In The Labyrinth

Ryan Joo Rui An

The mathematician does not study mathematics because it is useful; he studies it because he delights in it and he delights in it because it is beautiful.

— Henri Poincaré (1854–1912)
French mathematician and theoretical physicist

This is (still!) an incomplete draft. Please send corrections and comments to `ryanjooruian18@gmail.com`, or pull-request at `https://github.com/Ryanjoo18/latex`.

Last updated April 19, 2024.

Preface

The origin of the title “In The Labyrinth” comes from the following anecdote:

> Many students struggle with mathematics because it feels like wandering aimless in a labyrinth, unsure what to do to excel in Mathematics.

This book is my attempt to simplify and clarify topics in Mathematics at the undergraduate (and hopefully, someday, graduate) level.

About the author

At this moment of writing, I am a high school student working on my A Level studies in Singapore. I have about 11 years of participating in Mathematics competitions, including three years of experience in mental arithmetic and the rest few years in Mathematics Olympiad.

About this book

This book mainly serves as my notes when studying Mathematics at the university level. Feel free to refer to it too.

Acknowledgements

I am indebted to countless people for this work. Here is a partial (surely incomplete) list.

- Lecture notes by the University of Oxford, which can be found here.
- Lecture notes on MIT OpenCourseWare, which can be found here.
- The authors of all the books I have referred to when writing this book.

Problem Solving

In his book “How to Solve It”, George Pólya outlined the following problem solving cycle:

1. Understand the problem

   Ask yourself the following questions:

   - What are you asked to find or show?
   - Can you restate the question in your own words?
   - What part/information of the question is (un)important?
   - Can you think of a picture or a diagram that might help you understand the question?
   - Is there enough information to enable you to find a solution?
   - Do you understand all the words used in stating the question?
   - Do you need to ask a question to get the answer?
   - Why might this problem be difficult/easy?

2. Devise a plan

   Possible strategies:

   - Draw pictures or diagrams.
   - Eliminate possibilities
   - Be systematic.
   - Solve a simpler version of the problem & consider special cases.
   - Guess and check. Trial and error. Guess and test.
   - Look for a pattern or patterns.
   - Make a list / write down keywords.

3. Carry out the plan

   Answering the question

   - Try to use the strategy chosen in step 2.
   - If this strategy does not work, try another one.

4. Check and expand

   Look back reviewing and checking your results. Ask yourself the following questions:

   - Did you answer the question? Is your result reasonable?
   - What would change if you change the question a bit?
   - Is there a better/more interesting version of the question?
   - Is there another way of doing the problem which may be simpler?
   - Can the question or method be generalised to be useful for future problems?

Contents

Acronyms

H2FM H2 Further Mathematics. 37

H3M H3 Mathematics. 29, 32, 33, 36, 41–43, 65

Part I

Preliminaries

1 Mathematical Reasoning and Logic

§1.1 Logical statements and notation

It is useful to be familiar with the following terminology.

- A **definition** is a precise and unambiguous description of the meaning of a mathematical term. It characterises the meaning of a word by giving all the properties and only those properties that must be true.
- A **theorem** is a true mathematical statement that can be proven mathematically. In a mathematical paper, the term theorem is often reserved for the most important results.
- A **lemma** is a minor result whose sole purpose is to help in proving a theorem. It is a stepping stone on the path to proving a theorem. Very occasionally lemmas can take on a life of their own.
- A **corollary** is a result in which the (usually short) proof relies heavily on a given theorem. We often say that "this is a corollary of Theorem A".
- A **proposition** is a proven and often interesting result, but generally less important than a theorem.
- A **conjecture** is a statement that is unproved, but is believed to be true.
- An **axiom** is a statement that is assumed to be true without proof. These are the basic building blocks from which all theorems are proven.
- An **identity** is a mathematical expression giving the equality of two (often variable) quantities.
- A **paradox** is a statement that can be shown, using a given set of axioms and definitions, to be both true and false. Paradoxes are often used to show the inconsistencies in a flawed theory.

§1.1.1 Notation

A **proposition** is a sentence which has exactly one truth value, i.e. it is either true or false, but not both and not neither. A proposition is denoted by uppercase letters such as P and Q. If the proposition P depends on a variable x, it is sometimes helpful to denote it by $P(x)$.

We can so some algebra on propositions, which include

(i) **equivalence**, denoted by $P \equiv Q$, which means P and Q are logically equivalent statements;

(ii) **conjunction**, denoted by $P \wedge Q$, which means "P and Q";

(iii) **disjunction**, denoted by $P \vee Q$, which means "P or Q";

(iv) **negation**, denoted by $\neg P$, which means "not P".

Here are some useful properties when handling logical statements. You can easily prove all of them using truth tables.

- Double negation law:
$$P \equiv \neg(\neg P)$$
- Commutative property:
$$P \wedge Q \equiv Q \wedge P, \quad P \vee Q \equiv Q \vee P$$
- Associative property for conjunction:
$$(P \wedge Q) \wedge R \equiv P \wedge (Q \wedge R)$$
- Associative property for disjunction:
$$(P \vee Q) \vee R \equiv P \vee (Q \vee R)$$
- Distributive property for conjunction across disjunction:
$$P \wedge (Q \vee R) \equiv (P \wedge Q) \vee (P \wedge Q)$$
- Distributive property for disjunction across conjunction:
$$P \vee (Q \wedge R) \equiv (P \vee Q) \wedge (P \vee R)$$
- **De Morgan's Laws**:
$$\neg(P \vee Q) \equiv (\neg P \wedge \neg Q)$$
$$\neg(P \wedge Q) \equiv (\neg P \vee \neg Q)$$

Exercise 1.1.1

Assume that x is a fixed real number. What is the negation of the statement $1 < x < 2$?

Solution. The negation of $1 < x < 2$ is "it is not the case that $1 < x < 2$". However this is not useful.

Note that $1 < x < 2$ means $1 < x$ and $x < 2$. Let $P : 1 < x$ and $Q : x < 2$. Then the statement $1 < x < 2$ is $P \wedge Q$.

By De Morgan's Laws, we have $\neg(P \wedge Q) \equiv \neg P \vee \neg Q$.

The *Trichotomy Axiom of real numbers* states that given fixed real numbers a and b, exactly one of the statements $a < b, a = b, b < a$ is true. Hence $\neg P \equiv \neg(1 < x) \equiv (x \leq 1)$ and $\neg Q \equiv \neg(x < 2) \equiv (x \geq 2)$.

Thus
$$\neg(1 < x < 2) \equiv \neg(P \wedge Q) \equiv \neg P \vee \neg Q \equiv (1 \geq x) \vee (x \geq 2).$$

Therefore the negation of $1 < x < 2$ is logically equivalent to the statement $x \leq 1$ or $x \geq 2$. □

Exercise 1.1.2

Assume that n is a fixed positive integer. Find a useful denial of the statement

$$n = 2 \text{ or } n \text{ is odd}.$$

Solution. Using De Morgan's Laws,

$$\begin{aligned} \neg[(n = 2) \vee (n \text{ is odd})] &\equiv \neg(n = 2) \wedge \neg(n \text{ is odd}) \\ &\equiv (n \neq 2) \wedge (n \text{ is even}) \end{aligned}$$

where we are using the fact that every integer is either even or odd, but not both.

Thus a useful denial of the given statement is: n is an even integer other than 2. □

§1.1.2 If, only if, $\implies$

Implication is denoted by $P \implies Q$, which means "P implies Q", i.e. if P holds then Q also holds. It is equivalent to saying "If P then Q". The only case when $P \implies Q$ is false is when the hypothesis P is true and the conclusion Q is false.

$P \implies Q$ is known as a **conditional statement**. P is known as the **hypothesis**, Q is known as the **conclusion**.

Statements of this form are probably the most common, although they may sometimes appear quite differently. The following all mean the same thing:

(i) if P then Q;

(ii) P implies Q;

(iii) P only if Q;

(iv) P is a sufficient condition for Q;

(v) Q is a necessary condition for P.

The **converse** of $P \implies Q$ is given by $Q \implies P$; both are not logically equivalent.

The **inverse** of $P \implies Q$ is given by $\neg P \implies \neg Q$, i.e. the hypothesis and conclusion of the statement are both negated.

The **contrapositive** of $P \implies Q$ is given by $\neg Q \implies \neg P$; both are logically equivalent.

How to prove: To prove $P \implies Q$, start by assuming that P holds and try to deduce through some logical steps that Q holds too. Alternatively, start by assuming that Q does not hold and show that P does not hold (that is, we prove the contrapositive).

§1.1.3 If and only if, iff, $\iff$

Bidirectional implication is denoted by $P \iff Q$, which means both $P \implies Q$ and $Q \implies P$. We can read this as "P if and only if Q". The letters "iff" are also commonly used to stand for 'if and only if'.

$$P \iff Q \equiv (P \implies Q) \wedge (Q \implies P)$$

$P \iff Q$ is true exactly when P and Q have the same truth value.

$P \iff Q$ is known as a **biconditional statement**.

These statements are usually best thought of separately as 'if' and 'only if' statements.

How to prove: To prove $P \iff Q$, prove the statement in both directions, i.e. prove both $P \implies Q$ and $Q \implies P$. Remember to make very clear, both to yourself and in your written proof, which direction you are doing.

§1.1.4 Quantifiers

The **universal quantifier** is denoted by $\forall$, which means "for all" or "for every". An universal statement has the form $\forall x \in X, P(x)$.

The **existential quantifier** is denoted by $\exists$, which means "there exists". An existential statement has the form $\exists x \in X, P(x)$, where X is known as the **domain**.

These are versions of De Morgan's laws for quantifiers:

$$\neg \forall x \in X, P(x) \equiv \exists x \in X, \neg P(x)$$

$$\neg \exists x \in X, P(x) \equiv \forall x \in X, \neg P(x)$$

Exercise 1.1.3

Find a useful denial of the statement

$$\text{for all real numbers } x, \text{ if } x > 2, \text{ then } x^2 > 4$$

Solution. In logical notation, this statement is $(\forall x \in \mathbb{R})[x > 2 \implies x^2 > 4]$.

$$\begin{aligned} \neg\{(\forall x \in \mathbb{R})[x > 2 \implies x^2 > 4]\} &\equiv (\exists x \in \mathbb{R})\neg[x > 2 \implies x^2 > 4] \\ &\equiv (\exists x \in \mathbb{R})\neg[(x \le 2) \vee (x^2 > 4)] \\ &\equiv (\exists x \in \mathbb{R})[(x > 2) \wedge (x^2 \le 4)] \end{aligned}$$

Therefore a useful denial of the statement is:

$$\text{there exists a real number } x \text{ such that } x > 2 \text{ and } x^2 \le 4.$$

□

Exercise 1.1.4

Negate surjectivity.

Solution. If $f : X \to Y$ is not surjective, then it means that there exists $y \in Y$ not in the image of X, i.e. for all x in X we have $f(x) \ne y$.

$$\begin{aligned} \neg \forall y \in Y, \exists x \in X, f(x) = y &\iff \exists y \in Y, \neg(\exists x \in X, f(x) = y) \\ &\iff \exists y \in Y, \forall x \in X, \neg(f(x) = y) \\ &\iff \exists y \in Y, \forall x \in X, f(x) \ne y \end{aligned}$$

□

How to prove: To prove a statement of the form $\forall x \in X$ s.t. $P(x)$', start the proof with 'Let $x \in X$.' or 'Suppose $x \in X$ is given.' to address the quantifier with an arbitrary x; provided no other assumptions about x are made during the course of proving $P(x)$, this will prove the statement for all $x \in X$.

How to prove: To prove a statement of the form $\exists x \in X$ s.t. $P(x)$, there is not such a clear steer about how to continue: you may need to show the existence of an x with the right properties; you may need to demonstrate logically that such an x must exist because of some earlier assumption, or it may be that you can show constructively how to find one; or you may be able to prove by contradiction, supposing that there is no such x and consequently arriving at some inconsistency.

Remark. Read from left to right, and as new elements or statements are introduced they are allowed to depend on previously introduced elements but cannot depend on things that are yet to be mentioned.

Remark. To avoid confusion, it is a good idea to keep to the convention that the quantifiers come first, before any statement to which they relate.

§1.2 Proofs

§1.2.1 Direct Proof

A direct proof of $P \implies Q$ is a series of valid arguments that start with the hypothesis P and end with the conclusion Q. It may be that we can start from P and work directly to Q, or it may be that we make use of P along the way.

§1.2.2 Proof by Contrapositive

To prove $P \implies Q$, we can instead prove $\neg Q \implies \neg P$.

Exercise 1.2.1

For every integer a, prove that if $3a^2+1$ is even, then a is odd.

Proof. We prove this by contrapositive.

Suppose a is not odd. So $a = 2k$ for some integer k. Then

$$3a^2+1 = 3(2k)^2+1 = 2(6k^2)+1.$$

Since $3a^2+1 = 2q+1$ for some integer q, hence $3a^2+1$ is odd. □

Exercise 1.2.2

For $m \in \mathbb{Z}$, prove that if $3 \mid m^2$ then $3 \mid m$.

Proof. We prove this by contrapositive.

Suppose $3 \nmid m$. We shall prove $3 \nmid m^2$.

Case 1: $m = 3k+1$

Then $m^2 = (3k+1)^2 = 3(3k^2+2k)+1$ so m^2 has remainder 1 when divided by 3, hence $3 \nmid m^2$.

Case 2: $m = 3k+2$

This case shall be left as an exercise. □

§1.2.3 Disproof by Counterexample

Providing a counterexample is the best method for refuting, or dispoving, a conjecture.

In seeking counterexamples, it is a good idea to keep the cases you consider simple, rather than searching randomly. It is often helpful to consider “extreme” cases; for example, something is zero, a set is empty, or a function is constant.

The counterexample must make the hypothesis a true statement, and the conclusion a false statement.

§1.2.4 Proof by Cases

You can sometimes prove a statement by:

1. Dividing the situation into cases which exhaust all the possibilities; and
2. Showing that the statement follows in all cases.

Remark. It is important to cover all the possibilities.

§1.2.5 Proof by Contradiction

To prove P by contradiction, suppose that P is false, i.e. $\neg P$. Similarly, to prove $P \implies Q$ by contradiction, suppose that Q is false, i.e. $P \wedge \neg Q$.

Then show through some logical reasoning that this leads to a contradiction or inconsistency. We may arrive at something that contradicts the hypothesis P, or something that contradicts the initial supposition that Q is not true, or we may arrive at something that we know to be universally false.

Exercise 1.2.3: Irrationality of $\sqrt{2}$

Prove that $\sqrt{2}$ is irrational.

Proof. We prove by contradiction. Suppose otherwise, that $\sqrt{2}$ is rational. Using the definition of rational numbers, we can write it as $\sqrt{2} = \frac{a}{b}$ for some $a, b \in \mathbb{Z}, b \neq 0$.

