\delta u \, \delta v$ in S. Note that the small change $u \mapsto u + \delta u$ produces a change

$$\mathbf{x}(u + \delta u, v) - \mathbf{x}(u, v) \approx \frac{\partial \mathbf{x}}{\partial u} \delta u$$

Similarly, changing v, we have

$$\mathbf{x}(u, v + \delta v) - \mathbf{x}(u, v) \approx \frac{\partial \mathbf{x}}{\partial v} \delta v$$

So the patch of area $\delta u \, \delta v$ in D corresponds (to first order) to a parallelogram of area

$$\left| \frac{\partial \mathbf{x}}{\partial u} \times \frac{\partial \mathbf{x}}{\partial v} \right| \delta u \, \delta v$$

This leads us to define the scalar area element and the vector area element as follows:

$$\mathrm{d}S = \left| \frac{\partial \mathbf{x}}{\partial u} \times \frac{\partial \mathbf{x}}{\partial v} \right| \mathrm{d}u \, \mathrm{d}v$$

$$\mathrm{d}\mathbf{S} = \frac{\partial \mathbf{x}}{\partial u} \times \frac{\partial \mathbf{x}}{\partial v} \mathrm{d}u \, \mathrm{d}v = \hat{\mathbf{n}} \, \mathrm{d}S$$

So for instance the area of S is given by

$$\int_S \mathrm{d}S = \iint_D \left| \frac{\partial \mathbf{x}}{\partial u} \times \frac{\partial \mathbf{x}}{\partial v} \right| \mathrm{d}u \, \mathrm{d}v$$

As an example, consider the hemisphere of radius R.

$$S = \left\{ \mathbf{x} = \mathbf{x}(\theta, \phi) = \begin{pmatrix} R \sin\theta \cos\phi \\ R \sin\theta \sin\phi \\ R \cos\theta \end{pmatrix} = R\mathbf{e}_r : \ 0 \le \theta \le \frac{\pi}{2}, \ 0 \le \phi \le 2\pi \right\}$$

So

$$\frac{\partial \mathbf{x}}{\partial \theta} = \begin{pmatrix} R \cos\theta \cos\phi \\ R \cos\theta \sin\phi \\ -R \sin\theta \end{pmatrix} = R\mathbf{e}_\theta$$

$$\frac{\partial \mathbf{x}}{\partial \phi} = \begin{pmatrix} -R \sin\theta \sin\phi \\ R \sin\theta \cos\phi \\ 0 \end{pmatrix} = R \sin\theta \mathbf{e}_\phi$$

Hence

$$\mathrm{d}S = R^2 \sin\theta \left|\mathbf{e}_\theta \times \mathbf{e}_\phi\right| \mathrm{d}\theta\, \mathrm{d}\phi = R^2 \sin\theta\, \mathrm{d}\theta\, \mathrm{d}\phi$$

So the surface area of the hemisphere is

$$\int_{\theta=0}^{\frac{\pi}{2}} \mathrm{d}\theta \int_{\phi=0}^{2\pi} \mathrm{d}\phi\ R^2 \sin\theta = 2\pi R^2$$

Here is another example. Suppose the velocity of a fluid is $\mathbf{u}(\mathbf{x})$. Given a surface S, we might like to calculate how much fluid passes through it per unit time. On a small patch δS on S, the fluid passing through the small patch would be $(u \cdot \delta\mathbf{S})\,\delta t$ in time δt, where $\delta\mathbf{S}$ is the normal direction to the area δS. Over the whole surface, the amount that passes over S in δt is

$$\delta t \int_S \mathbf{u} \cdot \mathrm{d}\mathbf{S}$$

This kind of integral is called a 'flux integral'.

5.3. Choice of parametrisation of surfaces

Let $\mathbf{x} = \mathbf{x}(u, v)$ and $\mathbf{x} = \tilde{\mathbf{x}}(\tilde{u}, \tilde{v})$ be two different parametrisations of S with $(u, v) \in D$ and $(\tilde{u}, \tilde{v}) \in \tilde{D}'$. Since every coordinate in S has a pre-image in both D and D', there must be a relationship

$$\mathbf{x}(u, v) = \tilde{\mathbf{x}}(\tilde{u}(u, v), \tilde{v}(u, v))$$

By the chain rule,

$$\begin{aligned}
\frac{\partial \mathbf{x}}{\partial u} \times \frac{\partial \mathbf{x}}{\partial v} &= \left(\frac{\partial \tilde{\mathbf{x}}}{\partial \tilde{u}}\frac{\partial \tilde{u}}{\partial u} + \frac{\partial \tilde{\mathbf{x}}}{\partial \tilde{v}}\frac{\partial \tilde{v}}{\partial u}\right) \times \left(\frac{\partial \tilde{\mathbf{x}}}{\partial \tilde{u}}\frac{\partial \tilde{u}}{\partial v} + \frac{\partial \tilde{\mathbf{x}}}{\partial \tilde{v}}\frac{\partial \tilde{v}}{\partial v}\right) \\
&= \left(\frac{\partial \tilde{u}}{\partial u}\frac{\partial \tilde{v}}{\partial v} - \frac{\partial \tilde{u}}{\partial v}\frac{\partial \tilde{v}}{\partial u}\right)\left(\frac{\partial \tilde{\mathbf{x}}}{\partial \tilde{u}} \times \frac{\partial \tilde{\mathbf{x}}}{\partial \tilde{v}}\right) \\
&= \frac{\partial(\tilde{u}, \tilde{v})}{\partial(u, v)}\left(\frac{\partial \tilde{\mathbf{x}}}{\partial \tilde{u}} \times \frac{\partial \tilde{\mathbf{x}}}{\partial \tilde{v}}\right)
\end{aligned}$$

Hence,

$$\begin{aligned}
\int_S f\, \mathrm{d}S &= \iint_{\tilde{D}} f(\tilde{\mathbf{x}}(\tilde{u}, \tilde{v}))\left|\frac{\partial \tilde{\mathbf{x}}}{\partial \tilde{u}} \times \frac{\partial \tilde{\mathbf{x}}}{\partial \tilde{v}}\right| \mathrm{d}\tilde{u}\, \mathrm{d}\tilde{v} \\
&= \iint_D f(\mathbf{x}(u, v))\left|\frac{\partial \tilde{\mathbf{x}}}{\partial \tilde{u}} \times \frac{\partial \tilde{\mathbf{x}}}{\partial \tilde{v}}\right| \left|\frac{\partial(\tilde{u}, \tilde{v})}{\partial(u, v)}\right| \mathrm{d}u\, \mathrm{d}v \\
&= \iint_D f(\mathbf{x}(u, v))\left|\frac{\partial \mathbf{x}}{\partial u} \times \frac{\partial \mathbf{x}}{\partial v}\right| \mathrm{d}u\, \mathrm{d}v
\end{aligned}$$

So the result of the integral over the surface is independent of the choice of parametrisation.

6. Differential operators

6.1. Divergence, curl, and Laplacian

Recall the gradient operator ∇, which is defined in Cartesian coordinates as

$$\nabla = \mathbf{e}_i \frac{\partial}{\partial x_i}$$

For a vector field $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^3$, we define the divergence of $\mathbf{F}$ by

$$\nabla \cdot \mathbf{F}$$

In Cartesian coordinates,

$$\nabla \cdot \mathbf{F} = \left(\mathbf{e}_i \frac{\partial}{\partial x_i} \right) \cdot (F_j \mathbf{e}_j) = \frac{\partial F_i}{\partial x_i}$$

Note that the divergence of a vector field is a scalar field. We define the curl of $\mathbf{F}$ to be

$$\nabla \times \mathbf{F}$$

In Cartesian coordinates,

$$\nabla \times \mathbf{F} = \left(\mathbf{e}_j \frac{\partial}{\partial x_j} \right) \times (F_k \mathbf{e}_k) = e_j \times \left[\frac{\partial}{\partial x_j} (F_k \mathbf{e}_k) \right] = (\mathbf{e}_j \times \mathbf{e}_k) \frac{\partial F_k}{\partial x_j} = \varepsilon_{ijk} \frac{\partial F_k}{\partial x_j} \mathbf{e}_i$$

Hence (just in Cartesian coordinates):

$$[\nabla \times \mathbf{F}]_i = \varepsilon_{ijk} \frac{\partial F_k}{\partial x_j}$$

The curl of a vector field is another vector field. In terms of a 'formal' determinant, we can write

$$\nabla \times \mathbf{F} = \det \begin{pmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ \partial / \partial x_1 & \partial / \partial x_2 & \partial / \partial x_2 \\ F_1 & F_2 & F_3 \end{pmatrix}$$

We cannot trivially generalise the curl operator to spaces that do not have three spatial dimensions. Finally, we define the Laplacian of a scalar field $f : \mathbb{R}^3 \to \mathbb{R}$ as

$$\nabla^2 f \coloneqq \nabla \cdot \nabla f$$

In Cartesian coordinates, $[\nabla f]_i = \partial f / \partial x_i$, so

$$\nabla^2 f = \frac{\partial^2 f}{\partial x_i \partial x_i}$$

6.2. Explanation of divergence and curl

Consider

$$\mathbf{F}(\mathbf{x}) = \mathbf{x}$$

Using Cartesian coordinates,

$$\nabla \cdot \mathbf{F} = \frac{\partial}{\partial x_i} x_i = \delta_{ii} = 3$$

$$[\nabla \times \mathbf{F}]_i = \varepsilon_{ijk} \frac{\partial}{\partial x_j} x_k = \varepsilon_{ijk} \delta_{kj} = \varepsilon_{ijj} = 0$$

A positive divergence at a point indicates that the vector field is generally pointing away from that point. If thought of as a fluid, the point acts as a 'source' of fluid. A negative divergence indicates that the vector field is pointing towards that point, so it acts like a 'sink'. If a vector field has zero divergence, it can be thought of as representing the velocity of an incompressible fluid. The curl measures the local rotation of the vector field (or the related 'fluid') in a given direction. If the vector field was going anticlockwise in the $\mathbf{e}_1$-$\mathbf{e}_2$ plane, then the component of the curl in the $\mathbf{e}_3$ direction would be positive. If there is no local rotation, then the component is zero.

6.3. Identities

Proposition. For f, g scalar fields, $\mathbf{F}, \mathbf{G}$ vector fields, the following identities hold.

- $\nabla(fg) = (\nabla f)g + (\nabla g)f$
- $\nabla \cdot (f\mathbf{F}) = (\nabla f) \cdot \mathbf{F} + (\nabla \cdot \mathbf{F})f$
- $\nabla \times (f\mathbf{F}) = (\nabla f) \times \mathbf{F} + (\nabla \times \mathbf{F})f$
- $\nabla(\mathbf{F} \cdot \mathbf{G}) = \mathbf{F} \times (\nabla \times \mathbf{G}) + \mathbf{G} \times (\nabla \times \mathbf{F}) + (\mathbf{F} \cdot \nabla)\mathbf{G} + (\mathbf{G} \cdot \nabla)\mathbf{F}$
- $\nabla \times (\mathbf{F} \times \mathbf{G}) = \mathbf{F}(\nabla \cdot \mathbf{G}) - \mathbf{G}(\nabla \cdot \mathbf{F}) + (\mathbf{G} \cdot \nabla)\mathbf{F} - (\mathbf{F} \cdot \nabla)\mathbf{G}$
- $\nabla \cdot (\mathbf{F} \times \mathbf{G}) = (\nabla \times \mathbf{F}) \cdot \mathbf{G} - \mathbf{F} \cdot (\nabla \times \mathbf{G})$

Note, for example, that we can compute the dot product between vector fields and operators:

$$[(\mathbf{F} \cdot \nabla)\mathbf{G}]_i = \left(F_j \frac{\partial}{\partial x_j}\right) G_i = F_j \frac{\partial G_i}{\partial x_j}$$

Specifically, $\mathbf{F} \cdot \nabla$ is a differential operator, and $\nabla \cdot \mathbf{F}$ is a scalar field; they are not the same thing.

Proof. We will only prove the fifth one for now, as all the proofs are similar. The identities hold in any coordinate system, so we will choose the Cartesian coordinate system since the

basis vectors are the same everywhere.

$$\begin{aligned}
[\nabla \times (\mathbf{F} \times \mathbf{G})]_i &= \varepsilon_{ijk} \frac{\partial}{\partial x_j} (\mathbf{F} \times \mathbf{G})_k \\
&= \varepsilon_{ijk} \frac{\partial}{\partial x_j} \varepsilon_{klm} F_l G_m \\
&= \varepsilon_{ijk} \varepsilon_{klm} \frac{\partial}{\partial x_j} F_l G_m \\
&= \varepsilon_{ijk} \varepsilon_{klm} \left(F_l \frac{\partial G_m}{\partial x_j} + G_l \frac{\partial F_l}{\partial x_j} \right) \\
&= (\delta_{il}\delta_{jm} - \delta_{im}\delta_{jl}) \left(F_l \frac{\partial G_m}{\partial x_j} + G_l \frac{\partial F_l}{\partial x_j} \right) \\
&= F_i \frac{\partial G_j}{\partial x_j} - F_j \frac{\partial G_i}{\partial x_j} + G_j \frac{\partial F_i}{\partial x_j} - G_i \frac{\partial F_j}{\partial x_j} \\
&= [\mathbf{F}(\nabla \cdot \mathbf{G})]_i - [(\mathbf{F} \cdot \nabla)\mathbf{G}]_i + [(\mathbf{G} \cdot \nabla)\mathbf{F}]_i - [(\nabla \cdot \mathbf{F})\mathbf{G}]_i
\end{aligned}$$

□

6.4. Definitions in orthogonal curvilinear coordinate systems

For a general set of orthogonal curvilinear coordinates, divergence is defined by

$$\nabla \cdot \mathbf{F} = \left(\mathbf{e}_u \frac{1}{h_u} \frac{\partial}{\partial u} + \mathbf{e}_v \frac{1}{h_v} \frac{\partial}{\partial v} + \mathbf{e}_w \frac{1}{h_w} \frac{\partial}{\partial w} \right) \cdot (F_u \mathbf{e}_u + F_v \mathbf{e}_v + F_w \mathbf{e}_w)$$

We would get terms like

$$\begin{aligned}
\left(\mathbf{e}_u \frac{1}{h_u} \frac{\partial}{\partial u} \right) \cdot (F_v \mathbf{e}_v) &= \frac{1}{h_u} \mathbf{e}_u \cdot \left[\frac{\partial}{\partial u} (F_v \mathbf{e}_v) \right] \\
&= \frac{1}{h_u} \mathbf{e}_u \cdot \left[\frac{\partial F_v}{\partial u} \mathbf{e}_v + \frac{\partial \mathbf{e}_v}{\partial u} F_v \right] \\
&= \frac{F_v}{h_u} \left(\mathbf{e}_u \cdot \frac{\partial \mathbf{e}_v}{\partial u} \right)
\end{aligned}$$

We can combine all such terms and then derive that

$$\nabla \cdot \mathbf{F} = \frac{1}{h_u h_v h_w} \left[\frac{\partial}{\partial u} (h_v h_w F_u) + \frac{\partial}{\partial v} (h_u h_w F_v) + \frac{\partial}{\partial w} (h_u h_v F_w) \right]$$

$$\nabla \times \mathbf{F} = \frac{1}{h_u h_v h_w} \begin{vmatrix} h_u \mathbf{e}_u & h_v \mathbf{e}_v & h_w \mathbf{e}_w \\ \partial/\partial u & \partial/\partial v & \partial/\partial w \\ h_u F_u & h_v F_v & h_w F_w \end{vmatrix}$$

$$\nabla^2 f = \frac{1}{h_u h_v h_w} \left[\frac{\partial}{\partial u} \left(\frac{h_v h_w}{h_u} \frac{\partial f}{\partial u} \right) + \frac{\partial}{\partial v} \left(\frac{h_u h_w}{h_v} \frac{\partial f}{\partial v} \right) + \frac{\partial}{\partial w} \left(\frac{h_u h_v}{h_w} \frac{\partial f}{\partial w} \right) \right]$$

For cylindrical polar coordinates (ρ, ϕ, z), we have $(h_\rho, h_\phi, h_z) = (1, \rho, 1)$ and hence

$$\nabla \cdot \mathbf{F} = \frac{1}{\rho}\frac{\partial}{\partial \rho}(\rho F_\rho) + \frac{1}{\rho}\frac{\partial F_\phi}{\partial \phi} + \frac{\partial F_z}{\partial z}$$

$$\nabla \times \mathbf{F} = \frac{1}{\rho}\begin{vmatrix} \mathbf{e}_\rho & \rho\mathbf{e}_\phi & \mathbf{e}_z \\ \partial/\partial\rho & \partial/\partial\phi & \partial/\partial z \\ F_\rho & \rho F_\phi & F_z \end{vmatrix}$$

$$\nabla^2 f = \frac{1}{\rho}\frac{\partial}{\partial \rho}\left(\rho\frac{\partial f}{\partial \rho}\right) + \frac{1}{\rho^2}\frac{\partial^2 f}{\partial \phi^2} + \frac{\partial^2 f}{\partial z^2}$$

For spherical polar coordinates (r, θ, ϕ), we have $(h_r, h_\theta, h_\phi) = (1, r, r\sin\theta)$ and hence

$$\nabla \cdot \mathbf{F} = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2 F_r) + \frac{1}{r\sin\theta}\frac{\partial}{\partial \theta}(\sin\theta\, F_\theta) + \frac{1}{r\sin\theta}\frac{\partial F_\phi}{\partial \phi}$$

$$\nabla \times \mathbf{F} = \frac{1}{r^2\sin\theta}\begin{vmatrix} \mathbf{e}_r & r\mathbf{e}_\theta & r\sin\theta\,\mathbf{e}_\phi \\ \partial/\partial r & \partial/\partial\theta & \partial/\partial\phi \\ F_r & rF_\theta & r\sin\theta\, F_\phi \end{vmatrix}$$

$$\nabla^2 f = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial f}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta\frac{\partial f}{\partial \theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 f}{\partial \phi^2}$$

6.5. Laplacian of a vector field

The Laplacian of a vector field might be expected to be something like $\nabla \cdot (\nabla \mathbf{F})$. However, we have not defined the gradient of a vector field. In Cartesian coordinates, it would make sense that

$$\nabla^2 \mathbf{F} = \nabla^2(F_i \mathbf{e}_i) = (\nabla^2 F_i)\mathbf{e}_i \qquad (\dagger)$$

If this is the case, we can show then that, in Cartesian coordinates,

$$\nabla^2 \mathbf{F} = \nabla(\nabla \cdot \mathbf{F}) - \nabla \times (\nabla \times \mathbf{F})$$

In other words, in Cartesian coordinates,

$$[\nabla^2 \mathbf{F}]_i = \frac{\partial^2 F_i}{\partial x_j \partial x_j} = \nabla^2(F_i)$$

Since the right hand side of (†) is well-defined in any orthogonal curvilinear coordinate system, we will use it as a definition.

6.6. Relations between differential operators

Proposition. For a scalar field f and a vector field $\mathbf{F}$,

$$\nabla \times \nabla f = \mathbf{0}$$

and

$$\nabla \cdot \nabla \times \mathbf{F} = 0$$

In other words, curl ∘ grad gives zero, and div ∘ curl gives zero.

Proof. We will use Cartesian coordinates for simplicity.

$$\begin{aligned}[\nabla \times \nabla f]_i &= \varepsilon_{ijk} \frac{\partial}{\partial x_j} \left(\frac{\partial f}{\partial x_k} \right) \\ &= \varepsilon_{ijk} \frac{\partial^2 f}{\partial x_j \partial x_k}\end{aligned}$$

ε_{ijk} is antisymmetric in j and k, but $\frac{\partial^2 f}{\partial x_j \partial x_k}$ is symmetric in j and k. Hence the result is zero. Further,

$$\begin{aligned}\nabla \cdot \nabla \times \mathbf{F} &= \frac{\partial}{\partial x_i} \varepsilon_{ijk} \frac{\partial}{\partial x_j} F_k \\ &= \varepsilon_{ijk} \frac{\partial^2 F_k}{\partial x_i \partial x_j}\end{aligned}$$

Once again the ε term is antisymmetric and the partial derivative is symmetric, so the result follows. □

6.7. Irrotational and solenoidal forces

As a short aside, 'simply connected' means that any loop in a space can be 'shrunk' to any point within that space. It can also be referred to as '1-connected' since the loop is a one-dimensional manifold. For example, $\mathbb{R}^3$ is 1-connected, but $\mathbb{R}^3$ with the z-axis removed is not 1-connected; a loop around this axis cannot be shrunk to a point away from the axis.

We can write that a space is '2-connected' if it is 1-connected and any 2-manifold (surface) can be shrunk to any point within the space. Certainly $\mathbb{R}^3$ is 2-connected, but for example $\mathbb{R}^3$ without the origin is not 2-connected. The space is certainly 1-connected, but it is not 2-connected because a surface around the origin cannot be shrunk to a point away from the origin.

Recall that $\mathbf{F}$ is conservative if we can write $\mathbf{F} = \nabla f$. We say that $\mathbf{F}$ is irrotational if $\nabla \times \mathbf{F} = \mathbf{0}$. Hence, any conservative function is irrotational. The converse is true if the domain of $\mathbf{F}$ is 1-connected. We say that $\mathbf{F}$ is solenoidal if $\nabla \cdot \mathbf{F} = 0$. If there exists a vector potential $\mathbf{A}$ for $\mathbf{F}$, i.e. $\mathbf{F} = \nabla \times \mathbf{A}$, then $\mathbf{F}$ is solenoidal. The converse is true if the domain of $\mathbf{F}$ is 2-connected.

7. Integral theorems

7.1. Green's theorem

Proposition. If $P = P(x, y)$ and $Q = Q(x, y)$ are continuously differentiable on a planar domain $A \cup \partial A$ (A and its boundary), and ∂A is piecewise smooth, then

$$\oint_{\partial A} P\,\mathrm{d}x + Q\,\mathrm{d}y = \iint_A \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)\mathrm{d}x\,\mathrm{d}y$$

where the orientation of ∂A is such that A lies to the left while traversing ∂A.

Note that it is easy to arrive at this result for a rectangle. In this case,

$$\begin{aligned}\iint_A \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)\mathrm{d}x\,\mathrm{d}y &= \int_c^d \mathrm{d}y \int_a^b \mathrm{d}x\,\frac{\partial Q}{\partial x} - \int_a^b \mathrm{d}x \int_x^d \mathrm{d}y\,\frac{\partial P}{\partial y} \\ &= \int_c^d [Q(b, y) - Q(a, y)]\,\mathrm{d}y + \int_a^b [P(x, c) - P(x, d)]\,\mathrm{d}x \\ &= \oint_{\partial A} P\,\mathrm{d}x + Q\,\mathrm{d}y\end{aligned}$$

It then intuitively follows that we can approximate a surface with a set of small rectangles, and then the theorem should hold. As an example, let

$$P = -\frac{1}{2}y; \quad Q = \frac{1}{2}x$$

Then the area of some region is given by

$$\begin{aligned}\iint_A \mathrm{d}x\,\mathrm{d}y &= \iint_A \left(\frac{1}{2} + \frac{1}{2}\right)\mathrm{d}x\,\mathrm{d}y \\ &= \iint_A \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)\mathrm{d}x\,\mathrm{d}y \\ &= \frac{1}{2}\oint_{\partial A} x\,\mathrm{d}y - y\,\mathrm{d}x\end{aligned}$$

So letting A be the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 1$, we can parametrise ∂A by

$$[0, 2\pi] \ni t \mapsto \begin{pmatrix} a\cos t \\ b\sin t \end{pmatrix}$$

Hence the area is

$$\frac{1}{2}\int_0^{2\pi} \left(ab\cos^2 t + ab\sin^2 t\right)\mathrm{d}t = \pi ab$$

7.2. Stokes' theorem

Proposition. If $\mathbf{F}(\mathbf{x})$ is a continuously differentiable vector field, and S is an orientable, piecewise regular surface with a piecewise smooth boundary ∂S, then

$$\int_S (\nabla \times \mathbf{F}) \cdot \mathrm{d}\mathbf{S} = \oint_{\partial S} \mathbf{F} \cdot \mathrm{d}\mathbf{x}$$

This can be thought of as a generalisation to the fundamental theorem of calculus. From the fundamental theorem, we know that the integral of a differentiated function over an interval I is just the original function evaluated at the boundary ∂I. Likewise, Stokes' theorem states that the integral of the curl of a function (just another differential operator) over a surface S is just the original function evaluated at the boundary of the surface ∂S. In the one-dimensional fundamental theorem of calculus, we say that the function 'evaluated over the boundary' is simply the function applied to the final point, minus the function applied to the initial point; we are in some sense considering every point on the boundary ∂I. But in the case of ∂S being a curve, we must integrate around the curve boundary, since without an integral we can't consider infinitely many boundary points.

Note that for a surface to be 'orientable', it simply means that it has two sides, an inside and an outside. There must be a consistent choice of normal at each point. For example, a sphere is orientable, but a Möbius strip is not orientable.

Example. Let S be a cap of a sphere:

$$S = \left\{ \mathbf{x}(\theta, \phi) = \begin{pmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{pmatrix} \equiv \mathbf{e}_r; 0 \le \theta \le \alpha; 0 \le \phi < 2\pi \right\}$$

Now, let

$$\mathbf{F}(\mathbf{x}) = \begin{pmatrix} -x^2 y \\ 0 \\ 0 \end{pmatrix} \implies \nabla \times \mathbf{F} = \begin{pmatrix} 0 \\ 0 \\ x^2 \end{pmatrix}$$

On S,

$$\mathrm{d}\mathbf{S} = \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\theta} \times \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\phi} \,\mathrm{d}\theta\,\mathrm{d}\phi = \mathbf{e}_r \sin\theta \,\mathrm{d}\theta\,\mathrm{d}\phi$$

Note that since

$$x^2 \mathbf{e}_z \cdot \mathbf{e}_r = (\sin\theta\cos\phi)^2 \cos\theta$$

on S, we can compute

$$\int_S (\nabla \times \mathbf{F}) \cdot \mathrm{d}\mathbf{S} = \int_{\phi=0}^{2\pi} \left(\int_{\theta=0}^{\alpha} (\sin\theta\cos\phi)^2 \cos\theta \sin\theta \,\mathrm{d}\theta \right) \mathrm{d}\phi = \frac{\pi}{4} \sin^4 \alpha$$

We can instead compute the integral over the boundary. By Stokes' theorem, the results should match. ∂S is described by

$$[0, 2\pi] \ni t \mapsto \begin{pmatrix} \sin\alpha\cos t \\ \sin\alpha\sin t \\ \cos\alpha \end{pmatrix}$$

Then

$$\mathrm{d}\mathbf{x} = \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\,\mathrm{d}t = \sin\alpha \begin{pmatrix} -\sin t \\ \cos t \\ 0 \end{pmatrix} \mathrm{d}t$$

We can show that

$$\oint_{\partial S} \mathbf{F}\cdot\mathrm{d}\mathbf{S} = \sin^4\alpha \int_0^{2\pi} \left(-\cos^2 t \sin t\right)(-\sin t)\,\mathrm{d}t = \frac{\pi}{4}\sin^4\alpha$$

7.3. Stokes' theorem on closed surfaces

If S is an orientable, closed surface, and $\mathbf{F}$ is continuously differentiable, then

$$\int_S (\nabla\times\mathbf{F})\cdot\mathrm{d}\mathbf{S} = 0$$

This is clear since $\partial S = \varnothing$.

7.4. Zero circulation and irrotationality

Proposition. If $\mathbf{F}$ is continuously differentiable, and for every loop C we have that

$$\oint_C \mathbf{F}\cdot\mathrm{d}\mathbf{x} = 0$$

then $\nabla\times\mathbf{F} = \mathbf{0}$. In other words, $\mathbf{F}$ is irrotational if and only if $\mathbf{F}$ has zero circulation around all closed loops.

Note that the backward implication is trivial. If $\mathbf{F}$ has zero circulation around all loops, we can define that loop to be the boundary of some surface, and so the integral of the curl vanishes.

Proof. Suppose that the result is false; there exists a unit vector $\hat{\mathbf{k}}$ such that $\hat{\mathbf{k}}\cdot(\nabla\times\mathbf{F}(\mathbf{x}_0)) = \varepsilon > 0$ for some $\mathbf{x}_0$. By continuity, for a sufficiently small $\delta > 0$,

$$\hat{\mathbf{k}}\cdot(\nabla\times\mathbf{F}(\mathbf{x}_0)) > \frac{1}{2}\varepsilon; \quad \text{for } |\mathbf{x}-\mathbf{x}_0| < \delta$$

Now, we can take a loop in this ball $\{\mathbf{x} : |\mathbf{x}-\mathbf{x}_0| < \delta\}$ that lies entirely in a plane with normal $\hat{\mathbf{k}}$. Let this small loop's enclosed surface be S, with boundary ∂S. Then

$$0 = \oint_{\partial S} \mathbf{F}\cdot\mathrm{d}\mathbf{x} = \int_S \nabla\times\mathbf{F}\cdot\hat{\mathbf{k}}\,\mathrm{d}S > \frac{1}{2}\varepsilon\int \mathrm{d}S > 0$$

which is a contradiction. □

7.5. Intuition for curl as infinitesimal circulation

Let S_ε denote a region contained inside a disc of radius $\varepsilon > 0$, centred at $\mathbf{x}_0$ with normal $\hat{\mathbf{k}}$.

$$\begin{aligned}\int_{S_\varepsilon} \nabla \times \mathbf{F} \cdot \mathrm{d}\mathbf{S} &= \int_{S_\varepsilon} (\nabla \times \mathbf{F}(\mathbf{x}) - \nabla \times \mathbf{F}(\mathbf{x}_0)) \cdot \mathrm{d}\mathbf{S} + \int_{S_\varepsilon} \nabla \times \mathbf{F}(\mathbf{x}_0) \cdot \hat{\mathbf{k}} \, \mathrm{d}S \\ &= \underbrace{\int_{S_\varepsilon} (\nabla \times \mathbf{F}(\mathbf{x}) - \nabla \times \mathbf{F}(\mathbf{x}_0)) \cdot \mathrm{d}\mathbf{S}}_{o(\text{area}(S_\varepsilon))} + \nabla \times \mathbf{F}(\mathbf{x}_0) \cdot \hat{\mathbf{k}} \underbrace{\int_{S_\varepsilon} \mathrm{d}S}_{\text{area}(S_\varepsilon)}\end{aligned}$$

As ε shrinks, the first integral tends to zero faster than the second term. Hence,

$$\int_{S_\varepsilon} \nabla \times \mathbf{F} \cdot \mathrm{d}\mathbf{S} = \nabla \times \mathbf{F}(\mathbf{x}_0) \cdot \hat{\mathbf{k}} \cdot \text{area}(S_\varepsilon) + o(\text{area}(S_\varepsilon))$$

We can then see, by Stokes' theorem, that

$$\nabla \times \mathbf{F}(\mathbf{x}_0) \cdot \hat{\mathbf{k}} = \lim_{\varepsilon \to 0} \frac{1}{\text{area}(S_\varepsilon)} \oint_{\partial S_\varepsilon} \mathbf{F} \cdot \mathrm{d}\mathbf{x}$$

So the curl of $\mathbf{F}$ at $\mathbf{x}_0$ in the direction $\hat{\mathbf{k}}$ is the infinitesimal circulation around $\mathbf{x}_0$, per unit area.