We also assume that $\frac{a}{b}$ is simplified to lowest terms, since that can obviously be done with any fraction. Notice that in order for $\frac{a}{b}$ to be in simplest terms, both a and b cannot be even; one or both must be odd, otherwise we could simplify the fraction further.

Squaring both sides gives us

$$a^2 = 2b^2.$$

Since RHS is even, LHS must also be even. Hence it follows that a is even. Let $a = 2k$ where $k \in \mathbb{Z}$. Substituting $a = 2k$ into the above equation and simplifying it gives us

$$b^2 = 2k^2.$$

This means that b^2 is even, from which follows again that b is even.

This is a contradiction, as we started out assuming that $\frac{a}{b}$ was simplified to lowest terms, and now it turns out that a and b both would be even. Hence proven. □

Exercise 1.2.4

For any integer n, prove that there is no integer $a > 1$ such that $a \mid n$ and $a \mid (n+1)$.

Proof. Suppose there is an integer n and integer $a > 1$ such that $a \mid n$ and $a \mid (n+1)$.

Then $n = ak$ and $n + 1 = ah$ for some integers k and h.

$$ak + 1 = ah \implies 1 = a(h - k) \implies a \mid 1 \implies a = \pm 1$$

This contradicts $a > 1$.

Hence we conclude that, for any n, there is no integer $a > 1$ such that $a \mid n$ and $a \mid (n+1)$. □

§1.2.6 Proof of Uniqueness

$\exists!$ means “there exists a unique”.

To prove uniqueness, we can do one of the following:

- Assume $\exists x, y \in S$ such that $P(x) \wedge P(y)$ is true and show $x = y$.
- Argue by assuming that $\exists x, y \in S$ are distinct such that $P(x) \wedge P(y)$, then derive a contradiction.

To prove uniqueness and existence, we also need to show that $\exists x \in S$ s.t. $P(x)$ is true.

§1.2.7 Proof of Existence

To prove existential statements, we can adopt two approaches:

1. Constructive proof (direct proof)
2. Non-constructive proof (indirect proof)

Constructive Proof

To prove statements of the form $\exists x \in X$ s.t. $P(x)$, find or construct *a specific example* for x. To prove statements of the form $\forall y \in Y$, $\exists x \in X$ s.t. $P(x,y)$, construct example for x *in terms of* y (since x is dependent on y).

In both cases, you have to justify that your example x

1. belongs to the domain X, and
2. satisfies the condition P.

Exercise 1.2.5

Prove that we can find 100 consecutive positive integers which are all composite numbers.

Proof. We can prove this existential statement via constructive proof.

Our goal is to find integers $n, n+1, n+2, \dots, n+99$, all of which are composite.

Take $n = 101! + 2$. Then n has a factor of 2 and hence is composite. Similarly, $n + k = 101! + (k+2)$ has a factor $k+2$ and hence is composite for $k = 1, 2, \dots, 99$.

Hence the existential statement is proven. □

Exercise 1.2.6

Prove that for all rational numbers p and q with $p < q$, there is a rational number x such that $p < x < q$.

Proof. We prove this by construction. Our goal is to find such a rational x *in terms of* p *and* q.

We take the average. Let $x = \dfrac{p+q}{2}$ which is a rational number.

Since $p < q$,

$$x = \frac{p+q}{2} < \frac{q+q}{2} = q \implies x < q$$

Similarly,

$$x = \frac{p+q}{2} > \frac{p+p}{2} = p \implies p < x$$

Hence we have shown the existence of rational number x such that $p < x < q$.

Remark. For this type of question, there are two parts to prove: firstly, x satisfies the given statement; secondly, x is within the domain (for this question we do not have to prove x is rational since $\mathbb{Q}$ is closed under addition).

□

Exercise 1.2.7

Prove that for all rational numbers p and q with $p < q$, there is an irrational number r such that $p < r < q$.

Proof. We prove this by construction. Similarly, our goal is to find an irrational r in terms of p and q. Note that we cannot simply take $r = \frac{p+q}{2}$; a simple counterexample is the case $p = -1, q = 1$ where $r = 0$ is clearly not irrational.

Since p lies in between p and q, let $r = p + c$ where $0 < c < q - p$. Since $c < q - p$, we have $c = \frac{q-p}{k}$ for some $k > 1$; to make c irrational, we take k to be irrational.

Take $r = p + \frac{q-p}{\sqrt{2}}$. We need to show r is irrational and $p < r < q$.

Part 1: $p < r < q$

Since $q < p$, $r = p + (\text{positive number}) > p$. On the other hand, $\frac{q-p}{\sqrt{2}} < q - p$ so $r < p + (q - p) = q$.

Part 2: r is irrational

We prove by contradiction. Suppose r is rational. We have $\sqrt{2} = \frac{q-p}{r-p}$. Since p, q, r are all rational (and $r - p \neq 0$), RHS is rational. This implies that LHS is rational, i.e. $\sqrt{2}$ is rational, a contradiction. □

Non-constructive Proof

Use when specific examples are not easy or not possible to find or construct. Make arguments why such objects have to exist. May need to use proof by contradiction. Use definition, axioms or results that involve existential statements.

Exercise 1.2.8

Prove that every integer greater than 1 is divisible by a prime.

Proof. If n is prime, then we are done as $n \mid n$.

If n is not prime, then n is composite. So n has a divisor d_1 such that $1 < d_1 < n$. If d_1 is prime then we are done as $d_1 \mid n$. If d_1 is not prime then d_1 is composite, has divisor d_2 such that $1 < d_2 < n$.

If d_2 is prime, then we are done as $d_2 \mid d_1$ and $d_1 \mid n$ imply $d_2 \mid n$. If d_2 is not prime then d_2 is composite, has divisor d_3 such that $1 < d_3 < d_2$.

Continuing in this manner after k times, we will get

$$1 < d_k < d_{k-1} < \cdots < d_2 < d_1 < n$$

where $d_i \mid n$ for all i.

This process must stop after finite steps, as there can only be a finite number of d_i's between 1 and n. On the other hand, the process will stop only if there is a d_i which is a prime.

Hence we conclude that there must be a divisor d_i of n that is prime. □

Remark. This proof is also known as *proof by infinite descent*, a method which relies on the well-ordering principle of the positive integers.

Exercise 1.2.9

Prove that the equation $x^2 + y^2 = 3z^2$ has no solutions (x, y, z) in integers where $z \neq 0$.

Proof. Suppose we have a solution (x, y, z). Without loss of generality, we may assume that $z > 0$. By the least integer principle, we may also assume that our solution has z minimal. Taking remainders modulo 3, we see that

$$x^2 + y^2 \equiv 0 \pmod{3}$$

Recalling that squares may only be congruent to 0 or 1 modulo 3, we conclude that

$$x^2 \equiv y^2 \equiv 0 \implies x \equiv y \equiv 0 \pmod{3}$$

Writing $x = 3a$ and $y = 3b$ we obtain

$$9a^2 + 9b^2 = 3z^2 \implies 3(a^2 + b^2) = z^2 \implies 3 \mid z^2 \implies 3 \mid z$$

Now let $z = 3c$ and cancel 3's to obtain

$$a^2 + b^2 = 3c^2.$$

We have therefore constructed another solution $(a, b, c) = \left(\frac{x}{3}, \frac{y}{3}, \frac{z}{3}\right)$ to the original equation. However $0 < c < z$ contradicts the minimality of z. □

Exercise 1.2.10

An odd prime p may be written as a sum of two squares if and only $p \equiv 1 \pmod 4$.

Proof. We again use the method of descent, though this time *constructively*.

$(\implies)$ If $p = x^2 + y^2$, then both x and y are non-zero modulo p. Taking Legendre symbols, we see that

$$1 = \left(\frac{x^2}{p}\right) = \left(\frac{-y^2}{p}\right) = \left(\frac{-1}{p}\right) \implies p \equiv 1 \pmod 4$$

$(\impliedby)$ Suppose that p is a prime congruent to 1 modulo 4. We must show that there exist integers x, y such that $x^2 + y^2 = p$. We do this by descent:

1. Modulo p, the congruence $x^2 + 1 \equiv 0$ has a solution x since -1 is a quadratic residue. By taking $y = 1$, we may therefore assume the existence of a solution to an equation $x^2 + y^2 = mp$ for some integer $1 \le m < p$. If $m = 1$ we are done. Otherwise ...

2. Define
$$\begin{cases} u \equiv x \pmod m \\ v \equiv y \pmod m \end{cases} \qquad \text{such that } |u|, |v| \le \frac{m}{2}.$$
Since $xu + yv$, $xv - yu$ and $u^2 + v^2$ are all divisible by m, we may divide the identity
$$(u^2 + v^2)(x^2 + y^2) = (xu + yv)^2 + (xv - yu)^2$$
by m^2 to obtain an equation in integers:
$$kp = \left(\frac{xu + yv}{m}\right)^2 + \left(\frac{xv - yu}{m}\right)^2 \quad \text{where } k = \frac{u^2 + v^2}{m} \le \frac{m}{2}$$

3. We have therefore constructed an integer solution to $X^2 + Y^2 = kp$ with $k < m$. If $k \ge 2$, simply repeat the process from step 2: by descent, we must eventually reach $k = 1$.

□

§1.2.8 Pigeonhole Principle

Theorem 1.2.1 (Pigeonhole Principle (naive))**.** If m objects are placed into n boxes and $m > n$, then at least one box must contain more than one object.

Theorem 1.2.2 (Pigeonhole Principle (general))**.** If more than $k \cdot n$ objects are placed into n boxes, then at least one box must contain more than k objects.

§1.2.9 Proof by Mathematical Induction

Induction is an extremely powerful method of proof used throughout mathematics. It deals with infinite families of statements which come in the form of lists. The idea behind induction is in showing how each statement follows from the previous one on the list – all that remains is to kick off this logical chain reaction from some starting point.

Theorem 1.2.3 (Principle of Mathematical Induction (PMI))**.** Let $P(n)$ be a family of statements indexed by $\mathbb{Z}^+$. Suppose that

(i) (**base case**) $P(1)$ is true and

(ii) (**inductive step**) for all $k \in \mathbb{Z}^+$, $P(k) \implies P(k+1)$.

Then $P(n)$ is true for all $n \in \mathbb{Z}^+$.

Using logic notation, this is written as

$$\{P(1) \wedge (\forall n \in \mathbb{Z}^+)[P(k) \implies P(k+1)]\} \implies (\forall n \in \mathbb{Z}^+)P(n)$$

Induction is often visualised like toppling dominoes. The inductive step (ii) corresponds to placing each domino sufficiently close that it will be hit when the previous one falls over, and base case (i) corresponds to knocking over the first one.

$$P(1) \implies P(2) \implies \cdots \implies P(k) \implies P(k+1) \implies \cdots$$

Exercise 1.2.11

Prove that for any $n \in \mathbb{Z}^+$,

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

Proof. Let $P(n)$ be the statement $\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$.
Clearly $P(1)$ holds because for $n = 1$, the sum on the LHS is 1 and the expression on the RHS is also 1.
Now suppose $P(n)$ holds. Then we have

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

Adding $n+1$ to both sides,

$$\begin{aligned}\sum_{k=1}^{n+1} k &= \frac{n(n+1)}{2} + (n+1) \\ &= \frac{(n+1)(n+2)}{2} \\ &= \frac{(n+1)[(n+1)+1]}{2}\end{aligned}$$

thus $P(n+1)$ is true.
By PMI, $P(n)$ is true for all $n \in \mathbb{Z}^+$. □

Remark. Do not write $P(n) = \frac{n(n+1)}{2}$, as $P(n)$ is a statement, not an expression (which does not have truth values).

A corollary of induction is if the family of statements holds for $n \geq N$, rather than necessarily $n \geq 0$:

Corollary 1.2.1. Let N be an integer and let $P(n)$ be a family of statements indexed by integers $n \geq N$. Suppose that

(i) (**base case**) $P(N)$ is true and

(ii) (**inductive step**) for all $k \geq N$, $P(k) \implies P(k+1)$.

Then $P(n)$ is true for all $n \geq N$.

Proof. This follows directly by applying the above theorem to the statement $Q(n) = P(n+N)$ for $n \in N$. □

Strong Induction

Another variant on induction is when the inductive step relies on some earlier case(s) but not necessarily the immediately previous case. This is known as **strong induction**:

Theorem 1.2.4 (Strong Form of Induction)**.** Let $P(n)$ be a family of statements indexed by the natural numbers. Suppose that

(i) (**base case**) $P(1)$ is true and

(ii) (**inductive step**) for all $m \in \mathbb{Z}^+$, if for integers k with $1 \leq k \leq m$, $P(k)$ is true then $P(m+1)$ is true.

Then $P(n)$ is true for all $n \in \mathbb{N}$.

Using logic notation, this is written as

$$\{P(1) \wedge (\forall m \in \mathbb{Z}^+)[P(1) \wedge P(2) \wedge \cdots \wedge P(m) \implies P(m+1)]\} \implies (\forall n \in \mathbb{Z}^+)P(n)$$

Proof. We can this it to an instance of "normal" induction by defining a related family of statements $Q(n)$.

Let $Q(n)$ be the statement "$P(k)$ holds for $k = 0, 1, \ldots, n$". Then the conditions for the strong form are equivalent to

(i) $Q(0)$ holds and

(ii) for any n, if $Q(n)$ is true then $Q(n+1)$ is also true.

It follows by induction that $Q(n)$ holds for all n, and hence $P(n)$ holds for all n. □

The following example illustrates how the strong form of induction can be useful:

Example 1.2.1: Fundamental Theorem of Arithmetic

Every natural number greater than 1 may be expressed as a product of one or more prime numbers.

Proof. Let $P(n)$ be the statement that n may be expressed as a product of prime numbers.

Clearly $P(2)$ holds, since 2 is itself prime.

Let $n \geq 2$ be a natural number and suppose that $P(m)$ holds for all $m < n$.

- If n is prime then it is trivially the product of the single prime number n.
- If n is not prime, then there must exist some $r, s > 1$ such that $n = rs$. By the inductive hypothesis, each of r and s can be written as a product of primes, and therefore $n = rs$ is also a product of primes.

Thus, whether n is prime or not, we have have that $P(n)$ holds. By strong induction, $P(n)$ is true for all natural numbers. That is, every natural number greater than 1 may be expressed as a product of one or more primes. □

Cauchy Induction

Theorem 1.2.5 (Cauchy Induction)**.** Let $P(n)$ be a family of statements indexed by $\mathbb{Z}^+_{\geq 2}$. Suppose that

(i) (**base case**) $P(2)$ is true and

(ii) (**inductive step**) for all $k \in \mathbb{Z}^+$, $P(k) \implies P(2k)$ and $P(k) \implies (k-1)$.

Then $P(n)$ is true for all $n \in \mathbb{Z}^+_{\geq 2}$.

Exercise 1.2.12

Using Cauchy Induction, prove the AM–GM Inequality for n variables, which states that for positive reals $a_1, a_2, \ldots a_n$,

$$\frac{a_1 + a_2 + \cdots + a_n}{n} \geq \sqrt[n]{a_1 a_2 \cdots a_n}.$$

Proof. Let $P(n)$ be $\frac{a_1+a_2+\cdots+a_n}{n} \geq \sqrt[n]{a_1 a_2 \cdots a_n}$.
Base case $P(2)$ is true because

$$\frac{a_1 + a_2}{2} \geq \sqrt{a_1 a_2} \iff (a_1 + a_2)^2 \geq 4a_1 a_2 \iff (a_1 - a_2)^2 \geq 0$$

Next we show that $P(n) \implies P(2n)$, i.e. if AM–GM holds for n variables, it also holds for $2n$ variables:

$$\begin{aligned}
\frac{a_1 + a_2 + \cdots + a_{2n}}{2n} &= \frac{\frac{a_1+a_2+\cdots+a_n}{n} + \frac{a_{n+1}+a_{n+2}+\cdots+a_{2n}}{n}}{2} \\
\frac{\frac{a_1+a_2+\cdots+a_n}{n} + \frac{a_{n+1}+a_{n+2}+\cdots+a_{2n}}{n}}{2} &\geq \frac{\sqrt[n]{a_1 a_2 \cdots a_n} + \sqrt[n]{a_{n+1} a_{n+2} \cdots a_{2n}}}{2} \\
\frac{\sqrt[n]{a_1 a_2 \cdots a_n} + \sqrt[n]{a_{n+1} a_{n+2} \cdots a_{2n}}}{2} &\geq \sqrt{\sqrt[n]{a_1 a_2 \cdots a_n}\,\sqrt[n]{a_{n+1} a_{n+2} \cdots a_{2n}}} \\
\sqrt{\sqrt[n]{a_1 a_2 \cdots a_n}\,\sqrt[n]{a_{n+1} a_{n+2} \cdots a_{2n}}} &= \sqrt[2n]{a_1 a_2 \cdots a_{2n}}
\end{aligned}$$

The first inequality follows from n-variable AM–GM, which is true by assumption, and the second inequality follows from 2-variable AM–GM, which is proven above.

Finally we show that $P(n) \implies P(n-1)$, i.e. if AM–GM holds for n variables, it also holds for $n-1$ variables. By n-variable AM–GM, $\frac{a_1+a_2+\cdots+a_n}{n} \geq \sqrt[n]{a_1 a_2 \cdots a_n}$ Let $a_n = \frac{a_1+a_2+\cdots+a_{n-1}}{n-1}$ Then we have

$$\frac{a_1 + a_2 + \cdots + a_{n-1} + \frac{a_1+a_2+\cdots+a_{n-1}}{n-1}}{n} = \frac{a_1 + a_2 + \cdots + a_{n-1}}{n-1}$$

So,

$$\begin{aligned}
\frac{a_1 + a_2 + \cdots + a_{n-1}}{n-1} &\geq \sqrt[n]{a_1 a_2 \cdots a_{n-1} \cdot \frac{a_1 + a_2 + \cdots + a_{n-1}}{n-1}} \\
\Rightarrow \left(\frac{a_1 + a_2 + \cdots + a_{n-1}}{n-1}\right)^n &\geq a_1 a_2 \cdots a_{n-1} \cdot \frac{a_1 + a_2 + \cdots + a_{n-1}}{n-1} \\
\Rightarrow \left(\frac{a_1 + a_2 + \cdots + a_{n-1}}{n-1}\right)^{n-1} &\geq a_1 a_2 \cdots a_{n-1} \\
\Rightarrow \frac{a_1 + a_2 + \cdots + a_{n-1}}{n-1} &\geq \sqrt[n-1]{a_1 a_2 \cdots a_{n-1}}
\end{aligned}$$

By Cauchy Induction, this proves the AM–GM inequality for n variables. □

Other Variations

Apart from proving $P(n)$ indexed by $\mathbb{Z}^+$, we can also use PMI to prove statements of the form

- $(\forall n \in \mathbb{Z}) P(n)$
  Base case: $P(0)$
  Inductive step: $(\forall k \in \mathbb{Z}_{\geq 0}) P(k) \implies P(k+1)$ and $(\forall k \in \mathbb{Z}_{\leq 0}) P(k) \implies P(k-1)$

$$\cdots \impliedby P(-n) \impliedby \cdots \impliedby P(-1) \impliedby P(0) \implies P(1) \implies \cdots \implies P(n) \implies \cdots$$

- $(\forall n \in \mathbb{Q}) P(n)$
  Base case: $P(0)$
  Inductive step: $P(x) \implies P(-x)$ and $P\left(\frac{a}{b}\right) \implies P\left(\frac{a+1}{b}\right)$ and $P\left(\frac{a}{b}\right) \implies P\left(\frac{a}{b+1}\right)$

§1.2.10 Symmetry Principle

§1.2.11 Combinatorial Arguments and Proofs

Exercises

Some of the exercise problems here are from the "Number and Proofs" topic of H3 Mathematics, so the reader is assumed to have some basic knowledge in Number Theory, in particular modular arithmetic.

Problem 1. Let a, b be integers, not both 0. Prove that $\gcd(a+b, a-b) \le \gcd(2a, 2b)$.

Proof. Direct proof.

Let $e = \gcd(a+b, a-b)$. Then $e \mid (a+b)$ and $e \mid (a-b)$. So

$$e \mid (a+b)+(a-b) \implies e \mid 2a$$

and

$$e \mid (a+b)-(a-b) \implies e \mid 2b$$

This implies e is a common divisor of $2a$ and $2b$. So $e \le \gcd(2a, 2b)$. □

Problem 2 (Division Algorithm)**.** Let c and d be integers, not both 0. If q and r are integers such as $c = dq + r$, then $\gcd(c, d) = \gcd(d, r)$.

Proof. Let $m = \gcd(c, d)$ and $n = \gcd(d, r)$. To prove $m = n$, we will show $m \le n$ and $n \le m$.

(i) Show $n \le m$

Since $n = \gcd(d, r)$, $n \mid d$ and $n \mid r$. There exists integers x and y such that $d = nx$ and $r = ny$.

From $c = dq + r$, we have $c = (nx)q + ny = n(xq + y)$ thus $n \mid c$. n is a common divisor of c and d, so $n \le \gcd(c, d)$. Hence $n \le m$.

(ii) Show $m \le n$

This is left as an exercise.

□

Problem 3 (Euclid's Lemma)**.** Let a, b, c be any integers. If $a \mid bc$ and $\gcd(a, b) = 1$, then $a \mid c$.

Proof. Since $a \mid bc$, $bc = ak$ for some $k \in \mathbb{Z}$.

Since $\gcd(a, b) = 1$,

$$\begin{aligned} ax + by &= 1 \quad \text{for some } x, y \in \mathbb{Z} \\ cax + cby &= c \\ acx + aky &= c \\ a(cx + ky) &= c \end{aligned}$$

thus $a \mid c$. □

Problem 4. Let a and b be integers, not both 0. Show that $\gcd(a, b)$ is the smallest possible positive linear combination of a and b. (i.e. There is no positive integer $c < \gcd(a, b)$ such that $c = ax + by$ for some integers x and y.)

Proof. Prove by contradiction.

Suppose there is a positive integer $c < \gcd(a, b)$ such that $c = ax + by$ for some integers x and y.

Let $d = \gcd(a, b)$. Then $d \mid a$ and $d \mid b$, and hence $d \mid ax + by$. This means $d \mid c$.

Since c is positive, this implies $\gcd(a, b) = d \le c$. This contradicts $c < \gcd(a, b)$.

Hence we conclude that there is no positive integer $c < \gcd(a, b)$ such that $c = ax + by$ for some integers x and y. □

Problem 5. Use the Unique Factorisation Theorem to prove that, if a positive integer n is not a perfect square, then $\sqrt{n}$ is irrational.

[The Unique Factorisation Theorem states that every integer $n > 1$ has a unique standard factored form, i.e. there is exactly one way to express $n = p_1^{k_1} p_2^{k_2} \cdots p_t^{k_t}$ where $p_1 < p_2 < \cdots < p_t$ are distinct primes and $k_1, k_2, \ldots, k_t$ are some positive integers.]

Proof. Prove by contradiction.

Suppose n is not a perfect square and $\sqrt{n}$ is rational.

Then $\sqrt{n} = \frac{a}{b}$ for some integers a and b. Squaring both sides and clearing denominator gives

$$nb^2 = a^2. \tag{$*$}$$

Consider the standard factored forms of n, a and b:

$$n = p_1^{k_1} p_2^{k_2} \cdots p_t^{k_t}$$

$$a = q_1^{e_1} q_2^{e_2} \cdots q_u^{e_u} \implies a^2 = q_1^{2e_1} q_2^{2e_2} \cdots q_u^{2e_u}$$

$$b = r_1^{f_1} r_2^{f_2} \cdots r_v^{f_v} \implies b^2 = r_1^{2f_1} r_2^{2f_2} \cdots r_v^{2f_v}$$

i.e. the powers of primes in the standard factored form of a^2 and b^2 are all even integers.

This means the powers k_i of primes p_i in the standard factored form of n are also even by Unique Factorisation Theorem (UFT):

Note that all p_i appear in the standard factored form of a^2 with even power $2c_i$, because of $(*)$. By UFT, p_i must also appear in the standard factored form of nb^2 with the same even power $2c_i$.

If $p_i \nmid b$, then $k_i = 2c_i$ which is even. If $p_i \mid b$, then p_i will appear in b^2 with even power $2d_i$. So $k_i + 2d_i = 2c_i$, and hence $k_i = 2(c_i - d_i)$, which is again even.

Hence $n = p_1^{k_1} p_2^{k_2} \cdots p_t^{k_t} = \left(p_1^{\frac{k_1}{2}} p_2^{\frac{k_2}{2}} \cdots p_t^{\frac{k_t}{2}} \right)^2$.

Since $\frac{k_i}{2}$ are all integers, $p_1^{\frac{k_1}{2}} p_2^{\frac{k_2}{2}} \cdots p_t^{\frac{k_t}{2}}$ is an integer and n is a perfect square. This contradicts the given hypothesis that n is not a perfect square.

So we conclude that when a positive integer n is not a perfect square, then $\sqrt{n}$ is irrational. □

Problem 6 (Sieve of Eratosthenes)**.** If $p > 1$ is an integer and $n \mid p$ for each integer n for which $2 \le n \le \sqrt{p}$, then p is prime.

Proof. Prove by contrapositive.

Suppose that p is not prime, so it factors as $p = mn$ for $1 < m, n < p$.

Observe that it is not the case that both $m > \sqrt{p}$ and $n > \sqrt{p}$, because if this were true the inequalities would multiply to give $mn > \sqrt{p}\sqrt{p} = p$, which contradicts $p = mn$.

Therefore $m \le \sqrt{p}$ or $n \le \sqrt{p}$. Without loss of generality, say $n \le \sqrt{p}$. Then the equation $p = mn$ gives $n \mid p$, with $1 < n \le \sqrt{p}$. Hence it is not true that $n \nmid p$ for each integer n for which $2 \le n \le \sqrt{p}$. □

Problem 7 (H3M Specimen Q3, Fermat's Little Theorem)**.**

(i) Let p be an odd prime and let a be an integer not divisible by p.

  (a) Let T be the set of remainders for $a, 2a, \ldots, (p-1)a$, when divided by p. Show that $T = \{1, 2, \ldots, p-1\}$.

  (b) Hence prove that $a^{p-1} \equiv 1 \pmod{p}$.

(ii) Let x and y be two integers such that $x^5 + y^5$ is divisible by 5. Prove that $x^5 + y^5$ is divisible by 25.

Proof.

(i) (a) Let $S = \{1, 2, 3, \ldots, p-1\}$, the set of all non-zero positive remainders obtained when integers are divided by p.

Known fact: $p \nmid k$ for all $k \in S = \{1, 2, 3, \ldots, p-1\}$.

Given that T is the set of remainders when $a, 2a, 3a, \ldots, (p-1)a$ are divided by p.

Clearly, $T \subseteq S \cup \{0\}$.

Claim 1: $0 \notin T$.

Proof. Prove by contradiction.

Suppose $0 \in T$. Then $p \mid ka$ for some $k \in S = \{1, 2, 3, \ldots, p-1\}$. Since p is prime and $p \nmid a$, we apply Euclid's Lemma to conclude that $p \mid k$, which contracts Fact 1.]

Thus $T \subseteq S$. □

Claim 2: $T = S$ itself.

Proof. Prove by contradiction.

Suppose, on the contrary, that $T \neq S$.

Then, $T \subset S$ (i.e. T is a proper subset of S).

Since the sets are finite sets, $n(T) < n(S) = p - 1$. By the Pigeonhole Principle, there are (at least) two distinct ia and ja (from the list of $p-1$ terms: $a, 2a, 3a, \ldots, (p-1)a$ – the "pigeons"), where $1 \leq i \neq j \leq p-1$ that share the same remainder when divided p. The "holes" are the elements in T; here we get less holes: $n(T) < p - 1$ based on our (wrong) assumption.

$$\begin{aligned} ia &\equiv ja \pmod{p} \\ ia - ja &\equiv 0 \pmod{p} \\ (i-j)a &\equiv 0 \pmod{p} \end{aligned}$$

We can cancel a on both sides due to Euclid's lemma. Hence $i \equiv j \pmod{p}$.

Since both i and j belong to S, having them share the same remainder when divided by p means that they are actually the same. Thus $i = j$. This contradicts our initial choice of distinct ia and ja.

Hence $T = S = \{1, 2, 3, \ldots, p-1\}$. □

(b) Let

$$\begin{aligned} a \cdot 1 &\equiv r_1 \pmod{p} \\ a \cdot 2 &\equiv r_2 \pmod{p} \\ a \cdot 3 &\equiv r_3 \pmod{p} \\ &\vdots \\ a \cdot (p-1) &\equiv r_{p-1} \pmod{p} \end{aligned}$$

where $r_1, r_2, r_3, \ldots, r_{p-1}$ are distinct elements of $T = S = \{1, 2, 3, \ldots, p-1\}$.