7.6. Gauss' divergence theorem

Proposition. If $\mathbf{F}(\mathbf{x})$ is a continuously differentiable vector field, and V is a volume with a piecewise regular boundary ∂V, then

$$\int_V \nabla \cdot \mathbf{F} \, \mathrm{d}V = \int_{\partial V} \mathbf{F} \cdot \mathrm{d}\mathbf{S}$$

where the normal of ∂V points out of V.

There is also a two-dimensional version. If D is a planar region with a piecewise smooth boundary ∂D,

$$\int_D \nabla \cdot \mathbf{F} \, \mathrm{d}A = \oint_{\partial D} \mathbf{F} \cdot \mathbf{n} \, \mathrm{d}s$$

where the $\mathrm{d}s$ represents arc length, and where $\mathbf{n}$ points out of D.

Example. Let V be a cylinder, defined in cylindrical polar coordinates (ρ, ϕ, z) as

$$V = \{(\rho, \phi, z) : \ 0 \le \rho \le R, -h \le z \le h, 0 \le \phi < 2\pi\}$$

Let us label the boundary on the top S_+, the boundary on the bottom S_-, and the rest of the boundary S_R:

$$S_\pm = \{(\rho, \phi, z) : \ 0 \le \rho \le R, z = \pm h, 0 \le \phi < 2\pi\}$$
$$S_R = \{(\rho, \phi, z) : \ \rho = R, -h \le z \le h, 0 \le \phi < 2\pi\}$$

Consider $\mathbf{F}(\mathbf{x}) = \mathbf{x}$, hence $\nabla \cdot \mathbf{F} = 3$.

$$\int_V \nabla \cdot \mathbf{F} \, dV = 3 \int_V dV = 6\pi R^2 h$$

Using instead the divergence theorem,

$$\int_V \nabla \cdot \mathbf{F} \, dV = \int_{\partial V} \mathbf{F} \cdot d\mathbf{S}$$

On S_R, $d\mathbf{S} = \mathbf{e}_\rho R \, d\phi \, dz$, and $\mathbf{x} \cdot \mathbf{e}_\rho = R$. So we have the flux integral

$$\int_{S_R} \mathbf{F} \cdot d\mathbf{S} = \int_{z=-h}^{h} \int_{\phi=0}^{2\pi} R^2 \, d\phi \, dz = 4\pi R^2 h$$

On $S_\pm$, $d\mathbf{S} = \pm \mathbf{e}_z \rho \, d\rho \, d\phi$, and $\mathbf{x} \cdot \mathbf{e}_z = \pm h$. Hence

$$\int_{S_\pm} \mathbf{F} \cdot d\mathbf{S} = \int_{\phi=0}^{2\pi} \int_{\rho=0}^{R} h\rho \, d\rho \, d\phi = \pi R^2 h$$

The total is $6\pi R^2 h$ as expected.

Proposition. If $\mathbf{F}$ is continuously differentiable, and for every closed surface S we have

$$\int_S \mathbf{F} \cdot d\mathbf{S} = 0$$

then $\nabla \cdot \mathbf{F} = 0$.

Proof. Suppose that the result is false; $\nabla \cdot \mathbf{F}(\mathbf{x}_0) = \varepsilon > 0$. By continuity, for some sufficiently small $\delta > 0$ we have

$$\nabla \cdot \mathbf{F}(\mathbf{x}) > \frac{1}{2}\varepsilon \text{ for } |\mathbf{x}_0 - \mathbf{x}| < \delta$$

Now, we can choose a volume V inside the ball $|\mathbf{x}_0 - \mathbf{x}| < \delta$, and then by assumption, applying the divergence theorem,

$$0 = \int_{\partial V} \mathbf{F} \cdot d\mathbf{S} = \int_V \nabla \cdot \mathbf{F} \, dV > 0$$

which is a contradiction. We then conclude that if the vector field has zero net flux through any closed surface, it is solenoidal. □

7.7. Intuition for divergence as infinitesimal flux

Let V_ε be a volume in $\mathbb{R}^3$, contained inside a ball of radius $\varepsilon > 0$, centred at a point $\mathbf{x}_0$. Then

$$\int_{V_\varepsilon} \nabla \cdot \mathbf{F}\, \mathrm{d}V = \mathrm{vol}(V_\varepsilon) \nabla \cdot \mathbf{F}(\mathbf{x}_0) + \underbrace{\int_{V_\varepsilon} [\nabla \cdot \mathbf{F}(\mathbf{x}) - \nabla \cdot \mathbf{F}(\mathbf{x}_0)]\, \mathrm{d}V}_{o(\mathrm{vol}(V_\varepsilon))}$$

Dividing both sides by the volume of V_ε, and taking $\varepsilon \to 0$, we can apply the divergence theorem to get

$$\nabla \cdot \mathbf{F}(\mathbf{x}_0) = \lim_{\varepsilon \to 0} \frac{1}{\mathrm{vol}(V_\varepsilon)} \int_{\partial V_\varepsilon} \mathbf{F} \cdot \mathrm{d}\mathbf{S}$$

The divergence of $\mathbf{F}$ measures the infinitesimal flux per unit volume. If the flux is moving 'outward' at this point, $\nabla \cdot \mathbf{F} > 0$, and vice versa.

7.8. Conservation laws

Many equations in mathematical physics can be represented using density $\rho(\mathbf{x}, t)$ and a vector field $\mathbf{J}(\mathbf{x}, t)$, as follows.

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0 \qquad (\dagger)$$

This kind of equation is called a 'conservation law'. Suppose both ρ and $|\mathbf{J}|$ decrease rapidly as $|\mathbf{x}| \to \infty$. We will define the charge Q by

$$Q = \int_{\mathbb{R}^3} \rho(\mathbf{x}, t)\, \mathrm{d}V$$

We have conservation of charge;

$$\begin{aligned}
\frac{\mathrm{d}Q}{\mathrm{d}t} &= \int_{\mathbb{R}^3} \frac{\partial \rho}{\partial t}\, \mathrm{d}V \\
&= -\int_{\mathbb{R}^3} \nabla \cdot \mathbf{J}\, \mathrm{d}V \\
&= -\lim_{R \to \infty} \int_{|\mathbf{x}| \le R} \nabla \cdot \mathbf{J}\, \mathrm{d}V \\
&= -\lim_{R \to \infty} \int_{|\mathbf{x}| = R} \mathbf{J} \cdot \mathrm{d}\mathbf{S} \\
&= 0
\end{aligned}$$

as $\mathbf{J}$ decreases rapidly to zero as $|\mathbf{x}| \to \infty$. So $(\dagger)$ gives conservation of charge.

7.9. Proof of divergence theorem

Proof. Suppose first that

$$\mathbf{F} = F_z(x, y, z)\mathbf{e}_z$$

The divergence theorem states that

$$\int_V \underbrace{\frac{\partial F_z}{\partial z}}_{\nabla\cdot\mathbf{F}} \mathrm{d}V = \int_{\partial V} F_z \mathbf{e}_z \cdot \mathrm{d}\mathbf{S} \tag{†}$$

We would like to show that these two are really the same. First, let us simplify the problem to a convex volume V, such that we can split the boundary into two halves, one with normals in the positive z direction (S_+) and one with normals in the negative z direction (S_-). Then $\partial V = S_+ \cup S_-$. Project the volume into the x-y plane, and call this region A. This planar region is then the shape of the 'cut' between the S_+ and S_- halves. We can write

$$S_\pm = \left\{ \mathbf{x}(x, y) = \begin{pmatrix} x \\ y \\ g_\pm(x, y) \end{pmatrix} : (x, y) \in A \right\}$$

We can then say

$$\begin{aligned} \int_V \frac{\partial F_z}{\partial z} \mathrm{d}V &= \iint_A \left[\int_{z=g_-(x,y)}^{g_+(x,y)} \frac{\partial F_z}{\partial z} \mathrm{d}z \right] \mathrm{d}x\,\mathrm{d}y \\ &= \iint_A [F_z(x, y, g_+(x, y)) - F_z(x, y, g_-(x, y))]\,\mathrm{d}x\,\mathrm{d}y \end{aligned}$$

To calculate right hand side of (†), we need $\mathrm{d}\mathbf{S}$:

$$\begin{aligned} \mathrm{d}\mathbf{S} &= \frac{\partial \mathbf{x}}{\partial x} \times \frac{\partial \mathbf{x}}{\partial y} \mathrm{d}x\,\mathrm{d}y \\ &= \begin{pmatrix} -\partial g_\pm/\partial x \\ -\partial g_\pm/\partial y \\ 1 \end{pmatrix} \mathrm{d}x\,\mathrm{d}y \end{aligned}$$

Since we want the normal to point 'out' of V, on $S_\pm$ we have

$$\mathrm{d}\mathbf{S}\Big|_{S_\pm} = \pm \begin{pmatrix} -\partial g_\pm/\partial x \\ -\partial g_\pm/\partial y \\ 1 \end{pmatrix} \mathrm{d}x\,\mathrm{d}y$$

Therefore,

$$\begin{aligned} \int_{\partial V} \mathbf{F} \cdot \mathrm{d}\mathbf{S} &= \left[\int_{S_+} + \int_{S_-} \right] F_z \mathbf{e}_z \cdot \mathrm{d}\mathbf{S} \\ &= \iint_A F_z(x, y, g_+(x, y))\,\mathrm{d}x\,\mathrm{d}y - \iint_A F_z(x, y, g_-(x, y))\,\mathrm{d}x\,\mathrm{d}y \end{aligned}$$

which matches the expression we found for the left hand side of (†) above. In the same way, we can show that

$$\int_V \frac{\partial F_x}{\partial x}\,\mathrm{d}V = \int_{\partial V} F_x \mathbf{e}_x \cdot \mathrm{d}\mathbf{S}$$
$$\int_V \frac{\partial F_y}{\partial y}\,\mathrm{d}V = \int_{\partial V} F_y \mathbf{e}_y \cdot \mathrm{d}\mathbf{S}$$

and because the integrals are linear, we can compute their sum to find

$$\int_V \nabla \cdot \mathbf{F}\,\mathrm{d}V = \int_{\partial V} \mathbf{F} \cdot \mathrm{d}\mathbf{S}$$

which is exactly the divergence theorem. □

7.10. Proof of Green's theorem

We can use the two-dimensional divergence theorem to prove Green's theorem.

Proof. Let

$$\mathbf{F} = \begin{pmatrix} Q(x,y) \\ -P(x,y) \end{pmatrix}$$

Then

$$\iint_A \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)\mathrm{d}x\,\mathrm{d}y = \int_A \nabla \cdot \mathbf{F}\,\mathrm{d}A = \oint_{\partial A} \mathbf{F} \cdot \mathbf{n}\,\mathrm{d}s$$

If ∂A is parametrised with respect to arc length, this means that the unit tangent vector is

$$\mathbf{t} = \begin{pmatrix} x'(s) \\ y'(s) \end{pmatrix}$$

then the normal vector is

$$\mathbf{n} = \begin{pmatrix} y'(s) \\ -x'(s) \end{pmatrix}$$

Therefore,

$$\oint_{\partial A} \mathbf{F} \cdot \mathbf{n}\,\mathrm{d}s = \oint_{\partial A} \begin{pmatrix} Q \\ -P \end{pmatrix} \cdot \begin{pmatrix} y'(s) \\ -x'(s) \end{pmatrix} \mathrm{d}s = \oint_{\partial A} P\frac{\mathrm{d}x}{\mathrm{d}s}\,\mathrm{d}s + Q\frac{\mathrm{d}y}{\mathrm{d}s}\,\mathrm{d}s = \oint_{\partial A} P\,\mathrm{d}x + Q\,\mathrm{d}y$$

□

7.11. Proof of Stokes' theorem

We can now use Green's theorem to derive Stokes' theorem.

Proof. Consider a regular surface

$$S = \{\mathbf{x} = \mathbf{x}(u, v) : (u, v) \in A\}$$

Then the boundary is

$$\partial S = \{\mathbf{x} = \mathbf{x}(u, v) : (u, v) \in \partial A\}$$

Green's theorem gives

$$\oint_{\partial A} P\,\mathrm{d}u + Q\,\mathrm{d}v = \iint_A \left(\frac{\partial Q}{\partial u} - \frac{\partial P}{\partial v}\right)\mathrm{d}u\,\mathrm{d}v \tag{†}$$

We will now set

$$P(u, v) = \mathbf{F}(\mathbf{x}(u, v)) \cdot \frac{\partial \mathbf{x}}{\partial u}; \quad Q(u, v) = \mathbf{F}(\mathbf{x}(u, v)) \cdot \frac{\partial \mathbf{x}}{\partial v}$$

Then

$$P\,\mathrm{d}u + Q\,\mathrm{d}v = \mathbf{F}(\mathbf{x}(u, v)) \cdot \left(\frac{\partial \mathbf{x}}{\partial u}\,\mathrm{d}u + \frac{\partial}{\partial \mathbf{v}}\,\mathrm{d}v\right) = \mathbf{F}(\mathbf{x}(u, v)) \cdot \mathrm{d}\mathbf{x}(u, v)$$

And so we can compute the left hand side of (†):

$$\oint_{\partial A} P\,\mathrm{d}u + Q\,\mathrm{d}v = \oint_{\partial S} \mathbf{F} \cdot \mathrm{d}\mathbf{x}$$

For the right hand side, we must first compute some derivatives.

$$Q = F_i(\mathbf{x}(u, v))\frac{\partial x_i}{\partial v} \implies \frac{\partial Q}{\partial u} = \frac{\partial x_j}{\partial u}\frac{\partial F_i}{\partial x_j}\frac{\partial x_i}{\partial v} + F_i\frac{\partial^2 x_i}{\partial u \partial v}$$

$$P = F_i(\mathbf{x}(u, v))\frac{\partial x_i}{\partial u} \implies \frac{\partial Q}{\partial v} = \frac{\partial x_j}{\partial v}\frac{\partial F_i}{\partial x_j}\frac{\partial x_i}{\partial u} + F_i\frac{\partial^2 x_i}{\partial v \partial u}$$

Hence

$$\begin{aligned}
\frac{\partial Q}{\partial u} - \frac{\partial P}{\partial v} &= \left(\frac{\partial x_i}{\partial v}\frac{\partial x_j}{\partial u} - \frac{\partial x_i}{\partial u}\frac{\partial x_j}{\partial v}\right)\frac{\partial F_i}{\partial x_j} \\
&= \left(\delta_{ip}\delta_{jq} - \delta_{iq}\delta_{jp}\right)\frac{\partial F_i}{\partial x_j}\frac{\partial x_p}{\partial v}\frac{\partial x_q}{\partial u} \\
&= \varepsilon_{ijk}\varepsilon_{pqk}\frac{\partial F_i}{\partial x_j}\frac{\partial x_p}{\partial v}\frac{\partial x_q}{\partial u} \\
&= [-\nabla \times \mathbf{F}]_k \left(-\frac{\partial \mathbf{x}}{\partial u} \times \frac{\partial \mathbf{x}}{\partial v}\right)_k \\
&= (\nabla \times \mathbf{F}) \cdot \left(\frac{\partial \mathbf{x}}{\partial u} \times \frac{\partial \mathbf{x}}{\partial v}\right)
\end{aligned}$$

Therefore,

$$\iint_A \left(\frac{\partial Q}{\partial u} - \frac{\partial P}{\partial v}\right) \mathrm{d}u\, \mathrm{d}v = \iint_A (\nabla \times \mathbf{F}) \cdot \left(\frac{\partial \mathbf{x}}{\partial u} \times \frac{\partial \mathbf{x}}{\partial v}\right) = \iint_S (\nabla \times \mathbf{F}) \cdot \mathrm{d}\mathbf{S}$$

which gives Stokes' theorem as required. □

8. Maxwell's equations

8.1. Introduction and the equations

We will denote the magnetic field by $\mathbf{B}(\mathbf{x}, t)$, and the electric field by $\mathbf{E}(\mathbf{x}, t)$. These fields will depend on the current density $\mathbf{J}(\mathbf{x}, t)$, the electric current per unit area, and the charge density $\rho(\mathbf{x}, t)$, the electric charge per unit volume.

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0} \tag{1}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{2}$$

$$\nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0 \tag{3}$$

$$\nabla \times \mathbf{B} - \mu_0 \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} = \mu_0 \mathbf{J} \tag{4}$$

The constants ε_0 and μ_0 denote the permittivity and permeability of free space, which obey

$$\frac{1}{\mu_0 \varepsilon_0} = c^2$$

where c is the speed of light, $299\,792\,458\,\mathrm{m\,s^{-1}}$. Note that if we take the divergence of equation (4), we find

$$\begin{aligned} \mu_0 \varepsilon_0 \frac{\partial}{\partial t} (\nabla \cdot \mathbf{E}) + \mu_0 \nabla \cdot \mathbf{J} &= 0 \\ (1) \implies \mu_0 \varepsilon_0 \frac{\partial}{\partial t} \frac{\rho}{\varepsilon_0} + \mu_0 \nabla \cdot \mathbf{J} &= 0 \\ \frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} &= 0 \end{aligned}$$

which is a conservation law for charge.

8.2. Integral formulations of Maxwell's equations

Integrating (1) over some volume V, and applying the divergence theorem, gives

$$\begin{aligned} \nabla \cdot \mathbf{E} &= \frac{\rho}{\varepsilon_0} \\ \int_V \nabla \cdot \mathbf{E} \, \mathrm{d}V &= \frac{1}{\varepsilon_0} \int_V \rho \, \mathrm{d}V \\ \int_{\partial V} \mathbf{E} \cdot \mathrm{d}\mathbf{S} &= \frac{Q}{\varepsilon_0} \end{aligned}$$

where Q is the total charge in V. This is known as Gauss' law. For magnetic fields, we can integrate (2):

$$\int_V \nabla \cdot \mathbf{B}\,\mathrm{d}V = \int_{\partial V} \mathbf{B} \cdot \mathrm{d}\mathbf{S} = 0$$

Hence there is no net magnetic flux over any closed surface ∂V. This implies that we cannot have a magnetic field with only a north pole or only a south pole. Integrating (3) over a surface, and applying Stokes' theorem, gives

$$\begin{aligned}
\nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} &= 0 \\
\int_S \left(\nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t}\right) \cdot \mathrm{d}\mathbf{S} &= 0 \\
\oint_{\partial S} \mathbf{E} \cdot \mathrm{d}\mathbf{x} + \int_S \frac{\partial \mathbf{B}}{\partial t} \cdot \mathrm{d}\mathbf{S} &= 0 \\
\oint_{\partial S} \mathbf{E} \cdot \mathrm{d}\mathbf{x} &= -\frac{\mathrm{d}}{\mathrm{d}t} \int_S \mathbf{B} \cdot \mathrm{d}\mathbf{S}
\end{aligned}$$

So a change in the magnetic flux through a surface S induces a circulation in $\mathbf{E}$ about the boundary. Integrating (4) over a surface, again using Stokes' theorem, we have

$$\begin{aligned}
\int_S \left(\nabla \times \mathbf{B} - \mu_0 \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t}\right) \cdot \mathrm{d}\mathbf{S} &= \int_S \mu_0 \mathbf{J} \cdot \mathrm{d}\mathbf{S} \\
\oint_{\partial S} \mathbf{B} \cdot \mathrm{d}\mathbf{x} &= \int_S \mu_0 \mathbf{J} \cdot \mathrm{d}\mathbf{S} + \int_S \mu_0 \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} \cdot \mathrm{d}\mathbf{S} \\
\oint_{\partial S} \mathbf{B} \cdot \mathrm{d}\mathbf{x} &= \mu_0 \int_S \mathbf{J} \cdot \mathrm{d}\mathbf{S} + \mu_0 \varepsilon_0 \frac{\mathrm{d}}{\mathrm{d}t} \int_S \mathbf{E} \cdot \mathrm{d}\mathbf{S}
\end{aligned}$$

So if an electric current flows through a wire, this generates a circulation of the magnetic field around the wire.

8.3. Electromagnetic waves

In empty space, $\rho = 0$ and $\mathbf{J} = \mathbf{0}$. Maxwell's equations show that

$$\nabla \cdot \mathbf{E} = 0 \tag{1}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{2}$$

$$\nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0 \tag{3}$$

$$\nabla \times \mathbf{B} - \mu_0 \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} = 0 \tag{4}$$

Recall that the Laplacian of a vector field $\mathbf{F}$ is

$$\nabla^2 \mathbf{F} = \nabla(\nabla \cdot \mathbf{F}) - \nabla \times (\nabla \times \mathbf{F})$$

We can deduce that

$$\begin{aligned}
\nabla^2\mathbf{E} &= \nabla(\nabla\cdot\mathbf{E}) - \nabla\times(\nabla\times\mathbf{E}) \\
&= \nabla(0) - \nabla\times\left(-\frac{\partial\mathbf{B}}{\partial t}\right) \\
&= \nabla\times\left(\frac{\partial\mathbf{B}}{\partial t}\right) \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\nabla\times\mathbf{B} \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\mu_0\varepsilon_0\frac{\partial\mathbf{E}}{\partial t} \\
&= \frac{1}{c^2}\frac{\partial^2\mathbf{E}}{\partial t^2} \\
\therefore \nabla^2\mathbf{E} - \frac{1}{c^2}\frac{\partial^2\mathbf{E}}{\partial t^2} &= \mathbf{0}
\end{aligned}$$

which is the wave equation for waves travelling at speed c. Hence, in a vacuum, the electric field propagates at speed c. Similarly, for the magnetic field,

$$\begin{aligned}
\nabla^2\mathbf{B} &= \nabla(\nabla\cdot\mathbf{B}) - \nabla\times(\nabla\times\mathbf{B}) \\
&= \nabla(0) - \nabla\times\left(\mu_0\varepsilon_0\frac{\partial\mathbf{E}}{\partial t}\right) \\
&= -\mu_0\varepsilon_0\frac{\mathrm{d}}{\mathrm{d}t}\nabla\times\mathbf{E} \\
&= \mu_0\varepsilon_0\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial\mathbf{B}}{\partial t} \\
&= \frac{1}{c^2}\frac{\partial^2\mathbf{B}}{\partial t^2} \\
\therefore \nabla^2\mathbf{B} - \frac{1}{c^2}\frac{\partial^2\mathbf{B}}{\partial t^2} &= \mathbf{0}
\end{aligned}$$

Hence the magnetic field also propagates at speed c. So in general, we can say that electromagnetic waves always travel at speed c in a vacuum.

8.4. Electrostatics and magnetostatics

Suppose that all fields and source terms are independent of t. Then Maxwell's equations decouple into

$$\nabla\cdot\mathbf{E} = \frac{\rho}{\varepsilon_0} \tag{1}$$

$$\nabla\cdot\mathbf{B} = 0 \tag{2}$$

$$\nabla\times\mathbf{E} = 0 \tag{3}$$

$$\nabla\times\mathbf{B} = \mu_0\mathbf{J} \tag{4}$$

which gives one system of equations for **E**, and one for **B**. When considering the whole of $\mathbb{R}^3$, which is 2-connected, then equations (2) and (3) imply

$$\mathbf{E} = -\nabla\phi; \quad \mathbf{B} = \nabla \times \mathbf{A}$$

where ϕ is the electric potential, and **A** is the magnetic potential. Substituting into the other two equations, we have

$$\begin{aligned} (1) \implies -\nabla \cdot \nabla\phi &= \frac{\rho}{\varepsilon_0} \\ -\nabla^2\phi &= \frac{\rho}{\varepsilon_0} \end{aligned}$$

and

$$(4) \implies \nabla \times (\nabla \times \mathbf{A}) = \mu_0 \mathbf{J}$$

9. Poisson's and Laplace's equations

9.1. The boundary value problem

Many problems in mathematical physics can be reduced to the form

$$\nabla^2 \phi = F$$

This is called Poisson's equation. In the case that $F \equiv 0$, this is called Laplace's equation. We are interested in solving this equation on $\Omega \subseteq \mathbb{R}^n$ for $n = 2, 3$. This is too general to solve at the moment, so we will need to supply boundary conditions, which are very common in physical problems. In other words, ϕ will be known on $\partial\Omega$, or as $|\mathbf{x}| \to \infty$ if $\Omega = \mathbb{R}^n$. For instance, the Dirichlet problem is

$$\nabla^2 \phi = F \text{ inside } \Omega; \quad \phi = f \text{ on } \partial\Omega$$

The Neumann problem is

$$\nabla^2 \phi = F \text{ inside } \Omega; \quad \frac{\partial \phi}{\partial \mathbf{n}} = g \text{ on } \partial\Omega$$

where $\mathbf{n}$ is the normal to the surface, and $\frac{\partial \phi}{\partial \mathbf{n}} \coloneqq \mathbf{n} \cdot \nabla \phi$. As a further restriction, we must interpret the boundary conditions in an 'appropriate' manner; we assume that ϕ (or $\frac{\partial \phi}{\partial \mathbf{n}}$) approaches the behaviour at the boundary continuously as $\mathbf{x} \to \partial\Omega$. More precisely, ϕ and $\nabla\phi$ are continuous on $\Omega \cup \partial\Omega$. Note that if we are solving some equation $\nabla^2 \phi = 0$ in Ω, we must be certain that ϕ is actually well-defined on the entire set. As a worked example, consider

$$\nabla^2 \phi = r \text{ inside } \{r < a\}; \quad \phi = 1 \text{ on } \{r = a\}$$

We might guess that the solution is of the form $\phi(r)$. We can use the formula

$$\nabla^2 \phi = \frac{1}{r^2} \frac{\mathrm{d}}{\mathrm{d}r} \left(r^2 \frac{\mathrm{d}\phi}{\mathrm{d}r} \right)$$

to get

$$r^3 = \frac{\mathrm{d}}{\mathrm{d}r} \left(r^2 \frac{\mathrm{d}\phi}{\mathrm{d}r} \right) \text{ inside } \{r < a\}; \quad \phi(a) = 1$$

The general solution to the first part is

$$\phi(r) = A + \frac{B}{r} + \frac{1}{12} r^3$$

The $\frac{B}{r}$ term is *not* well-defined inside $\{r < a\}$, therefore $B = 0$ to eliminate the problematic term. By the second part, we can solve for A:

$$1 = \phi(a) = A + \frac{1}{12} a^3 \implies A = 1 - \frac{1}{12} a^3$$

Hence the solution is

$$\phi(r) = 1 + \frac{1}{12} \left(r^3 - a^3 \right)$$

9.2. Uniqueness of solutions

When solving Poisson's or Laplace's equation, we want to ensure that the solution we find is unique. If it is unique, then we can apply similar logic to solving differential equations, where we can guess the form of an equation and then derive the solution from that, and we don't need to worry about solutions that do not have this form. Consider a generic linear problem

$$L\phi = F \text{ in } \Omega; \quad B\phi = f \text{ on } \partial\Omega \tag{†}$$

where L and B are linear differential operators. If ϕ_1 and ϕ_2 are both solutions to (†), then consider $\psi = \phi_1 - \phi_2$. By linearity,

$$L\psi = L\phi_1 - L\phi_2 = F - F = 0 \text{ in } \Omega$$

and

$$B\psi = B\phi_1 - B\phi_2 = f - f = 0 \text{ on } \partial\Omega$$

If we can show that the only solution to these new equations is $\psi = 0$, we must conclude that $\phi_1 = \phi_2$, which means that there is only one solution to (†). Hence the solution to a linear problem is unique if and only if the only solution to the homogeneous problem is zero.

Proposition. The solution to the Dirichlet problem is unique. The solution to the Neumann problem is unique up to the addition of an arbitrary constant.

Proof. Let $\psi = \phi_1 - \phi_2$ be the difference between two solutions. In the Dirichlet case, we want to show that $\psi = 0$, and in the Neumann case, we want to show that ψ is an arbitrary constant. We know that

$$\nabla^2\psi = 0 \text{ in } \Omega; \quad B\psi = 0 \text{ on } \partial\Omega$$

where $B\psi = \psi$ in the Dirichlet problem, or $B\psi = \frac{\partial\psi}{\partial\mathbf{n}}$ in the Neumann problem. Consider the non-negative functional

$$I[\psi] = \int_\Omega |\nabla\psi|^2 \, \mathrm{d}V \geq 0$$

Clearly,

$$I[\psi] = 0 \iff \nabla\psi = 0 \text{ everywhere in } \Omega$$

Now, note that we can apply the divergence theorem to get

$$\begin{aligned} I[\psi] &= \int_\Omega |\nabla\psi|^2 \,\mathrm{d}V \\ &= \int_\Omega \nabla\psi \cdot \nabla\psi \,\mathrm{d}V \\ &= \int_\Omega \big(\nabla\cdot(\psi\nabla\psi) - \psi\nabla^2\psi\big)\,\mathrm{d}V \\ &= \int_\Omega \nabla\cdot(\psi\nabla\psi)\,\mathrm{d}V \\ &= \int_{\partial\Omega} \psi\nabla\psi\cdot\mathrm{d}\mathbf{S} \\ &= \int_{\partial\Omega} \psi\nabla\psi\cdot\mathbf{n}\,\mathrm{d}S \\ &= \int_{\partial\Omega} \psi\frac{\mathrm{d}\psi}{\mathrm{d}\mathbf{n}}\,\mathrm{d}S \end{aligned}$$

In the Dirichlet case, $I[\psi] = 0$ since $\psi = 0$ on the boundary. In the Neumann case, $I[\psi] = 0$ as well, since $\frac{\mathrm{d}\psi}{\mathrm{d}\mathbf{n}} = 0$. Hence, in either case, $\nabla\psi = 0$ everywhere in Ω. Therefore, ψ is a constant throughout Ω. In the Dirichlet case, we know that $\psi = 0$ on the boundary, hence $\psi = 0$ everywhere as it is continuous. However, in the Neumann problem, no such deduction can be made. □

Example. Here is an example from electrostatics. Consider the charge density ρ defined by

$$\rho(\mathbf{x}) = \begin{cases} 0 & r < a \\ F(r) & r \geq a \end{cases}$$

We can show that there is no electric field in the region $r < a$. We know that the electric potential ϕ will satisfy

$$\nabla^2\phi = \frac{-\rho(\mathbf{x})}{\varepsilon_0} = 0 \text{ if } r < a$$

By symmetry, we will try a ϕ of the form $\phi(r)$. Hence, $\phi(a)$ is constant on the boundary $r = a$. Note that the unique solution to

$$\nabla^2\phi = 0 \text{ for } r < a; \quad \phi = \text{constant on } r = a$$

is exactly that ϕ is constant everywhere. Hence

$$\mathbf{E} = -\nabla\psi = 0 \text{ throughout } r < a$$

This can be viewed as a version of Newton's shell theorem.

9.3. Gauss' flux method for spherically symmetric sources

Suppose the source term (the F on the right hand side of Poisson's equation) is spherically symmetric, so F is a function of $r = |\mathbf{x}|$. Assuming we are trying to solve the equation for $\Omega = \mathbb{R}^3$, we can rewrite the problem as

$$\nabla \cdot \nabla \phi = F \tag{$*$}$$

Since the right hand side only depends on r, the same is true of the left hand side. So we might guess a ϕ of the form $\phi(r)$. In which case, we can compute

$$\nabla \phi = \phi'(r)\mathbf{e}_r$$

Using Gauss' flux method, we will integrate $(*)$ over some spherical region $|\mathbf{x}| < R$, and use the divergence theorem.

$$\int_{|\mathbf{x}|<R} \nabla \cdot \nabla \phi \, \mathrm{d}V = \int_{|\mathbf{x}|=R} \nabla \phi \cdot \mathrm{d}\mathbf{S} = \int_{|\mathbf{x}|<R} F(r) \, \mathrm{d}V$$

Thinking of the source term F as some kind of density, for instance charge density or mass density, the right hand side can be thought of as the total amount of charge or mass inside the ball. We will call this term $Q(R)$.

$$\int_{|\mathbf{x}|=R} \nabla \phi \cdot \mathrm{d}\mathbf{S} = Q(R)$$

Recall that on a sphere of radius R, $\mathrm{d}\mathbf{S} = \mathbf{e}_r R^2 \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi$. Therefore, on the boundary $|\mathbf{x}| = R$,

$$\nabla \phi \cdot \mathrm{d}\mathbf{S} = \phi'(r)\mathbf{e}_r \cdot \mathbf{e}_r R^2 \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi = \phi'(r) R^2 \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi = \phi'(r) \, \mathrm{d}S$$

Hence,

$$Q(R) = \int_{|\mathbf{x}|=R} \phi'(r) \, \mathrm{d}S$$

But $\phi'(r)$ is a constant on the surface we are integrating over. Therefore,

$$Q(R) = \phi'(R) \int_{|\mathbf{x}|=R} \mathrm{d}S = 4\pi R^2 \phi'(R)$$

In summary,

$$\phi'(R) = \frac{Q(R)}{4\pi R^2} \implies \nabla \phi = \frac{Q(R)}{4\pi R^2}\mathbf{e}_r$$

Example. Recall the first of Maxwell's equations:

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}$$

If we are dealing with electrostatics, the curl of $\mathbf{E}$ is zero. Hence $\mathbf{E} = -\nabla\phi$, so

$$\nabla^2\phi = -\frac{\rho}{\varepsilon_0}$$

Consider a charge density ρ of the form

$$\rho(r) = \begin{cases} \rho_0, & 0 \le r \le a \\ 0, & r > a \end{cases}$$

By the previous result,

$$\phi'(r) = \frac{1}{4\pi\varepsilon_0}\frac{Q(r)}{r^2}$$

where

$$Q(r) = \int_{|\mathbf{x}|\le R} \rho(r)\,\mathrm{d}V$$

Note, if $R > a$ then $Q(R) = Q(a)$, which we will denote Q for the total charge. Hence, we have the following solution:

$$\mathbf{E}(\mathbf{x}) = \begin{cases} \frac{1}{4\pi\varepsilon_0}\frac{Q(r)}{r^2}\mathbf{e}_r, & r \le a \\ \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}\mathbf{e}_r, & r > a \end{cases}$$

If we take $a \to 0$, but keeping Q fixed, this represents a point charge. Then

$$\mathbf{E}(\mathbf{x}) = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}\mathbf{e}_r$$

In this case, the charge density ρ is

$$\rho(\mathbf{x}) = Q\delta(\mathbf{x})$$

where δ is the Dirac delta function.

9.4. Cylindrical symmetry

Suppose instead that the source term F is cylindrically symmetric, so F is a function of ρ, the distance from the z axis. Similarly as before, we can guess that ϕ is a function only of ρ. We can integrate $\nabla \cdot \nabla\phi = F(\rho)$ over a cylinder V of radius R and height a.

$$\nabla\phi = \phi'(\rho)\mathbf{e}_\rho$$

Hence,

$$\int_V \nabla \cdot \nabla\phi\,\mathrm{d}V = \int_V F(\rho)\,\mathrm{d}V$$

The left hand side becomes

$$\int_{\partial V} \nabla\phi \cdot \mathrm{d}\mathbf{S}$$

On the top circle, the normal $\mathbf{n}$ would be in the $\mathbf{e}_z$ direction, and on the bottom circle, $\mathbf{n}$ would be in the $-\mathbf{e}_z$ direction. On the curved surface, $\mathbf{n}$ would be in the $\mathbf{e}_\rho$ direction. Note that since $\nabla\phi$ only has a component in the $\mathbf{e}_\rho$ direction, on both the top and bottom circles will provide no contribution to the final result for this boundary integral. $\mathrm{d}\mathbf{S} = R\,\mathrm{d}\phi\,\mathrm{d}z\,\mathbf{e}_\rho$, hence

$$\int_{\partial V} \nabla\phi \cdot \mathrm{d}\mathbf{S} = \int_{\phi=0}^{2\pi}\int_{z=z_0}^{z_0+a} \phi'(R)R\,\mathrm{d}\phi\,\mathrm{d}z = 2\pi\int_{z=z_0}^{z_0+a} \phi'(R)R\,\mathrm{d}z = 2\pi aR\phi'(R)$$

Substituting into the above equation gives

$$\phi'(R) = \frac{1}{2\pi aR}\int_V F(\rho)\,\mathrm{d}V$$

Note that the integral $\int_V F(\rho)\,\mathrm{d}V$ is given by

$$\int_V F(\rho)\,\mathrm{d}V = \int_{\phi=0}^{2\pi}\mathrm{d}\phi\int_{z=z_0}^{z_0+a}\mathrm{d}z\int_{\rho=0}^{R}\mathrm{d}\rho\,F(\rho)\rho = 2\pi a\int_0^R F(\rho)\rho\,\mathrm{d}\rho$$

In conclusion,

$$\phi'(\rho) = \frac{1}{\rho}\int_0^\rho sF(s)\,\mathrm{d}s$$

Example. Consider a line of charge density λ per unit length along an infinitesimally thick wire. We could proceed analogously to the last example before, by considering a cylinder with positive radius a, using Gauss' flux method, and then letting $a \to 0$. However, we will use a different method. Let $F(\rho)$ be the desired charge density. So if we integrate $F(\rho)$ over any cylinder C of length 1, we should retrieve the value λ.

$$\begin{aligned}\lambda = \int_C F(\rho)\,\mathrm{d}V &= \int_{z=z_0}^{z_0+1}\mathrm{d}z\int_{\phi=0}^{2\pi}\mathrm{d}\phi\int_{\rho=0}^{R}\mathrm{d}\rho\,\rho F(\rho)\\ &= 2\pi\int_0^R \mathrm{d}\rho\,\rho F(\rho)\end{aligned}$$

By inspection, F must have the form of a delta function, so $F(\rho) = \lambda\delta(\rho)\frac{1}{2\pi\rho}$. Hence the corresponding electric potential ϕ is given by

$$\phi'(\rho) = -\frac{1}{\varepsilon_0\rho}\int_0^\rho \lambda\delta(s)\frac{1}{2\pi}\,\mathrm{d}s = \frac{-\lambda}{2\pi\varepsilon_0\rho}$$

Hence,

$$E(\mathbf{x}) = \frac{1}{2\pi\varepsilon_0}\frac{\mathbf{e}_\rho}{\rho}$$

9.5. Superposition principle

Consider a linear operator L. If we have solutions $L\psi_n = F_n$ for $n = 1, 2, \ldots$, then we have $L\left(\sum_n \psi_n\right) = \sum_n F_n$ by linearity. In other words, we can superimpose solutions. We can often break up a forcing term into several smaller, simpler components, and if L is a linear differential operator we can solve for these components separately. For example, we can consider the electric potential due to a pair of point charges Q_a at $\mathbf{x} = \mathbf{a}$, and Q_b at $\mathbf{x} = \mathbf{b}$. The charge density would be

$$\rho(\mathbf{x}) = Q_a \delta(\mathbf{x} - \mathbf{a}) + Q_b \delta(\mathbf{x} - \mathbf{b})$$

For one point charge, we know that the electric potential obeys

$$-\nabla^2 \phi = \frac{Q_a}{\varepsilon_0} \delta(\mathbf{x} - \mathbf{a})$$

Hence,

$$\phi(\mathbf{x}) = \frac{Q_a}{4\pi\varepsilon_0} \frac{1}{|\mathbf{x} - \mathbf{a}|}$$

Then by the superposition principle, for two particles,

$$\phi(\mathbf{x}) = \frac{Q_a}{4\pi\varepsilon_0} \frac{1}{|\mathbf{x} - \mathbf{a}|} + \frac{Q_b}{4\pi\varepsilon_0} \frac{1}{|\mathbf{x} - \mathbf{b}|}$$

Now, consider the electric potential outside a ball of radius $|\mathbf{x}| < R$ of uniform charge density ρ_0. Suppose that the ball has several balls removed from its interior. These 'subtracted' balls have the form

$$|\mathbf{x} - \mathbf{a}_i| < R_i; \quad i = 1, \ldots, N$$

We further require that the balls lay inside the main ball, and do not intersect:

$$|\mathbf{a}_i| + R_i < R; \quad |\mathbf{a}_i - \mathbf{a}_j| > R_i + R_j$$

We can use the superposition principle to represent each hole as a ball of uniform charge density $-\rho_0$. So the effective potential in $|\mathbf{x}| > R$ (outside the ball) from each hole is

$$\phi(x) = -\frac{Q_i}{4\pi\varepsilon_0} \frac{1}{|\mathbf{x} - \mathbf{a}_i|}; \quad Q_i = \frac{4}{3}\pi R_i^3 \rho_0$$

Hence, the total potential from the ball and its holes is

$$\phi(x) = \frac{Q}{4\pi\varepsilon_0} \frac{1}{|\mathbf{x}|} - \sum_i \frac{Q_i}{4\pi\varepsilon_0} \frac{1}{|\mathbf{x} - \mathbf{a}_i|}$$

9.6. Integral solutions

We know that the electric potential due to a point charge at **a** is proportional to the inverse of the distance to the particle. We can think of a generic distribution of charge density as an infinite collection of superimposed particles, which leads us to consider an integral form for a superposition.

$$\int_{\mathbb{R}^3} \frac{F(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}\,\mathrm{d}V(\mathbf{y})$$

where F is the forcing term.

Proposition. Suppose $F \to 0$ 'rapidly' as $|\mathbf{x}| \to \infty$. The unique solution to the Dirichlet problem

$$\begin{cases} \nabla^2\phi = F & \mathbf{x} \in \mathbb{R}^3 \\ |\phi| \to 0 & |\mathbf{x}| \to \infty \end{cases}$$

is given by

$$\phi(\mathbf{x}) = -\frac{1}{4\pi}\int_{\mathbb{R}^3} \frac{F(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}\,\mathrm{d}V(\mathbf{y})$$

This result is another way of saying that

$$\nabla^2\left(\frac{-1}{4\pi|\mathbf{x}|}\right) = \delta(\mathbf{x})$$

since by differentiating with respect to x under the integral sign,

$$\begin{aligned} \nabla^2\left(\frac{-1}{4\pi}\int_{\mathbb{R}^3} \frac{F(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}\,\mathrm{d}V(\mathbf{y})\right) &= \frac{-1}{4\pi}\int_{\mathbb{R}^3} F(\mathbf{y})\nabla^2\left(\frac{1}{|\mathbf{x}-\mathbf{y}|}\right)\mathrm{d}V(\mathbf{y}) \\ &= \int_{\mathbb{R}^3} F(\mathbf{y})\delta(\mathbf{x}-\mathbf{y})\,\mathrm{d}V(\mathbf{y}) \\ &= F(\mathbf{x}) \end{aligned}$$

so it is sufficient to prove that this Laplacian identity holds. A full proof will not be given here, but here is some intuition to guide the idea. Note that for $r \neq 0$,

$$\begin{aligned} \nabla^2\left(\frac{1}{r}\right) &= \frac{\partial^2}{\partial x_i \partial x_i}\left(\frac{1}{r}\right) \\ &= \frac{\partial}{\partial x_i}\left(\frac{-x_i}{r^3}\right) \\ &= \frac{-\delta_{ii}}{r^3} + \frac{3x_i x_i}{r^5} \\ &= \frac{-3}{r^3} + \frac{3r^2}{r^5} \\ &= 0 \end{aligned}$$

So certainly $\nabla^2\left(-\frac{1}{4\pi|\mathbf{x}|}\right) = \delta(\mathbf{x})$ for $\mathbf{x} \neq 0$. Assuming that the divergence theorem holds for delta functions, for any ball $|\mathbf{x}| < R$ we would also have

$$\begin{aligned}\int_{|\mathbf{x}|<R} \nabla^2\left(\frac{1}{|\mathbf{x}|}\right) \mathrm{d}V &= \int_{|\mathbf{x}|=R} \nabla\left(\frac{1}{|\mathbf{x}|}\right) \cdot \mathrm{d}\mathbf{S} \\ &= \int_{\theta=0}^{\pi} \mathrm{d}\theta \int_{\phi=0}^{2\pi} \mathrm{d}\phi \left(\frac{-\mathbf{e}_r}{R^2}\right) \cdot \mathbf{e}_r R^2 \sin\theta \\ &= \int_{\theta=0}^{\pi} \mathrm{d}\theta \int_{\phi=0}^{2\pi} \mathrm{d}\phi \left(\frac{-1}{R^2}\right) R^2 \sin\theta \\ &= -4\pi\end{aligned}$$

So for any $R > 0$,

$$\int_{|\mathbf{x}|<R} \nabla^2\left(\frac{-1}{4\pi|\mathbf{x}|}\right) \mathrm{d}V = 1 = \int_{|\mathbf{x}|<R} \delta(\mathbf{x})\, \mathrm{d}V$$

So we might conclude that this Laplacian operator really does give the Dirac delta function.

9.7. Harmonic functions

Harmonic functions are solutions to Laplace's equation,

$$\nabla^2 \phi = 0$$

Proposition. If ϕ is harmonic on $\Omega \subset \mathbb{R}^3$, then

$$\phi(\mathbf{a}) = \frac{1}{4\pi r^2} \int_{|\mathbf{x}-\mathbf{a}|=r} \phi(\mathbf{x})\, \mathrm{d}S \qquad (*)$$

for $\mathbf{a} \in \Omega$, and r sufficiently small such that all $\mathbf{x}$ are in Ω.

This is known as the 'mean value' property; it essentially shows that the value of ϕ at any given point $\mathbf{a}$ is the average of ϕ on the surface of any ball around $\mathbf{a}$.

Proof. Let $F(r)$ denote the right hand side of $(*)$, $\frac{1}{4\pi r^2} \int_{|\mathbf{x}-\mathbf{a}|=r} \phi(\mathbf{x})\, \mathrm{d}S$. Then,

$$F(r) = \frac{1}{4\pi r^2} \int_{|\mathbf{x}|=r} \phi(\mathbf{a}+\mathbf{x})\, \mathrm{d}S$$

We can parametrise this sphere using spherical polar coordinates, giving

$$\begin{aligned}F(r) &= \frac{1}{4\pi r^2} \int_{\phi=0}^{2\pi} \left[\int_{\theta=0}^{\pi} \phi(\mathbf{a}+r\mathbf{e}_r) r^2 \sin\theta\, \mathrm{d}\theta\right] \mathrm{d}\phi \\ &= \frac{1}{4\pi} \int_{\phi=0}^{2\pi} \left[\int_{\theta=0}^{\pi} \phi(\mathbf{a}+r\mathbf{e}_r) \sin\theta\, \mathrm{d}\theta\right] \mathrm{d}\phi \qquad (\dagger)\end{aligned}$$

Differentiating with respect to r, using $\frac{\mathrm{d}}{\mathrm{d}r}\phi(\mathbf{a}+r\mathbf{e}_r) = \mathbf{e}_r \cdot \nabla\phi(\mathbf{a}+r\mathbf{e}_r)$,

$$\begin{aligned}
F'(r) &= \frac{1}{4\pi}\int_{\phi=0}^{2\pi}\int_{\theta=0}^{\pi} \mathbf{e}_r \cdot \nabla\phi(\mathbf{a}+r\mathbf{e}_r)\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \\
&= \frac{1}{4\pi r^2}\int_{\phi=0}^{2\pi}\int_{\theta=0}^{\pi} \mathbf{e}_r \cdot \nabla\phi(\mathbf{a}+r\mathbf{e}_r) r^2\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \\
&= \frac{1}{4\pi r^2}\int_{|\mathbf{x}|=r} \mathbf{e}_r \cdot \nabla\phi(\mathbf{a}+r\mathbf{e}_r)\,\mathrm{d}S \\
&= \frac{1}{4\pi r^2}\int_{|\mathbf{x}|=r} \nabla\phi(\mathbf{a}+\mathbf{x})\cdot\mathrm{d}\mathbf{S} \\
&= \frac{1}{4\pi r^2}\int_{|\mathbf{x}-\mathbf{a}|=r} \nabla\phi\cdot\mathrm{d}\mathbf{S} \\
&= \frac{1}{4\pi r^2}\int_{|\mathbf{x}-\mathbf{a}|<r} \nabla\cdot\nabla\phi\,\mathrm{d}V \\
&= \frac{1}{4\pi r^2}\int_{|\mathbf{x}-\mathbf{a}|<r} \nabla^2\phi\,\mathrm{d}V \\
&= \frac{1}{4\pi r^2}\int_{|\mathbf{x}-\mathbf{a}|<r} 0\,\mathrm{d}V \\
&= 0
\end{aligned}$$

Now, note from (†) that if $r \to 0$, then $F(r) \to \phi(\mathbf{a})$, and the result follows. □

9.8. Intuitive explanation of Laplacian

We can use the central idea of the above proof to examine what the Laplacian operator is really doing.

Proposition. For any smooth function $\phi : \mathbb{R}^3 \to \mathbb{R}$,

$$\nabla^2\phi(\mathbf{a}) = \lim_{r\to 0}\frac{6}{r^2}\left[\frac{1}{4\pi r^2}\int_{|\mathbf{x}-\mathbf{a}|=r}\phi(\mathbf{x})\,\mathrm{d}S\right] - \phi(\mathbf{a})$$

In particular, if ϕ satisfies the mean value property, then it is harmonic.

In some sense, the Laplacian is measuring how the value of ϕ at a point differs from its average over a small sphere centred at this point.

Proof. Consider a function $G(r)$ defined by

$$G(r) = \frac{1}{4\pi r^2}\int_{|\mathbf{x}-\mathbf{a}|=r}\phi(\mathbf{x})\,\mathrm{d}S - \phi(\mathbf{a})$$

G measures the extent to which ϕ differs from its average. From the previous proof,

$$G(r) = F(a) - \phi(\mathbf{a}) \implies G'(r) = F'(r)$$

So,

$$G'(r) = \frac{1}{4\pi r^2} \int_{|\mathbf{x}-\mathbf{a}|<r} \nabla^2 \phi \, \mathrm{d}V$$

Now, note that as $r \to 0$,

$$\begin{aligned}\int_{|\mathbf{x}-\mathbf{a}|<r} \nabla^2\phi(\mathbf{x}) \, \mathrm{d}V &= \nabla^2\phi(\mathbf{a}) \int_{|\mathbf{x}-\mathbf{a}|<r} \mathrm{d}V + \int_{|\mathbf{x}-\mathbf{a}|<r} \big(\nabla^2\phi(\mathbf{x}) - \nabla^2\phi(\mathbf{a})\big) \, \mathrm{d}V \\ &= \frac{4\pi}{3r^3}\nabla^2\phi(\mathbf{a}) + o(r^3)\end{aligned}$$

Now, as $r \to 0$,

$$G'(r) = \frac{1}{4\pi r^2}\left[\frac{4\pi}{3r^3}\nabla^2\phi(\mathbf{a}) + o(r^3)\right] = \frac{r}{3}\nabla^2\phi(\mathbf{a}) + o(r)$$

Comparing this to the Taylor expansion,

$$G'(r) = G'(0) + rG''(0) + o(r)$$

So certainly, $G'(0) = 0$ since there is no constant term in $G'(r)$. Further, $G''(0) = \frac{1}{3}\nabla^2\phi(\mathbf{a})$. Now,

$$G(r) = G(0) + rG'(0) + \frac{r^2}{2}G''(0) + o(r^2)$$

We know that $G(0) = F(0) - \phi(\mathbf{a}) = 0$, hence

$$G(r) = \frac{1}{6}\nabla^2\phi(\mathbf{a})r^2 + o(r^2) \implies \nabla^2\phi(\mathbf{a}) = \lim_{r\to 0}\frac{6}{r^2}G(r)$$

which gives the result as required. □

9.9. Non-existence of maximum points

Proposition. If ϕ is harmonic on some volume $\Omega \subset \mathbb{R}^3$, then ϕ cannot have a maximum point at any interior point on Ω, unless ϕ is constant.

Proof. Suppose that there exists a maximum point at $\mathbf{a} \in \Omega$. Then $\phi(\mathbf{a}) \geq \phi(\mathbf{x})$ for all $\mathbf{x} \in \Omega$. Then,

$$\phi(\mathbf{a}) \geq \phi(\mathbf{x}) \text{ on } |\mathbf{x} - \mathbf{a}| \leq \varepsilon$$

for some ε small enough such that the ball is inside Ω. By the mean value property,

$$\phi(\mathbf{a}) = \frac{1}{4\pi\varepsilon^2}\int_{|\mathbf{x}-\mathbf{a}|=\varepsilon} \phi(\mathbf{x}) \, \mathrm{d}S$$

Hence,

$$0 = \frac{1}{4\pi\varepsilon^2} \int_{|\mathbf{x}-\mathbf{a}|=\varepsilon} (\phi(\mathbf{a}) - \phi(\mathbf{x})) \, \mathrm{d}S$$

Note that the integrand is always non-negative, so in order for the integral to equal zero, the integrand must be zero everywhere on the ball. So $\phi(\mathbf{a}) = \phi(\mathbf{x})$. Since ε was arbitrary, we can shrink the ball to a smaller ball around the same point, so $\phi(\mathbf{a}) = \phi(\mathbf{x})$ for all $\mathbf{x}$ such that $|\mathbf{x} - \mathbf{a}| \leq \varepsilon$. Hence, ϕ is locally constant.

Now, given any other point $\mathbf{y}$, we can introduce a finite sequence of overlapping balls such that the centre of the $(n+1)$th ball is contained inside the nth ball, and where the first ball is centred at $\mathbf{a}$ and the last ball is centred at $\mathbf{y}$. Inductively, the function is constant on each such ball. Hence ϕ is actually constant everywhere, since $\mathbf{y}$ was arbitrarily chosen. □

Corollary. If ϕ is harmonic on Ω, then for $\mathbf{x} \in \Omega$,

$$\phi(\mathbf{x}) \leq \max_{\mathbf{y} \in \partial\Omega} \phi(\mathbf{y})$$

This is called the maximum principle.

10. Cartesian tensors

Throughout this section on tensors, we deal exclusively with Cartesian coordinate systems.

10.1. Intuitive description of vectors and changes of basis

Consider a right-handed orthonormal basis $\{\mathbf{e}_i\}$ for $\mathbb{R}^3$, with respect to some fixed Cartesian coordinate axes. We can write a vector using this basis as

$$\mathbf{x} = x_i \mathbf{e}_i$$

Note that the vector $\mathbf{x}$ and the components x_i are not the same; the components only give the vector when in combination with the given basis vectors $\{\mathbf{e}_i\}$. If we instead use $\{\mathbf{e}'_i\}$, then the same position vector $\mathbf{x}$ would be written as a linear combination $x'_i \mathbf{e}'_i$. Hence,

$$x_j \mathbf{e}_j = x'_j \mathbf{e}'_j \tag{$*$}$$

Since the $\{\mathbf{e}_j\}$ and $\{\mathbf{e}'_j\}$ are orthonormal,

$$\mathbf{e}_i \cdot \mathbf{e}_j = \delta_{ij}; \quad \mathbf{e}'_i \cdot \mathbf{e}'_j = \delta_{ij}$$

From $(*)$,

$$x'_i = \delta_{ij} x'_j = (\mathbf{e}'_i \cdot \mathbf{e}'_j) x'_j = \mathbf{e}'_i \cdot (\mathbf{e}'_j x'_j) = \mathbf{e}'_i \cdot (\mathbf{e}_j x_j) = (\mathbf{e}'_i \cdot \mathbf{e}_j) x_j$$

So let

$$R_{ij} = \mathbf{e}'_i \cdot \mathbf{e}_j$$

Then

$$x'_i = R_{ij} x_j$$

Alternatively,

$$x_i = \delta_{ij} x_j = (\mathbf{e}_i \cdot \mathbf{e}_j) x_j = \mathbf{e}_i \cdot (\mathbf{e}_j x_j) = \mathbf{e}_i \cdot (\mathbf{e}'_j x'_j) = (\mathbf{e}_i \cdot \mathbf{e}'_j) x'_j$$

And therefore, we get

$$x_i = R_{ji} x'_j = R_{ki} x'_k \implies x_j = R_{kj} x'_k$$

Combining the two results, we have

$$x'_i R_{ij} x_j = R_{ij} R_{kj} x'_k$$

Therefore,

$$(\delta_{ik} - R_{ij} R_{kj}) x'_k = 0$$

Since this is true for all vectors $\mathbf{x}$, we get

$$R_{ij} R_{kj} = \delta_{ik}$$

So if R is a matrix with entries R_{ij}, then

$$RR^\intercal = I$$

So the R_{ij} are the components of an orthogonal matrix. Further, since

$$x_j \mathbf{e}_j = x'_i \mathbf{e}'_i = R_{ij} x_j \mathbf{e}'_i$$

holds for all x_j, we also have

$$\mathbf{e}_j = R_{ij} \mathbf{e}'_i$$

and since both $\{\mathbf{e}_i\}$ and $\{\mathbf{e}'_i\}$ are right handed, we have

$$1 = \mathbf{e}_1 \cdot (\mathbf{e}_2 \times \mathbf{e}_3) = R_{i1} R_{j2} R_{k3} \mathbf{e}'_i \cdot (\mathbf{e}'_j \times \mathbf{e}'_k) = R_{i1} R_{j2} R_{k3} \varepsilon_{ijk} = \det R$$

Hence R is orthogonal, and has determinant 1. Hence R is a rotation matrix. If we transform from a right-handed orthonormal set of basis vectors $\{\mathbf{e}_i\}$ to another basis $\{\mathbf{e}'_i\}$, then the components of a vector $\mathbf{v}$ transform according to $v'_i = R_{ij} v_j$. We call objects whose components transform in this way 'rank 1 tensors', or more commonly, 'vectors'. The basis vectors themselves transform according to $\mathbf{e}'_j = R_{ij} \mathbf{e}_i$.

10.2. Intuitive description of scalars and scalar products

Consider the dot product between two vectors, $\sigma = \mathbf{a} \cdot \mathbf{b}$. This should ideally be independent of the set of basis vectors chosen to describe $\mathbf{a}$ and $\mathbf{b}$. So with a basis $\{\mathbf{e}_i\}$, we have

$$\sigma = a_i b_j \delta_{ij} = a_i b_i$$

If instead we use a different set of basis vectors $\{\mathbf{e}_j\}$, we define

$$\sigma' = a'_i b'_i$$

We can use $a'_i = R_{ip} a_p$ and $b'_i = R_{iq} b_q$ to give

$$\sigma' = R_{ip} R_{iq} a_p b_q = \delta_{pq} a_p b_q = a_i b_i = \sigma$$

Since the sets of basis vectors are related by R, σ is unchanged under changes of coordinates. We call objects which are invariant under transformations like this 'rank 0 tensors', or 'scalars'.

10.3. Intuitive description of linear maps

Let $\mathbf{n} \in \mathbb{R}^3$ be a fixed unit vector, and we define a linear map

$$T : \mathbf{x} \to \mathbf{y} = T(\mathbf{x}) = \mathbf{x} - (\mathbf{x} \cdot \mathbf{n})\mathbf{n}$$

This T is the orthogonal projection into the plane normal to $\mathbf{n}$. Using a set of basis vectors $\{\mathbf{e}_i\}$, we get

$$y_i\mathbf{e}_i = T(x_j\mathbf{e}_j) = x_jT(\mathbf{e}_j) = x_j(\mathbf{e}_j - n_in_j\mathbf{e}_i) = (\delta_{ij} - n_in_j)x_j\mathbf{e}_i$$

Hence,

$$y_i = (\delta_{ij} - n_in_j)x_j$$

So we will set

$$T_{ij} = \delta_{ij} - n_in_j \implies y_i = T_{ij}x_j$$

We call the T_{ij} the *components* of the linear map T with respect to the basis vectors $\mathbf{e}_i$. Consider a different set of basis vectors $\{\mathbf{e}'_i\}$.

$$y'_i = (\delta_{ij} - n'_in'_j)x'_j; \quad T'_{ij} = \delta_{ij} - n'_in'_j$$

Using $n'_i = R_{ij}n_j$, noting that R is orthogonal, we have

$$T'_{ij} = \delta_{ij} - R_{ip}n_jR_{jq}n_q = R_{ip}R_{jq}(\delta_{pq} - n_pn_q) = R_{ip}R_{jq}T_{pq}$$

So the components of a linear map transform according to two multiplications:

$$T'_{ij} = R_{ip}R_{jq}T_{pq}$$

We call such objects 'rank 2 tensors'.

10.4. Definition

Definition. An object whose components $T_{ij\ldots k}$ transform according to

$$T'_{ij\ldots k} = R_{ip}R_{jq}\ldots R_{kr}T_{pq\ldots r}$$

is called a (Cartesian) tensor of rank n if T has n indices, where $R_{ij} = \mathbf{e}'_i \cdot \mathbf{e}_j$ are the components of an orthogonal matrix, so $R_{ip}R_{jp} = \delta_{ij}$.

For example, if u_i, v_j, w_k are the components of n vectors, then

$$T_{ij\ldots k} = u_iv_j\ldots w_k$$

define the components of a tensor of rank n.

Proof. We can transform each vector individually.

$$T'_{ij\ldots k} = u'_iv'_j\ldots w'_k = R_{ip}u_pR_{jq}v_q\ldots R_{kr}w_r = R_{ip}R_{jq}R_{kr}T_{ij\ldots k}$$

as expected. □

10.5. Kronecker δ and Levi-Civita ε

As another example, consider the Kronecker δ. It was previously defined without reference to any basis by

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

So $\delta'_{ij} = \delta_{ij}$ by definition. Note that

$$R_{ip}R_{jq}\delta_{pq} = R_{iq}R_{jq} = \delta_{ij} = \delta'_{ij}$$

hence δ transforms like a rank 2 tensor, so it is indeed a rank 2 tensor. Now, consider the Levi-Civita symbol ε. It is defined without reference to any basis as

$$\varepsilon_{ijk} = \begin{cases} +1 & (i\ j\ k) \text{ even} \\ -1 & (i\ j\ k) \text{ odd} \\ 0 & \text{otherwise} \end{cases}$$

Note that $\varepsilon'_{ijk} = \varepsilon_{ijk}$, and

$$R_{ip}R_{jq}R_{kr}\varepsilon_{pqr} = \det R \cdot \varepsilon_{ijk} = \varepsilon_{ijk}$$

Hence ε is a rank 3 tensor.

10.6. Electrical conductivity tensor

Experiments suggest that there is a linear relationship between the current $\mathbf{J}$ produced in a conductive medium and the electric field $\mathbf{E}$ that it is exposed to. Hence $\mathbf{J} = \sigma\mathbf{E}$, or $J_i = \sigma_{ij}E_j$. σ_{ij} is called the 'electrical conductivity tensor'. It really is a rank 2 tensor, indeed

$$\begin{aligned} J'_i &= \sigma'_{ij}E'_j \\ R_{ip}J_p &= \sigma'_{ij}E'_j \\ R_{ip}\sigma_{pq}E_q &= \sigma'_{ij}E'_j \end{aligned}$$

Since R is orthogonal,

$$E'_j = R_{jq}E_q \iff E_q = R_{jq}E'_j$$

Hence,

$$R_{ip}R_{jq}\sigma_{pq}E'_j = \sigma'_{ij}E'_j$$

Since this is true for all choices of E_j,

$$R_{ip}R_{jq}\sigma_{pq} = \sigma'_{ij}$$

So it really is a rank 2 tensor.