So multiplying the LHS and RHS respectively of these congruence equations,

$$a^{p-1}(p-1)! \equiv r_1 r_2 r_3 \cdots r_{p-1} \pmod{p}$$

Since $r_1, r_2, r_3, \ldots, r_{p-1}$ is just a rearrangement of $1, 2, 3, \ldots, p-1$,

$$a^{p-1}(p-1)! \equiv 1 \cdot 2 \cdot 3 \cdots (p-1) \pmod{p}$$

or

$$a^{p-1}(p-1)! \equiv (p-1)! \pmod{p}$$

But $p \nmid (p-1)!$ so by Euclid's lemma,

$$a^{p-1} \equiv 1 \pmod{p}$$

as desired.

(ii) Prove by cases

Given 5 divides $x^5 + y^5$.

Case 1: either x or y is divisible by 5

WLOG, assume $5 \mid x$. Then $x = 5k$ for some integer k.

Then $x^5 = (5k)^5 = 5^2(5^3k^5) = 25t$ so $25 \mid x^5$.

Since we can write $y^5 = (x^5 + y^5) - x^5$, $5 \mid y^5$ so $5 \mid y$. We can then similarly show that $25 \mid y^5$.

Hence $25 \mid x^5 + y^5$.

Case 2: both x and y are not divisible by 5

Since 5 is a prime, by Fermat's Little Theorem, $x^5 \equiv x \pmod 5$ and $y^5 \equiv y \pmod 5$, so $x^5+y^5 \equiv x+y \pmod 5$.

Since $5 \mid x^5 + y^5$, we have also $5 \mid x + y$, i.e. $x + y = 5k$ for some integer k. We rewrite $y = 5k - x$.

Then by binomial expansion,

$$y^5 = (5k - x)^5 = \sum_{i=0}^{5} \binom{5}{i} (5k)^{5-i}(-x)^i$$

which gives $y^5 \equiv (-x)^5 \pmod{25}$ as all the other terms are divisible by 25.

Hence $x^5 + y^5 \equiv 0 \pmod{25}$.

□

Problem 8 (Euclid's proof)**.** There are infinitely many primes.

Proof. Prove by contradiction.

Suppose otherwise, that the list of primes is finite. Let $p_1, \dots, p_r$ be our finite list of primes. We want to show this is not the full list of the primes.

Consider the number

$$N = p_1 \cdots p_r + 1.$$

Since $N > 1$, it has a prime factor p. The prime p cannot be any of $p_1, \dots, p_r$ since N has remainder 1 when divided by each p_i. Therefore p is a prime not on our list, so the set of primes cannot be finite. □

Problem 9. If n is an integer, prove that 3 divides $n^3 - n$.

Proof. Prove by cases. This is done by partitioning $\mathbb{Z}$ according to remainders when divided by d (i.e. equivalence classes).

We prove the three cases: $n = 3k$, $n = 3k + 1$, and $n = 3k + 2$.

Case 1: $n = 3k$ for some integer k

Then

$$n^3 - n = (3k)^3 - (3k) = 3(9k^3 - k).$$

Since $9k^3 - k$ is an integer, $3 \mid n^3 - n$.

Case 2: $n = 3k + 1$ for some integer k

Then

$$n^3 - n = (3k + 1)^3 - (3k + 1) = 3(9k^3 + 9k^2 + 2k).$$

Since $9k^3 + 9k^2 + 2k$ is an integer, $3 \mid n^3 - n$.

Case 3: $n = 3k + 2$ for some integer k

The proof is similar and shall be left as an exercise. □

Problem 10 (H3M 2017 Q3).

(a) Consider integer solutions of the equation

$$1591x + 3913y = 9331.$$

Show that there is no solution with x prime.

(b) Let a, b, r and s be integers such that

$$ra + sb = 1.$$

(i) Prove that, if a and b are both factors of an integer n, then ab is a factor of n.

(ii) Given that any integers u and v, prove by construction that there is an integer x such that both

$$x \equiv u \pmod{a} \quad \text{and} \quad x \equiv v \pmod{b}.$$

Proof.

(a) First we find $\gcd(1591, 3913)$ using the Euclidean Algorithm.

$$\begin{aligned} 3913 &= 2 \times 1591 + 731 \\ 1591 &= 2 \times 731 + 129 \\ 731 &= 5 \times 129 + 86 \\ 129 &= 1 \times 86 + 43 \\ 86 &= 2 \times 43 + 0 \end{aligned}$$

Thus $\gcd(1591, 3913) = 43$. By Bezout's Lemma, there are integer solutions for $1591x + 3913y = 43$. Since $43 \mid 9331$, multiplying both sides by some constant, there are also integer solutions for $1591x + 3913y = 9331$.

To prove by contradiction, we assume that x is prime, and there exists some integer y such that $1591x + 3913y = 9331$. Dividing both sides by 43,

$$37x + 91y = 217. \tag{$\star$}$$

Observe that $7 \mid 91y$ and $7 \mid 217$, so $7 \mid 37x$.

Since $\gcd(7, 37) = 1$ so $7 \mid x$. By our assumption, x is a prime so $x = 7$.

Substituting $x = 7$ into $(\star)$, we get $y = -\dfrac{6}{13}$, which contradicts y being an integer.

Hence we conclude that x cannot be a prime.

(b) (i) If a and b are both factors of n, then we have $n = pa$ and $n = qb$ for some integers p and q. Given $ra + sb = 1$, we have

$$\begin{aligned} rna + snb &= n \\ r(qb)a + s(pa)b &= n \\ (rq + sp)ab &= n \end{aligned}$$

and hence ab is a factor of n.

(ii) Prove by construction.
Given that $ra + sb = 1$. Multiplying both sides by $v - u$ gives

$$\begin{aligned} ra(v-u) + sb(v-u) &= v - u \\ ra(v-u) + u &= sb(u-v) + v \end{aligned}$$

We define $x = ra(v-u) + u = sb(u-v) + v$. Then $x \equiv u \pmod{a}$ and $x \equiv v \pmod{b}$.

Remark. The above proof shows the *existence* of solution by a construction.

□

Problem 11 (H3M 2021)**.** Let $Q = \{1, 2, \dots, p-1\}$ for some prime p, and let there be N integers in Q whose cubes are congruent to 1 modulo p.

(a) Use the pigeonhole principle to prove that for each integer $x \in Q$ there is precisely one integer $y \in Q$ such that $xy \equiv 1 \pmod{p}$.

(b) Explain why the number of choices of integers $x, y, z \in Q$ such that $xyz \equiv 1 \pmod{p}$ is $(p-1)^2$.

(c) Use the principle of inclusion and exclusion to prove that the number of choices of three different integers $x, y, z \in Q$ such that $xyz \equiv 1 \pmod{p}$ is $(p-1)(p-4) + 2N$.

(d) Hence prove that $N \equiv (p-1)^2 \pmod{3}$.

(e) Given that $p \equiv 1 \pmod{3}$, prove that there is an integer $x \in Q$ such that $x^2 + x + 1 \equiv 0 \pmod{p}$.

Proof.

(a) Note that, by Quotient Remainder Theorem, every integer not divisible by p is congruent to an integer in Q modulo p, and no two integers in Q are congruent to each other modulo p.

We have two parts to prove: existence and uniqueness of inverse modulo p

Existence: prove by contradiction

Suppose there is an $x \in Q$ such that for all $y \in Q$, $xy \not\equiv 1 \pmod{p}$.

There are $p-1$ possible $y \in Q$, but there are less than $p-1$ possible $xy \in Q$ (since $xy \equiv 1 \pmod{p}$ is excluded).

By Pigeonhole Principle, there are two different $y_1, y_2 \in Q$ such that $xy_1 \equiv xy_2 (\not\equiv 1) \pmod{p}$. Then

$$p \mid xy_1 - xy_2 \implies p \mid x(y_1 - y_2) \implies p \mid y_1 - y_2 \implies y_1 \equiv y_2 \pmod{p} \implies y_1 = y_2$$

which is a contradiction. Hence every $x \in Q$ has a $y \in Q$ such that $xy \equiv 1 \pmod{p}$.

Uniqueness: prove by contradiction

Suppose there are two different $y_1, y_2 \in Q$ such that $xy_1 \equiv xy_2 (\equiv 1) \pmod{p}$.

The rest is similar to the above, and thus left as an exercise to the reader.

(b) Use combinatorics.

There are $p-1$ ways each to choose x and y.

By (a), there is only 1 way to choose $z \in Q$, the modular inverse of xy mod p, such that $(xy)z \equiv 1 \pmod{p}$.

Hence there is a total number of $(p-1)^2$ choices of x, y, z such that $xyz \equiv 1 \pmod{p}$.

(c) Let U contain all (x, y, z) such that $xyz \equiv 1 \pmod{p}$, A is a subset of U such that $x \equiv y \pmod{p}$, B is a subset of U such that $x \equiv z \pmod{p}$, C is a subset of U such that $y \equiv z \pmod{p}$.

Note that $A \cap B = A \cap C = B \cap C = A \cap B \cap C$ are all subsets of U such that $x \equiv y \equiv z \pmod{p}$, i.e. this subset of U contains all (x, x, x) such that $x^3 \equiv 1 \pmod{p}$.

We have $|U| = (p-1)^2$ from (b), $|A| = |B| = |C| = p-1$, and $|A \cap B \cap C| = N$.

By principle of inclusion and exclusion,

$$|A \cup B \cup C| = 3(p-1) - 2N.$$

To find the complement of $A \cup B \cup C$,

$$|U - (A \cup B \cup C)| = (p-1)^2 - \big(3(p-1) - 2N\big) = (p-1)(p-4) + 2N.$$

(d) From (c), the number of choices of three different $x, y, z \in Q$ such that $xyz \equiv 1 \pmod{p}$ is $(p-1)(p-4) + 2N$.

Since the number of combinations of such x, y, z are symmetrical, this number is divisible by 3. That is,

$$\begin{aligned}(p-1)(p-4) + 2N &\equiv 0 \pmod{3}\\ (p-1)(p-1) - N &\equiv 0 \pmod{3}\\ (p-1)^2 &\equiv N \pmod{3}\end{aligned}$$

(e) From (d), $N \equiv (p-1)^2 \equiv 0 \pmod 3 \implies 3 \mid N \implies N \geq 3$.
There are at least 3 different x such that $x^3 \equiv 1 \pmod p$. Choose such an $x \in Q$ such that $x \neq 1$.

$$x^3 - 1 \equiv 0 \pmod p$$

Factorising this gives

$$(x-1)(x^2+x+1) \equiv 0 \pmod p$$

Hence

$$p \mid (x-1)(x^2+x+1)$$

Since $\nmid x - 1$ as $x \in \{1, 2, \ldots, p-1\}$,

$$p \mid x^2 + x + 1$$

thus $x^2 + x + 1 \equiv 0 \pmod p$

□

Problem 12. Prove that for every pair of irrational numbers p and q such that $p < q$, there is an irrational x such that $p < x < q$.

Proof. Consider the average of p and q: $p < \frac{p+q}{2} < q$.

If $\frac{p+q}{2}$ is irrational, take $x = \frac{p+q}{2}$ and we are done.

If $\frac{p+q}{2}$ is rational, call it r, take the average of p and r: $p < \frac{p+r}{2} < r < q$. Since p is irrational and r is rational, $\frac{p+r}{2}$ is irrational. In this case, we take $x = \frac{3p+q}{4}$. □

Problem 13. Given n real numbers $a_1, a_2, \ldots, a_n$. Show that there exists an a_i $(1 \le i \le n)$ such that a_i is greater than or equal to the mean (average) value of the n numbers.

Proof. Prove by contradiction.

Let $\bar{a}$ denote the mean value of the n given numbers. Suppose $a_i < \bar{a}$ for all a_i. Then

$$\bar{a} = \frac{a_1 + a_2 + \cdots + a_n}{n} < \frac{\bar{a} + \bar{a} + \cdots + \bar{a}}{n} = \frac{n\bar{a}}{n} = \bar{a}.$$

We derive $\bar{a} < \bar{a}$, which is a contradiction.

Hence there must be some a_i such that $a_i > \bar{a}$. □

Problem 14. Prove that the following statement is false: there is an irrational number a such that for all irrational number b, ab is rational.

Thought process: prove the negation of the statement: for every irrational number a, there is an irrational number b such that ab is irrational.

Proving technique: constructive proof (note that we can consider multiple cases and construct more than one b)

Proof. Given an irrational number a, let us consider $\frac{\sqrt{2}}{a}$.

Case 1: $\frac{\sqrt{2}}{a}$ is irrational.

Take $b = \frac{\sqrt{2}}{a}$. Then $ab = \sqrt{2}$ which is irrational.

Case 2: $\frac{\sqrt{2}}{a}$ is rational.

Then the reciprocal $\frac{a}{\sqrt{2}}$. Since $\sqrt{6}$ is irrational, the product $\left(\frac{a}{\sqrt{2}}\right)\sqrt{6} = a\sqrt{3}$ is irrational. Take $b = \sqrt{3}$, which is irrational. Then $ab = a\sqrt{3}$ which is irrational. □

Problem 15. Prove that there are infinitely many prime numbers that are congruent to 3 modulo 4.

Proof. Prove by contradiction.

Suppose there are only finitely many primes that are congruent to 3 modulo 4. Let $p_1, p_2, \ldots, p_m$ be the list of all the primes that are congruent to 3 modulo 4.

We construct an integer M by $M = (p_1 p_2 \cdots p_m)^2 + 2$.

We have the following observation:

(i) $M \equiv 3 \mod 4$.

(ii) Every p_i divides $M - 2$.

(iii) None of the p_i divides M. [Otherwise, together with (ii), this will imply p_i divides 2, which is impossible.]

(iv) M is not a prime number. [Otherwise, by (i), M is a prime number congruent to 3 modulo 4. But $M \neq p_i$ for all $1 \leq i \leq m$. This contradicts the assumption that $p_1, p_2, \ldots, p_m$ are all the prime numbers congruent to 3 modulo 4.]

From the above discussion, we know that M is a composite number by (iv). So it has a prime factorization $M = q_1 q_2 \cdots q_k$.

Since M is odd, all these prime factors q_j must be odd, and hence q_j must be congruent to either 1 or 3 modulo 4.

By (iii), q_j cannot be any of the p_i. So all q_j must be congruent to 1 modulo 4. Then M, which is the product of q_j, must also be congruent to 1 modulo 4.

This contradicts (i) that M is congruent to 3 modulo 4.

Hence we conclude that there must be infinitely many primes that are congruent to 3 modulo 4. □

Problem 16. Prove that, for any positive integer n, there is a perfect square m^2 (m is an integer) such that $n \leq m^2 \leq 2n$.

Proof. Prove by contradiction.