10.7. Indexed objects without tensor transformation properties

It is possible to construct objects with indices that do not transform as tensors. For example, given a Cartesian right handed basis $\{\mathbf{e}_i\}$, we can define an arbitrary array of numbers with components A_{ij}, and set $A'_{ij} = 0$ in all other bases $\{\mathbf{e}'_i\}$. Clearly this array of numbers does not transform like a tensor.

10.8. Operations on tensors

Let $A_{ij\ldots k}, B_{ij\ldots k}$ be rank n tensors, we define

$$(A + B)_{ij\ldots k} = A_{ij\ldots k} + B_{ij\ldots k}$$

$A + B$ is also a rank n tensor, by linearity. Further,

$$(\alpha A)_{ij\ldots k} = \alpha A_{ij\ldots k}$$

αA is also a rank n tensor. We also define the *tensor product* between a rank m tensor $U_{ij\ldots k}$ and a rank n tensor $V_{pq\ldots r}$ as

$$(U \otimes V)_{ij\ldots kpq\ldots r} = U_{ij\ldots k} V_{pq\ldots r}$$

Now, $U \otimes V$ is a rank $m + n$ tensor. Indeed,

$$(U\otimes V)_{i\ldots jp\ldots q} = U'_{i\ldots j} V'_{p\ldots q} = R_{ia} \ldots R_{jb} U_{a\ldots b} R_{pc} \ldots R_{qd} V_{c\ldots d} = R_{ia} \ldots R_{jb} R_{pc} \ldots R_{qd} (U\otimes V)_{a\ldots bc\ldots d}$$

Further, given a rank $n \geq 2$ tensor $T_{ijk\ldots\ell}$, we can define a tensor of rank $n-2$ by *contracting* on a pair of indices. For instance, contracting on i and j is defined by

$$\delta_{ij} T_{ijk\ldots\ell} = T_{iik\ldots\ell}$$

This is really a tensor of rank $n - 2$:

$$T'_{iik\ldots\ell} = R_{ip} R_{iq} R_{kr} \ldots R_{\ell s} T_{pqr\ldots s} = \delta_{pq} R_{kr} \ldots R_{\ell s} T_{pqr\ldots s} = R_{kr} \ldots R_{\ell s} T_{ppr\ldots s}$$

10.9. Symmetric and antisymmetric tensors

We say that $T_{ij\ldots k}$ is symmetric in (i, j) if

$$T_{ij\ldots k} = T_{ji\ldots k}$$

This really is a well-defined property of the *tensor*, not its coordinates. In a different coordinate frame,

$$T'_{ij\ldots k} = R_{ip} R_{jq} \ldots R_{kr} T_{pq\ldots r} = R_{ip} R_{jq} \ldots R_{kr} T_{qp\ldots r} = T'_{ji\ldots k}$$

Similarly, we say that $A_{ij\ldots k}$ is antisymmetric in (i, j) if

$$A_{ij\ldots k} = -A_{ji\ldots k}$$

which similarly is invariant of the choice of basis. We say that a tensor is *totally* (anti-) symmetric if it is (anti-) symmetric in all pairs of indices. For example, the δ_{ij} rank 2 tensor and $a_i a_j a_k$ rank 3 tensor (where $\mathbf{a}$ is a vector) are totally symmetric tensors. The Levi-Civita alternating tensor ε is totally antisymmetric.

In fact, in three dimensions, ε is the only totally antisymmetric tensor (up to scaling), and there are no nonzero higher-rank antisymmetric tensors. Indeed, if $T_{ij\ldots k}$ is totally antisymmetric and has rank n, then $T_{ij\ldots k} = 0$ if any two indices are the same. But if we have more than three indices, by the pigeonhole principle we must have two matching indices (provided we are working in three dimensions). If $n = 3$, then there are only $3! = 6$ choices of components that give a nonzero value of T_{ijk}, and by antisymmetry, $T_{123} = T_{231} = T_{312} = \lambda$ and by antisymmetry $T_{213} = T_{132} = T_{321} = -\lambda$ which defines the ε symbol.

11. Tensor calculus

11.1. Introduction

A vector field assigns a vector $\mathbf{v}$ to every position $\mathbf{x} \in \mathbb{R}^3$. A scalar field assigns a scalar ϕ to every position. We generalise this notion to a *tensor field* of rank n, written $T_{ij\dots k}(\mathbf{x})$, which assigns a rank n tensor to every point $\mathbf{x}$. Recall that

$$x_i' = R_{ij}x_j \iff x_j = R_{ij}x_i'$$

Differentiating both sides with respect to x_k', we get

$$\frac{\partial x_j}{\partial x_k'} = R_{ij}\frac{\partial x_i'}{\partial x_k'} = R_{ij}\delta_{ik} = R_{kj}$$

By the chain rule, we then have

$$\frac{\partial}{\partial x_i'} = \frac{\partial x_j}{\partial x_i'}\frac{\partial}{\partial x_j} = R_{ij}\frac{\partial}{\partial x_j}$$

Informally, we can say that $\frac{\partial}{\partial x_i'}$ transforms like a rank 1 tensor.

Proposition. If $T_{i\dots j}$ is a tensor field of rank n, then

$$\underbrace{\frac{\partial}{\partial x_p}\cdots\frac{\partial}{\partial x_q}}_{m \text{ terms}} T_{i\dots j}(\mathbf{x})$$

is a tensor field of rank $n + m$.

Proof. We check the transformation under a change of basis. Let the above expression be $A_{p\dots qi\dots j}$. Then

$$\begin{aligned} A'_{p\dots qi\dots j} &= \frac{\partial}{\partial x_p'}\cdots\frac{\partial}{\partial x_q'}T'_{i\dots j}(\mathbf{x}) \\ &= R_{pa}\frac{\partial}{\partial x_a}\cdots R_{qb}\frac{\partial}{\partial x_b}R_{ic}\dots R_{jd}T_{c\dots d}(\mathbf{x}) \\ &= R_{pa}\dots R_{qb}R_{ic}\dots R_{jd}A_{a\dots bc\dots d} \end{aligned}$$

□

Note that this only works in Cartesian coordinates, since the R matrices are constant here. In a general coordinate system, this is not the case, and we cannot move the change of basis matrices outside the derivatives in this case.

11.2. Differential operators producing tensor fields

If ϕ is a scalar field, then

$$[\nabla\phi]_i = \frac{\partial\phi}{\partial x_i}$$

Hence $\nabla\phi$ is a rank 1 tensor field, which is a vector field. If $\mathbf{v}$ is a vector field,

$$\nabla\cdot\mathbf{v} = \frac{\partial v_i}{\partial x_i}$$

which is a rank 0 tensor field since it is a contraction of $\frac{\partial \mathbf{v}_i}{\partial x_j}$. Alternatively, from first principles,

$$\frac{\partial v_i'}{\partial x_i'} = R_{ip}\frac{\partial}{\partial x_p}R_{iq}v_q = R_{ip}R_{iq}\frac{\partial v_q}{\partial x_p} = \delta_{pq}\frac{\partial v_q}{\partial x_p} = \frac{\partial v_i}{\partial x_i}$$

hence the divergence of a vector field really is a scalar field.

$$[\nabla\times\mathbf{v}]_i = \varepsilon_{ijk}\frac{\partial v_k}{\partial x_j}$$

From first principles we can show that

$$\begin{aligned}
\varepsilon_{ijk}'\frac{\partial v_k'}{\partial x_j'} &= R_{ia}R_{jb}R_{kc}\varepsilon_{abc}R_{jp}\frac{\partial}{\partial x_p}R_{kq}v_q \\
&= R_{ia}\varepsilon_{abc}R_{jb}R_{jp}R_{kc}R_{kq}\frac{\partial v_q}{\partial x_p} \\
&= R_{ia}\varepsilon_{abc}\delta_{bp}\delta_{cq}\frac{\partial v_q}{\partial x_p} \\
&= R_{ia}\varepsilon_{abc}\frac{\partial v_c}{\partial x_b}
\end{aligned}$$

which is the transformation law for a rank 1 tensor, so the curl of a vector field is a vector field.

11.3. Divergence theorem with tensor fields

Proposition. For a tensor field $T_{ij\ldots k\ldots\ell}(\mathbf{x})$, we have

$$\int_V \frac{\partial}{\partial x_k}T_{ij\ldots k\ldots\ell}\,\mathrm{d}V = \int_{\partial V} T_{ij\ldots k\ldots\ell}n_k\,\mathrm{d}S$$

Proof. Consider the vector field

$$v_k = a_i b_j \ldots c_\ell T_{ij\ldots k\ldots\ell}$$

where the $a_i, b_j, \dots, c_\ell$ are the components of some constant vectors. Applying the divergence theorem to this vector field, we have

$$\int_V \frac{\partial v_k}{\partial x_k} \, \mathrm{d}V = \int_{\partial V} v_k n_k \, \mathrm{d}S$$

$$a_i b_j \dots c_\ell \int_V \frac{\partial}{\partial x_k} T_{ij\dots k\dots \ell} \, \mathrm{d}V = a_i b_j \dots c_\ell \int_{\partial V} T_{ij\dots k\dots \ell} n_k \, \mathrm{d}S$$

Since this is true for any choice of vectors $a_i, b_i, \dots, c_i$, the result follows. □

12. Properties of tensors

12.1. Symmetry and antisymmetry

Observe for a rank 2 tensor that

$$T_{ij} = \frac{1}{2}\left(T_{ij} + T_{ji}\right) + \frac{1}{2}\left(T_{ij} - Tji\right) \equiv S_{ij} + A_{ij}$$

where the S_{ij} are the symmetric components, and the A_{ij} are the antisymmetric components of the tensor. Note that the symmetric part S_{ij} has six independent components (the main diagonal and everything above it), and the antisymmetric part A_{ij} has three independent components (everything above the main diagonal) since the main diagonal is zero. So the number of independent components of the symmetric part and the antisymmetric part add up to the number of independent components of a general rank 2 tensor in $\mathbb{R}^3$ (nine). Intuitively, we might think that the information contained in A_{ij} could be represented as some vector, since it has the same amount of independent components.

Proposition. Every rank 2 tensor T_{ij} can be decomposed uniquely into

$$T_{ij} = S_{ij} + \varepsilon_{ijk}\omega_k$$

where

$$\omega_i = \frac{1}{2}\varepsilon_{ijk}T_{jk}$$

and S_{ij} is symmetric.

Proof. From above, we can find $S_{ij} = \frac{1}{2}\left(T_{ij} + T_{ji}\right)$. We now just need to show that

$$\varepsilon_{ijk}\omega_k = \frac{1}{2}\left(T_{ij} - Tji\right)$$

We can see that

$$\begin{aligned}
\varepsilon_{ijk}\omega_k &= \frac{1}{2}\varepsilon_{ijk}\varepsilon_{k\ell m}T_{\ell m} \\
&= \frac{1}{2}(\delta_{i\ell}\delta_{jm} - \delta_{im}\delta_{j\ell})T_{\ell m} \\
&= \frac{1}{2}(T_{ij} - T_{ji})
\end{aligned}$$

To show uniqueness, we now suppose that

$$T_{ij} = S_{ij} + A_{ij} = \tilde{S}_{ij} + \widetilde{A}_{ij} = \widetilde{T}_{ij}$$

If we take the symmetric part of both sides (i.e. $T_{ij} + T_{ji} = \widetilde{T}_{ij} + \widetilde{T}_{ji}$), we get $S_{ij} = \tilde{S}_{ij}$. Likewise, we have $A_{ij} = \widetilde{A}_{ij}$ by eliminating the equal symmetric parts. □

As an example, consider an elastic body. Each point $\mathbf{x}$ in such a body will undergo a small displacement $\mathbf{u}(\mathbf{x})$ when applied to some force. Consider nearby points $\mathbf{x} + \delta\mathbf{x}$ and $\mathbf{x}$ that were initially separated by δx. They will become separated by

$$(\mathbf{x} + \delta\mathbf{x} + \mathbf{u}(\mathbf{x} + \delta\mathbf{x})) - (\mathbf{x} + \mathbf{u}(\mathbf{x})) = \delta\mathbf{x} + \mathbf{u}(\mathbf{x} + \delta\mathbf{x}) - \mathbf{u}(\mathbf{x})$$

So the change in displacement is

$$\mathbf{u}(\mathbf{x} + \delta\mathbf{x}) - \mathbf{u}(\mathbf{x})$$

This value gives us an idea of how much deformation the body is subjected to. Assuming this is a smooth deformation, we have

$$u_i(\mathbf{x} + \delta\mathbf{x}) - u_i(\mathbf{x}) = \frac{\partial u_i}{\partial x_j}\delta_{x_j} + o(\delta\mathbf{x})$$

We then decompose $\frac{\partial u_i}{\partial x_j}$ as follows.

$$\frac{\partial u_i}{\partial x_j} = e_{ij} + \varepsilon_{ijk}\omega_k$$

where the e_{ij} is the symmetric part, and the $\varepsilon_{ijk}\omega_k$ is the antisymmetric part. In particular,

$$e_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)$$

is called the linear strain tensor. Considering the other tensor,

$$\omega_k = \frac{1}{2}\varepsilon_{ijk}\frac{\partial u_j}{\partial x_k} = \frac{-1}{2}(\nabla \times \mathbf{u})_i$$

Then,

$$u_i(\mathbf{x} + \delta\mathbf{x}) - u_i(\mathbf{x}) = e_{ij}\delta x_j + [\delta\mathbf{x} \times \boldsymbol{\omega}]_i + o(\delta\mathbf{x})$$

So the antisymmetric part corresponds to a rotation, and is irrelevant for describing the deformation of the internals of the body. So by separating the symmetric and antisymmetric parts, we can in fact remove the antisymmetric part from the equation in order to study just the linear strain.

Example. As another example, let us consider the inertia tensor, which is a common rank 2 tensor. Suppose a body with density $\rho(\mathbf{x})$ occupies a volume $V \subset \mathbb{R}^3$, where each point in the body is rotating with constant angular velocity $\boldsymbol{\omega}$ about an axis through the origin. The

velocity of a point $\mathbf{x} \in V$ is given by $\mathbf{v} = \boldsymbol{\omega} \times \mathbf{x}$. Hence, the total angular momentum is

$$\begin{aligned}
\mathbf{L} &= \int_V \rho(\mathbf{x})(\mathbf{x} \times \mathbf{v})\,\mathrm{d}V \\
&= \int_V \rho(\mathbf{x})(\mathbf{x} \times (\boldsymbol{\omega} \times \mathbf{x}))\,\mathrm{d}V \\
L_i &= \int_{\mathcal{V}} \rho(\mathbf{x})(x_k x_k \omega_i - x_i x_j \omega_j)\,\mathrm{d}V \\
&= \int_{\mathcal{V}} \rho(\mathbf{x})(x_k x_k \delta_{ij} \omega_j - x_i x_j \omega_j)\,\mathrm{d}V \\
&= I_{ij}\omega_j
\end{aligned}$$

where I_{ij} is the inertia tensor defined by

$$I_{ij} = \int_{\mathcal{V}} \rho(\mathbf{x})(x_k x_k \delta_{ij} - x_i x_j)\,\mathrm{d}V$$

and where

$$\mathcal{V} = \{x_i : x_i \mathbf{e}_i \in V\}$$

If we had used a different basis, we would have found

$$\begin{aligned}
I'_{ij} &= \int_{\mathcal{V}'} \rho(\mathbf{x})(x'_k x'_k \delta_{ij} - x'_i x'_j)\,\mathrm{d}V \\
&= R_{ip}R_{jq}\int_{\mathcal{V}} \rho(\mathbf{x})(x_k x_k \delta_{pq} - x_p x_q)\,\mathrm{d}V \\
&= R_{ip}R_{jq}I_{pq}
\end{aligned}$$

So it really is a rank 2 tensor. As an example, consider the ellipsoid

$$V = \left\{\mathbf{x} : \frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} + \frac{x_3^2}{c^2} \leq 1\right\}$$

with uniform density ρ_0. Then the mass is given by

$$M = \frac{4}{3}\pi\rho_0 abc$$

Then the inertia tensor with respect to this set of basis vectors is given by

$$I_{ij} = \int_{\mathcal{V}} \rho(\mathbf{x})(x_k x_k \delta_{ij} - x_i x_j)\,\mathrm{d}V$$

To help with these integrals, we make the following parametrisation into scaled spherical coordinates:

$$\begin{cases} x_1 = ar\cos\phi\sin\theta \\ x_2 = br\sin\phi\sin\theta \\ x_3 = cr\cos\theta \end{cases} \qquad \phi \in [0, 2\pi), \theta \in [0, \pi], r \in [0, 1]$$

Note that if $i \neq j$, then by symmetry we have

$$\int_V \rho_0 x_i x_j \,\mathrm{d}V = 0$$

Further,

$$\begin{aligned}
I_{11} &= \rho_0 \int_V x_2^2 + x_3^2 \,\mathrm{d}V \\
&= \rho_0 abc \int_{\phi=0}^{2\pi} \mathrm{d}\phi \int_{\theta=0}^{\pi} \mathrm{d}\theta \int_{r=0}^{1} \mathrm{d}r\, r^2 (b^2 \sin^2\phi \sin^2\theta + c^2 \cos^2\theta) r^2 \sin\theta \\
&= \rho_0 \frac{abc}{5} \int_0^{\pi} (\pi b^2 \sin^2\theta + 2\pi c^2 \cos^2\theta) \sin\theta \,\mathrm{d}\theta \\
&= \frac{3M}{20} \int_0^{\pi} (b^2 \sin^2\theta + (2c^2 - b^2)\cos^2\theta \sin\theta) \,\mathrm{d}\theta \\
&= \frac{3M}{20}\left(2b^2 + \frac{2}{3}(2c^2 - b^2)\right) \\
&= \frac{M}{5}(b^2 + c^2)
\end{aligned}$$

So by symmetry,

$$I_{22} = \frac{M}{5}(a^2 + c^2); \quad I_{33} = \frac{M}{5}(a^2 + b^2)$$

Hence,

$$I_{ij} = \frac{M}{5}\begin{pmatrix} b^2 + c^2 & 0 & 0 \\ 0 & a^2 + c^2 & 0 \\ 0 & 0 & a^2 + b^2 \end{pmatrix}$$

In particular, if $a = b = c$,

$$I_{ij} = \frac{2M}{5}\delta_{ij}$$

Proposition. If T_{ij} is symmetric, then there exists a basis $\{\mathbf{e}_i\}$ for which T_{ij} only has nonzero entries on the diagonal. The coordinate axes of this basis are called the principal axes of the tensor.

Proof. Recall that for a real symmetric matrix M, we can diagonalise it using an orthogonal transformation with determinant 1. The change of basis formula for a matrix is exactly that for a rank 2 tensor, so we can always choose such a change of basis to give a diagonal matrix. □

12.2. Isotropic tensors

Definition. A tensor is isotropic if it is invariant under changes with respect to the choice of Cartesian coordinate axes.

$$T'_{ij\ldots k} = R_{ip}R_{jq}\ldots R_{kr}T_{pq\ldots r} = T_{ij\ldots k}$$

for any choice of rotation R.

Note that by definition, every scalar is isotropic. The Kronecker and Levi-Civita tensors are also isotropic, as we saw above.

12.3. Classifying isotropic tensors in three dimensions

Proposition. The isotropic tensors on $\mathbb{R}^3$, ordered by rank, are exactly (up to the multiplication of a multiplicative scalar)

Rank 0: all tensors

Rank 1: no nonzero tensors

Rank 2: the Kronecker δ

Rank 3: the Levi-Civita ε

Rank 4: $\alpha\delta_{ij}\delta_{k\ell} + \beta\delta_{ik}\delta_{j\ell} + \gamma\delta_{i\ell}\delta_{jk}$ where α, β, γ are scalars

and for ranks higher than 4, they are a linear combination of products of δ and ε terms, for instance $\delta_{ij}\varepsilon_{k\ell m}$.

Proof. This is a non-rigorous sketch proof.

Rank 0: By definition, such tensors do not transform components under a change of basis.

Rank 1: Let v_i be the components of an isotropic vector of rank 1. Then, for any R, we must have

$$v_i = R_{ij}v_j$$

Let R be a rotation by π about the z axis, so

$$R = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Hence,

$$v_1 = -v_1; \quad v_2 = -v_2; \quad v_3 = v_3$$

Hence, $v_1 = 0, v_2 = 0$. Alternatively, let

$$R = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

Then clearly $v_3 = -v_3 = 0$. Hence the only tensor with this property is the zero tensor.

Rank 2: If T_{ij} are the components of an isotropic tensor of rank 2, then for all choices of R, we have

$$T_{ij} = R_{ip}R_{jq}T_{pq}$$

Let R be a rotation by $\frac{\pi}{2}$ about each axis, so for example in the z direction,

$$R = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

So $T_{13} = R_{1p}R_{3q}T_{pq} = R_{12}R_{33}T_{23} = T_{23}$. Analogously we find, $T_{23} = -T_{13}$. Hence, $T_{13} = T_{23} = 0$. Further, $T_{11} = R_{1p}R_{1q}T_{pq} = R_{12}R_{12}T_{22} = T_{22}$. So by symmetry,

$$T_{11} = T_{22} = T_{33}; \quad T_{13} = T_{23} = T_{12} = T_{31} = T_{32} = T_{21} = 0$$

which is exactly the δ tensor, up to a scale factor.

Rank 3: For rank 3 tensors, we can use the same idea, but with more indices.

□

12.4. Integrals with isotropic tensors

Consider an integral of the form

$$T_{ij\dots k} = \int_{|\mathbf{x}|<R} f(r)x_i x_j \dots x_k \,\mathrm{d}V$$

where $x_k x_k = r^2$, and $\mathrm{d}V(x) = \mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_3$. Note that $f(r)$ and $\{\mathbf{x} : |\mathbf{x}| < R\}$ are invariant under rotation. Since $|J|$ under a rotation is 1, we have

$$\begin{aligned} T'_{ij\dots k} &= \int_{|\mathbf{x}|<R} f(r)x'_i x'_j \dots x'_k \,\mathrm{d}x'_1\,\mathrm{d}x'_2\,\mathrm{d}x'_3 \\ &= \int_{|\mathbf{x}|<R} f(r)R_{ip}x_p R_{jq}x_q \dots R_{kr}x_r \,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_3 \end{aligned}$$

We will now make the substitution

$$y_i = R_{ij}x_j; \quad \mathrm{d}V = \mathrm{d}y_1\,\mathrm{d}y_2\,\mathrm{d}y_3$$

Hence,

$$\begin{aligned} T'_{ij\dots k} &= \int_{|\mathbf{x}|<R} f(r)y_i y_j \dots y_k \,\mathrm{d}V(\mathbf{y}) \\ &= \int_{|\mathbf{x}|<R} f(r)x_i x_j \dots x_k \,\mathrm{d}V(\mathbf{x}) \\ &= T_{ij\dots k} \end{aligned}$$

Hence such an integral always yields an isotropic tensor. If we take $R \to \infty$, this corresponds to an integral over $\mathbb{R}^3$. As an example, consider

$$T_{ij} = \int_{\mathbb{R}^3} e^{-r^5} x_i x_j \, \mathrm{d}V$$

Then T_{ij} is isotropic, hence $T_{ij} = \alpha \delta_{ij}$. Contracting on (i, j) to find α, we get

$$\begin{aligned}
\alpha \delta_{ii} &= 3\alpha \\
&= \int_{\mathbb{R}^3} e^{-r^5} r^2 \, \mathrm{d}V \\
&= 4\pi \int_0^\infty e^{-r^5} r^2 r^2 \, \mathrm{d}r \\
&= 4\pi \int_0^\infty e^{-r^5} r^4 \, \mathrm{d}r \\
&= \frac{4}{5}\pi
\end{aligned}$$

Hence,

$$T_{ij} = \frac{4}{15}\pi \delta_{ij}$$

As another example, consider the inertia tensor I_{ij} of a ball of radius R, uniform density ρ_0, and mass $M = \frac{4\pi}{3} R^3 \rho_0$. Recall that

$$I_{ij} = \int_{|\mathbf{x}|<R} \rho_0 (x_k x_k \delta_{ij} - x_i x_j) \, \mathrm{d}V$$

Both terms give an isotropic result, so the sum I_{ij} is isotropic. Contracting on (i, j), we have

$$\begin{aligned}
\alpha \delta_{ii} &= 3\alpha \\
&= \int_{|\mathbf{x}|<R} \rho_0 (r^2 \delta_{ii} - x_i x_i) \, \mathrm{d}V \\
&= \int_{|\mathbf{x}|<R} \rho_0 (3r^2 - r^2) \, \mathrm{d}V \\
&= \int_{|\mathbf{x}|<R} \rho_0 2r^2 \, \mathrm{d}V \\
&= 4\pi \int_0^R \rho_0 2r^4 \, \mathrm{d}r \\
&= \frac{4\pi}{3} \rho_0 R^3 \left(\frac{3}{R^3} \cdot 2 \cdot \frac{R^5}{5} \right) \\
&= \frac{6MR^2}{5}
\end{aligned}$$

Hence,

$$I_{ij} = \frac{2MR^2}{5}\delta_{ij}$$

12.5. Bilinear and multilinear maps as tensors

For a tensor T_{ij}, consider the bilinear map $t : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$ defined by

$$t(\mathbf{a}, \mathbf{b}) = T_{ij} a_i b_j$$

The left hand side really is well defined, since the right hand side does not depend on the choice of basis vectors. Conversely, suppose we have a bilinear map t. Then, for a given basis $\{\mathbf{e}_i\}$, this defines an array T_{ij} by

$$t(\mathbf{a}, \mathbf{b}) = t(a_i \mathbf{e}_i, b_j \mathbf{e}_j) = a_i b_j t(\mathbf{e}_i, \mathbf{e}_j) = a_i b_j T_{ij}$$

Changing basis with $\mathbf{e}'_i = R_{ip}\mathbf{e}_p$, we find

$$T'_{ij} = t(\mathbf{e}'_i, \mathbf{e}'_j) = t(R_{ip}\mathbf{e}_p, R_{jq}\mathbf{e}_q) = R_{ip}R_{jq}t(\mathbf{e}_p, \mathbf{e}_q)$$

hence this T_{ij} really is a rank 2 tensor. So there is a bijection between bilinear maps and rank 2 tensors. In particular, if the map

$$(\mathbf{a}, \mathbf{b}) \mapsto T_{ij} a_i b_j$$

is a bilinear map, and independent of basis, then T_{ij} *must* be the components of a rank 2 tensor. The same proof applies for higher-rank tensors.

12.6. Quotient theorem

Recall from earlier that the conductivity tensor σ_{ij} satisfying $J_i = \sigma_{ij}E_j$ was really a tensor, by using the definitions. The quotient theorem allows us to deduce similar results more generally. The name originates from the apparent 'quotient' of J_i by E_j to give σ_{ij}.

Proposition. Let $T_{i\dots jp\dots q}$ be an array of numbers defined in each Cartesian coordinate system, such that

$$v_{i\dots j} = T_{i\dots jp\dots q} u_{p\dots q}$$

and that $v_{i\dots j}$ is a tensor for all tensors $u_{p\dots q}$. Then $T_{i\dots jp\dots q}$ is a tensor.

Proof. We will first consider the special case $u_{p\dots q} = c_p \dots d_q$ for vectors $\mathbf{c}, \dots, \mathbf{d}$. Then by assumption,

$$v_{i\dots j} = T_{i\dots jp\dots q} c_p \dots d_q$$

is a tensor. In particular,

$$v_{i\dots j} a_i \dots b_j = T_{i\dots jp\dots q} a_i \dots b_j c_p \dots d_q$$

is a scalar, since the left hand side is just a contraction over all indices. Since the right hand side is invariant under a change in basis, this leads us to define the multilinear map

$$t(\mathbf{a}, \dots, \mathbf{b}, \mathbf{c}, \dots, \mathbf{d}) = T_{i\dots jp\dots q} a_i \dots b_j c_p \dots d_q$$

Hence $T_{i\dots jp\dots q}$ really is a tensor. □

As an example, consider the linear strain tensor

$$e_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)$$

where $\mathbf{u}(\mathbf{x})$ measures the change in displacement at $\mathbf{x}$. Experiments suggest that the internal stress tensor σ_{ij} experienced by a body under a deformation $\mathbf{u}(\mathbf{x})$ depends linearly on the strain e_{ij} at each point. Hence we might assume that there exists some array $c_{ijk\ell}$ such that

$$\sigma_{ij} = c_{ijk\ell} e_{k\ell}$$

However, we can't actually apply the quotient theorem here, since $e_{k\ell}$ cannot be *any* tensor, it can only be any *symmetric* tensor. See example sheet 4 for the resolution of this apparent problem: if $c_{ijk\ell} = c_{ij\ell k}$, then we can apply the quotient theorem. We call $c_{ijk\ell}$ the stiffness tensor, which is a property of the material being subjected to the force. Suppose that the material is isotropic, then we might guess that $c_{ijk\ell}$ should be isotropic. Hence,

$$c_{ijk\ell} = \alpha\delta_{ij}\delta_{k\ell} + \beta\delta_{ik}\delta_{j\ell} + \gamma\delta_{i\ell}\delta_{jk}$$

where α, β, γ are scalars. Putting this into the relationship between σ and e, we find

$$\sigma_{ij} = \alpha\delta_{ij}e_{kk} + \beta e_{ij} + \gamma e_{ji} = \lambda\delta_{ij}e_{kk} + 2\mu e_{ij}$$

which is a higher-dimensional analogue of Hooke's Law. We can in fact invert this. By contracting on (i, j) we find

$$\sigma_{ii} = 3\lambda e_{ii} + 2\mu e_{ii}$$

Hence,

$$e_{kk} = \frac{\sigma_{kk}}{3\lambda + 2\mu}$$

We then have

$$\sigma_{ij} = \lambda\delta_{ij}\frac{\sigma_{kk}}{3\lambda + 2\mu} + 2\mu e_{ij} \implies 2\mu e_{ij} = \sigma_{ij} - \sigma_{kk}\delta_{ij}\frac{\lambda}{3\lambda + 2\mu}$$

VIII. Analysis I

Lectured in Lent 2021 by Prof. G. Paternain

In this course, we rigorously define what it means for a sequence to approach a particular value; this is called a limit. Limits can be used to define things like derivatives and integrals, without appealing to concepts such as infinitesimals.

We begin by using limits to make sense of infinite summations, also called series. In general, a series may not have a sum (take $1 + 1 + \cdots$, for example), but many series do approach a value as we keep adding more terms (such as $1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots \to 2$). We prove various facts about when series converge to a value, and when they do not.

Then, we define what it means for a function to approach a value as we get closer to a particular input. We can use this to define the derivative of a function. As with series, derivatives may not exist (for example, $|x|$ at $x = 0$), so we need to be careful to restrict our analysis to differentiable functions. We can similarly make a rigorous definition of the integral, and show the fundamental theorem of calculus: under suitable assumptions, the derivative of the integral of a function is the original function.

Contents

1. Limits and convergence 522
1.1. Definition of limit 522
1.2. Fundamental axiom of the real numbers 522
1.3. Properties of limits 522
1.4. Harmonic series 523
1.5. Limits in the complex plane 523
1.6. The Bolzano–Weierstrass theorem 523
1.7. Cauchy sequences 524
2. Series 526
2.1. Definition 526
2.2. Geometric series 527
3. Convergence tests 528
3.1. Comparison test 528
3.2. Cauchy's root test 528
3.3. D'Alembert's ratio test 528
3.4. Cauchy's condensation test 529
3.5. Alternating series 530
4. Absolute convergence 532
4.1. Absolute convergence 532
4.2. Conditional convergence and rearrangement 532
5. Continuity 534
5.1. Definitions 534
5.2. Making continuous functions 534
5.3. Composition of continuous functions 535
6. Limit of a function 537
6.1. Definition 537
6.2. Properties 537
6.3. Intermediate value theorem 537
6.4. Bounds of a continuous function 538
6.5. Inverse functions 539
7. Differentiability 541
7.1. Definitions 541
7.2. Differentiation of sums and products 542
7.3. Differentiating polynomial terms 543
7.4. Chain rule 543
7.5. Rolle's theorem 544
7.6. Mean value theorem 544

7.7. Properties of a function from its derivative . . . 544
7.8. Inverse function theorem . . . 545
7.9. Derivative of rational powers . . . 546
7.10. Mean value theorem applied to limits . . . 546
7.11. Cauchy's mean value theorem . . . 546
8. Taylor's theorem . . . 548
8.1. Lagrange's and Cauchy's remainders . . . 548
8.2. Bounding error terms . . . 550
8.3. Binomial series . . . 550
9. Power series . . . 552
9.1. Complex differentiation . . . 552
9.2. Definition of power series . . . 552
9.3. Radius of convergence . . . 553
9.4. Infinite differentiability . . . 555
9.5. Defining standard functions . . . 558
9.6. Exponential and logarithmic functions . . . 559
9.7. Real numbered exponents . . . 560
9.8. Trigonometric functions . . . 561
9.9. Circle constants . . . 561
10. Integration . . . 563
10.1. Geometry of trigonometric functions . . . 563
10.2. Hyperbolic functions . . . 563
10.3. Defining the Riemann integral . . . 563
10.4. Determining integrability . . . 565
10.5. Monotonic and continuous functions . . . 565
10.6. Complicated integrable functions . . . 567
10.7. Properties of Riemann integral . . . 568
11. Fundamental theorem of calculus . . . 570
11.1. Breaking an interval . . . 570
11.2. Fundamental theorem of calculus . . . 571
12. Integration techniques . . . 573
12.1. Integration by parts . . . 573
12.2. Integration by substitution . . . 573
13. Integrals in Taylor's theorem . . . 574
13.1. Integral remainder form of Taylor's theorem . . . 574
13.2. Mean value theorem for integrals . . . 574
13.3. Deriving Lagrange's and Cauchy's remainders for Taylor's theorem 575
14. Uses of integration . . . 576
14.1. Improper integration . . . 576
14.2. Integral test for series convergence . . . 578
14.3. Piecewise continuous functions . . . 579

1. Limits and convergence

1.1. Definition of limit

Definition. We say that the sequence $a_n \to a$ as $n \to \infty$ if given $\varepsilon > 0$, $\exists N$ such that $|a_n - a| < \varepsilon$ for all $n \geq N$. Note that this N is actually a function of ε; we may need to choose a very large N if the ε provided is very small, for instance.

Definition. An increasing sequence is a sequence for which $a_n \leq a_{n+1}$, and a decreasing sequence is a sequence for which $a_n \geq a_{n+1}$. Such increasing and decreasing sequences are called monotone. A strictly increasing sequence or a strictly decreasing sequence simply strengthens the inequalities to not include the equality case.

1.2. Fundamental axiom of the real numbers

If we have some increasing sequence $a_n \in \mathbb{R}$, where $\exists A \in \mathbb{R}$ such that $\forall n \geq 1$, $a_n \leq A$, then $\exists a \in \mathbb{R}$ such that $a_n \to a$ as $n \to \infty$. This is also known as the 'least upper bound' axiom or property. This axiom applies equivalently to decreasing sequences of real numbers bounded below. We can also rephrase the axiom to state that every non-empty set of real numbers that is bounded above has a supremum.

Definition. We say that the supremum $\sup S$ of a non-empty, bounded above set S is K if

(i) $x \leq K$ for all $x \in S$

(ii) given $\varepsilon > 0$, $\exists x \in S$ such that $x > K - \varepsilon$

Note that the supremum (and hence the infimum) is unique.

1.3. Properties of limits

Lemma. The following properties about real sequences hold.

(i) The limit is unique. That is, if $a_n \to a$ and $a_n \to b$, then $a = b$.

(ii) If $a_n \to a$ as $n \to \infty$ and $n_1 < n_2 < \ldots$, then $a_{n_j} \to a$ as $j \to \infty$. In other words, subsequences converge to the same limit.

(iii) If $a_n = c$ for all n, then $a_n \to c$ as $n \to \infty$.

(iv) If $a_n \to a$ and $b_n \to b$, then $a_n + b_n \to a + b$.

(v) If $a_n \to a$ and $b_n \to b$, then $a_n b_n \to ab$.

(vi) If $a_n \to a$, $a_n \neq 0$ for all n, and $a \neq 0$, then $\frac{1}{a_n} \to \frac{1}{a}$.

(vii) If $a_n \to a$, and $a_n \leq A$ for all n, then $a \leq A$.

Proof. We prove the some of these statements here.

(i) Given $\varepsilon > 0$, $\exists n_1$ such that $|a_n - a| < \varepsilon$ for all $n \geq n_1$, and $\exists n_2$ such that $|a_n - b| < \varepsilon$ for all $n \geq n_2$. So let $N = \max(n_1, n_2)$, so both inequalities hold. Then for all $n \geq N$, using the triangle inequality, $|a - b| \leq |a_n - a| + |a_n - b| < 2\varepsilon$. So $a = b$.

(ii) Given $\varepsilon > 0$, $\exists N$ such that $|a_n - a| < \varepsilon$ for all $n \geq N$. Since $n_j \geq j$ (by induction), $\left|a_{n_j} - a\right| < \varepsilon$ for all $j \geq N$.

(v) $|a_n b_n - ab| \leq |a_n b_n - a_n b| + |a_n b - ab| = |a_n||b_n - b| + |b||a_n - a|$.

If $a_n \to a$, then given $\varepsilon > 0$, $\exists N_1$ such that $|a_n - a| < \varepsilon$ for all $n \geq N_1$. $(*)$

If $b_n \to b$, then given $\varepsilon > 0$, $\exists N_2$ such that $|b_n - b| < \varepsilon$ for all $n \geq N_2$.

Using $(*)$, if $n \geq N_1(1)$ (i.e. $\varepsilon = 1$), $|a_n - a| < 1$, so $|a_n| \leq |a| + 1$.

Therefore $|a_n b_n - ab| \leq \varepsilon(|a| + 1 + |b|)$ for all $n \geq N_3(\varepsilon) = \max\{N_1(1), N_1(\varepsilon), N_2(\varepsilon)\}$.

□

1.4. Harmonic series

Lemma. The sequence $\frac{1}{n}$ tends to zero as $n \to \infty$.

Proof. We know that $\frac{1}{n}$ is a decreasing sequence, and it is bounded below by zero. Hence it converges to a limit a. We will prove now that $a = 0$. $\frac{1}{2n} = \frac{1}{2} \cdot \frac{1}{n}$, and by property (v) above, $\frac{1}{2n}$ tends to $\frac{1}{2} \cdot a$. But $\frac{1}{2n}$ is a subsequence of $\frac{1}{n}$, and so by property (ii) it converges to a. So by property (i), $\frac{1}{2} \cdot a = a$ hence $a = 0$. □

1.5. Limits in the complex plane

Remark. The definition of the limit of a sequence makes perfect sense for $a_n \in \mathbb{C}$.

Definition. $a_n \to a$ if given $\varepsilon > 0$, $\exists N$ such that $\forall n \geq N$, $|a_n - a| < \varepsilon$.

From this definition, it is easy to check that properties (i)–(vi) hold for complex numbers.

However, property (vii) makes no sense in the world of the complex numbers since they do not have an ordering.

1.6. The Bolzano–Weierstrass theorem

Theorem. If x_n is a sequence of real numbers, and there exists some k such that $|x_n| \leq k$ for all n, then we can find $n_1 < n_2 < n_3 < n_4 < \ldots$ and $x \in \mathbb{R}$ such that $x_{n_j} \to x$ as $j \to \infty$. In other words, any bounded sequence has a convergent subsequence.

Remark. This theorem does not state anything about the uniqueness of such a subsequence; indeed, there could exist many subsequences that have possibly different limits. For example, $x_n = (-1)^n$ gives $x_{2n+1} \to -1$ and $x_{2n} \to 1$.

Proof. Let $[a_1, b_1]$ be the range of the sequence, i.e. $[-k, k]$. Then let the midpoint $c_1 = \frac{a_1+b_1}{2}$. Consider the following alternatives:

(i) $x_n \in [a_1, c]$ for infinitely many values of n.

(ii) $x_n \in [c, b_1]$ for infinitely many values of n.

Note that cases 1 and 2 could hold at the same time. If case 1 holds, we set $a_2 = a_1$ and $b_2 = c$. If case 1 fails, then case 2 must hold, so we can set $a_2 = c$ and $b_2 = b_1$. We have now constructed a subsequence whose range is half as large as the original sequence, and it contains infinitely many values of x_n.

We can proceed inductively to construct sequences a_n, b_n such that $x_m \in [a_n, b_n]$ for infinitely many values of m. This is known as a 'bisection method'. By construction, $a_{n-1} \leq a_n \leq b_n \leq b_{n-1}$. Since we are dividing by two each time,

$$b_n - a_n = \frac{1}{2}(b_{n-1} - a_{n-1}) \tag{$*$}$$

Note that a_n is a bounded, increasing sequence; and b_n is a bounded, decreasing sequence. By the Fundamental Axiom of the Real Numbers, a_n and b_n converge to limits $a \in [a_1, b_1]$ and $b \in [a_1, b_1]$. Using $(*)$, $b - a = \frac{b-a}{2} \implies b = a$.

Since $x_m \in [a_n, b_n]$ for infinitely many values of m, having chosen n_j such that $x_{n_j} \in [a_j, b_j]$, there is $n_{j+1} > n_j$ such that $x_{n_{j+1}} \in [a_{j+1}, b_{j+1}]$. Informally, this works because we have an unlimited supply of such x values. Hence

$$a_j \leq x_{n_j} \leq b_j$$

So this $x_{n_j} \to a$, so we have constructed a convergent subsequence. □

1.7. Cauchy sequences

Definition. A sequence a_n is called a Cauchy sequence if given $\varepsilon > 0$ there exists $N > 0$ such that $|a_n - a_m| < \varepsilon$ for all $n, m \geq N$. Informally, the terms of the sequence grow ever closer together such that there are infinitely many consecutive terms within a small region.

Lemma. If a sequence converges, it is a Cauchy sequence.

Proof. If $a_n \to a$, given $\varepsilon > 0$ then $\exists N$ such that $\forall n \geq N, |a_n - a| < \varepsilon$. Then take $m, n \geq N$, and we have

$$|a_n - a_m| \leq |a_n - a| + |a_m - a| < 2\varepsilon$$

□

Theorem. Every Cauchy sequence converges.

Proof. First, we note that if a_n is a Cauchy sequence then it is bounded. Let us take $\varepsilon = 1$, so $N = N(1)$ in the Cauchy property. Then

$$|a_n - a_m| < 1$$

for all $m, n \geq N(1)$. So by the triangle inequality,

$$|a_m| \leq |a_m - a_N| + |a_N| < 1 + |a_N|$$

So the sequence after this point is bounded by $1+|a_N|$. The remaining terms in the sequence are only finitely many, so we can compute the maximum of all of those terms along with $1 + |a_N|$ to produce a bound k for all n.

By the Bolzano–Weierstrass Theorem, this sequence a_n has a convergent subsequence $a_{n_j} \to a$. We want to prove that $a_n \to a$. Given $\varepsilon > 0$, there exists j_0 such that $\left|a_{n_j} - a\right| < \varepsilon$ for all $j \geq j_0$. Also, $\exists N(\varepsilon)$ such that $|a_m - a_n| < \varepsilon$ for all $m, n \geq N(\varepsilon)$. Combining these, we can take a j such that $n_j \geq \max\{N(\varepsilon), n_{j_0}\}$. Then, if $n \geq N(\varepsilon)$, using the triangle inequality,

$$|a_n - a| \leq \left|a_n - a_{n_j}\right| + \left|a_{n_j} - a\right| < 2\varepsilon$$

□

Therefore, on $\mathbb{R}$, a sequence is convergent if and only if it is a Cauchy sequence. This is sometimes referred to as the general principle of convergence, however this is a relatively old-fashioned name. This property is very useful, since we don't need to know what the limit actually is.

2. Series

2.1. Definition

Let a_n be a real or complex sequence. We say that $\sum_{j=1}^{\infty} a_j$ converges to s if the sequence of partial sums s_N converges to s as $N \to \infty$, i.e.

$$s_N = \sum_{j=1}^{N} a_j \to s$$

If the sequence of partial sums does not converge, then we say that the series diverges. Note that any problem on series can be turned into a problem on sequences, by considering their partial sums.

Lemma. (i) If $\sum_{j=1}^{\infty} a_j$ and $\sum_{j=1}^{\infty} b_j$ converge, then so does $\sum_{j=1}^{\infty}(\lambda a_j + \mu b_j)$, where $\lambda, \mu \in \mathbb{C}$.

(ii) Suppose $\exists N$ such that $a_j = b_j$ for all $j \geq N$. Then either $\sum_{j=1}^{\infty} a_j$ and $\sum_{j=1}^{\infty} b_j$ both converge, or they both diverge. In other words, the initial terms do not matter for considering convergence (but the sum will change).

Proof. (i) We have

$$\begin{aligned} s_N &= \sum_{j=1}^{N} (\lambda a_j + \mu b_j) \\ &= \sum_{j=1}^{N} \lambda a_j + \sum_{j=1}^{N} \mu b_j \\ &= \lambda c_N + \mu d_N \\ \therefore s_N &\to \lambda c + \mu d \end{aligned}$$

(ii) For any $n \geq N$, we have

$$\begin{aligned} s_N &= \sum_{j=1}^{n} a_j = \sum_{j=1}^{N-1} a_j + \sum_{j=n}^{N} a_j \\ d_N &= \sum_{j=1}^{n} b_j = \sum_{j=1}^{N-1} b_j + \sum_{j=n}^{N} b_j \end{aligned}$$

Taking the difference, we get

$$s_N - d_N = \sum_{j=1}^{N-1} a_j - \sum_{j=1}^{N-1} b_j$$

which is finite. So s_N converges if and only if d_N also converges.

□

2.2. Geometric series

Let $a_n = x^{n-1}$, where $n \geq 1$. Then

$$s_n = \sum_{j=1}^{n} a_j = 1 + x + x^2 + \cdots + x^{n-1}$$

Then

$$s_n = \begin{cases} \frac{1-x^n}{1-x} & \text{if } x \neq 1 \\ n & \text{if } x = 1 \end{cases}$$

This can be shown by observing that

$$xs_n = x + x^2 + \cdots + x^n = s_n - 1 + x^n \implies s_n(1-x) = 1 - x^n$$

If $|x| < 1$, then $x^n \to 0$ as $x \to \infty$. So $s_n \to \frac{1}{1-x}$. If $x > 1$, then $x^n \to \infty$ and so $s_n \to \infty$. If $x < -1$, s_n oscillates. For completeness, if $x = -1$, s_n oscillates between 0 and 1.

Note that the statement $s_n \to \infty$ means that given $a \in \mathbb{R}$, $\exists N$ such that $s_n > a$ for all $n \geq N$, and a similar statement holds for negative infinity (swapping the inequality). If s_n does not converge or tend to $\pm\infty$, we say that s_n oscillates.

Thus the geometric series converges if and only if $|x| < 1$. Note that to prove that $x^n \to 0$ if $|x| < 1$, we can consider the case $0 < x < 1$ and write $1/x = 1 + \delta$ for some positive δ. Then $x^n = \frac{1}{(1+\delta)^n} \leq \frac{1}{1+\delta n}$ from the binomial expansion, and this tends to zero as required.

Lemma. If $\sum_{j=1}^{\infty} a_j$ converges, then $\lim_{j\to\infty} a_j = 0$.

Proof. Given $s_n = \sum_{j=1}^{n} a_j$, we have $a_n = s_n - s_{n-1}$. If $s_n \to a$, then $a_n \to 0$ since s_{n-1} also tends to a. □

Remark. The converse is not true. For example, the harmonic series diverges, but the terms approach zero. Consider

$$\begin{aligned} s_{2n} &= s_n + \frac{1}{n+1} + \frac{1}{n+2} + \cdots + \frac{1}{2n} \\ &> s_n + \frac{1}{2n} + \frac{1}{2n} + \cdots + \frac{1}{2n} \\ &= s_n + \frac{1}{2} \end{aligned}$$

So as $n \to \infty$, if the sequence is convergent then the sequences s_n and s_{2n} tend to the same limit, but they clearly do not.

3. Convergence tests

3.1. Comparison test

In this section, we will let $a_n \in \mathbb{R}, a_n \geq 0$. In other words, all series contain only non-negative real terms.

Theorem. Suppose $0 \leq b_n \leq a_n$ for all n. If $\sum_{j=1}^{\infty} a_j$ converges, then $\sum_{j=1}^{\infty} b_j$ converges.

Proof. Let s_N be the Nth partial sum over the a_n, and let d_N be the Nth partial sum over the b_n. Since $b_n \leq a_n, d_N \leq s_N$. But $s_N \to s$, so $d_N \leq s_N \leq s$. So d_N is an increasing sequence that is bounded above by s, so it converges. □

For example, let us analyse the behaviour of the sum of the sequence $\frac{1}{n^2}$. Note that

$$\frac{1}{n^2} < \frac{1}{n(n-1)} = \frac{1}{n-1} - \frac{1}{n}$$

for $n \geq 2$. By the comparison test, it is sufficient to show that the series on the right hand side converges, in order to show that the original series converges.

$$\sum_{j=2}^{N} a_j = 1 - \frac{1}{N} \to 1$$

as required. So the original series tends to some value less than or equal to 2.

3.2. Cauchy's root test

Theorem. Suppose we have a sequence of non-negative terms a_n. Suppose that $a_n^{1/n} \to a$ as $n \to \infty$. Then if $a < 1$, the series $\sum a_n$ converges. If $a > 1$, the series $\sum a_n$ diverges.

Remark. Nothing can be said if $a = 1$. There is an example later of this fact.

Proof. If $a < 1$, let us choose an r such that $a < r < 1$. By the definition of the limit, $\exists N$ such that $\forall n \geq N$, $a_n^{1/n} < r$. This implies that $a_n < r^n$. The geometric series $\sum r^n$ converges. By comparison, the series a_n converges.

If $a > 1$, for all $n \geq N$, $a_n^{1/n} > 1$ which implies $a_n > 1$, thus $\sum a_n$ diverges, since a_n does not tend to zero. □

3.3. D'Alembert's ratio test

Theorem. Suppose $a_n > 0$, and $\frac{a_{n+1}}{a_n} \to \ell$. If $\ell < 1$, then the series $\sum a_n$ converges. If $\ell > 1$, then the series $\sum a_n$ diverges.

Remark. Like before, no conclusion can be drawn if $\ell = 1$.

Proof. Suppose $\ell < 1$. We can choose $\ell < r < 1$, $\exists N$ such that $\forall n \geq N$, $\frac{a_{n+1}}{a_n} < r$. Therefore $a_n < r^{n-N} a_N$. Hence, $a_n < k r^n$ where k is independent of n. Applying the comparison test, the series $\sum a_n$ must converge.

If $\ell > 1$, we can choose $\ell > r > 1$. Then $\exists N$ such that $\forall n \geq N$, $\frac{a_{n+1}}{a_n} > r$. As before, $a_n > r^{n-N} a_N$. But the r^{n-N} diverges, so the original series diverges. □

Example. Consider $\sum_1^\infty \frac{n}{2^n}$. We have

$$\frac{a_{n+1}}{a_n} = \frac{(n+1)/2^{n+1}}{n/2^n} \to \frac{1}{2}$$

So we have convergence, by the ratio test. Now, consider $\sum_1^\infty \frac{1}{n}$ and $\sum_1^\infty \frac{1}{n^2}$. In both cases, the ratio test gives limit 1. So the ratio test is inconclusive if the limit is 1. Since $n^{1/n} \to 1$, the root test is also inconclusive when the limit is 1. To check this limit, we can write

$$n^{1/n} = 1 + \delta_n; \quad \delta_n > 0$$

$$n = (1+\delta_n)^n > \frac{n(n-1)}{2}\delta_n^2$$

using the binomial expansion.

$$\implies \delta_n^2 < \frac{2}{n-1} \implies \delta_n \to 0$$

The root test is a good candidate for series that contain powers of n, for example

$$\sum_1^\infty \left[\frac{n+1}{3n+5}\right]^n$$

In this instance, for example, we have convergence.

3.4. Cauchy's condensation test

Theorem. Let a_n be a decreasing sequence of positive terms. Then $\sum_1^\infty a_n$ converges if and only if $\sum_1^\infty 2^n a_{2^n}$ converges.

Proof. First, note that if a_n is decreasing, then

$$a_{2^k} \underset{(*)}{\leq} a_{2^{k-1}+i} \underset{(\dagger)}{\leq} a_{2^{k-1}}; \quad 1 \leq i \leq 2^{k-1}; \quad k \geq 1$$

Now let us assume that $\sum a_n$ converges to $A \in \mathbb{R}$. Then, by $(*)$,

$$\begin{aligned} 2^{n-1} a_{2^n} &= a_{2^n} + a_{2^n} + \cdots + a_{2^n} \\ &\leq a_{2^{n-1}+1} + a_{2^{n-1}+2} + \cdots + a_{2^n} \\ &= \sum_{m=2^{n-1}+1}^{2^n} a_m \end{aligned}$$

Thus,

$$\sum_{n=1}^{N} 2^{n-1} a_{2^n} \leq \sum_{n=1}^{N} \sum_{m=2^{n-1}+1}^{2^n} a_m = \sum_{n=2}^{2^N} a_m$$

Therefore,

$$\sum_{n=1}^{N} 2^n a_{2^n} \leq 2 \sum_{n=2}^{2^N} a_m \leq 2(A - a_1)$$

Thus $\sum_{n=1}^{N} 2^n a_{2^n}$ converges, since it is increasing and bounded above. For the converse, we will assume that $\sum 2^n a_{2^n}$ converges to B. Using (†),

$$\begin{aligned} \sum_{m=2^{n-1}}^{2^n} a_m &= a_{2^{n-1}} + a_{2^{n-1}+1} + \cdots + a_{2^n} \\ &\leq a_{2^{n-1}} + a_{2^{n-1}} + \cdots + a_{2^{n-1}} \\ &= 2^{n-1} a_{2^{n-1}} \end{aligned}$$

So we have

$$\sum_{m=2}^{2^N} a_m = \sum_{n=1}^{N} \sum_{m=2^{n-1}+1}^{2^n} a_m \leq \sum_{n=1}^{N} 2^{n-1} a_{2^{n-1}} \leq \frac{1}{2} B$$

Therefore, $\sum_{m=1}^{N} a_m$ is a bounded, increasing sequence and hence converges. □

Let us consider an example of this test. Consider the series definition of the Riemann zeta function

$$\zeta(k) = \sum_{n=1}^{\infty} \frac{1}{n^k}$$

For what $k \in \mathbb{R}, k > 0$ does this series converge? This is equivalent to asking if the following series converges.

$$\sum_{n=1}^{\infty} 2^n \left[\frac{1}{2^n}\right]^k = \sum_{n=1}^{\infty} \left(2^{1-k}\right)^n$$

Hence it converges if and only if $2^{1-k} < 1 \iff k > 1$.

3.5. Alternating series

An alternating series is a series where the sign on each term switches between positive and negative.

Theorem (Alternating Series Test)**.** If a_n decreases and tends to zero as $u \to \infty$, then the alternating series

$$\sum_{1}^{\infty} (-1)^{n+1} a_n$$

converges.

Proof. Let us consider the partial sum

$$s_n = a_1 - a_2 + a_3 - a_4 + \cdots + (-1)^{n+1} a_n$$

In particular,

$$s_{2n} = (a_1 - a_2) + (a_3 - a_4) + \cdots + (a_{2n-1} - a_{2n})$$

Since the sequence is decreasing, each parenthesised block is positive. Then $s_{2n} \geq s_{2n-2}$. We can also write the partial sum as

$$s_{2n} = a_1 - (a_2 - a_3) - (a_4 - a_5) - \cdots - (a_{2n-2} - a_{2n-1}) - a_{2n}$$

Each parenthesised block here is negative. So $s_{2n} \leq a_1$. So s_{2n} is increasing and bounded above, so it must converge. Now, note that

$$s_{2n+1} = s_{2n} + a_{2n+1} \to s_{2n}$$

since $a_{2n+1} \to 0$. So s_{2n+1} also converges, in fact to the same limit. Hence s_n converges to this same limit. □

4. Absolute convergence

4.1. Absolute convergence

Definition. Let $a_n \in \mathbb{C}$. Then if $\sum_{n=1}^{\infty} |a_n|$ converges, then the series is called absolutely convergent.

Remark. Since $|a_n| \geq 0$, we can use the previous tests to check for absolute convergence.

Theorem. Let $a_n \in \mathbb{C}$. If this series is absolutely convergent, it is convergent.

Proof. Suppose first that a_n is a sequence of real numbers. Then let

$$v_n = \begin{cases} a_n & \text{if } a_n \geq 0 \\ 0 & \text{if } a_n < 0 \end{cases}; \quad w_n = \begin{cases} 0 & \text{if } a_n \geq 0 \\ -a_n & \text{if } a_n < 0 \end{cases}$$

Hence,

$$v_n = \frac{|a_n| + a_n}{2}; \quad w_n = \frac{|a_n| - a_n}{2}$$

Clearly, $v_n, w_n \geq 0$, and $a_n = v_n - w_n$, and $|a_n| = v_n + w_n$. If $\sum |a_n|$ converges, then by comparison $\sum v_n$ and $\sum w_n$ also converge, and hence $\sum a_n$ converges. Now, let us consider the case where a_n is complex. Then we can write $a_n = x_n + iy_n$ where x_n, y_n are real sequences. Note that $|x_n|, |y_n| \leq |a_n|$. So by comparison x_n and y_n converge, so a_n converges. □

Here are some examples.

(i) The alternating harmonic series $\sum \frac{(-1)^n}{n}$ is convergent, but not absolutely convergent.

(ii) $\sum \frac{z^n}{2^n}$ is absolutely convergent when $|z| < |2|$, because it reduces to a real geometric series. If $|z| \geq 2$, then $|a_n| \geq 1$, so we do not have absolute convergence.

4.2. Conditional convergence and rearrangement

If the series is convergent but not absolutely convergent, it is called *conditionally* convergent. The sum to which a series converges depends on the order in which the terms are added.

Definition. Let σ be a bijection of the positive integers to itself, then

$$a'_n = a_{\sigma(n)}$$

is a rearrangement of a_n.

Theorem. If $\sum_1^{\infty} a_n$ is absolutely convergent, then every rearrangement of this series converges to the same value.

Proof. First, let us consider the real case. Let $\sum a'_n$ be a rearrangement of $\sum a_n$. Let $s_n = \sum_1^n a_n$, and $t_n = \sum_1^n a'_n$. Let s_n converge to s. Suppose first that $a_n \geq 0$. Then given any $n \in \mathbb{N}$, we can find some $q \in \mathbb{N}$ such that s_q contains every term of t_n. Since the $a_n \geq 0$,

$$t_n \leq s_q \leq s$$

As $n \to \infty$, the t_n is an increasing sequence bounded above, so it must tend to a limit t, where $t \leq s$. Note, however, that this argument is symmetric; we can equally derive that $s \leq t$. Therefore $s = t$.

Now, let us drop the condition that $a_n \geq 0$. We can now consider v_n, w_n from above:

$$v_n = \frac{|a_n| + a_n}{2}; \quad w_n = \frac{|a_n| - a_n}{2}$$

Since $\sum |a_n|$ converges, both $\sum v_n, \sum w_n$ converge. Since all $v_n, w_n \geq 0$, we can deduce that $\sum v_n = \sum v'_n$ and $\sum w'_n = \sum w_n$. The claim follows since $a_n = v_n - w_n$.

For the case $a_n \in \mathbb{C}$, we can write $a_n = x_n + iy_n$, noting that $|x_n|, |y_n| \geq |a_n|$. By comparison, the series $\sum x_n, \sum y_n$ are absolutely convergent, and by the previous case, $\sum x_n = \sum x'_n$ and $\sum y'_n = \sum y'_n$. Since $a'_n = x'_n + y'_n$, $\sum a_n = \sum a'_n$ as required. □

5. Continuity

5.1. Definitions

Let $E \subseteq \mathbb{C}$ be a non-empty set, and $f : E \to \mathbb{C}$ be any function, and let $a \in E$. Certainly, this includes the case in which f is a real-valued function and $E \subseteq \mathbb{R}$.

Definition. f is continuous at a if for every sequence $z_n \in E$ that converges to a, we have $f(z_n) \to f(a)$.

We can use an alternative definition:

Definition (ε-δ definition)**.** f is continuous at a if given $\varepsilon > 0$, $\exists \delta > 0$ such that for every $z \in E$, if $|z - a| < \delta$, then $|f(z) - f(a)| < \varepsilon$.

We will immediately prove that both definitions are equivalent. First, let us prove that the ε-δ definition implies the first definition.

Proof. We know that given $\varepsilon > 0, \exists \delta > 0$ such that for all $z \in E$, $|z - a| < \delta$ implies $|f(z) - f(a)| < \varepsilon$. Let $z_n \to a$, then by the definition of the limit of the sequence then there exists n_0 such that for all $n \geq n_0$ we have $|z_n - a| < \delta$. But this implies that $|f(z_n) - f(a)| < \varepsilon$, i.e. $f(z_n) \to f(a)$. □

We now prove the converse, that the first definition implies the second.

Proof. We know that for every sequence $z_n \in E$ that converges to a, $f(z_n) \to f(a)$. Suppose f is not continuous at a, according to the ε-δ definition. Then there exists some ε such that for all $\delta > 0$, there exists $z \in E$ such that $|z - a| < \delta$ but $|f(z) - f(a)| \geq \varepsilon$. So, let us construct a sequence of δ values to substitute into this definition. Let $\delta = 1/n$. Then the z_n given by this δ is such that $|z_n - a| < 1/n$ and $|f(z_n) - f(a)| \geq \varepsilon$. Clearly, $z_n \to a$, but $f(z_n)$ does not tend to $f(a)$ because the difference between the two is always greater than ε. This is a contradiction, since we assumed that f is continuous by the first definition. So f is continuous by the ε-δ definition. □

5.2. Making continuous functions

We can create new continuous functions from old ones by manipulating them in a number of ways.

Proposition. Let $g, f : E \to \mathbb{C}$ be continuous functions at a point $a \in E$. Then all of the functions

- $f(z) + g(z)$
- $f(z)g(z)$
- $\lambda f(z)$ for some constant λ

are all continuous. In addition, if $f(z) \neq 0$ everywhere in E, then $\frac{1}{f}$ is a continuous function at a.