Suppose otherwise, that $n > m^2$ and $(m+1)^2 > 2n$ so that there is no square between n and $2n$, then

$$(m+1)^2 > 2n > 2m^2.$$

Since we are dealing with integers and the inequalities are strict, we get

$$(m+1)^2 \geq 2m^2 + 2$$

which simplifies to

$$0 \geq m^2 - 2m + 1 = (m-1)^2$$

The only value for which this is possible is $m = 1$, but you can eliminate that easily enough. □

Problem 17 (H3M Specimen)**.** For any positive integer n, if one square is removed from a $2^n \times 2^n$ checkerboard, the remaining squares can be completely covered by triominoes (an L-shaped domino consisting of three squares).

Proof. Prove by induction.

Base case: $P(1)$ is clearly true.

Inductive step: $P(k) \implies P(k+1)$ is true for all k, i.e. if a $2^k \times 2^k$ checkerboard with a square removed can be completely covered by triominoes, then a $2^{k+1} \times 2^{k+1}$ checkerboard with a square removed can be completely covered by triominoes.

(i) Divide the $2^{k+1} \times 2^{k+1}$ checkerboard into four $2^k \times 2^k$ sub-boards.

(ii) One of the sub-boards include the removed square.

(iii) WLOG, assume the top left sub-board has the removed square.

(iv) By induction hypothesis, this sub-board can be covered by triominoes.

(v) For the top right sub-board, we cover it with trominoes with a remaining square at the bottom left corner.

(vi) For the bottom right sub-board, we cover it with trominoes with a remaining square at the top left corner.

(vii) For the bottom left sub-board, we cover it with trominoes with a remaining square at the top right corner.

(viii) The remaining three squares from (v) to (vii) are connected and can be covered by one triomino.

□

Remark. Although it is easy to visualise this by drawing it out, always produce a written proof.

Problem 18. Prove that for every positive integer $n \geq 4$,

$$n! > 2^n.$$

Proof. Let $P(n) : n! > 2^n$

Base case: $P(4)$

LHS: $4! = 4 \times 3 \times 2 \times 1 = 24$, RHS: $2^4 = 16 < 24$

So $P(4)$ is true.

Inductive step: $P(k) \implies P(k+1)$ for all $k \in \mathbb{Z}^+_{\geq 4}$

$$\begin{aligned} k! &> 2^k \\ (k+1)k! &> 2^k(k+1) \\ &> 2^k 2 \quad \text{since from } k \geq 4,\ k+1 \geq 5 > 2 \\ &= 2^{k+1} \end{aligned}$$

hence proven $P(k) \implies P(k+1)$ for integers $k \geq 4$.

By PMI, we have proven $P(n)$ for all integers $n \geq 4$. □

Problem 19 (H2FM TJC 2023)**.** Prove by mathematical induction, for $n \geq 2$,

$$\sqrt[n]{n} < 2 - \frac{1}{n}.$$

Proof. Let $P(n) : \sqrt[n]{n} < 2 - \frac{1}{n}$ for $n \geq 2$.

Base case: $P(2)$

When $n = 2$, $\sqrt{2} = 1.41\cdots < 2 - \frac{1}{2} = 1.5$ which is true. Hence $P(2)$ is true.

Inductive step: $P(k) \implies P(k+1)$ for all $k \in \mathbb{Z}^+_{\geq 2}$

Assume $P(k)$ is true for $k \geq 2, k \in \mathbb{Z}^+$, i.e.

$$\sqrt[k]{k} < 2 - \frac{1}{k} \implies k < \left(2 - \frac{1}{k}\right)^k$$

We want to prove that $P(k+1)$ is true, i.e.

$$k + 1 < \left(2 - \frac{1}{k+1}\right)^{k+1}$$

Since $k > 2$, we have

$$\begin{aligned} \left(2 - \frac{1}{k+1}\right)^{k+1} &> \left(2 - \frac{1}{k}\right)^{k+1} \quad \because k > 2 \\ &= \left(2 - \frac{1}{k}\right)^k \left(2 - \frac{1}{k}\right) \\ &> k\left(2 - \frac{1}{k}\right) \quad \text{[by inductive hypothesis]} \\ &= 2k - 1 = k + k - 1 > k - 1 \because k > 2 \end{aligned}$$

Hence $P(k+1)$ is true.

Since $P(2)$ is true and $P(k) \implies P(k+1)$, by mathematical induction $P(n)$ is true. □

Problem 20. Prove that for all integers $n \geq 3$,

$$\left(1+\frac{1}{n}\right)^n < n$$

Proof. **Base case:** $P(3)$

On the LHS, $\left(1+\frac{1}{3}\right)^3 = \frac{64}{27} = 2\frac{10}{27} < 3$. Hence $P(3)$ is true.

Inductive step: $P(k) \implies P(k+1)$ for all $k \in \mathbb{Z}^+_{\geq 3}$

Our inductive hypothesis is

$$\left(1+\frac{1}{k}\right)^k < k$$

Multiplying both sides by $\left(1+\frac{1}{k}\right)$ (to get a $k+1$ in the power),

$$\left(1+\frac{1}{k}\right)^k\left(1+\frac{1}{k}\right) = \left(1+\frac{1}{k}\right)^{k+1} < k\left(1+\frac{1}{k}\right) = k+1$$

Since $k < k+1 \iff \frac{1}{k} > \frac{1}{k+1}$,

$$\left(1+\frac{1}{k}\right)^{k+1} > \left(1+\frac{1}{k+1}\right)^{k+1}$$

The rest of the proof follows easily. □

A sequence of integers F_i, where integer $1 \leq i \leq n$, is called the *Fibonacci sequence* if and only if it is defined recursively by $F_1 = 1$, $F_2 = 1$, $F_n = F_{n-1} + F_{n-2}$ for $n > 2$.

Problem 21. Let F_i be the Fibonacci sequence. Prove that $3 \nmid n$ if and only if F_n is odd.

Proof. **Forward direction:** $3 \nmid n \implies F_n$ is odd

Backward direction: F_n is odd $\implies 3 \nmid n$ (We prove the contrapositive: $3 \mid n \implies F_n$ is even)

Hence we only need to prove the following via PMI:

- $(\forall n \in \mathbb{Z}^+ \text{ and } 3 \nmid n), F_n$ is odd
  Base case: $P(1), P(2)$
  Inductive step: $P(k) \implies P(k+3)$ for all $k \geq 1$
- $(\forall n \in \mathbb{Z}^+ \text{ and } 3 \mid n), F_n$ is even
  Base case: $P(3)$
  Inductive step: $P(k) \implies P(k+3)$ for all $k \geq 3$

[Note that we have partitioned the domain into two.]

Hence to show $\forall n \in \mathbb{Z}^+ \; P(n)$,

Base case: $P(1), P(2), P(3)$

Inductive step: $\forall k \in \mathbb{Z}^+ \; P(k) \implies P(k+3)$ □

Problem 22. Let a_i where integer $1 \le i \le n$ be a sequence of integers defined recursively by initial conditions $a_1 = 1$, $a_2 = 1$, $a_3 = 3$ and the recurrence relation $a_n = a_{n-1} + a_{n-2} + a_{n-3}$ for $n > 3$.

For all $n \in \mathbb{Z}^+$, prove that

$$a_n \le 2^{n-1}.$$

Proof. Let $P(n) : a_n \le 2^{n-1}$.

Given the recurrence relation, it could be possible to use $P(k), P(k+1), P(k+2)$ to prove $P(k+3)$ for all $k \in \mathbb{Z}^+$.

Base case: $P(1), P(2), P(3)$

$P(1) : a_1 = 1 \le 2^{1-1} = 1$ is true.

$P(2) : a_2 = 1 \le 2^{2-1} = 2$ is true.

$P(3) : a_3 = 3 \le 2^{3-1} = 4$ is true.

Inductive step: $P(k) \wedge P(k+1) \wedge P(k+2) \implies P(k+3)$ for all $k \in \mathbb{Z}^+$

By inductive hypothesis, for $k \in \mathbb{Z}^+$ we have $a_k \le 2^k, a_{k+1} \le 2^{k+1}, a_{k+2} \le 2^{k+2}$.

$$\begin{aligned}
a_{k+3} &= a_k + a_{k+1} + a_{k+2} \quad \text{[start from recurrence relation]} \\
&\le 2^k + 2^{k+1} + 2^{k+2} \quad \text{[use inductive hypothesis]} \\
&= 2^k(1 + 2 + 2^2) \\
&< 2^k(2^3) \quad \text{[approximation, since } 1 + 2 + 2^2 < 2^3\text{]} \\
&= 2^{k+3}
\end{aligned}$$

which is precisely $P(k+3) : a_{k+3} \le 2^{k+3}$. □

Problem 23 (Bézout's lemma)**.** Let a and b be integers, not both 0. Prove that $\gcd(a,b) = ax_0 + by_0$ for some integers x_0 and y_0.

Solution. Given a and b, consider the set

$$S = \{z \in \mathbb{Z} \mid z > 0; \exists x, y \in \mathbb{Z}, z = ax + by\}.$$

S satisfies the conditions of well-ordering principle, and hence has a smallest element $c = ax_0 + by_0$. We want to show that (i) c is a common divisor of a and b; (ii) $c = \gcd(a,b)$.

(i) c is a common divisor of a and b

Suppose $c \nmid a$. By quotient-remainder theorem, $a = cq + r$ where $0 < r < c$.

Then

$$a = (ax_0 + by_0)q + r \implies r = a - (ax_0 + by_0)q \implies r = a(1 - x_0q) - b(y_0q)$$

So r is an element in S, and $r < c$. This contradicts the minimality of c in S. Hence $c \mid a$. Then $a = (ax_0 + by_0)q + r$.

Similarly, $c \mid b$.

(ii) $c = \gcd(a,b)$

Suppose otherwise, that c is not the greatest common divisor of a and b.

Let there exists some $d > c$ which satisfies $d \mid a$ and $d \mid b$.

Then $d \mid (ax + by)$ for any x and y. So d divides all elements in S. In particular, $d \mid c$, which means $d \le c$, a contradiction.

Hence $c = \gcd(a,b)$.

This concludes the proof that $\gcd(a,b) = ax_0 + by_0$ for some integers x_0 and y_0. □

Problem 24 (H3M 2017 Q8)**.** The Fibonacci sequence is defined recursively by $F_{n+1} = F_n + F_{n-1}$ and $F_1 = 1, F_2 = 1$.

(i) Find the periods of Fibonacci sequences modulo 3 and 4.

(ii) For any positive integer m, show that we can find two pairs (F_j, F_{j+1}) and (F_k, F_{k+1}) which are the same modulo m with $1 \le j < k \le m^2 + 1$.

(iii) For m, j and k as in (ii), explain why the Fibonacci sequence modulo m is periodic with period dividing $k - j$.

(iv) For any positive integer m, prove that there is a Fibonacci number which is a multiple of m.

Solution.

(i) Modulo 3: $1, 1, 2, 0, 2, 2, 1, 0, 1, 1, \ldots$ has period 8.

Modulo 4: $1, 1, 2, 3, 1, 0, 1, 1, \ldots$ has period 6.

(ii) Modulo m, there are m possible values $0, 1, 2, \ldots, m - 1$. So there are exactly m^2 possible distinct pairs (a, b).

If we consider $m^2 + 1$ pairs of (F_i, F_{i+1}) modulo m where $1 \le i \le m^2 + 1$, we can find two pairs (F_j, F_{j+1}) and (F_k, F_{k+1}) which are the same modulo m, by Pigeonhole Principle.

(iii) This is the same as showing $F_{j+n} \equiv F_{k+n} \pmod{m}$ for all non negative integer n.

We prove using mathematical induction.

Basis step: $P(0)$ and $P(1)$

$$F_j \equiv F_k \pmod{m} \quad F_{j+1} \equiv F_{k+1} \pmod{m}$$

Inductive step: $P(q-1) \land P(q) \implies P(q+1)$ for all $q \ge 1$

Given $F_{j+q-1} \equiv F_{k+q-1} \pmod{m}$ and $F_{j+q} \equiv F_{k+q} \pmod{m}$. Then $F_{j+q-1} + F_{j+q} \equiv F_{k+q-1} + F_{k+q} \pmod{m}$ so $F_{j+q+1} \equiv F_{k+q+1} \pmod{m}$.

By mathematical induction, the sequence repeats itself after $k - j$ terms. This implies the period of the sequence divides $k - j$.

(iv) For any positive m, by part (iii), the Fibonacci sequence modulo m is periodic. That is, (F_1, F_2) is congruent to (F_i, F_{i+1}) modulo m for some $i > 2$:

$$F_i \equiv F_1 \equiv 1 \pmod{m} \quad F_{i+1} \equiv F_2 \equiv 1 \pmod{m}$$

Then $F_{i-1} = F_{i+1} - F_i \equiv 1 - 1 \equiv 0 \pmod{m}$, which means $m \mid F_{i-1}$.

We have proven that there is a Fibonacci number which is a multiple of m.

□

Problem 25 (H3M 2018 Q5)**.** A $p \times q$ chessboard can be tessellated with $a \times b$ tiles.

A unit square (x, y) is shaded if and only if $x \equiv y \pmod a$.

(i) Explain why the following are necessary conditions for such a tessellation

  (a) ab is a factor of pq.

  (b) p and q can be written in the form $ma + nb$ where m and n are non-negative integers.

  (c) The $p \times q$ chessboard has $\frac{pq}{a}$ shaded squares.

(ii) Let t be the smaller of r and s such that

$$\begin{aligned} p &\equiv r \pmod a \quad 0 \le r < a \\ q &\equiv s \pmod a \quad 0 \le s < a \end{aligned}$$

  (a) Explain why the number of shaded squares in the $p \times q$ chessboard is $\frac{pq - rs}{a} + t$.

  (b) Hence prove that for a tessellation, either $a \mid p$ or $a \mid q$.

Solution.

(i) (a) A $p \times q$ chessboard has pq squares, a $a \times b$ tile has ab squares.

Suppose k tiles are used to tessellate the board. Then $pq = kab$. Hence $ab \mid pq$.

  (b) p and q are the height and base of the $p \times q$ chessboard respectively, a and b are the height and base of each $a \times b$ tile respectively. Each tile can be places horizontally or vertically in the tessellation.

If we tessellate the board at the bottom from left to right with m vertical and n horizontal tiles, there will be $ma + nb$ squares at the bottom row of the board. Each row of the board is made up of q squares. So we get $q = ma + nb$.

Similarly, if we tessellate the board on the left from bottom to top, we will get $p = sa + tb$ (with s horizontal and t vertical tiles).

(ii) (a)

□

Problem 26 (H3M 2018 Q6, Dirichlet's approximation theorem)**.** Let x be any positive real numbers and n be any positive integer. Prove that there are integers a and b with $1 \le b \le n$, such that

$$\left| x - \frac{a}{b} \right| < \frac{1}{bn}.$$

Solution. For any real number y, we write $y = \lfloor y \rfloor + \{y\}$, where $\lfloor y \rfloor$ denotes the integer part of y and $\{y\}$ denotes the fractional part of y, $0 \le \{y\} < 1$.