Proof. Using the first definition, this is obvious using the fact that limits of sequences behave analogously. □

Trivially, the function $f(z) = z$ is continuous. From this, we can derive that every polynomial is continuous at every point in $\mathbb{C}$. Note that we say that f is continuous on the entire set E if it is continuous at every point $a \in E$.

5.3. Composition of continuous functions

Theorem. Let $f : A \to \mathbb{C}$ and $g : B \to \mathbb{C}$ where $A, B \subseteq \mathbb{C}$ be two functions that can be composed, i.e. $f(A) \subseteq B$. If f is continuous at $a \in A$ and g is continuous at $f(a) \in B$, then $g \circ f : A \to \mathbb{C}$ is continuous at a.

Proof. Take any sequence $z_n \to a$. By assumption, $f(z_n) \to f(a)$. Now, let us define a new sequence $w_n = f(z_n)$. Then $w_n \in B$ and $w_n \to f(a)$. Thus, $g(f(z_n)) = g(w_n) \to g(f(a))$ by continuity, as required. □

Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} \sin\left(\frac{1}{x}\right) & x \neq 0 \\ 0 & x = 0 \end{cases}$$

This is assuming the knowledge of $\sin(x)$ being a continuous function $\mathbb{R} \to \mathbb{R}$, which we will prove later. So $f(x)$ is certainly continuous at every point on $\mathbb{R}$ excluding 0, since it is the composition of two continuous functions. We can prove it is discontinuous at $x = 0$ by providing a sequence, for example

$$\frac{1}{x_n} = \left(2n + \frac{1}{2}\right)\pi$$

Then $x_n \to 0$, and $f(x_n) = 1$. But $f(0) \neq 1$, so it is discontinuous. Let us modify the example as follows.

$$f(x) = \begin{cases} x \sin\left(\frac{1}{x}\right) & x \neq 0 \\ 0 & x = 0 \end{cases}$$

We can prove that this sequence is continuous at 0. For an arbitrary sequence $x_n \to 0$, then $|f(x_n)| \leq |x_n|$ because $|\sin x| \leq 1$. So $f(x_n)$ is bounded by x_n, which tends to zero, so $f(x_n)$ tends to zero as required. Now for a final example, let

$$f(x) = \begin{cases} 1 & x \in \mathbb{Q} \\ 0 & x \notin \mathbb{Q} \end{cases}$$

This is discontinuous at every point. If $x \in \mathbb{Q}$, take a sequence $x_n \to x$ with all x_n irrational, then $f(x_n) = 0$ but $f(x) = 1$. Similarly, if $x \notin \mathbb{Q}$, take a sequence $x_n \to x$ with all x_n rational, then $f(x_n) = 1$ but $f(x) = 0$.

6. Limit of a function

6.1. Definition

Let $f : E \subseteq \mathbb{C} \to \mathbb{C}$. We would like to define what is meant by $\lim_{z\to a} f(z)$, even when $a \notin E$. Further, if we have a set with an isolated point, for example $E = \{0\} \cup [1, 2]$, it does not make sense to talk about limits tending to 0 since there are no points in E close to 0.

Definition. Let $E \subseteq \mathbb{C}$, $a \in \mathbb{C}$. a is a limit point of E if for any $\delta > 0$, there exists $z \in E$ such that $0 < |z - a| < \delta$.

First, note that a is a limit point if and only if there exists a sequence $z_n \in E$ such that $z_n \to a$, but notably $z_n \neq a$ for all n.

Definition. Let $f : E \subseteq \mathbb{C} \to \mathbb{C}$, and let $a \in \mathbb{C}$ be a limit point of E. We say that $f \to \ell$ as $z \to a$, if given $\varepsilon > 0$ there exists $\delta > 0$ such that whenever $0 < |z - a| < \delta$ and $z \in E$, $|f(z) - \ell| < \varepsilon$. Equivalently, $f(z_n) \to \ell$ for every sequence $z_n \in E$, such that $z_n \to a$ but $z_n \neq a$.

Therefore if $a \in E$ is a limit point, then $\lim_{z\to a} f(z) = f(a)$ if and only if f is continuous at a. If $a \in E$ is isolated (not a limit point) then f at a is trivially continuous, since there are no points near a but a itself.

6.2. Properties

The limit of a function has very similar properties when compared to the limit of a sequence.

(i) It is unique. $f(z) \to A$, $f(z) \to B$ implies $A = B$.

(ii) $f(z) \to A$, $g(z) \to B$ implies

  (a) $f(z) + g(z) \to A + B$

  (b) $f(z) \cdot g(z) \to AB$

  (c) If $B \neq 0$, $\frac{f(z)}{g(z)} \to \frac{A}{B}$

6.3. Intermediate value theorem

Theorem. Let $f : [a, b] \to \mathbb{R}$ be a continuous function where $f(a) \neq f(b)$. Then f takes all values in the interval $[f(a), f(b)]$.

Proof. Without loss of generality, let us assume $f(a) < f(b)$. Let us take an η such that $f(a) < \eta < f(b)$. We want to prove that there exists some value $c \in [a, b]$ with $f(c) = \eta$. Let s be the set of points defined by

$$s = \{x \in [a, b] : f(x) < \eta\}$$

$a \in s$ therefore the set s is non-empty. The set is also clearly bounded above by b. So there is a supremum of this set, say $\sup s = c$ where $c \leq b$. This point c can be visualised as the last point at which $y = f(x)$ crosses the line $y = c$. We intend to show that the function at this rightmost point is η.

By the definition of the supremum, given n there exists $x_n \in s$ such that $c - \frac{1}{n} < x_n \leq c$. So the sequence x_n tends to c. We know that $f(x_n) < \eta$ for all x_n by definition of the set s. By the continuity of f, $f(x_n) \to f(c)$. Thus,

$$f(c) \leq \eta \tag{$*$}$$

Now, let us consider the fact that $c \neq b$. If $c = b$, then $f(b) \leq \eta$ which is a contradiction since $\eta < f(b)$. So for a large n, we can ensure that $c + \frac{1}{n} \in [a, b]$. So by continuity of the function, $f(c + \frac{1}{n}) \to f(c)$. But since $c + \frac{1}{n} > c$, then necessarily $f(c + \frac{1}{n}) \geq \eta$ because c is the supremum of s. Thus

$$f(c) \geq \eta$$

Combining this with $(*)$ we get $f(c) = \eta$. □

This theorem is very useful for finding zeroes and fixed points. For example, we can prove the existence of the Nth root of a positive real number y. Let

$$f(x) = x^N$$

Then f is certainly continuous on the interval $[0, 1 + y]$, since

$$0 = f(0) < y < (1 + y)^N = f(1 + y)$$

By the intermediate value theorem, there exists a point $c \in (0, 1+y)$ such that $f(c) = c^N = y$. So c is a positive Nth root of y. We can also prove the uniqueness of such a point. Suppose $d^N = y$ with $d > 0$ and $d \neq c$. Without loss of generality, suppose $d < c$. Then $d^N < c^N$ so $d^N \neq y$, which is a contradiction.

6.4. Bounds of a continuous function

Theorem. Let $f : [a, b] \to \mathbb{R}$ be continuous. Then the function is bounded, i.e. there exists $k \in \mathbb{R}$ such that $|f(x)| \leq k$ for every point $x \in [a, b]$.

Proof. Suppose that such a function f is not bounded. Then in particular, given any integer $n \geq 1$, there exists $x_n \in [a, b]$ such that $|f(x_n)| > n$. By the Bolzano–Weierstrass theorem, the sequence x_n, which is bounded by $a \leq x_n \leq b$, has a convergent subsequence $x_{n_j} \to x$, such that $x \in [a, b]$. Then by continuity of f, $f(x_{n_j}) \to f(x)$. But $\left|f(x_{n_j})\right| > n_j \to \infty$. This is a contradiction. □

We can actually improve this statement.

Theorem. Suppose $f : [a,b] \to \mathbb{R}$ is a continuous function. Then there exist $x_1, x_2 \in [a,b]$ such that

$$f(x_1) \leq f(x) \leq f(x_2)$$

for all $x \in [a,b]$. In other words, a continuous function on a closed bounded interval is bounded and attains its bounds.

Proof. Let $A = \{f(x) : x \in [a,b]\}$ be the image of $[a,b]$ under f. By the above theorem, A is bounded. It is also non-empty, hence it has a supremum $M = \sup A$ (and analogously an infimum $\inf A$, whose proof is almost identical). Then by the definition of the supremum, given an integer $n \geq 1$ there exists $x_n \in [a,b]$ such that $M - \frac{1}{n} < f(x_n) \leq M$. By the Bolzano–Weierstrass theorem, there exists a convergent subsequence $x_{n_j} \to x \in [a,b]$. Since $f(x_{n_j}) \to M$, then by continuity, $f(x) = M$. □

Here is an alternative proof of the same theorem.

Proof. As before, let A be the image of f, and M be the supremum of A. Suppose there is no $x_2 \in [a,b]$ such that $f(x_2) = M$. Then let $g(x) = \frac{1}{M - f(x)}$ for $x \in [a,b]$. Since there exists no x such that $M = f(x)$, $g(x)$ is continuous since we are never dividing by zero. So g is bounded. So by the previous theorem, there is some $k > 0$ such that $g(x) \leq k$ for all $x \in [a,b]$. This means that $f(x) \leq M - \frac{1}{k}$ on $[a,b]$ for this k, but this cannot happen since M is the supremum. □

Note that these theorems are certainly false if the interval is not closed: consider the counter-example $(0,1]$ and the function $x \mapsto x^{-1}$.

6.5. Inverse functions

Definition. f is increasing for $x \in [a,b]$ if $f(x_1) \leq f(x_2)$ for all $x_1 \leq x_2 \in [a,b]$. If $f(x_1) < f(x_2)$ then the function is strictly increasing. A function may be called decreasing or strictly decreasing analogously.

Definition. A function f is called monotone if it is either increasing or decreasing.

Theorem. Let $f : [a,b] \to \mathbb{R}$ be continuous and strictly increasing for $x \in [a,b]$. Let $c = f(a)$, $d = f(b)$. Then $f : [a,b] \to [c,d]$ is bijective, and the inverse $g := f^{-1} : [c,d] \to [a,b]$ is continuous and strictly increasing.

A similar theorem holds for strictly decreasing functions.

Proof. Let $c < k < d$. From the intermediate value theorem, there exists h such that $f(h) = k$. This h must be unique since the function is strictly increasing. Then we can define $g(k) = h$, giving us an inverse $g : [c,d] \to [a,b]$ for f.

First, note that g is strictly increasing. Indeed, for $y_1 < y_2$ then $y_1 = f(x_1), y_2 = f(x_2)$. This means that if $x_2 \geq x_1$, then since f is increasing $y_2 \leq y_1$ which is a contradiction.

Now, note that g is continuous. Indeed, given $\varepsilon > 0$, we can let $k_1 = f(h - \varepsilon)$ and $k_2 = f(h + \varepsilon)$. If f is strictly increasing, then $k_1 < k < k_2$. Then $h - \varepsilon < g(y) < h + \varepsilon$. So let $\delta = \min(k_2 - k, k - k_1)$ where $k \in (c, d)$, establishing continuity as claimed. □

7. Differentiability

7.1. Definitions

Let $f : E \subseteq \mathbb{C} \to \mathbb{C}$. Mostly we will take E to be an interval in the real numbers, or a disc in the complex plane.

Definition. Let $x \in E$ be a point such that there exists a sequence $x_n \in E$ with $x_n \neq x$, but $x_n \to x$, i.e. x is a limit point. f is said to be differentiable at x with derivative $f'(x)$ if

$$\lim_{y \to x} \frac{f(y) - f(x)}{y - x} = f'(x)$$

If f is differentiable at each point in E, we say that f is differentiable on E.

Remark. One interpretation of the definition is to write it in the form

$$\varepsilon(h) \coloneqq f(x+h) - f(x) - hf'(x); \quad \lim_{h \to 0} \frac{\varepsilon(h)}{h} = 0$$

so ε is $o(h)$. Hence,

$$f(x+h) = f(x) + hf'(x) + \varepsilon(h)$$

We could have made an alternative definition for differentiability. f is differentiable at x if there exists A and ε such that

$$f(x+h) = f(x) + hA + \varepsilon(h) \text{ where } \lim_{h \to 0} \frac{\varepsilon(h)}{h} = 0$$

If such an A exists, then it is unique, since A is the limit

$$A = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

We could have alternatively written the definition as

$$f(x+h) = f(x) + hf'(x) + h\varepsilon_f(h) \text{ where } \lim_{h \to 0} \varepsilon_f(h) = 0$$

or perhaps

$$f(x) = f(a) + (x-a)f'(a) + (x-a)\varepsilon_f(x) \text{ where } \lim_{x \to a} \varepsilon_f(x) = 0$$

Note further that if f is differentiable at x, f is certainly continuous at x. This follows from the fact that $\varepsilon(h) \to 0$, and hence $f(x+h) \to f(x)$ as $h \to 0$.

As an example, let us consider $f(x) = |x|$ for $f : \mathbb{R} \to \mathbb{R}$. Is the function at the point $x = 0$ differentiable? If $x > 0$, we have $f'(x) = 1$, but if $x < 0$, we have $f'(x) = -1$. These results can be checked directly using the definitions above. But we have produced two sequences for $h \to 0$ which give different values, so the derivative is not defined here.

7.2. Differentiation of sums and products

Proposition. (i) If $f(x) = c$ for all $x \in E$, then f is differentiable with $f'(x) = 0$.

(ii) If f and g are differentiable at x, then so is $f + g$, where $(f + g)'(x) = f'(x) + g'(x)$.

(iii) If f and g are differentiable at x, then so is fg, where $(fg)'(x) = f'(x)g(x) + g'(x)f(x)$.

(iv) If f is differentiable at x and $f(x) \neq 0$, then so is $\frac{1}{f}$, where $(\frac{1}{f})'(x) = \frac{-f'(x)}{(f(x))^2}$.

Proof. (i) $\lim_{h\to 0} \frac{c-c}{h} = 0$ as required.

(ii) Since all relevant limits are well-defined,

$$\lim_{h\to 0} \frac{f(x+h) + g(x+h) - f(x) - g(x)}{h} = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} + \lim_{h\to 0} \frac{g(x+h) - g(x)}{h} = f'(x) + g'(x$$

(iii) Let $\phi(x) = f(x)g(x)$. Then, since f is continuous at x,

$$\begin{aligned}
\lim_{h\to 0} \frac{\phi(x+h) - \phi(x)}{h} &= \lim_{h\to 0} \frac{f(x+h)g(x+h) - f(x)g(x)}{h} \\
&= \lim_{h\to 0} f(x+h)\frac{g(x+h) - g(x)}{h} + g(x)\frac{f(x+h) - f(x)}{h} \\
&= \lim_{h\to 0} f(x)\frac{g(x+h) - g(x)}{h} + g(x)\frac{f(x+h) - f(x)}{h} \\
&= f(x)g'(x) + g(x)f'(x)
\end{aligned}$$

(iv) Let $\phi(x) = \frac{1}{f(x)}$. Then,

$$\begin{aligned}
\lim_{h\to 0} \frac{\phi(x+h) - \phi(x)}{h} &= \lim_{h\to 0} \frac{\frac{1}{f(x+h)} - \frac{1}{f(x)}}{h} \\
&= \lim_{h\to 0} \frac{f(x) - f(x+h)}{hf(x)f(x+h)} \\
&= \frac{-f'(x)}{f(x)f(x)}
\end{aligned}$$

□

Remark. From (iii) and (iv), we can immediately find the quotient rule,

$$\left(\frac{f(x)}{g(x)}\right)' = \frac{g(x)f'(x) - f(x)g'(x)}{(g(x))^2}$$

7.3. Differentiating polynomial terms

As an example of the differentiability properties we saw last lecture, we can find the derivative of $f(x) = x^n$ for $n \in \mathbb{Z}$, $n > 0$. If $n = 1$, clearly $f'(x) = 1$. We can show inductively that $f'(x) = nx^{n-1}$. Indeed,

$$\begin{aligned}(x^n)' &= x \cdot (x^{n-1})' + (x)' \cdot x^{n-1} \\ &= (n-1)x^{n-1} + x^{n-1} \\ &= nx^{n-1}\end{aligned}$$

We can now take $f(x) = x^{-n}$. Using the reciprocal law,

$$\begin{aligned}f'(x) &= \frac{-(x^n)'}{(x^n)^2} \\ &= \frac{-nx^{n-1}}{x^{2n}} \\ &= -nx^{-n-1}\end{aligned}$$

7.4. Chain rule

Theorem. Let $f : U \to \mathbb{C}$ be such that $f(x) \in V$ for all $x \in U$. If f is differentiable at $a \in U$, and $g : V \to \mathbb{C}$ is differentiable at $f(a) \in V$, then $g \circ f$ is differentiable at a with

$$gf'(a) = f'(a)g'(f(a))$$

Proof. We know that we can write

$$f(x) = f(a) + (x-a)f'(a) + \varepsilon_f(x)(x-a)$$

where $\lim_{x\to a} \varepsilon_f(x) = 0$. Further,

$$g(y) = g(b) + (y-b)g'(b) + \varepsilon_g(y)(y-b)$$

where $\lim_{y\to b} \varepsilon_g(y) = 0$, and $b = f(a)$. We will set $\varepsilon_f(a) = 0$ and $\varepsilon_g(b) = 0$, so they are continuous at $x = a$ and $y = b$, so that everything is well-defined when we begin to compose the functions. Now, $y = f(x)$, so

$$\begin{aligned}g(f(x)) &= g(b) + (f(x)-b)g'(b) + \varepsilon_g(f(x))(f(x)-b) \\ &= g(f(a)) + \big[(x-a)f'(a) + \varepsilon_f(x)(x-a)\big]\big[g'(b) + \varepsilon_g(f(x))\big] \\ &= g(f(a)) + (x-a)f'(a)g'(b) + (x-a)\underbrace{\big[\varepsilon_f(x)g'(b) + \varepsilon_g(f(x))\big(f'(a) + \varepsilon_f(x)\big)\big]}_{\sigma(x)}\end{aligned}$$

Now, we just need to show that $\lim_{x\to a} \sigma(x) = 0$ in order to prove the theorem. Clearly

$$\sigma(x) = \underbrace{\varepsilon_f(x)}_{\to 0} g'(b) + \underbrace{\varepsilon_g(f(x))}_{\to 0}\big(f'(a) + \varepsilon_f(x)\big)$$

Hence $\sigma(x) \to 0$ as required. □

7.5. Rolle's theorem

Theorem. Let $f : [a, b] \to \mathbb{R}$ be a continuous function on $[a, b]$ and differentiable on (a, b). If $f(a) = f(b)$, then there exists $c \in (a, b)$ such that $f'(c) = 0$.

Proof. Let M be the maximum point and m be the minimum point of the function. Recall that in Lecture 8 we proved that any function achieves its bounds. Let $k = f(a)$. If $M = m = k$, then f must be a constant, and clearly $f'(c) = 0$ for every value $c \in (a, b)$. Otherwise, either $M > k$ or $m < k$. Suppose $M > k$ (the proof is very similar if $m < k$). Then there exists some value $c \in (a, b)$ such that $f(c) = M$. We would like to show that $f'(c) = 0$, so let us suppose that $f'(c) \neq 0$. If $f'(c) > 0$, then there are values $d > c$ where $f(d) > f(c)$. Indeed,

$$f(h + c) - f(c) = h\,[f'(c) + \varepsilon(h)]$$

For a small, positive h, this value is positive. This contradicts the fact that M is the maximum. Similarly, if $f'(c) < 0$ there are values $d < c$ with $f(d) > f(c)$. Hence $f'(c) = 0$. □

7.6. Mean value theorem

We can make a small change to Rolle's theorem and obtain the mean value theorem.

Theorem. Let $f : [a, b] \to \mathbb{R}$ be a continuous function on $[a, b]$ and differentiable on (a, b). Then there exists $c \in (a, b)$ such that

$$f(b) - f(a) = f'(c)(b - a)$$

Proof. Let ϕ be a function defined by $\phi(x) = f(x) - kx$, choosing a k such that $\phi(a) = \phi(b)$. We can find that

$$f(b) - bk = f(a) - ak \implies k = \frac{f(b) - f(a)}{b - a}$$

By Rolle's theorem, there exists $c \in (a, b)$ such that $\phi'(c) = 0$. Now, note that $f'(x) = \phi'(x) + k$, hence there exists c such that $f'(c) = k$. □

Remark. We will often rewrite the mean value theorem as follows.

$$f(a + h) = f(a) + hf'(a + \theta h)$$

where $\theta \in (0, 1)$. Note, however, that θ is a function of h, so if we begin to shrink h then θ may change.

7.7. Properties of a function from its derivative

We can deduce certain facts about a function by observing the properties its derivative exhibits. These results are mostly trivial corollaries to the mean value theorem, proven in the last lecture.

Corollary. Let $f : [a, b] \to \mathbb{R}$ be continuous, and differentiable on (a, b). Then we have

(i) If $f'(x) > 0$ for all $x \in (a, b)$, then f is strictly increasing on $[a, b]$;

(ii) If $f'(x) \geq 0$ for all $x \in (a, b)$, then f is increasing on $[a, b]$;

(iii) If $f'(x) = 0$ for all $x \in (a, b)$, then f is constant on $[a, b]$.

Part (iii) of this corollary is essentially solving the most simple differential equation; we are showing that the only possible solutions to this equation are the constant functions. Note that similar statements about decreasing functions hold.

Proof. (i) We have $f(y) - f(x) = f'(c)(y - x)$ for some $c \in (x, y)$. If $f'(c) > 0$, then $f(y) - f(x) > 0$.

(ii) Analogously to before, $f(y) - f(x) = f'(c)(y - x)$ for some $c \in (x, y)$. If $f'(c) \geq 0$, then $f(y) - f(x) \geq 0$.

(iii) By the mean value theorem on $[a, x]$, if $f'(c) = 0$, then $f(x) - f(a) = 0$.

□

7.8. Inverse function theorem

Theorem. Let $f : [a, b] \to \mathbb{R}$ be a continuous function on $[a, b]$ and differentiable on (a, b), with $f'(x) > 0$ everywhere on (a, b). Let $f(a) = c, f(b) = d$. Then the function $f : [a, b] \to [c, d]$ is bijective, and $f^{-1} : [c, d] \to [a, b]$ is differentiable on (c, d) with

$$\left(f^{-1}\right)'(x) = \frac{1}{f'\left(f^{-1}(x)\right)}$$

Note, in lecture 8 it was proven that a continuous strictly increasing function has a continuous inverse. This strengthens that claim to include the differentiability property if the original function was differentiable.

Proof. We know from lecture 8 that there exists $g : [c, d] \to [a, b]$ which is a strictly increasing continuous function, which is the inverse of f. We must now show that g is differentiable and that its derivative has the required form as stated in the claim. Now, let $y = f(x)$. Given $k \neq 0$, let h be given by

$$y + k = f(x + h)$$

Alternatively, written in terms of g,

$$x + h = g(y + k)$$

So clearly $h \neq 0$. Since g is continuous, if $k \to 0$ then $h \to 0$. Then

$$\begin{aligned} \frac{g(y+k)-g(y)}{k} &= \frac{x+h-x}{f(x+h)-y} \\ &= \frac{h}{f(x+h)-f(x)} \\ \therefore \lim_{k\to 0} \frac{g(y+k)-g(y)}{k} &= \lim_{h\to 0} \frac{h}{f(x+h)-f(x)} \\ &= \frac{1}{f'(x)} \end{aligned}$$

as required. □

7.9. Derivative of rational powers

First, let $g(x) = x^{1/q}$ for some positive integer q. We can find that $f(x) = x^q$ has the derivative $f'(x) = qx^{q-1}$. By the inverse function theorem, $g'(x) = \frac{1}{q}x^{1/q-1}$. Now, if $g(x) = x^{p/q}$, where p is an integer and q is a positive integer, then by the chain rule $g'(x) = \frac{p}{q}x^{p/q-1}$ which matches the expected result.

7.10. Mean value theorem applied to limits

Suppose $f, g : [a, b] \to \mathbb{R}$ are continuous, and differentiable on (a, b). Suppose further that $g(a) \neq g(b)$. The mean value theorem can be applied to both functions, and will give two points $s, t \in (a, b)$ such that

$$\frac{f(b)-f(a)}{g(b)-g(a)} = \frac{(b-a)f'(s)}{(b-a)g'(t)} = \frac{f'(s)}{g'(t)}$$

This gives us a way to simplify a limit of the form of the left hand side (as $b \to a$) by instead considering the right hand side. We can apply Cauchy's mean value theorem, seen in the next lecture.

7.11. Cauchy's mean value theorem

Theorem. If $f, g : [a, b] \to \mathbb{R}$ are continuous, and differentiable on (a, b), there exists $t \in (a, b)$ such that

$$(f(b)-f(a))g'(t) = f'(t)(g(b)-g(a))$$

We can recover the normal mean value theorem from Cauchy's generalisation by taking $g(x) = x$.

Proof. Let

$$\phi(x) = \begin{vmatrix} 1 & 1 & 1 \\ f(a) & f(x) & f(b) \\ g(a) & g(x) & g(b) \end{vmatrix}$$

Certainly $\phi(x)$ is continuous on $[a, b]$ and differentiable on (a, b), by using previous results. Also, $\phi(a) = \phi(b) = 0$ by observing the linear dependence of the columns. By Rolle's theorem, there exists $t \in (a, b)$ such that $\phi'(t) = 0$. We can expand $\phi'(t)$ and this will show the required result.

$$\phi'(x) = f'(x)g(b) - g'(x)f(b) + f(a)g'(x) - g(a)f'(x) = f'(x)[g(b) - g(a)] + g'(x)[f(a) - f(b)]$$

□

Example (l'Hôpital's rule)**.** The derivation of l'Hôpital's rule is on an example sheet, so here we will consider only a special case of it, using Cauchy's mean value theorem.

$$\ell = \lim_{x \to 0} \frac{e^x - 1}{\sin x}$$

We can write

$$\ell = \lim_{x \to 0} \frac{e^x - e^0}{\sin x - \sin 0} = \frac{e^t}{\cos t}$$

for some $t \in (0, x)$. So as $x \to 0$, $t \to 0$ and hence

$$\frac{e^t}{\cos t} \to 1$$

8. Taylor's theorem

8.1. Lagrange's and Cauchy's remainders

Theorem (Taylor's Theorem with Lagrange's Remainder)**.** Suppose f and its derivatives up to order $n-1$ are continuous in $[a, a+h]$, and $f^{(n)}$ exists for $x \in (a, a+h)$. Then

$$f(a+h) = f(a) + hf'(a) + \frac{h^2}{2!}f''(a) + \cdots + \frac{h^{n-1}}{(n-1)!}f^{(n-1)}(a) + \frac{h^n}{n!}f^{(n)}(a+\theta h)$$

where $\theta \in (0,1)$.

Note that for $n = 1$, this is exactly the mean value theorem, so this can be seen as an nth order extension of the mean value theorem. We commonly write R_n for the final error term $\frac{h^n}{n!}f^{(n)}(a+\theta h)$. This is known as Lagrange's form of the remainder.

Proof. For $0 \le t \le h$, we define

$$\phi(t) = f(a+t) - f(a) - tf'(a) - \cdots - \frac{t^{n-1}}{(n-1)!}f^{(n-1)}(a) - \frac{t^n}{n!}B$$

where we choose B suitably such that $\phi(h) = 0$. (Recall that in the proof of the mean value theorem, we used $f(x) - kx$ and picked k suitably such that this allowed the use of Rolle's theorem. This is entirely analogous, but generalised to the nth derivative). Note that

$$\phi(0) = \phi'(0) = \cdots = \phi^{(n-1)}(0) = 0$$

We can use Rolle's theorem inductively n times. Since $\phi(0) = \phi(h) = 0$, there is a point $0 < h_1 < h$ such that $\phi'(h_1) = 0$. Since $\phi'(0) = \phi'(h_1) = 0$, there is a point $0 < h_2 < h_1$ such that $\phi''(h_2) = 0$. This continues until we find a point $0 < h_n < h$ such that $\phi^{(n)}(h_n) = 0$. Hence $h_n = \theta h$ for some $0 < \theta < 1$. Now, $\phi^{(n)}(t) = f^{(n)}(a+t) - B$. We can see now that $B = f^{(n)}(a+\theta h)$, which gives the required result. □

We can prove an alternative version of Taylor's theorem with a different error term.

Theorem (Taylor's Theorem with Cauchy's Remainder)**.** Suppose (equivalently to before) f and its derivatives up to order $n-1$ are continuous in $[a, a+h]$, and $f^{(n)}$ exists for $x \in (a, a+h)$. Then

$$f(a+h) = f(a) + hf'(a) + \frac{h^2}{2!}f''(a) + \cdots + \frac{h^{n-1}}{(n-1)!}f^{(n-1)}(a) + R_n$$

where

$$R_n = \frac{(1-\theta)^{n-1}h^n f^{(n)}(a+\theta h)}{(n-1)!}$$

for $\theta \in (0,1)$.

Proof. For simplicity, in this proof we let $a = 0$, although the same argument applies when $a \neq 0$. Let us define

$$F(t) = f(h) - f(t) - (h-t)f'(t) - \cdots - \frac{(h-t)^{n-1} f^{(n-1)}(t)}{(n-1)!}$$

for $t \in [0, h]$. Then

$$\begin{aligned} F'(t) &= -f'(t) + f'(t) - (h-t)f''(t) + (h-t)f''(t) - \frac{1}{2}(h-t)^2 f'''(t) + \frac{1}{2}(h-t)^2 f'''(t) \\ &\quad - \cdots - \frac{(h-t)^{n-1}}{(n-1)!} f^{(n)}(t) \\ &= -\frac{(h-t)^{n-1}}{(n-1)!} f^{(n)}(t) \end{aligned}$$

Let

$$\phi(t) = F(t) - \left[\frac{h-t}{h}\right]^p F(0)$$

where $p \in \mathbb{N}$ and $1 \leq p \leq n$. Then

$$\phi(0) = \phi(h) = 0$$

By Rolle's theorem, there exists $\theta \in (0, 1)$ such that

$$\phi'(\theta h) = 0$$

We can compute ϕ' to find

$$\phi'(\theta h) = F'(\theta h) + \frac{p(1-\theta)^{p-1}}{h} F(0) = 0$$

Substituting everything back into F gives

$$0 = \frac{-h^{n-1}(1-\theta)^{n-1}}{(n-1)!} f^{(n)}(\theta h) + \frac{p(1-\theta)^{p-1}}{h} \left[f(h) - f(0) - h'(0) - \cdots - \frac{h^{n-1}}{(n-1)!} f^{(n-1)}(0) \right]$$

Hence

$$f(h) = f(0) + hf'(0) + \frac{h^2}{2!} f''(0) + \cdots + \frac{h^{n-1}}{(n-1)!} f^{(n-1)}(0) + \underbrace{\frac{h^n (1-\theta)^{n-1} f^{(n)}(\theta h)}{(n-1)! \cdot p(1-\theta)^{p-1}}}_{R_n}$$

By letting $p = n$, we get Lagrange's remainder. If $p = 1$, we get Cauchy's remainder. □

8.2. Bounding error terms

Recall that Lagrange's remainder is

$$R_n = \frac{h^n}{n!} f^{(n)}(a + \theta h)$$

and Cauchy's remainder is

$$R_n = \frac{(1-\theta)^{n-1} h^n f^{(n)}(a + \theta h)}{(n-1)!}$$

and that we can write

$$f(h) = P_{n-1}(h) + R_n$$

where P_{n-1} is the Taylor polynomial to $(n-1)$th order. To get a Taylor series for a function f, we need to prove that the R_n tend to zero as $n \to \infty$. In general, this requires estimates for the R_n and it could take a lot of effort to prove whether this limit is zero or not. Note also that the theorems deducing the remainder terms work equally well in an interval $[a+h, a]$ where $h < 0$.