We divide the interval $[0,1)$ into n smaller intervals of measure $\frac{1}{n}$. Consider $\{x\}, \{2x\}, \dots, \{nx\}$. Let I_i denote the interval $\left[\frac{i-1}{n}, \frac{i}{n}\right]$, where $1 \le i \le n$.

We now consider two cases:

Case 1: Some $\{kx\}$ falls in I_1

Then $kx - \lfloor kx \rfloor = \{kx\} < \frac{1}{n}$.

Dividing both sides by k,

$$\left| x - \frac{\lfloor kx \rfloor}{k} \right| < \frac{1}{kn}.$$

By taking $a = \lfloor kx \rfloor$ and $b = k$, we have the inequality.

Case 2: None of $\{kx\}$ falls in I_1

This means all $\{kx\}$ fall into $I_2, I_3, \dots, I_n$. By Pigeonhole Principle, at least two $\{kx\}$ fall in the same I_i.

Let $\frac{i-1}{n} \le \{px\} < \frac{i}{n}$ and $\frac{i-1}{n} \le \{qx\} < \frac{i}{n}$. Then

$$\begin{aligned}
\left| \{px\} - \{qx\} \right| &< \frac{1}{n} \\
\left| (px - \lfloor px \rfloor) - (qx - \lfloor qx \rfloor) \right| &< \frac{1}{n} \\
\left| (px - qx) - (\lfloor px \rfloor - \lfloor qx \rfloor) \right| &< \frac{1}{n} \\
\left| (p-q)x - (\lfloor px \rfloor - \lfloor qx \rfloor) \right| &< \frac{1}{n}
\end{aligned}$$

Dividing both sides by $p - q$,

$$\left| x - \frac{(\lfloor px \rfloor - \lfloor qx \rfloor)}{p-q} \right| < \frac{1}{(p-q)n}.$$

WLOG assume $p > q$. Then $1 \le p-q < n$. By taking $a = \lfloor px \rfloor - \lfloor qx \rfloor$ and $b = p-q$, we have the inequality. □

Problem 27 (Wilson's Theorem)**.** Let p be a prime number. Prove that $(p-1)!+1$ is divisible by p.

Proof. We first prove the uniqueness of inverse modulo p: for each $x \in Q = \{1, 2, \ldots, p-1\}$ for some prime p, there is precisely one integer y such that $xy \equiv 1 \pmod{p}$.

Proof. Suppose otherwise, that there are two distinct inverses for x modulo p; that is, $xy_1 \equiv 1 \pmod{p}$ and $xy_2 \equiv 1 \pmod{p}$. Then $x(y_1 - y_2) \equiv 0 \pmod{p}$. Since $x \nmid p$, by Euclid's lemma, $y_1 \equiv y_2 \pmod{p}$ so $y_1 = y_2 + kp$ for some integer k. But we know that $0 \leq y_1, y_2 < p$, so $kp = y_1 - y_2$, $0 \leq kp < p$ thus $k = 0$. Hence $y_1 = y_2$. □

If $y \neq x$, we can pair up elements of Q such that their product is congruent to 1 modulo p.

If $y = x$, then $x^2 \equiv 1 \pmod{p}$. Thus

$$p \mid x^2 = 1 \implies p \mid (x+1)(x-1) \implies p \mid x+1 \text{ or } p \mid x-1 \implies x \equiv \pm 1 \pmod{p}$$

which is $1^2 \equiv 1 \pmod{p}$ and $(p-1)^2 \equiv 1 \pmod{p}$. So aside 1 and $p-1$, all other elements can be paired up. Hence,

$$\begin{aligned}(p-1)!+1 &\equiv 1(p-1)+1 \pmod{p} \\ &\equiv p-1+1 \pmod{p} \\ &\equiv p \pmod{p}\end{aligned}$$

Hence $(p-1)!+1$ is divisible by p. □

Problem 28. For $m, n \in \mathbb{N}$, prove that

$$F_{n+m+1} = F_n F_m + F_{n+1} F_{m+1}.$$

Proof. For $n \in \mathbb{N}$, take $P(n) : F_{n+m+1} = F_n F_m + F_{n+1} F_{m+1}$ for all $m \in \mathbb{N}$ in the cases $k = n$ and $k = n + 1$. So we are using induction to progress through n and dealing with m simultaneously at each stage.

To verify $P(0)$, we note that

$$F_{m+1} = F_0 F_m + F_1 F_{m+1}$$

and

$$F_{m+2} = F_1 F_m + F_2 F_{m+1}$$

for all m, as $F_0 = 0$ and $F_1 = F_2 = 1$.

For the inductive step we assume $P(n)$, i.e. that for all $m \in \mathbb{N}$, Fn+m+1 = FnFm + Fn+1Fm+1, Fn+m+2 = Fn+1Fm + Fn+2Fm+1. To prove $P(n+1)$ it remains to show that for all $m \in \mathbb{N}$,

$$F_{n+m+3} = F_{n+2} F_m + F_{n+3} F_{m+1}.$$

From our $P(n)$ assumptions and the definition of the Fibonacci numbers, LHS of (5) = Fn+m+3 = Fn+m+2 + Fn+m+1 = FnFm + Fn+1Fm+1 + Fn+1Fm + Fn+2Fm+1 = (Fn + Fn+1) Fm + (Fn+1 + Fn+2) Fm+1 = Fn+2Fm + Fn+3Fm+1 = RHS of (5). □

2 Set Theory

§2.1 Basics

§2.1.1 Notation

You should, by now, be familiar with the following definitions and notation:

- A **set** S can be loosely defined as a collection of objects.
- For a set S, we write $x \in S$ to mean that x is an **element** of S, and $x \notin S$ if otherwise.
- A set can be defined in terms of some property $P(x)$ that the elements $x \in S$ satisfy, denoted by the following **set builder notation**:
$$\{x \in S \mid P(x)\}$$
- Some basic sets (of numbers) you should be familiar with:
  - $\mathbb{N} = \{0, 1, 2, 3, \dots\}$ denotes the natural numbers (non-negative integers).
  - $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$ denotes the integers.
  - $\mathbb{Q} = \{\frac{p}{q} \mid p, q \in \mathbb{Z}, q \neq 0\}$ denotes the rational numbers.
  - $\mathbb{R}$ denotes the real numbers, which can be expressed in terms of decimal expansion.
  - $\mathbb{C} = \{x + yi \mid x, y \in \mathbb{R}\}$ denotes the of complex numbers.
- The **empty set** is the set with no elements, denoted by $\varnothing$.
- A is a **subset** of B if every element of A is in B, denoted by $A \subseteq B$.
$$A \subseteq B \iff \forall x, x \in A \implies x \in B$$

  $\subseteq$ is transitive, i.e. if $A \subseteq B$ and $B \subseteq C$, then $A \subseteq C$.

  Proof. Let $x \in A$. Since $A \subseteq B$ and $x \in A$, $x \in B$. Since $B \subseteq C$ and $x \in B$, $x \in C$. Hence $A \subseteq C$. □

  A is a **proper subset** of B if $A \subseteq B$ and $A \neq B$, denoted by $A \subset B$.
  Using this definition, we have the relationship
$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}$$
- A and B are **equal** if and only if they contain the same elements, denoted by $A = B$.
  To prove that A and B are equal, we simply need to prove that $A \subseteq B$ and $A \subseteq B$.

Proof. We have

$$\begin{aligned} A = B &\iff (\forall x)[x \in A \iff x \in B] \\ &\iff (\forall x)[(x \in A \implies x \in B) \wedge (x \in B \implies x \in A)] \\ &\iff \{(\forall x)[x \in A \implies x \in B]\} \wedge (\forall x)[x \in B \implies x \in A)] \\ &\iff (A \subseteq B) \wedge (B \subseteq A) \end{aligned}$$

□

- Some frequently occurring subsets of the real numbers are known as **intervals**, which can be visualised as sections of the real line:
  - Open interval
  $$(a,b) = \{x \in \mathbb{R} \mid a < x < b\}$$
  - Closed interval
  $$[a,b] = \{x \in \mathbb{R} \mid a \le x < b\}$$
  - Half open interval
  $$(a,b] = \{x \in \mathbb{R} \mid a < x \le b\}$$
- The **power set** $\mathcal{P}(A)$ of A is the set of all subsets of A (including the set itself and the empty set).
- An **ordered pair** is denoted by (a,b), where the order of the elements matters. Two pairs (a_1,b_1) and (a_2,b_2) are equal if and only if $a_1 = a_2$ and $b_1 = b_2$.

  Similarly, we have ordered triples (a,b,c), quadruples (a,b,c,d) and so on. If there are n elements it is called an n-tuple.
- The **Cartesian product** of sets A and B, denoted by $A \times B$, is the set of all ordered pairs with the first element of the pair coming from A and the second from B:

  $$A \times B = \{(a,b) \mid a \in A, b \in B\} \tag{2.1}$$

  More generally, we define $A_1 \times A_2 \times \cdots \times A_n$ to be the set of all ordered n-tuples $(a_1,a_2,\dots,a_n)$, where $a_i \in A_i$ for $1 \le i \le n$. If all the A_i are the same, we write the product as A^n.

  Example 2.1.1. $\mathbb{R}^2$ is the Euclidean plane, $\mathbb{R}^3$ is the Euclidean space, and $\mathbb{R}^n$ is the n-dimensional Euclidean space.

  $$\begin{aligned} \mathbb{R} \times \mathbb{R} = \mathbb{R}^2 &= \{(x,y) \mid x,y \in \mathbb{R}\} \\ \mathbb{R} \times \mathbb{R} \times \mathbb{R} = \mathbb{R}^3 &= \{(x,y,z) \mid x,y,z \in \mathbb{R}\} \\ \mathbb{R}^n &= \{(x_1,x_2,\dots,x_n) \mid x_1,x_2,\dots,x_n \in \mathbb{R}\} \end{aligned}$$

§2.1.2 Algebra of Sets

Given $A \subset S$ and $B \subset S$.

- The **union** $A \cup B$ is the set consisting of elements that are in A or B (or both):

  $$A \cup B = \{x \in S \mid x \in A \vee x \in B\}$$

- The **intersection** $A \cap B$ is the set consisting of elements that are in both A and B:

  $$A \cap B = \{x \in S \mid x \in A \wedge x \in B\}$$

  A and B are **disjoint** if both sets have no element in common:

  $$A \cap B = \varnothing$$

More generally, we can take unions and intersections of arbitrary numbers of sets, even infinitely many. If we have a family of subsets $\{A_i \mid i \in I\}$, where I is an **indexing set**, we write

$$\bigcup_{i\in I} A_i = \{x \mid \exists i \in I\,(x \in A_i)\}$$

and

$$\bigcap_{i\in I} A_i = \{x \mid \forall i \in I\,(x \in A_i)\}$$

- The **complement** of A, denoted by A^c or A', is the set containing elements that are not in A:

  $$A^c = \{x \in S \mid x \notin A\}$$

- The **set difference**, or complement of B in A, denoted by $A \smallsetminus B$, is the subset consisting of those elements that are in A and not in B:

  $$A \smallsetminus B = \{x \in A \mid x \notin B\}$$

  Note that $A \smallsetminus B = A \cap B^c$.

Proposition 2.1.1 (Double Inclusion)**.** Let $A \subset S$ and $B \subset S$. Then

$$A = B \iff A \subseteq B \text{ and } B \subseteq A \tag{2.2}$$

Proof. We prove both directions.

Forward direction:

If $A = B$, then every element in A is an element in B, so certainly $A \subseteq B$, and similarly $B \subseteq A$.

Backward direction:

Suppose $A \subseteq B$, and $B \subseteq A$. Then for every element $x \in S$, if $x \in A$ then $A \subseteq B$ implies that $x \in B$, and if $x \notin A$ then $B \subseteq A$ means $x \notin B$. So $x \in A$ if and only if $x \in B$, and therefore $A = B$. □

Proposition 2.1.2 (Distributive Laws)**.** Let $A \subset S$, $B \subset S$ and $C \subset S$. Then

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C) \tag{2.3}$$

$$(A \cap B) \cap C = (A \cup C) \cap (B \cup C) \tag{2.4}$$

Proof. For the first one, suppose x is in the LHS, that is $x \in A \cup (B \cap C)$. This means that $x \in A$ or $x \in B \cap C$ (or both). Thus either $x \in A$ or x is in both B and C (or x is in all three sets). If $x \in A$ then $x \in A \cup B$ and $x \in A \cup C$, and therefore x is in the RHS. If x is in both B and C then similarly x is in both $A \cup B$ and $A \cup C$. Thus every element of the LHS is in the RHS, which means we have shown $A \cup (B \cap C) \subseteq (A \cup B) \cap (A \cup C)$.

Conversely suppose that $x \in (A \cup B) \cap (A \cup C)$. Then x is in both $A \cup B$ and $A \cup C$. Thus either $x \in A$ or, if $x \notin A$, then $x \in B$ and $x \in C$. Thus $x \in A \cup (B \cap C)$. Hence $(A \cup B) \cap (A \cup C) \subseteq A \cup (B \cap C)$.

By double inclusion, $(A \cup B) \cap (A \cup C) = A \cup (B \cap C)$.

The proof of the second one follows similarly and is left as an exercise. □

Proposition 2.1.3 (De Morgan's Laws)**.** Let $A \subset S$ and $B \subset S$. Then

$$(A \cup B)^c = A^c \cap B^c \tag{2.5}$$

$$(A \cap B)^c = A^c \cup B^c \tag{2.6}$$

Proof. For the first one, suppose $x \in (A \cup B)^c$. Then x is not in either A or B. Thus $x \in A^c$ and $x \in B^c$, and therefore $x \in A^c \cap B^c$.

Conversely, suppose $x \in A^c \cap B^c$. Then $x \notin A$ and $x \notin B$, so x is in neither A nor B, and therefore $x \in (A \cup B)^c$.

By double inclusion, the first result holds. The second result follows similarly and is left as an exercise. □

De Morgan's laws extend naturally to any number of sets, so if $\{A_i \mid i \in I\}$ is a family of subsets of S, then

$$\left(\bigcap_{i\in I} A_i\right)^c = \bigcup_{i\in I} A_i^c \quad \text{and} \quad \left(\bigcup_{i\in I} A_i\right)^c = \bigcap_{i\in I} A_i^c$$

Exercise 2.1.1

Prove the following:

1. $(\bigcup_{i\in I} A_i) \cup B = \bigcup_{i\in I}(A_i \cup B)$
2. $(\bigcap_{i\in I} A_i) \cup B = \bigcap_{i\in I}(A_i \cup B)$
3. $(\bigcup_{i\in I} A_i) \cup \big(\bigcup_{j\in J} B_j\big) = \bigcup_{(i,j)\in I\times J}(A_i \cup B_j)$
4. $(\bigcap_{i\in I} A_i) \cup \big(\bigcap_{j\in J} B_j\big) = \bigcap_{(i,j)\in I\times J}(A_i \cup B_j)$

Exercise 2.1.2

Let $S \subset A \times B$. Express the set A_S of all elements of A which appear as the first entry in at least one of the elements in S.
(A_S here may be called the projection of S onto A.)

§2.1.3 Cardinality

Informally, the **cardinality** of S, denoted $|S|$, is a measure of its "size". We will be able to give a nicer definition of cardinality later, once we have discussed bijections, but the following provides a recursive definition of the cardinality for a finite set:

Definition 2.1.1 (Finiteness and the cardinality of a finite set)**.** The empty set $\varnothing$ is finite, with $|\varnothing| = 0$.

S is **finite** with $|S| = n + 1$, if there exists $s \in S$ such that $|S \setminus \{s\}| = n$ for some $n \in \mathbb{Z}^+$. We call $|S|$ the **cardinality** of S.

Any set that is not finite is said to be **infinite**.

It is not hard to see that this means that if $S = \{x_1, x_2, \ldots, x_n\}$, and $x_i \neq x_j$ whenever $i \neq j$, then $|S| = n$. Conversely, if $|S| = n$ then S is a set with n elements.

Proposition 2.1.4. Let A and B be finite sets. Then $|A \cup B| = |A| + |B| - |A \cap B|$.

Proof. The proof is left as an exercise. □

Proposition 2.1.5 (Subsets of a finite set)**.** If a set A is finite with $|A| = n$, then its power set has $|\mathcal{P}(A)| = 2^n$.

Proof. We use induction. For the initial step, note that if $|A| = 0$ then $A = \varnothing$ has no elements, so there is a single subset $\varnothing$, and therefore $|\mathcal{P}(A)| = 1 = 2^0$.

Now suppose that $n \geq 0$ and that $|P(S)| = 2^n$ for any set S with $|S| = n$. Let A be any set with $|A| = n+1$. By definition, this means that there is an element a and a set $A_0 = A \setminus \{a\}$ with $|A_0| = n$. Any subset of A must either contain the element a or not, so we can partition $\mathcal{P}(A) = P(A_0) \cup \{S \cup \{a\} \mid S \in P(A_0)\}$. These two sets are disjoint, and each of them has cardinality $|P(A_0)| = 2^n$ by the inductive hypothesis. Hence $|\mathcal{P}(A)| = 2^n + 2^n = 2^{n+1}$.

Thus, by induction, the result holds for all n. □

Another way to see this is through combinatorics: Consider the process of creating a subset. We can do this systematically by going through each of the $|A|$ elements in A and making the yes/no decision whether to put it in the subset. Since there are $|A|$ such choices, that yields $2^{|A|}$ different combinations of elements and therefore $2^{|A|}$ different subsets.

§2.2 Relations

§2.2.1 Definition

Definition 2.2.1. R is a **relation** between A and B if and only if R is a subset of the Cartesian product $A \times B$, i.e. $R \subseteq A \times B$.

$a \in A$ and $b \in B$ are **related** if $(a, b) \in R$, denoted by aRb.

Remark. A relation is a set of ordered pairs.

Visually speaking, a relation is uniquely determined by a simple bipartite graph over A and B. On the bipartite graph, this is usually represented by an edge between a and b.

Definition 2.2.2. A **binary relation** in A is a relation between A and itself, i.e. $R \subseteq A \times A$.

A and B are the **domain** and **range** of R respectively, denoted by $\operatorname{dom} R$ and $\operatorname{ran} R$ respectively, if and only if $A \times B$ is the smallest Cartesian product of which R is a subset.

Example 2.2.1. Given $R = \{(1, a), (1, b), (2, b), (3, b)\}$, then $\operatorname{dom} R = \{1, 2, 3\}$ and $\operatorname{ran} R = \{a, b\}$.

In many cases we do not actually use R to write the relation because there is some other conventional notation:

Example 2.2.2.

- The “less than or equal to” relation $\le$ on the set of real numbers is $\{(x, y) \in \mathbb{R}^2 \mid x \le y\}$. We write $x \le y$ if (x, y) is in this set.
- The “divides” relation $\mid$ on $\mathbb{N}$ is $\{(m, n) \in \mathbb{N}^2 : m \text{ divides } n\}$. We write $m \mid n$ if (m, n) is in this set.
- For a set S, the “subset” relation $\subseteq$ on $\mathcal{P}(S)$ is $\{(A, B) \in \mathcal{P}(S)^2 \mid A \subseteq B\}$. We write $A \subseteq B$ if (A, B) is in this set.

§2.2.2 Properties of relations

Let A be a set, R a relation on A, and $x, y, z \in A$. We say that

- R is **reflexive** if xRx for all $x \in A$;
- R is **symmetric** if $xRy \implies yRx$;
- R is **anti-symmetric** if xRy and $yRx \implies x = y$;
- R is **transitive** if xRy and $yRz \implies xRz$.

Example 2.2.3 (Less than or equal to)**.** The relation $\leq$ on R is reflexive, anti-symmetric, and transitive, but not symmetric.

Definition 2.2.3. Any relation on A that is reflexive, anti-symmetric, and transitive is called a **partial order**, denoted by $\leq$. It is called a **total order** if for every $x, y \in A$, either xRy or yRx (or both).

Example 2.2.4 (Less than)**.** The relation $<$ on R is not reflexive, symmetric, or anti-symmetric, but it is transitive.

Example 2.2.5 (Not equal to)**.** The relation $\neq$ on R is not reflexive, anti-symmetric or transitive, but it is symmetric.

Exercise 2.2.1: Congruence modulo n

Let $n \geq 2$ be an integer, and define R on $\mathbb{Z}$ by saying aRb if and only if $a - b$ is a multiple of n. Prove that R is reflexive, symmetric and transitive.

Proof.

- Reflexivity: For any $a \in \mathbb{Z}$ we have aRa as 0 is a multiple of n.
- Symmetry: If aRb then $a - b = kn$ for some integer k. So $b - a = -kn$, and hence bRa.
- Transitivity: If aRb and bRc then $a - b = kn$ and $b - c = ln$ for integers k, l. So then $a - c = (a - b) + (b - c) = (k + l)n$, and hence aRc.

□

§2.2.3 Equivalence relations, equivalence classes, and partitions

One important type of relation is an equivalence relation. An equivalence relation is a way of saying two objects are, in some particular sense, “the same”.

Definition 2.2.4. A binary relation R on A is an **equivalence relation** if it is reflexive, symmetric and transitive.

Notation. We use the symbol $\sim$ to denote the equivalence relation R in $A \times A$: whenever $(a, b) \in R$ we denote $a \sim b$.

An equivalence relation provides a way of grouping together elements that can be viewed as being the same:

Definition 2.2.5. Given an equivalence relation $\sim$ on a set A, and given $x \in A$, the **equivalence class** of x, denoted $[x]$, is the subset

$$[x] = \{y \in A \mid y \sim x\}$$

Example 2.2.6 (Congruence modulo n)**.** For the equivalence relation of congruence modulo n, the equivalence class of 1 is the set $1 = \{\dots, -n+1, 1, n+1, 2n+1, \dots\}$; that is, all the integers that are congruent to 1 modulo n.

Properties of equivalence classes:

- Every two equivalence classes are disjoint
- The union of equivalence classes form the entire set

You can translate these properties into the point of view from the elements: Every element belongs to one and only one equivalence class.

- No element belongs to two distinct classes
- All elements belong to an equivalence class

Definition 2.2.6. The **set of equivalence classes** (quotient sets) are the set of all equivalence classes, denoted by $A/\sim$.

Grouping the elements of a set into equivalence classes provides a partition of the set, which we define as follows:

Definition 2.2.7. A **partition** of a set A is a collection of subsets $\{A_i \subseteq A \mid i \in I\}$, where I is an indexing set, with the property that

(i) $A_i \neq \varnothing$ for all $i \in I$ (that is, all the subsets are non-empty),

(ii) $\bigcup_{i \in I} Ai = A$ (that is, every member of A lies in one of the subsets),

(iii) $A_i \cap A_j = \varnothing$ for every $i \neq j$ (that is, the subsets are disjoint).

The subsets are called the parts of the partition.

Example 2.2.7 (Odd and even natural numbers)**.** $\{\{n \in \mathbb{N} \mid n \text{ is divisible by } 2\}, \{n \in \mathbb{N} \mid n{+}1 \text{ is divisible by } 2\}\}$ forms a partition of the natural numbers, into evens and odds.

§2.3 Functions

§2.3.1 Definition

Definition 2.3.1. Given two sets X and Y, a **function** f from X to Y is a mapping of every element of X to some element of Y, denoted by $f : X \to Y$.

X and Y are known as the **domain** and **codomain** of f respectively.

Remark. The definition requires that a unique element of the codomain is assigned for every element of the domain. For example, for a function $f : \mathbb{R} \to \mathbb{R}$, the assignment $f(x) = \frac{1}{x}$ is not sufficient as it fails at $x = 0$. Similarly, $f(x) = y$ where $y^2 = x$ fails because $f(x)$ is undefined for $x < 0$, and for $x > 0$ it does not return a unique value; in such cases, we say the the function is **ill-defined**. We are interested in the opposite; functions that are **well-defined**.

Definition 2.3.2. Given a function $f : X \to Y$, the **image** (or range) of f is

$$f(X) = \{f(x) \mid x \in X\} \subseteq Y$$

More generally, given $A \subseteq X$, the image of A under f is

$$f(A) = \{f(x) \mid x \in A\} \subseteq Y$$

Given $B \subseteq Y$, the **pre-image** of B under f is

$$f^{-1}(B) = \{x \mid f(x) \in B\} \subseteq X$$

Remark. Beware the potentially confusing notation: for $x \in X$, $f(x)$ is a single element of Y, but for $A \subseteq X$, $f(A)$ is a set (a subset of Y). Note also that $f^{-1}(B)$ should be read as "the pre-image of B" and not as "f-inverse of B"; the pre-image is defined even if no inverse function exists (in which case f^{-1} on its own has no meaning; we discuss invertibility of a function below).

Exercise 2.3.1

Prove the following statements:

(a) $f(A \cup B) = f(A) \cup f(B)$

(b) $f(A_1 \cup \cdots \cup A_n) = f(A_1) \cup \cdots \cup f(A_n)$

(c) $f(\bigcup_{\lambda \in A} A_\lambda) = \bigcup_{\lambda \in A} f(A_\lambda)$

(d) $f(A \cap B) \subset f(A) \cap f(B)$

(e) $f^{-1}(f(A)) \supset A$

(f) $f(f^{-1}(A)) \subset A$

(g) $f^{-1}(A \cup B) = f^{-1}(A) \cup f^{-1}(B)$

(h) $f^{-1}(A \cap B) = f^{-1}(A) \cap f^{-1}(B)$

(i) $f^{-1}(A_1 \cup \cdots \cup A_n) = f^{-1}(A_1) \cup \cdots \cup f^{-1}(A_n)$

(j) $f^{-1}(\bigcup_{\lambda \in A} A_\lambda) = \bigcup_{\lambda \in A} f^{-1}(A_\lambda)$

If a function is defined on some larger domain than we care about, it may be helpful to restrict the domain:

Definition 2.3.3 (Restriction)**.** Given a function $f : X \to Y$ and a subset $A \subseteq X$, the **restriction** of f to A is the map $f|_A : A \to Y$ defined by $f|_A(x) = f(x)$ for all $x \in A$.

The restriction is almost the same function as the original f – just the domain has changed.

Another rather trivial but nevertheless important function is the identity map:

Definition 2.3.4 (Identity map)**.** Given a set X, the **identity** $\mathrm{id}_X : X \to X$ is defined by $\mathrm{id}_X(x) = x$ for all $x \in X$.

Notation. If the domain is unambiguous, the subscript may be removed.

§2.3.2 Injectivity, Surjectivity, Bijectivity

Definition 2.3.5. $f : X \to Y$ is **injective** if each element of Y has at most one element of X that maps to it.

$$\forall x_1, x_2 \in X,\ f(x_1) = f(x_2) \implies x_1 = x_2$$

Proposition 2.3.1. If $f : X \to Y$ is injective and $g : Y \to Z$ is injective, then $g \circ f : X \to Z$ is injective.

Proof. Let $f : X \to Y$ and $g : Y \to Z$ be arbitrary injective functions. We want prove that the function $g \circ f : X \to Z$ is also injective.

To do so, we will prove $\forall x, x' \in X$ that

$$(g \circ f)(x) = (g \circ f)(x') \implies x = x'$$

Suppose that $(g \circ f)(x) = (g \circ f)(x')$. Expanding out the definition of $g \circ f$, this means that $g(f(x)) = g(f(x'))$.

Since g is injective and $g(f(x)) = g(f(x'))$, we know $f(x) = f(x')$.

Similarly, since f is injective and $f(x) = f(x')$, we know that $x = x'$, as required. □

Proposition 2.3.2. f is injective if and only if for any set Z and any functions $g_1, g_2 : Z \to X$ we have $f \circ g_1 = f \circ g_2 \implies g_1 = g_2$.

Proof. **Forward direction:**

If f is injective, we ultimately wish to show that $g_1 = g_2$, so in order to do this we consider all possible inputs $z \in Z$, hoping to show that $g_1(z) = g_2(z)$.

But this is quite simple because we are given that $f \circ g_1 = f \circ g_2$ and that f is injective, so

$$f \circ g_1(z) = f \circ g_2(z) \implies g_1(z) = g_2(z)$$

Backward direction:

We specifically pick $Z = \{1\}$, basically some random one-element set.

Then $\forall x, y \in X$, we define

$$g_1 : Z \to X, g_1(1) = x$$
$$g_2 : Z \to Y, g_2(1) = y$$

Then

$$f(x) = f(y) \implies f(g_1(1)) = f(g_2(1)) \implies g_1(1) = g_2(1) \implies x = y$$

□

Definition 2.3.6. $f : X \to Y$ is **surjective** if every element of Y is mapped to at least one element of X.

$$\forall y \in Y,\ \exists x \in X \text{ s.t. } f(x) = y$$

Proposition 2.3.3. If $f : X \to Y$ is surjective and $g : Y \to Z$ is surjective, then $g \circ f : X \to Z$ is surjective.

Proof. Let $f : X \to Y$ and $g : Y \to Z$ be arbitrary surjective functions. We want to prove that the function $g \circ f : X \to Z$ is subjective.

To do so, we want to prove that for any $z \in Z$, there is some $x \in X$ such that $(g \circ f)(x) = z$. Equivalently, we want to prove that for any $z \in Z$, there is some $x \in X$ such that $g(f(x)) = z$.