8.3. Binomial series

Proposition. Let

$$f(x) = (1+x)^r$$

for some $r \in \mathbb{Q}$. If $|x| < 1$, then

$$f(x) = 1 + \binom{r}{1} x + \cdots + \binom{r}{n} x^n + \ldots$$

where

$$\binom{r}{n} = \frac{r(r-1)\cdots(r-n+1)}{n!}$$

Proof. Clearly,

$$f^{(n)}(x) = r(r-1)\cdots(r-n+1)(1+x)^{r-n}$$

These coefficients correspond exactly with that of the Taylor polynomial. If $r \in \mathbb{N}$, then $f^{(r+1)}(x) \equiv 0$, so clearly the R_n are zero as $n \to \infty$. In general, using Lagrange's form of the remainder,

$$R_n = \frac{x^n}{n!} f^{(n)}(\theta x) = \binom{r}{n} \frac{x^n}{(1+\theta x)^{n-r}}$$

Note that in principle, θ depends both on x and n. For $0 < x < 1$, $(1+\theta x)^{n-r} > 1$ for $n > r$. Now observe that the series given by

$$\sum \binom{r}{n} x^n$$

is absolutely convergent for $|x| < 1$. Indeed, we can apply the ratio test and find that

$$\left|\frac{a_{n+1}}{a_n}\right| = \left|\frac{(r-n)x}{n+1}\right|$$

which tends to $|x|$ as $n \to \infty$. In particular therefore, the terms $\binom{r}{n}x^n$ tend to zero for $|x| < 1$. Hence for $n > r$ and $0 < x < 1$, we have

$$|R_n| \leq \left|\binom{r}{n}x^n\right| \to 0$$

So the claim is proven in the range $0 \leq x < 1$. If $x < 0$, then the step when we compare $(1+\theta x)^{n-r}$ with 1 breaks down. Let us instead use the Cauchy form of the remainder to bypass this step.

$$R_n = \frac{(1-\theta)^{n-1}x^n f^{(n)}(\theta x)}{(n-1)!} = \frac{(1-\theta)^{n-1}r(r-1)\cdots(r-n+1)(1+\theta x)^{r-n}x^n}{(n-1)!}$$

By regrouping terms, we get

$$R_n = \frac{r(r-1)\cdots(r-n+1)}{(n-1)!}\cdot\frac{(1-\theta)^{n-1}}{(1+\theta x)^{n-r}}x^n = r\binom{r-1}{n-1}x^n(1+\theta x)^{r-1}\left(\underbrace{\frac{1-\theta}{1+\theta x}}_{<1}\right)^{n-1}$$

Hence

$$|R_n| \leq \left|r\binom{r-1}{n-1}x^n\right|(1+\theta x)^{r-1}$$

This will then tend to zero, after a bit more effort; we can bound the $(1+\theta x)^{r-1}$ term by the maximum of 1 and $(1+x)^{r-1}$, which is independent of n, and then the result will follow. □

9. Power series

9.1. Complex differentiation

The complex derivative and the real derivative have the same core properties, for instance linearity, the product rule and the chain rule. However, the complex derivative is significantly more restrictive than the real derivative, since we can approach a point in any number of directions. If we can find a function that is complex differentiable with this restriction, we actually get a whole array of features for free. As an example of this restriction, consider the function $f(z) = \overline{z}$. This function is actually nowhere differentiable. If it were differentiable, then any sequence tending to z would yield the same limit when substituted into the definition of the derivative. Consider first the sequence

$$z_n = z + \frac{1}{n} \to z$$

Then

$$\frac{f(z_n) - f(z)}{z_n - z} = \frac{\overline{z} + \frac{1}{n} - \overline{z}}{z + \frac{1}{n} - z} = 1$$

Now consider the sequence

$$z_n = z + \frac{i}{n} \to z$$

Then

$$\frac{f(z_n) - f(z)}{z_n - z} = \frac{\overline{z} - \frac{i}{n} - \overline{z}}{z + \frac{i}{n} - z} = -1$$

Hence $f(z)$ is nowhere differentiable. On the other hand, the real function $f(x, y) = (x, -y)$ is clearly real differentiable, since it is linear; but in the complex world the function $z \mapsto \overline{z}$ is not linear.

9.2. Definition of power series

A power series is a series of the form

$$\sum_{n=0}^{\infty} a_n z^n$$

where $z \in \mathbb{C}$, and the a_n is a given sequence of complex numbers. We can also take a power series of the form

$$\sum_{n=0}^{\infty} a_n (z - z_0)^n$$

but for simplicity we will take $z_0 = 0$ in all of the analysis we will conduct on power series.

9.3. Radius of convergence

Lemma. If the series

$$\sum_{n=0}^{\infty} a_n z_1^n$$

converges for some point z_1, and $|z| < |z_1|$, then the series

$$\sum_{n=0}^{\infty} a_n z^n$$

also converges absolutely.

Proof. Since $\sum_{n=0}^{\infty} a_n z_1^n$ converges, $a_n z_1^n \to 0$. Thus the sequence $a_n z_1^n$ is bounded by some $k > 0$, i.e. for all n, $|a_n z_1^n| < k$. Then

$$|a_n z^n| \leq k \left| \frac{z}{z_1} \right|^n$$

Since the geometric series $\sum_0^{\infty} \left| \frac{z}{z_1} \right|^n$ converges, the lemma follows by comparison. □

Using this lemma, we can find that there exists a radius inside which any given power series converges absolutely. This radius might be zero, and it might be infinite.

Theorem. Any power series either

(i) converges absolutely for all z, or

(ii) converges absolutely for all z where $|z| < R$ and diverges for all z where $|z| > R$, or

(iii) converges for $z = 0$ only.

The circle $|z| = R$ is called the circle of convergence, and R is called the radius of convergence. Note that this theorem does not make any claim about the behaviour *on* the circle of convergence, just the behaviour inside it.

Proof. Let

$$S = \left\{ x \in \mathbb{R} : \ x \geq 0, \sum_0^{\infty} a_n x^n \text{ converges} \right\}$$

Clearly, $0 \in S$. By the above lemma, if $x_1 \in S$, then $[0, x_1] \subseteq S$. If $S = [0, \infty)$, then we have case (i) above due to the lemma.

If $S \neq [0, \infty)$, there exists a supremum $0 \leq R = \sup S < \infty$.

We must now just deal with case (ii), which is $R > 0$. For all z_1 with $|z_1| < R$ there exists R_0 such that $|z_1| < R_0 < R$, and absolute convergence follows using the lemma. If $|z_1| > R$, there exists R_0 such that $|z_1| > R_0 > R$. If the series with z_1 converges, then by the lemma the same would be true for R_0. But R_0 does not converge, so this is a contradiction. □

Lemma. If

$$\left|\frac{a_{n+1}}{a_n}\right| \to \ell$$

as $n \to \infty$, then $R = \frac{1}{\ell}$.

Proof. By the ratio test, we have absolute convergence if

$$\left|\frac{a_{n+1}}{a_n}\frac{z^{n+1}}{z^n}\right| < 1$$

So we have absolute convergence if $|z| < \frac{1}{\ell}$ and divergence if $|z| > \frac{1}{\ell}$ as required. □

Lemma. If

$$|a_n^{1/n}| \to \ell$$

as $n \to \infty$, then $R = \frac{1}{\ell}$.

This can be shown similarly using the root test.

Example. (i) Consider the series $\sum_0^\infty \frac{z^n}{n!}$. Using the ratio test, the series converges absolutely everywhere.

(ii) The geometric series $\sum_0^\infty z^n$ gives $R = 1$ by the ratio test. In this case, $|z| = 1$ gives divergence.

(iii) The series $\sum_0^\infty n! z^n$ has $R = 0$, which again can be seen using the ratio test.

(iv) Consider $\sum_1^\infty \frac{z^n}{n}$. This also has $R = 1$ by the ratio test. Note that the series diverges for $z = 1$ since we get the harmonic series. However, it converges when $z = -1$ by the alternating series test. To work out the behaviour at other points on the circle of convergence, we could consider the series $\sum_1^\infty \frac{z^n}{n}(1 - z)$, which converges exactly when the original series does. The partial sums are

$$\begin{aligned} S_N &= \sum_1^N \left[\frac{z_n - z^{n+1}}{n}\right] \\ &= \sum_1^N \frac{z^n}{n} - \sum_1^N \frac{z^{n+1}}{n} \\ &= \sum_1^N \frac{z^n}{n} - \sum_2^{N+1} \frac{z^n}{n-1} \\ &= z - \frac{z^{N+1}}{N+1} + \sum_2^{N+1} \frac{-z^n}{n(n-1)} \end{aligned}$$

If $|z| = 1$, then the term $\frac{z^{N+1}}{N+1}$ will vanish as $N \to \infty$. If $z \neq 1$, the term $\sum_2^{N+1} \frac{-z^n}{n(n-1)}$ converges as $N \to \infty$. So S_N does indeed converge for $|z| = 1$, $z \neq 1$.

(v) Now, consider $\sum_1^\infty \frac{z^n}{n^2}$. This has $R = 1$ by the ratio test, but it converges for all z with $|z| = 1$.

(vi) If we have $\sum_0^\infty nz^n$, we have $R = 1$, but diverges for all $|z| = 1$.

In conclusion, we cannot determine the behaviour at the boundary in the general case. Inside the radius of convergence, power series will behave as if they were simply polynomials.

9.4. Infinite differentiability

Theorem. Let $f(z) = \sum_0^\infty a_n z^n$ have a radius of convergence R. Then f is complex differentiable at all points with $|z| < R$, with

$$f'(z) = \sum_1^\infty n a_n z^{n-1}$$

with the same radius of convergence as the original series.

This proof comprises the entire subsection. This whole subsection is non-examinable, but included for completeness. First, we will state two lemmas.

Lemma. If $\sum_0^\infty a_n z^n$ has radius of convergence R, then both series

$$\sum_1^\infty n a_n z^{n-1}$$

and

$$\sum_2^\infty n(n-1) a_n z^{n-2}$$

also have radius of convergence R.

Proof. Let R_0 be such that $0 < |z| < R_0 < R$. Since $a_0 R_0^n \to 0$, the sequence $a_0 R_0^n$ is bounded. In other words there exists a k such that $|a_n R_0^n| \leq k$ for all $n \geq 0$. Thus,

$$|a_n n z^{n-1}| = \frac{n}{|z|} |a_n R_0^n| \left|\frac{z}{R_0}\right|^n \leq \frac{kn}{|z|} \left|\frac{z}{R_0}\right|^n$$

But

$$\sum n \left|\frac{z}{R_0}\right|^n$$

converges by the ratio test, since the ratio is

$$\frac{n+1}{n} \left|\frac{z}{R_0}\right|^{n+1} \left|\frac{R_0}{z}\right|^n = \frac{n+1}{n} \left|\frac{z}{R_0}\right| \to \left|\frac{z}{R_0}\right| < 1$$

Hence, the original series $\sum_1^\infty na_n z^{n-1}$ is absolutely bounded above by a convergent series, and therefore is absolutely convergent. So it is known that the radius of convergence of this derivative series is *at least* R. Now, if $|z| > R$, the series diverges since $|a_n z^n|$ is unbounded, and hence $|na_n z^n|$ is also unbounded. The same proof applies to the series for the second derivative. □

We will need this 'second derivative' condition in order to talk about the remainder term after the first derivative, which is related to the second derivative.

Lemma. First, for all $2 \leq r \leq n$.

$$\binom{n}{r} \leq n(n-1)\binom{n-2}{r-2}$$

Further, for all $z \in \mathbb{C}$, $h \in \mathbb{C}$,

$$\left|(z+h)^n - z^n - nhz^{n-1}\right| \leq n(n-1)(|z| + |h|)^{n-2}|h|^2$$

Proof. For the first part, we can expand the definitions to get

$$\frac{\binom{n}{r}}{\binom{n-2}{r-2}} = \frac{n(n-1)}{r(r-1)} \leq n(n-1)$$

as required. For the second part, we can apply the binomial expansion to cancel the other two terms, and we get

$$\begin{aligned}
(z+h)^n - z^n - nhz^{n-1} &= \left(\sum_{r=0}^{n} \binom{n}{r} z^{n-r}h^r\right) - z^n - nhz^{n-1} \\
&= \sum_{r=2}^{n} \binom{n}{r} z^{n-r}h^r \\
\therefore \left|(z+h)^n - z^n - nhz^{n-1}\right| &= \left|\sum_{r=2}^{n} \binom{n}{r} z^{n-r}h^r\right| \\
&\leq \sum_{r=2}^{n} \left|\binom{n}{r} z^{n-r}h^r\right| \\
&= \sum_{r=2}^{n} \binom{n}{r} |z|^{n-r}|h|^r \\
&\leq n(n-1) \underbrace{\left[\sum_{r=2}^{n} \binom{n-2}{r-2} |z|^{n-r}|h|^{r-2}\right]}_{(|z|+|h|)^{n-2}} |h|^2 \\
&= n(n-1)(|z| + |h|)^{n-2}|h|^2
\end{aligned}$$

as required. □

Now, we can prove the original theorem.

Proof. By the first lemma, we may define $f'(z)$ to be

$$f'(z) = \sum_1^\infty n a_n z^{n-1}$$

We now just need to prove that

$$\lim_{h\to 0} I = 0; \quad I = \frac{f(z+h) - f(z) - hf'(z)}{h}$$

We can substitute the expressions we have found for each power series:

$$\begin{aligned} I &= \frac{\sum_0^\infty a_n(z+h)^n - \sum_0^\infty a_n z^n - h\sum_1^\infty n a_n z^{n-1}}{h} \\ &= \frac{1}{h}\sum_0^\infty \left[a_n(z+h)^n - a_n z^n - hna_n z^{n-1}\right] \\ &= \frac{1}{h}\sum_0^\infty a_n\left[(z+h)^n - z^n - hnz^{n-1}\right] \\ |I| &= \frac{1}{|h|}\left|\lim_{N\to\infty}\sum_0^N a_n\left[(z+h)^n - z^n - hnz^{n-1}\right]\right| \end{aligned}$$

Since the modulus function is continuous,

$$\begin{aligned} |I| &= \frac{1}{|h|}\lim_{N\to\infty}\left|\sum_0^N a_n\left[(z+h)^n - z^n - hnz^{n-1}\right]\right| \\ &\le \frac{1}{|h|}\lim_{N\to\infty}\sum_0^N \left|a_n\left[(z+h)^n - z^n - hnz^{n-1}\right]\right| \\ &= \frac{1}{|h|}\sum_0^\infty |a_n|\cdot\left|(z+h)^n - z^n - hnz^{n-1}\right| \end{aligned}$$

By the second part of the second lemma above,

$$\begin{aligned} |I| &\le \frac{1}{|h|}\sum_0^\infty |a_n|\cdot n(n-1)(|z|+|h|)^{n-2}|h|^2 \\ &= |h|\sum_0^\infty |a_n|\cdot n(n-1)(|z|+|h|)^{n-2} \end{aligned}$$

For $|h|$ small enough, $(|z| + |h|) < R$. Therefore, by the first lemma above,

$$\sum_0^\infty |a_n| \cdot n(n-1)(|z| + |h|)^{n-2}$$

converges to some $A(h)$. But $A(h)$ is monotonically decreasing, so

$$|I| \leq |h|A(h) \leq |h|A(r)$$

for some r such that $|z| + r < R$. We can now let $h \to 0$, giving

$$|I| \to 0$$

as required. □

9.5. Defining standard functions

We can now use this differentiability property to cleanly define the standard exponential, logarithmic and trigonometric functions. Let $e : \mathbb{C} \to \mathbb{C}$ be defined by

$$e(z) = \sum_0^\infty \frac{z^n}{n!}$$

We have already seen that it has infinite radius of convergence. Straight from the above theorem, e is infinitely differentiable everywhere, and it is its own derivative. Note that if a function $F : \mathbb{C} \to \mathbb{C}$ has $F'(z) = 0$ for all $z \in \mathbb{C}$, then F is constant. Indeed, consider $g(t) = F(tz) = u(t) + iv(t)$ where $t, u, v \in \mathbb{R}$. Then by the chain rule, $g'(t) = F'(tz)z = 0$ and hence $u'(t) + iv'(t) = 0$, giving $u'(t) = 0$ and $v'(t) = 0$ everywhere. We can now apply the real-valued case, showing that u and v (and hence F) are constant everywhere. Now, let $a, b \in \mathbb{C}$, and consider

$$F(z) = e(a + b - z)e(z)$$

Then

$$F'(z) = -e(a + b - z)e(z) + e(a + b - z)e(z) = 0$$

Hence $e(a + b - z)e(z)$ is constant for all z, hence

$$e(a + b - z)e(z) = e(a + b - 0)e(0) = e(a + b)$$

Since z is arbitrary, we can set $z = b$ to recover the familiar relation

$$e(a + b - b)e(b) = e(a + b) \implies e(a)e(b) = e(a + b)$$

9.6. Exponential and logarithmic functions

Last lecture, we covered the power series form of the exponential function $e : \mathbb{C} \to \mathbb{C}$. Note that if we input a real number, the output is also real. Hence, $e : \mathbb{R} \to \mathbb{R}$. This restricted definition of the function has the following properties.

Theorem. (i) $e : \mathbb{R} \to \mathbb{R}$ is everywhere differentiable, and $e'(x) = e(x)$.

(ii) $e(x + y) = e(x)e(y)$.

(iii) $e(x) > 0$.

(iv) e is strictly increasing.

(v) $e(x) \to \infty$ as $x \to \infty$, and $e(x) \to 0$ as $x \to -\infty$.

(vi) $e : \mathbb{R} \to (0, \infty)$ is a bijection.

Proof. The first two properties follow from the last lecture.

(iii) Clearly, $e(x) > 0$ for all $x \geq 0$ by considering the power series, which contains only positive terms for $x > 0$, and also $e(0) = 1$. Also, $e(0) = e(x - x) = e(x)e(-x)$, hence for all negative x, $e(x) > 0$.

(iv) Since $e'(x) = e(x)$, $e'(x) = e(x) > 0$ everywhere.

(v) By considering partial sums, if $x > 0$ we have $e(x) > 1 + x$, so if $x \to \infty$, $e(x) \to \infty$. When $x \to -\infty$, $e(x) = \frac{1}{e(x)} \to 0$.

(vi) Injectivity follows from being strictly increasing. For surjectivity, we need to show that given any $y \in (0, \infty)$ there exists some x such that $e(x) = y$. Due to property (v) above, we can certainly find real numbers a and b such that $e(a) < y < e(b)$. By the intermediate value theorem, there exists $x \in \mathbb{R}$ such that $e(x) = y$.

□

Remark. We have essentially proven that $e : (\mathbb{R}, +) \to ((0, \infty), \times)$ is a group isomorphism. This is exactly the same as showing that it is a bijection. Since e is a function, there exists an inverse function $\ell : ((0, \infty), \times) \to (\mathbb{R}, +)$.

Theorem. (i) $\ell : (0, \infty) \to \mathbb{R}$ is a bijection, and $\ell(e(x)) = x$ for all $x \in \mathbb{R}$, and $e(\ell(x)) = x$ for all $x \in (0, \infty)$.

(ii) ℓ is differentiable and its derivative is $\ell'(t) = \frac{1}{t}$.

(iii) $\ell(xy) = \ell(x) + \ell(y)$.

Proof. (i) This first property is obvious from the definition.

(ii) By the inverse function theorem, ℓ is differentiable everywhere and $\ell'(t) = \frac{1}{t}$ as required.

(iii) From IA Groups, if e is an isomorphism, so is its inverse.

□

9.7. Real numbered exponents

We will now define for $\alpha \in \mathbb{R}$ and $x > 0$ the function

$$r_\alpha(x) = e(\alpha \ell(x))$$

This can be taken as the definition of x raised to the power α.

Theorem. Suppose $x, y > 0$ and $\alpha, \beta \in \mathbb{R}$. Then

(i) $r_\alpha(xy) = r_\alpha(x) r_\alpha(y)$

(ii) $r_{\alpha+\beta}(x) = r_\alpha(x) r_\beta(x)$

(iii) $r_\alpha(r_\beta(x)) = r_{\alpha\beta}(x)$

(iv) $r_1(x) = x$, and $r_0(x) = 1$

Proof. (i) $r_\alpha(xy) = e(\alpha\ell(xy)) = e(\alpha\ell(x) + \alpha\ell(y)) = e(\alpha\ell(x))e(\alpha\ell(y)) = r_\alpha(x) r_\alpha(y)$

(ii) $r_{\alpha+\beta}(x) = e((\alpha+\beta)\ell(x)) = e(\alpha\ell(x) + \beta\ell(x)) = e(\alpha\ell(x))e(\beta\ell(x)) = r_\alpha(x) r_\beta(x)$

(iii) $r_\alpha(r_\beta(x)) = e(\alpha\ell[e(\beta\ell(x))]) = e(\alpha\beta\ell(x)) = r_{\alpha\beta}(x)$

(iv) $r_1(x) = e(\ell(x)) = x$, and $r_0(x) = e(0\ell(x)) = e(0) = 1$

□

Suppose we want to compute $r_n(x)$, where $n \in \mathbb{Z}$. Then $r_n(x) = r_{1+\cdots+1}(x) = x \cdots x$, so we have agreement between $r_n(x)$ and our previous definition of x^n. Similarly, since $r_1(x) r_{-1}(x) = 1$, we have $r_{-1}(x) = \frac{1}{x}$. Further, $r_{\frac{1}{q}}(x) = x^{\frac{1}{q}}$. Therefore, $r_{\frac{p}{q}}(x) = x^{\frac{p}{q}}$. So this definition is simply a more general definition for exponentiation by a real number.

From now, we will let $\exp(x) \equiv e(x)$, $\log(x) \equiv \ell(x)$, and $x^\alpha \equiv r_\alpha(x)$. In fact, $\exp(x) = e^x$ for a suitable number e, since $e(x) = e(x \log(e)) = r_x(e) = e^x$ where $e \coloneqq e(1) = \sum_0^\infty \frac{1}{n!}$.

Finally, we can compute the derivative of x^α using the chain rule.

$$(x^\alpha)' = \left(e^{\alpha \log x}\right)' = e^{\alpha \log x} \alpha \frac{1}{x} = \alpha x^\alpha x^{-1} = \alpha x^{\alpha-1}$$

as expected. Further, if $f(x) = a^x$, we can find

$$f'(x) = \left(e^{x \log a}\right)' = e^{x \log a} \log a = a^x \log a$$

9.8. Trigonometric functions

We define

$$\cos z = 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \frac{z^6}{6!} + \cdots = \sum_0^\infty \frac{(-1)^k z^{2k}}{(2k)!}$$
$$\sin z = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \cdots = \sum_0^\infty \frac{(-1)^k z^{2k+1}}{(2k+1)!}$$

Both power series have infinite radius of convergence, by the ratio test (the same proof from the exponential function can be used here). Hence $\cos z$ and $\sin z$ are differentiable, and $\frac{\mathrm{d}}{\mathrm{d}z}\cos z = -\sin z$ and $\frac{\mathrm{d}}{\mathrm{d}z}\sin z = \cos z$ as expected, by termwise differentiation. Further, we can deduce that

$$e^{iz} = \sum_0^\infty \frac{(iz)^n}{n!} = \sum_0^\infty \frac{(iz)^{2k}}{(2k)!} + \sum_0^\infty \frac{(iz)^{2k+1}}{(2k+1)!}$$

Note that

$$(iz)^{2k} = (-1)^k z^{2k}; \quad (iz)^{2k+1} = i(-1)^k z^{2k+1}$$

Hence,

$$e^{iz} = \cos z + i \sin z$$

Similarly,

$$e^{-iz} = \cos z - i \sin z$$

We can then write

$$\cos z = \frac{1}{2}\left(e^{iz} + e^{-iz}\right); \quad \sin z = \frac{1}{2i}\left(e^{iz} - e^{-iz}\right)$$

Many common trigonometric identities follow from this, such as the identity $\cos^2 z + \sin^2 z \equiv 1$. However, we have not deduced the period of the functions. Now, restricted to the real case, $\sin x, \cos x \in \mathbb{R}$, and the identity $\cos^2 z + \sin^2 z \equiv 1$ gives that $|\sin x| \leq 1$ and $|\cos x| \leq 1$ for all real x.

9.9. Circle constants

Proposition. There is a smallest positive number π such that

$$\cos \frac{\pi}{2} = 0$$

and we have $\sqrt{2} < \frac{\pi}{2} < \sqrt{3}$.

Proof. If $0 < x < 2$,

$$\sin x = \left(x - \frac{x^3}{3!}\right) + \left(\frac{x^5}{5!} - \frac{x^7}{7!}\right) + \cdots$$

For this range of values, each parenthesised block is positive, so $\sin x > 0$. So in this range,

$$\frac{\mathrm{d}}{\mathrm{d}x} \cos x < 0$$

Hence, $\cos x$ is a strictly decreasing function on this interval. Now,

$$\cos\sqrt{2} = 1 - \frac{\sqrt{2}^2}{2!} + \left(\frac{\sqrt{2}^4}{4!} - \frac{\sqrt{2}^6}{6!}\right) + \cdots > 0$$

since each bracketed block is positive.

$$\cos\sqrt{3} = 1 - \frac{\sqrt{3}^2}{2!} + \frac{\sqrt{3}^4}{4!} - \left(\frac{\sqrt{3}^6}{6!} - \frac{\sqrt{3}^8}{8!}\right) + \cdots < 0$$

since all the bracketed terms are positive, and being subtracted from a negative number. By the intermediate value theorem, the existence of such a π follows. □

Corollary. We have that $\sin\frac{\pi}{2} = 1$.

Proof. We know that $\cos^2\frac{\pi}{2} + \sin^2\frac{\pi}{2} = 1$, and $\sin\frac{\pi}{2} > 0$, so the result follows. □

Theorem. The following standard properties about the periodicity of trigonometric functions hold.

(i) $\sin\left(z + \frac{\pi}{2}\right) = \cos z$, and $\cos\left(z + \frac{\pi}{2}\right) = -\sin z$

(ii) $\sin(z + \pi) = -\sin z$, and $\cos(z + \pi) = -\cos z$

(iii) $\sin(z + 2\pi) = \sin z$, and $\cos(z + 2\pi) = \cos z$

The proofs are immediate from the angle addition formulae. This then implies that

$$e^{iz+2\pi i} = e^{iz}$$

Hence e^z is periodic with period $2\pi i$.

10. Integration

10.1. Geometry of trigonometric functions

Recall that given any two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^2$, we can define the dot (scalar) product by

$$\mathbf{x} \cdot \mathbf{y} = (x_1, x_2) \cdot (y_1, y_2) = x_1 y_1 + x_2 y_2$$

By the Cauchy–Schwarz inequality, we have

$$|\mathbf{x} \cdot \mathbf{y}| \leq \|\mathbf{x}\| \|\mathbf{y}\|$$

where we define the Euclidean norm in the normal way. Thus, for $\mathbf{x} \neq 0$, $\mathbf{y} \neq 0$, we have

$$-1 \leq \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} \leq 1$$

We now define the angle between two vectors $\mathbf{x}$ and $\mathbf{y}$ as exactly the unique number $\theta \in [0, \pi]$ such that

$$\cos\theta = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

10.2. Hyperbolic functions

We define the functions cosh and sinh as follows.

$$\cosh z = \frac{1}{2}(e^z + e^{-z})$$

$$\sinh z = \frac{1}{2}(e^z - e^{-z})$$

Hence

$$\cosh z = \cos(iz); \quad \sinh z = -i \sin(iz)$$

We can then show that

$$\frac{\mathrm{d}}{\mathrm{d}z} \cosh z = \sinh z; \quad \frac{\mathrm{d}}{\mathrm{d}z} \sinh z = \cosh z$$

and further,

$$\cosh^2 z - \sinh^2 z \equiv 1$$

10.3. Defining the Riemann integral

Definition. A *dissection* or *partition* $\mathcal{D}$ of $[a, b]$ is a finite subset of $[a, b]$ containing the end points a and b. We write

$$\mathcal{D} = \{x_0, x_1, \ldots, x_n\}$$

with $a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$.

Definition. We define the *upper sum* of a bounded function f associated with a partition $\mathcal{D}$ by

$$S(f,\mathcal{D}) = \sum_{j=1}^{n}(x_j - x_{j-1}) \sup_{x\in[x_{j-1},x_j]} f(x)$$

The *lower sum* is defined similarly,

$$s(f,\mathcal{D}) = \sum_{j=1}^{n}(x_j - x_{j-1}) \inf_{x\in[x_{j-1},x_j]} f(x)$$

Clearly then $S \geq s$ for all $\mathcal{D}$.

Lemma. If $\mathcal{D}$ and $\mathcal{D}'$ are dissections with $\mathcal{D}' \supseteq \mathcal{D}$ ($\mathcal{D}'$ is a refinement of $\mathcal{D}$), then

$$S(f,\mathcal{D}) \underset{\text{(i)}}{\geq} S(f,\mathcal{D}') \underset{\text{(ii)}}{\geq} s(f,\mathcal{D}') \underset{\text{(iii)}}{\geq} s(f,\mathcal{D})$$

Proof. Inequality (ii) is obvious, we have already shown this to be true. Now, suppose $\mathcal{D}'$ contains a single extra point y compared to $\mathcal{D}$, where $y \in (x_{r-1}, x_r)$. Clearly,

$$\sup_{x\in[x_{r-1},y]} f(x), \sup_{x\in[y,x_r]} f(x) \leq \sup_{x\in[x_{r-1},x_r]}$$

Then

$$(x_r - x_{r-1}) \sup_{x\in[x_{r-1},x_r]} f(x) \geq (y - x_{r-1}) \sup_{x\in[r_{r-1},y]} f(x) + (x_r - y) \sup_{x\in[y,r]} f(x)$$

Hence,

$$S(f,\mathcal{D}) \geq S(f,\mathcal{D}')$$

The same proof holds for inequality (iii), and inductively we can show that this works for any amount of extra points. □

Lemma. If $\mathcal{D}_1, \mathcal{D}_2$ are arbitrary dissections, then

$$S(f,\mathcal{D}_1) \geq S(f,\mathcal{D}_1 \cup \mathcal{D}_2) \geq s(f,\mathcal{D}_1 \cup \mathcal{D}_2) \geq s(f,\mathcal{D}_2)$$

In particular, $S(f,\mathcal{D}_1) \geq s(f,\mathcal{D}_2)$.

Proof. Let $\mathcal{D}' = \mathcal{D}_1 \cup \mathcal{D}_2$, which is a refinement of both $\mathcal{D}_1$ and $\mathcal{D}_2$, and apply the previous lemma. □

Definition. The *upper integral* of f is

$$I^*(f) = \inf_{\mathcal{D}} S(f,\mathcal{D})$$

Note that such an integral always exists, since the upper sums are always bounded below by an arbitrary lower sum. Hence the infimum does indeed exist and is finite. Similarly,

$$I_*(f) = \sup_{\mathcal{D}} s(f,\mathcal{D})$$

Then by the lemmas above, $I^\star(f) \geq I_\star(f)$, since $S(f, \mathcal{D}_2) \geq s(f, \mathcal{D}_1)$ for arbitrary dissections $\mathcal{D}_1$ and $\mathcal{D}_2$.

Definition. A bounded function $f : [a, b] \to \mathbb{R}$ is (Riemann) integrable if $I^\star(f) = I_\star(f)$. If this equality holds, we write

$$\int_a^b f(x)\,\mathrm{d}x = I^\star(f) = I_\star(f) = \int_a^b f$$

10.4. Determining integrability

Theorem. A function $f : [a, b] \to \mathbb{R}$ is Riemann integrable if and only if given $\varepsilon > 0$, there exists $\mathcal{D}$ such that

$$S(f, \mathcal{D}) - s(f, \mathcal{D}) < \varepsilon$$

Proof. For every dissection $\mathcal{D}$, we have that $0 \leq I^\star(f) - I_\star(f) \leq S(f, \mathcal{D}) - s(f, \mathcal{D})$. If the given condition holds, $0 \leq I^\star(f) - I_\star(f) \leq S(f, \mathcal{D}) - s(f, \mathcal{D}) < \varepsilon$ for all $\varepsilon > 0$. This immediately implies that f is Riemann integrable since the upper integral and the lower integral match. Conversely, if f is integrable, by the definition of the supremum and infimum, there are partitions $\mathcal{D}_1$ and $\mathcal{D}_2$ such that

$$\int_a^b f - \frac{\varepsilon}{2} = I_\star(f) - \frac{\varepsilon}{2} < s(f, \mathcal{D}_1)$$

Also,

$$\int_a^b f + \frac{\varepsilon}{2} = I^\star(f) + \frac{\varepsilon}{2} > S(f, \mathcal{D}_2)$$

From last lecture, we can use the fact that $\mathcal{D}_1 \cup \mathcal{D}_2$ is a refinement of both $\mathcal{D}_1$ and $\mathcal{D}_2$ to show that

$$S(f, \mathcal{D}_1 \cup \mathcal{D}_2) - s(f, \mathcal{D}_1 \cup \mathcal{D}_2) \leq S(f, \mathcal{D}_2) - s(f, \mathcal{D}_1)$$

Now,

$$S(f, \mathcal{D}_2) - s(f, \mathcal{D}_1) < \int_a^b f + \frac{\varepsilon}{2} - \int_a^b f + \frac{\varepsilon}{2} = \varepsilon$$

as required. □

10.5. Monotonic and continuous functions

We can use this theorem to show that monotonic and continuous functions are integrable. Note that monotonic and continuous functions (defined on a closed interval) are always bounded.

Theorem. Suppose a function $f : [a, b] \to \mathbb{R}$ is monotonic. Then f is integrable.

Proof. Suppose f is increasing. Then

$$\sup_{x\in[x_{j-1}-x_j]} f(x) = f(x_j)$$

and similarly

$$\inf_{x\in[x_{j-1}-x_j]} f(x) = f(x_{j-1})$$

Thus,

$$S(f,\mathcal{D}) - s(f,\mathcal{D}) = \sum_{j=1}^{n} (x_j - x_{j-1})\left[f(x_j) - f(x_{j-1})\right]$$

Let us choose the dissection

$$\mathcal{D} = \left\{a, a + \frac{b-a}{n}, a + 2\frac{b-a}{n} + \cdots + b\right\}$$

giving

$$x_j = a + j\frac{b-a}{n}$$

for $0 \leq j \leq n$. In this case,

$$S(f,\mathcal{D}) - s(f,\mathcal{D}) = \frac{b-a}{n}\sum_{j=1}^{n}\left[f(x_j) - f(x_{j-1})\right] = \frac{b-a}{n}\left[f(b) - f(a)\right] \to 0$$

so then using the above theorem, f is integrable. □

To prove an analogous result for continuous function, we must first prove the following lemma.

Lemma (Uniform Continuity)**.** Suppose a function $f : [a,b] \to \mathbb{R}$ is continuous. Then given $\varepsilon > 0$, $\exists\delta > 0$ such that if $|x-y| < \delta$, we have $|f(x) - f(y)| < \varepsilon$.

Note that in this lemma, we are saying that there exists such a δ that works for *every* pair of points within δ. The definition of continuity only provides a δ that depends on x, so this is stronger than the definition of continuity, and this property does not hold for all continuous functions.

Proof. Suppose there does not exist such a δ. Then there exists some $\varepsilon > 0$ such that for all $\delta > 0$ there exist $x, y \in [a,b]$ such that $|x-y| < \delta$ but $|f(x) - f(y)| \geq \varepsilon$. Let $\delta = \frac{1}{n}$. For this choice, we can find sequences x_n and y_n with $|x_n - y_n| < \frac{1}{n}$ but $|f(x_n) - f(y_n)| \geq \varepsilon$. By the Bolzano–Weierstrass theorem, since we are working in a closed bounded interval, the x_n and y_n have convergent subsequences that tend to c and d. Then by the triangle inequality,

$$\left|y_{n_k} - c\right| \leq \left|y_{n_k} - x_{n_k}\right| + \left|x_{n_k} - c\right| \to 0$$

So $c = d$. But $\left|f(x_{n_k}) - f(y_{n_k})\right| \geq \varepsilon$, and by continuity as $k \to \infty$, $|f(c) - f(c)| \geq \varepsilon$ which is a contradiction. □

Theorem. Suppose a function $f : [a, b] \to \mathbb{R}$ is continuous. Then f is integrable.

Proof. We know that given $\varepsilon > 0$, there exists $\delta > 0$ such that $|x - y| < \delta \implies |f(x) - f(y)| < \varepsilon$. So now, let

$$\mathcal{D} = \left\{a, a + \frac{b-a}{n}, a + 2\frac{b-a}{n} + \cdots + b\right\}$$

where n is chosen large enough such that $\frac{b-a}{n} < \delta$. Then, for any $x, y \in [x_{j-1}, x_j]$, we have that $|f(x) - f(y)| < \varepsilon$. We can now write

$$\max_{x \in [x_{j-1}, x_j]} f(x) - \min_{x \in [x_{j-1}, x_j]} f(x) = f(p) - f(q) < \varepsilon$$

Therefore, the upper sums and lower sums differ by at most $(b - a)\varepsilon$. Hence, f is integrable. □

10.6. Complicated integrable functions

In principle, many functions that are not continuous or monotonic can be integrated using the Riemann integral. For example, the function $f : [0, 1] \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} \frac{1}{q} & x = \frac{p}{q} \in (0, 1] \text{ in its lowest form} \\ 0 & \text{otherwise} \end{cases}$$

is Riemann integrable. We know that $s(f, \mathcal{D}) = 0$ for all $\mathcal{D}$, since any interval will contain irrational numbers. We will show that given $\varepsilon > 0$, there exists $\mathcal{D}$ such that $S(f, \mathcal{D}) < \varepsilon$. If this is true, then this function f really is Riemann integrable, with $\int f = 0$. We will choose $N \in \mathbb{N}$ such that $\frac{1}{N} < \frac{\varepsilon}{2}$. Then

$$S = \left\{x \in [0, 1] : f(x) \geq \frac{1}{N}\right\} = \left\{\frac{p}{q} : 1 \leq q \leq N, 1 \leq p \leq q\right\}$$

This set S is a finite set, hence

$$S = \{0, t_1, \ldots, t_R\}; \quad 0 < t_1 < \cdots < t_R = 1$$

Consider a dissection $\mathcal{D}$ such that

(1) Each t_k is in some interval $[x_{j-1}, x_j]$, and

(2) For all k, the unique interval containing t_k has length at most $\frac{\varepsilon}{2R}$.

Such a dissection can certainly be constructed. Then, in any interval that does not contain a t_k, f in this interval is less than $\frac{1}{N}$. In any interval that does contain a t_k, $f \geq \frac{1}{N}$ but $f < 1$ everywhere. Since there are R such intervals, each of which with length $\frac{\varepsilon}{2R}$, we have

$$S(f, \mathcal{D}) \leq \frac{1}{N} + \frac{\varepsilon}{2} < \varepsilon$$

10.7. Properties of Riemann integral

Consider functions f and g which are bounded and integrable on $[a, b]$.

(1) If $f \leq g$ on $[a, b]$, then $\int f \leq \int g$.

(2) $f + g$ is integrable on $[a, b]$, and $\int (f + g) = \int f + \int g$.

(3) For any constant k, kf is integrable, and $\int kf = k \int f$.

(4) $|f|$ is integrable, and $|\int f| \leq \int |f|$.

(5) fg is integrable.

Proof. We will see proofs for some of these properties.

(1) If $f \leq g$, then

$$\int f = I^\star(f) \leq S(f, \mathcal{D}) \leq S(g, \mathcal{D})$$

Hence,

$$\int f = I^\star(f) \leq I^\star(g) = \int g$$

(2) We have

$$\sup_{[x_{j-1}, x_j]} (f + g) \leq \sup_{[x_{j-1}, x_j]} f + \sup_{[x_{j-1}, x_j]} g$$

Therefore,

$$S(f + g, \mathcal{D}) \leq S(f, \mathcal{D}) + S(g, \mathcal{D})$$

Now, consider two dissections $\mathcal{D}_1, \mathcal{D}_2$. Now,

$$I^\star(f + g) \leq S(f + g, \mathcal{D}_1 \cup \mathcal{D}_2) \leq S(f, \mathcal{D}_1 \cup \mathcal{D}_2) + S(g, \mathcal{D}_1 \cup \mathcal{D}_2) \leq S(f, \mathcal{D}_1) + S(g, \mathcal{D}_2)$$

We can then fix $\mathcal{D}_1$ and take the infimum over $\mathcal{D}_2$ to get

$$I^\star(f + g) \leq S(f, \mathcal{D}_1) + I^\star(g)$$

Taking the infimum over $\mathcal{D}_1$ gives

$$I^\star(f + g) \leq I^\star(f) + I^\star(g) = \int f + \int g$$

A completely similar argument will show that

$$I_\star(f + g) \geq \int f + \int g$$

Combining this, $f + g$ must be integrable, since $I^\star(f + g) \geq I_\star(f + g)$. This integral is then exactly $\int f + \int g$.

(4) Consider first $f_+(x) = \max(f(x), 0)$. We want to show that f_+ is integrable. We can check that

$$\sup_{[x_{j-1},x_j]} f_+ - \inf_{[x_{j-1},x_j]} f_+ \leq \sup_{[x_{j-1},x_j]} f - \sup_{[x_{j-1},x_j]} f$$

We know that given $\varepsilon > 0$, there exists $\mathcal{D}$ such that

$$S(f, \mathcal{D}) - s(f, \mathcal{D}) < \varepsilon$$

Hence,

$$S(f_+, \mathcal{D}) - s(f_+, \mathcal{D}) \leq S(f, \mathcal{D}) - s(f, \mathcal{D}) < \varepsilon$$

Therefore f_+ is integrable. But $|f| = 2f_+ - f$, hence $|f|$ is integrable by properties (2) and (3). Since $-|f| \leq f \leq |f|$, we can use monotonicity from (1) to find that

$$\left| \int f \right| \leq \int |f|$$

as claimed.

(5) Let f be integrable and positive. Then we can check that

$$\sup_{[x_{j-1},x_j]} f^2 = \left(\underbrace{\sup_{[x_{j-1},x_j]} f}_{M_j} \right)^2; \quad \inf_{[x_{j-1},x_j]} f^2 = \left(\underbrace{\inf_{[x_{j-1},x_j]} f}_{m_j} \right)^2$$

Then,

$$\begin{aligned} S(f^2, \mathcal{D}) - s(f^2, \mathcal{D}) &= \sum_{j=1}^{n} (x_j - x_{j-1})(M_j^2 - m_j^2) \\ &= \sum_{j=1}^{n} (x_j - x_{j-1})(M_j - m_j)(M_j + m_j) \end{aligned}$$

The function f is bounded by some constant k, therefore the bracket $(M_j + m_j)$ is bounded by $2k$, which gives

$$S(f^2, \mathcal{D}) - s(f^2, \mathcal{D}) \leq 2k(S(f, \mathcal{D}) - s(f, \mathcal{D}))$$

So f^2 is integrable. Now, considering any f, $|f| \geq 0$ is a non-negative integrable function. Since $f^2 = |f^2|$, we deduce that f^2 is integrable for any integrable f. Finally, for fg, note that

$$4fg = (f + g)^2 - (f - g)^2$$

The right hand side is integrable, so the left hand side is integrable.

□

11. Fundamental theorem of calculus

11.1. Breaking an interval

Let f be integrable on $[a,b]$. If $a < c < b$, then f is integrable over $[a,c]$ and $[c,b]$, with

$$\int_a^b f = \int_a^c f + \int_c^b f$$

Conversely, if f is integrable on $[a,c]$ and $[c,b]$, then f is integrable over $[a,b]$ and the same equality holds for the combination of the integrals.

Proof. We first make two observations. First, if $\mathcal{D}_1$ is a dissection of $[a,c]$ and $\mathcal{D}_2$ is a dissection of $[c,b]$, then $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$ is a dissection of $[a,b]$, and

$$S(f, \mathcal{D}_1 \cup \mathcal{D}_2) = S\left(f\Big|_{[a,c]}, \mathcal{D}_1\right) + S\left(f\Big|_{[c,b]}, \mathcal{D}_2\right) \tag{$*$}$$

Also, if $\mathcal{D}$ is a dissection of $[a,b]$, then

$$S(f,\mathcal{D}) \geq S(f, \mathcal{D} \cup \{c\}) = S\left(f\Big|_{[a,c]}, \mathcal{D}_1\right) + S\left(f\Big|_{[c,b]}, \mathcal{D}_2\right) \tag{$\dagger$}$$

Now,

$$(*) \implies I^\star(f) \leq I^\star\left(f\Big|_{[a,c]}\right) + I^\star\left(f\Big|_{[c,b]}\right)$$

Further,

$$(\dagger) \implies I^\star(f) \geq I^\star\left(f\Big|_{[a,c]}\right) + I^\star\left(f\Big|_{[c,b]}\right)$$

Hence,

$$I^\star(f) = I^\star\left(f\Big|_{[a,c]}\right) + I^\star\left(f\Big|_{[c,b]}\right)$$

This argument also applies for the lower integral, therefore

$$0 \leq I^\star(f) - I_\star(f) = \underbrace{I^\star\left(f\Big|_{[a,c]}\right) - I_\star\left(f\Big|_{[a,c]}\right)}_{A} + \underbrace{I^\star\left(f\Big|_{[c,b]}\right) + I_\star\left(f\Big|_{[c,b]}\right)}_{B}$$

Note that $A, B \geq 0$. If f is integrable on $[a,c]$ and $[c,b]$, then $A = B = 0$, hence $I^\star(f) = I_\star(f)$ and it is integrable on $[a,b]$. If f is integrable on $[a,b]$, then we know $I^\star(f) = I_\star(f)$, so $A = B = 0$ so f is integrable on $[a,c]$ and $[c,b]$. □

Note that we take the following convention:

$$\int_a^b f = -\int_b^a f$$

and if $a = b$, then this value is zero. With this convention, if f is bounded with $|f| \leq k$, then

$$\left|\int_a^b f\right| \leq k|b-a|$$

11.2. Fundamental theorem of calculus

Suppose a function $f : [a,b] \to \mathbb{R}$ is bounded and integrable. Then since it is integrable on any sub-interval, we can define

$$F(x) = \int_a^x f(t)\,\mathrm{d}t$$

for $x \in [a,b]$.

Theorem. F is continuous.

Proof. We know that

$$F(x+h) - F(x) = \int_x^{x+h} f(t)\,\mathrm{d}t$$

We want this quantity to vanish as $h \to 0$. We find, given that f is bounded by k,

$$|F(x+h) - F(x)| = \left|\int_x^{x+h} f(t)\,\mathrm{d}t\right| \leq k|h|$$

So the result follows as $h \to 0$. □

Theorem. If in addition f is continuous at x, then F is differentiable, with $F'(x) = f(x)$.

Proof. Consider

$$\left|\frac{F(x+h)-F(x)}{h} - f(x)\right|$$

If this tends to zero, then the theorem holds.

$$\left|\frac{F(x+h)-F(x)}{h} - f(x)\right| = \frac{1}{|h|}\left|\int_x^{x+h} f(t)\,\mathrm{d}t - hf(x)\right| = \frac{1}{|h|}\left|\int_x^{x+h} [f(t)-f(x)]\,\mathrm{d}t\right|$$

Since f is continuous at x, given $\varepsilon > 0$, $\exists \delta > 0$ such that $|t-x| - \delta \implies |f(t)-f(x)| < \varepsilon$. If $|h| < \delta$, then the integrand is bounded by ε. Hence,

$$\left|\frac{F(x+h)-F(x)}{h} - f(x)\right| \leq \frac{1}{|h|}|h\varepsilon| = \varepsilon$$

So we can make this value as small as we like. So the theorem holds. □

For example, consider the function

$$f(x) = \begin{cases} -1 & x \in [-1, 0] \\ 1 & x \in (0, 1] \end{cases}$$

This is a bounded, integrable function, with

$$F(x) = -1 + |x|$$

Note that this F is not differentiable at $x = 0$.

Corollary. If $f = g'$ is a continuous function on $[a, b]$, then

$$\int_a^x f(t)\,\mathrm{d}t = g(x) - g(a)$$

is a differentiable function on $[a, b]$.

Proof. From above, $F - g$ has zero derivative in $[a, b]$, hence $F - g$ is constant. Since $F(a) = 0$, we get $F(x) = g(x) - g(a)$. □

Note that every continuous function f has an 'indefinite' integral (or 'antiderivative') written $\int f(x)\,\mathrm{d}x$, which is determined uniquely up to the addition of a constant. Note further that we have now essentially solved the differential equation

$$\begin{cases} y'(x) = f(x) \\ y(a) = y_0 \end{cases}$$

and shown that there is a unique solution to this ordinary differential equation.

12. Integration techniques

12.1. Integration by parts

We can use the fundamental theorem of calculus to deduce familiar integration techniques, such as integration by parts, and integration by substitution.

Corollary. Suppose f', g' exist and are continuous on $[a, b]$. Then

$$\int_a^b f'g = fg\Big|_a^b - \int_a^b fg'$$

Proof. By the product rule, we have

$$(fg)' = f'g + fg'$$

Then by the fundamental theorem of calculus,

$$\int_a^b (fg)' = fg\Big|_a^b = \int_a^b f'g + \int_a^b fg'$$

and the result follows. □

12.2. Integration by substitution

Corollary. Let $g : [\alpha, \beta] \to [a, b]$ with $g(\alpha) = a, g(\beta) = b$ and let g' exist and be continuous on $[\alpha, \beta]$. Let $f : [a, b] \to \mathbb{R}$ be continuous. Then

$$\int_a^b f(x)\,\mathrm{d}x = \int_\alpha^\beta f(g(t))g'(t)\,\mathrm{d}t$$

Proof. Let $F(x) = \int_a^x f(t)\,\mathrm{d}t$. Then let $h(t) = F(g(t))$. This is well defined since g takes values in $[a, b]$. Then,

$$\begin{aligned}
\int_\alpha^\beta f(g(t))g'(t)\,\mathrm{d}t &= \int_\alpha^\beta F'(g(t))g'(t)\,\mathrm{d}t \\
&= \int_\alpha^\beta h'(t)\,\mathrm{d}t \\
&= h(\beta) - h(\alpha) \\
&= F(b) - F(a) \\
&= F(b) \\
&= \int_a^b f(x)\,\mathrm{d}x
\end{aligned}$$

□

13. Integrals in Taylor's theorem

13.1. Integral remainder form of Taylor's theorem

Theorem. Let f such that $f^{(n)}(x)$ is continuous for $x \in [0, h]$. Then

$$f(h) = f(0) + \cdots + \frac{h^{n-1} f^{(n-1)}(0)}{(n-1)!} + R_n$$

where

$$R_n = \frac{h^n}{(n-1)!} \int_0^1 (1-t)^{n-1} f^{(n)}(th)\,\mathrm{d}t$$

Note that for this formulation of Taylor's theorem, we require continuity of $f^{(n)}(x)$, whereas with the previous remainders, the nth derivative need not be continuous.

Proof. First, by substituting $u = th$, we can see that it is sufficient to show

$$R_n = \frac{1}{(n-1)!} \int_0^h (h-u)^{n-1} f^{(n)}(u)\,\mathrm{d}u$$

Now, integrating by parts, we have

$$\begin{aligned} R_n &= \frac{-h^{n-1} f^{(n-1)}(0)}{(n-1)!} + \frac{1}{(n-2)!} \int_0^h (h-u)^{n-2} f^{(n-1)}(u)\,\mathrm{d}u \\ &= \frac{-h^{n-1} f^{(n-1)}(0)}{(n-1)!} + R_{n-1} \end{aligned}$$

Hence,

$$R_n = -\frac{h^{n-1} f^{(n-1)}(0)}{(n-1)!} - \frac{h^{n-2} f^{(n-2)}(0)}{(n-2)!} - \cdots - \underbrace{\int_0^h f'(u)\,\mathrm{d}u}_{f(h)-f(0)}$$

which is exactly all the other terms in the Taylor polynomial as required. □

13.2. Mean value theorem for integrals

Theorem. Let $f, g : [a, b] \to \mathbb{R}$ be continuous with $g(x) \neq 0$ for all $x \in (a, b)$. Then

$$\exists c \in (a, b) \text{ s.t. } \int_a^b f(x)g(x)\,\mathrm{d}x = f(c) \int_a^b g(x)\,\mathrm{d}x$$

Note that if we let $g(x) = 1$, we get

$$\int_a^b f(x)\,\mathrm{d}x = f(c)(b-a)$$

Proof. We will use Cauchy's mean value theorem to get this result. Let

$$F(x) = \int_a^x fg; \quad G(x) = \int_a^x g$$

Then there exists an intermediate point c such that

$$(F(b) - F(a))G'(c) = F'(c)(G(b) - G(a))$$

By the fundamental theorem of calculus,

$$\left(\int_a^b fg\right)g(c) = f(c)g(c)\left(\int_a^b g\right)$$

Now, since $g \neq 0$ everywhere,

$$\int_a^b fg = f(c)\int_a^b g$$

□

13.3. Deriving Lagrange's and Cauchy's remainders for Taylor's theorem

We can use this new mean value theorem to recover the other forms of the remainders in Taylor's theorem. We have

$$R_n = \frac{h^n}{(n-1)!}\int_0^1 (1-t)^{n-1} f^{(n)}(th)\,\mathrm{d}t$$

and we want to show that this is equal to

$$\frac{h^n}{n!}f^{(n)}(a+\theta h); \quad \frac{(1-\theta)^{n-1}h^n f^{(n)}(a+\theta h)}{(n-1)!}$$

First, let us apply the above mean value theorem with $g \equiv 1$ and the entire integrand in R_n as f. Then

$$R_n = \frac{h^n}{(n-1)!}\int_0^1 (1-t)^{n-1} f^{(n)}(th)\,\mathrm{d}t = \frac{h^n}{(n-1)!}(1-\theta)^{n-1} f^{(n)}(\theta h)$$

as required for Cauchy's remainder. To find Lagrange's remainder, we need to use the above mean value theorem with $g = (1-t)^{n-1}$, which is positive everywhere in $(0,1)$, and $f = f^{(n)}(th)$. Then

$$R_n = \frac{h^n}{(n-1)!}f^{(n)}(\theta h)\int_0^1 (1-t)^{n-1}\,\mathrm{d}t$$

This integral is simple to find by inspection:

$$R_n = \frac{h^n}{(n-1)!}f^{(n)}(\theta h)\frac{1}{n} = \frac{h^n}{n!}f^{(n)}(\theta h)$$

as required.

14. Uses of integration

14.1. Improper integration

Definition. Suppose $f : [a, \infty) \to \mathbb{R}$ is integrable (and therefore bounded) on every interval of the form $[a, R]$, and further, as $R \to \infty$, we have $\int_a^R f \to \ell$.

Then we say that the integral $\int_a^\infty f$ exists (or converges), and its value is ℓ. If this integral does not tend to a limit, we say that $\int_a^\infty f$ diverges.

We can similarly define the integral $\int_{-\infty}^a f$. If $\int_a^\infty f = \ell_1$ and $\int_{-\infty}^a f = \ell_2$, we can write

$$\int_{-\infty}^{\infty} f = \ell_1 + \ell_2$$

Note that this last condition is *not* the same as saying that $\lim_{R\to\infty} \int_{-R}^R f$ exists. For this two-sided improper integral to exist, we need the stronger condition that the one-sided improper integrals exist on either side. For example, consider $f(x) = x$. Clearly $\int_{-R}^R f = 0$, but this function is not improper integrable. For example, consider

$$\int_1^\infty \frac{\mathrm{d}x}{x^k}$$

This converges if and only if $k > 1$. Indeed, if $k \neq 1$,

$$\int_1^R \frac{\mathrm{d}x}{x^k} = \frac{x^{1-k}}{1-k}\bigg|_1^R = \frac{R^{1-k} - 1}{1-k}$$

which is clearly finite in the limit if and only if $k > 1$. If $k = 1$, then we can find

$$\int_1^R \frac{\mathrm{d}x}{x} = \log R \to \infty$$

as expected. Note the following observations.

(1) $\frac{1}{\sqrt{x}}$ is continuous (and bounded) on $[\delta, 1]$ for all $\delta > 0$, and

$$\int_\delta^1 \frac{\mathrm{d}x}{\sqrt{x}} = 2\sqrt{x}\bigg|_\delta^1 = 2 - 2\sqrt{\delta}$$

So as $\delta \to 0$, this integral tends to 2. This integral is defined, even though the value of the function at zero is unbounded. So we commonly write

$$\int_0^1 \frac{\mathrm{d}x}{\sqrt{x}} = 2$$

(2) Similarly, we write

$$\int_0^1 \frac{\mathrm{d}x}{x} = \lim_{\delta\to 0}\int_\delta^1 \frac{\mathrm{d}x}{x} = \lim_{\delta\to 0} \log x\Big|_\delta^1 = \lim_{\delta\to 0}(\log 1 - \log\delta)$$

Since this limit does not exist, the integral $\int_0^1 \frac{\mathrm{d}x}{x}$ does not exist.

(3) If $f \geq 0$ and $g \geq 0$ for $x \geq a$, and $f(x) \leq kg(x)$, where k is a constant for $x \geq a$, then

$$\int_a^\infty g \text{ converges } \implies \int_a^\infty f \text{ converges, and } \int_a^\infty f \leq \int_a^\infty g$$

This is similar to the comparison test for series. First, note that $\int_a^R f \leq k\int_a^R g$. Further, $\int_a^R f$ is an increasing function of R since $f \geq 0$, and bounded above, since $\int_a^\infty g$ converges. Let

$$\ell = \sup_{R\geq a}\int_a^R f < \infty$$

Then we want to show that $\lim_{R\to\infty}\int_a^R f = \ell$. Given $\varepsilon > 0$, by the definition of the supremum $\exists R_0$ such that

$$\int_a^{R_0} f \geq \ell - \varepsilon$$

Thus for all $R \geq R_0$ we have

$$\int_a^R f \geq \int_a^{R_0} f \geq \ell - \varepsilon$$

Hence,

$$0 \leq \ell - \int_a^R f \leq \varepsilon$$

As an example, consider the integral

$$\int_0^\infty \exp\left(\frac{-x^2}{2}\right)\mathrm{d}x$$

Now, for $x \geq 1$, we can bound the integrand by $\exp\left(-\frac{x}{2}\right)$, and the integral of this bound is clearly bounded. Hence the original integral converges.

(4) If $\sum a_n$ converges, then $a_n \to 0$. However, with improper integrals, this is not necessarily the case. Consider a convergent series a_n, where $0 < a_n < 1$ for all n. Then define the function f defined by

$$f(n+r) = \begin{cases} 1 & \text{if } r < a_n \\ 0 & \text{otherwise} \end{cases}$$

where the input x is split into the integer part n and the remainder r. This function is essentially a sequence of rectangles of height 1 and width a_n, spaced so that each rectangle starts at an integer value of x. Clearly, we have

$$\int_0^n f = \sum_0^n a_n$$

where n is an integer. So the integral converges, but the integrand does not tend to zero.

14.2. Integral test for series convergence

Theorem. Let $f(x)$ be a positive decreasing function for $x \geq 1$. Then,

(1) The integral $\int_1^\infty f(x)\,\mathrm{d}x$ and the series $\sum_1^\infty f(x)$ both converge or diverge. (Note that such a function is always Riemann integrable on a closed interval since it is bounded and decreasing.)

(2) As $n \to \infty$, $\sum_{r=1}^n f(r) - \int_1^n f(x)\,\mathrm{d}x$ tends to a limit ℓ such that $0 \leq \ell \leq f(1)$.

Proof. If $n - 1 \leq x \leq n$, then

$$f(n-1) \geq f(x) \geq f(n)$$

Hence,

$$f(n-1) \geq \int_{n-1}^n f(x)\,\mathrm{d}x \geq f(n)$$

Adding up such integrals, we get

$$\sum_1^{n-1} f(r) \geq \int_1^n f(x)\,\mathrm{d}x \geq \sum_2^n f(r)$$

Then the first claim is obvious. For the second claim, let

$$\phi(n) = \sum_1^n f(r) - \int_1^n f(x)\,\mathrm{d}x$$

Then, using the inequalities established above,

$$\phi(n) - \phi(n-1) = f(n) - \int_{n-1}^n f(x)\,\mathrm{d}x \leq 0$$

So ϕ is a decreasing sequence. Further,

$$0 \leq \phi(n) \leq f(1)$$

ϕ is bounded, so it converges to some limit ℓ. □

Example. First, consider the sum $\sum_1^\infty \frac{1}{n^k}$. By the integral test, this converges if and only if $k > 1$. As a more complicated example, consider $\sum_2^\infty \frac{1}{n \log n}$. Let $f(x) = \frac{1}{x \log x}$, and

$$\int_2^R \frac{\mathrm{d}x}{x \log x} = \log(\log x)\Big|_2^R$$

which diverges, so by the integral test the series diverges.

Corollary (Euler–Mascheroni Constant)**.** As $n \to \infty$,

$$\sum_1^n \frac{1}{n} - \int_1^n \frac{1}{n} = 1 + \frac{1}{2} + \cdots + \frac{1}{n} - \log n \to \gamma$$

where $\gamma \in [0, 1]$. This is known as the Euler–Mascheroni constant. It is unknown whether γ is irrational.

14.3. Piecewise continuous functions

Definition. A function $f : [a, b] \to \mathbb{R}$ is piecewise continuous if there is a dissection $\mathcal{D}$ such that f is continuous on all intervals defined by this dissection, and that the one-sided limits

$$\lim_{x \to x_{j-1}^+} f(x); \quad \lim_{x \to x_{j-1}^-} f(x)$$

exist.

We can extend the class of Riemann integrable functions to include piecewise continuous functions as well. This is true since we use this dissection to construct the upper and lower sums. The one-sided limits are here to ensure that the function is bounded near these discontinuities. We might now ask how large the discontinuity set is allowed to be in order for f to still be Riemann integrable. As we have seen from examples before, it is possible to have a function which has countably many discontinuity points, but is still Riemann integrable.