Consider any $z \in Z$. Since $g : Y \to Z$ is surjective, there is some $y \in Y$ such that $g(y) = z$. Similarly, since $f : X \to Y$ is surjective, there is some $x \in X$ such that $f(x) = y$. This means that there is some $x \in X$ such that $g(f(x)) = g(y) = z$, as required. □

Proposition 2.3.4. f is surjective if and only if for any set Z and any functions $g_1, g_2 : Y \to Z$ we have $g_1 \circ f = g_2 \circ f \implies g_1 = g_2$.

Proof. **Forward direction:**

First we suppose that f is surjective. Again, we wish to show that $g_1 = g_2$, so we need to consider every possible input y in Y. Then, since f is injective, we can always pick $x \in X$ such that $f(x) = y$.

Then

$$g_1 \circ f = g_2 \circ f \implies g_1 \circ f(x) = g_2 \circ f(x) \implies g_1(y) = g_2(y)$$

On the other hand, if f is not surjective, then there exists $y \in Y$ such that for all $x \in X$ we have $f(x) \neq y$. We then aim to construct set Z and $g_1, g_2 : Y \to Z$ such that

(i) $g_1(y) \neq g_2(y)$

(ii) $\forall y' \neq y, g_1(y') = g_2(y')$

Because if this is satisfied, then $\forall x \in X$, since $f(x) \neq y$ we have from (ii) that $g_1(f(x)) = g_2(f(x))$; thus $g_1 \circ f = g_2 \circ f$, and yet from (i) we have $g_1 \neq g_2$.

Backward direction:

We construct $Z = Y \cup \{1, 2\}$ for some random $1, 2 \notin Y$.

Then we define

$$g_1 : Y \to Z, g_1(y) = 1, g_1(y') = y' \qquad g_2 : Y \to Z, g_2(y) = 2, g_2(y') = y'$$

Then when y is not in the image of f, these two functions will satisfy $g_1 \circ f = g_2 \circ f$ but not $g_1 = g_2$.

So conversely, if for any set Z and any functions $g_i : Y \to Z$ we have $g_1 \circ f = g_2 \circ f \implies g_1 = g_2$, such a value y that is in the codomain but not in the range of f cannot appear, and hence f must be surjective. □

Definition 2.3.7. $f : X \to Y$ is **bijective** if it is both injective and surjective: each element of Y is mapped to a unique element of X.

§2.3.3 Cardinality and Countable Sets

When do two sets have the same size? Cantor answered this question in the 1800s, stating that two sets have the same size when you can pair each element in one set with a unique element in the other.

Definition 2.3.8 (Cardinality)**.** X and Y have the same **cardinality** if there exists a bijection $f : X \to Y$, denoted by $|X| = |Y|$.

Definition 2.3.9 (Cardinality of finite sets)**.** The empty set $\varnothing$ is finite and has cardinality $|\varnothing| = 0$. A non-empty set A is said to be **finite** and have cardinality $|A| = n \in \mathbb{Z}^+$ if and only if there exists a bijection from A to the set $\{1, 2, \ldots, n\}$.

Remark. Note that for finite sets X and Y, a function $f : X \to Y$ can only be **injective** if $|Y| \geq |X|$, since for any injective function the number of elements in the image $f(X)$, is equal to the number of elements in the domain, and $f(X) \subseteq Y$. In other words, the codomain of an injective function cannot be smaller than the domain.[1]

Similarly, a function $f : X \to Y$ can only be **surjective** if $|Y| \leq |X|$. Hence if f is bijective, then $|X| = |Y|$; that is, the domain and codomain of a bijection have equal cardinality. (These results hold true for infinite sets too, though less obviously).

Theorem 2.3.1 (Cantor–Schröder–Bernstein)**.** If $|X| \leq |Y|$ and $|Y| \leq |X|$ then $|X| = |Y|$.

Definition 2.3.10 (Countably infinite)**.** A set X is **countably infinite** if it has the same cardinality as the set $\mathbb{Z}^+$.

Definition 2.3.11 (Countable)**.** A set X is **countable** if it is either finite or infinitely countable.

Example 2.3.1. The set $2\mathbb{Z}^+ = \{2n \mid n \in \mathbb{Z}^+\}$ is countably infinite, i.e. $|2\mathbb{Z}^+| = |\mathbb{Z}^+|$.

To prove this, define the function $f : \mathbb{Z}^+ \to 2\mathbb{Z}^+$ as $f(n) = 2n$. Then, f is injective – if $f(n) = f(m)$ then $2n = 2m \implies n = m$. Furthermore, f is surjective, as if $m \in 2\mathbb{Z}^+$ then $\exists n \in \mathbb{Z}^+$ such that $m = 2n = f(n)$.

Example 2.3.2. $\mathbb{Z}$ is countable since we have a bijection $f : \mathbb{Z}^+ \to \mathbb{Z}$ given by

$$f(k) = \begin{cases} \dfrac{k}{2} & \text{if } k \text{ is even} \\ \dfrac{1-k}{2} & \text{if } k \text{ is odd} \end{cases}$$

In other words, $f(1) = 0, f(2) = 1, f(3) = -1, f(4) = 2, f(5) = -2, f(6) = -3, \ldots$ where our function f stretches in both positive and negative directions.

Theorem 2.3.2 (Cantor)**.** For a set A, $|A| < |\mathcal{P}(A)|$.

Proof. Define the function $f : A \to \mathcal{P}(A)$ by $f(x) = \{x\}$. Then, f is injective as $\{x\} = \{y\} \implies x = y$. Thus $|A| \leq |\mathcal{P}(A)|$. To finish the proof now all we need to show is that $|A| \neq |\mathcal{P}(A)|$. We will do so through contradiction. Suppose that $|A| = |\mathcal{P}(A)|$. Then, there exists a surjection $g : A \to \mathcal{P}(A)$. We define the set B as

$$B := \{x \in A \mid x \notin g(x)\} \in \mathcal{P}(A)$$

Since g is surjective, there exists a $b \in A$ such that $g(b) = B$. There are two cases:

1. $b \in B$. Then $b \notin g(b) = B \implies b \notin B$.
2. $b \notin B$. Then $b \notin g(b) = B \implies b \in B$.

In either case we obtain a contradiction. Thus, g is not surjective so $|A| \neq |\mathcal{P}(A)|$. □

Corollary 2.3.1. For all $n \in \mathbb{N} \cup \{0\}$, $n < 2^n$.

Proof. This can be easily proven through induction. □

[1] This is also referred to as the pigeonhole principle: if n letters are placed in m pigeonholes and $n > m$, then at least one hole must contain more than one letter; the non-injective function in that case is the assignment of pigeonholes to letters.

Lemma 2.3.1. If X is a countable set and $B \subseteq A$ then B is countable.

Lemma 2.3.2. If $\{A_1, A_2, \dots\}$ is a collection of countably many countable sets then the set $\bigcup_{i=1}^{\infty} A_i$ is countable.

Lemma 2.3.3. If $\{A_1, A_2, \dots, A_n\}$ is a collection of finitely many countable sets then the set $A_1 \times \cdots \times A_n$ is countable.

§2.3.4 Composition of functions and invertibility

Definition 2.3.12. Given two functions $f : X \to Y$ and $g : Y \to Z$, the **composition** $g \circ f : X \to Z$ is defined by

$$(g \circ f)(x) = g(f(x)) \quad \forall x \in X.$$

The composition of functions is not commutative. However, composition is associative, as the following results shows:

Proposition 2.3.5 (Associativity)**.** Let $f : X \to Y$, $g : Y \to Z$, $h : Z \to W$ be three functions. Then

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

Proof. Let $x \in X$. Then, by the definition of composition, we have

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

□

The following proposition addresses the extent to which composition of functions preserves injectivity and surjectivity:

Proposition 2.3.6. Let $f : X \to Y$ and $g : Y \to Z$ be functions.

(i) If f and g are injective then so is $g \circ f$. Conversely, if $g \circ f$ is injective, then f is injective, but g need not be.

(ii) If f and g are surjective then so is $g \circ f$. Conversely, if $g \circ f$ is surjective, then g is surjective, but f need not be.

Proof. For the first part of (i), suppose $(g \circ f)(x_1) = (g \circ f)(x_2)$ for some $x_1, x_2 \in X$. From the injectivity of g we know that $g(f(x_1)) = g(f(x_2))$ implies $f(x_1) = f(x_2)$, and then from the injectivity of f we know that this implies $x_1 = x_2$. So $g \circ f$ is injective.

For the second part of (i), suppose $f(x_1) = f(x_2)$ for some $x_1, x_2 \in X$. Then applying g gives $g(f(x_1)) = g(f(x_2))$, and by the injectivity of $g \circ f$ this means $x_1 = x_2$. So f is injective. To see that g need not be injective, a counterexample is $X = Z = \{0\}, Y = \mathbb{R}$, with $f(0) = 0$ and $g(y) = 0$ for all $y \in \mathbb{R}$. □

Recalling that id_X is the identity map on X, we can define invertibility:

Definition 2.3.13. A function $f : X \to Y$ is **invertible** if there exists a function $g : Y \to X$ such that $g \circ f = \mathrm{id}_X$ and $f \circ g = \mathrm{id}_Y$. The function g is the **inverse** of f, denoted by $g = f^{-1}$.

Note that directly from the definition, if f is invertible then f^{-1} is also invertible, and $(f^{-1})^{-1} = f$.

An immediate concern we might have is whether there could be multiple such functions g, in which case the inverse f^{-1} would not be well-defined. This is resolved by the following result:

Proposition 2.3.7 (Uniqueness of inverse)**.** If $f : X \to Y$ is invertible then its inverse is unique.

Proof. Let g_1 and g_2 be two functions for which $g_i \circ f = \mathrm{id}_X$ and $f \circ g_i = \mathrm{id}_Y$. Using the fact that composition is associative, and the definition of the identity maps, we can write

$$g_1 = g_1 \circ \mathrm{id}_Y = g_1 \circ (f \circ g_2) = (g_1 \circ f) \circ g_2 = \mathrm{id}_X \circ g_2 = g_2$$

□

The following result shows how to invert the composition of invertible functions:

Proposition 2.3.8. Let $f : X \to Y$ and $g : Y \to Z$ be functions. If f and g are invertible, then $g \circ f$ is invertible, and $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.

Proof. Making repeated use of the fact that function composition is associative, and the definition of the inverses f^{-1} and g^{-1}, we note that

$$\begin{aligned}(f^{-1} \circ g^{-1}) \circ (g \circ f) &= ((f^{-1} \circ g^{-1}) \circ g) \circ f \\ &= (f^{-1} \circ (g^{-1} \circ g)) \circ f \\ &= (f^{-1} \circ \mathrm{id}_Y) \circ f \\ &= f^{-1} \circ f \\ &= \mathrm{id}_X\end{aligned}$$

and similarly,

$$\begin{aligned}(g \circ f) \circ (f^{-1} \circ g^{-1}) &= g \circ (f \circ (f^{-1} \circ g^{-1})) \\ &= g \circ ((f \circ f^{-1}) \circ g^{-1}) \\ &= g \circ (\mathrm{id}_Y \circ g^{-1}) \\ &= g \circ g^{-1} \\ &= \mathrm{id}_Z\end{aligned}$$

which shows that $f^{-1} \circ g^{-1}$ satisfies the properties required to be the inverse of $g \circ f$. □

The following result provides an important and useful criterion for invertibility:

Theorem 2.3.3. A function $f : X \to Y$ is invertible if and only if it is bijective.

Proof. We prove this in both directions.

Forward direction:

Suppose f is invertible, so it has an inverse $f^{-1} : Y \to X$. To show f is injective, suppose that for some $x_1, x_2 \in X$ we have $f(x_1) = f(x_2)$. Then applying f^{-1} to both sides and noting that by definition $f^{-1} \circ f = \mathrm{id}_X$, we see that $x_1 = f^{-1}(f(x_1)) = f^{-1}(f(x_2)) = x_2$. So f is injective. To show that f is surjective, let $y \in Y$, and note that $f^{-1}(y) \in X$ has the property that $f(f^{-1}(y)) = y$. So f is surjective. Therefore f is bijective.

Backward direction:

Suppose that f is bijective, we aim to show that there is a well-defined $g : Y \to X$ such that $g \circ f = \mathrm{id}_X$ and $f \circ g = \mathrm{id}_Y$. Since f is surjective, we know that for any $y \in Y$, there is an $x \in X$ such that $f(x) = y$. Furthermore, since f is injective, we know that this x is unique. So for each $y \in Y$ there is a unique $x \in X$ such that $f(x) = y$. This recipe provides a well-defined function $g(y) = x$, for which we have $g(f(x)) = x$ for any $x \in X$ and $f(g(y)) = y$ for any $y \in Y$. So g satisfies the property required to be an inverse of f and therefore f is invertible. □

It is also possible to define left-inverse and right-inverse functions as functions that partially satisfy the definition of the inverse:

Definition 2.3.14. A function $f : X \to Y$ is **left invertible** if there exists a function $g : Y \to X$ such that $g \circ f = \mathrm{id}_X$, and is **right invertible** if there exists a function $h : Y \to X$ such that $f \circ h = \mathrm{id}_Y$.

As may be somewht apparent from the previous proof, being left- and right-invertible is equivalent to being injective and surjective, respectively. We leave this as an exercise to show.

§2.3.5 Strictly Monotonic Functions / Increasing and Decreasing Functions

Definition 2.3.15. A function f is **strictly monotonic increasing** if for all $x_1, x_2 \in D_f$,

$$x_2 > x_1 \iff f(x_2) > f(x_1).$$

Definition 2.3.16. A function f is **strictly monotonic decreasing** if for all $x_1, x_2 \in D_f$,

$$x_2 > x_1 \iff f(x_2) < f(x_1).$$

Definition 2.3.17. A function f is **increasing** on an interval I if for all $x_1, x_2 \in I$ with $x_1 < x_2$, $f(x_1) \le f(x_2)$.

Definition 2.3.18. A function f is **decreasing** on an interval I if for all $x_1, x_2 \in I$ with $x_1 < x_2$, $f(x_1) \ge f(x_2)$.

Locate roots of an equation: suppose $f(x)$ is continuous in the interval $[a, b]$,

- If $f(a)$ and $f(b)$ have *opposite* signs, i.e. $f(a)f(b) < 0$, then there is an odd number of real roots (counting repeated) in $[a, b]$.

  Furthermore, if f is either strictly increasing or decreasing in $[a, b]$, then $f(x) = 0$ has *exactly one real root* in $[a, b]$.
- If $f(a)$ and $f(b)$ have *same* signs, i.e. $f(a)f(b) > 0$, then there is an even number of roots (counting repeated) in $[a, b]$.

§2.3.6 Convex and Concave Functions

Definition 2.3.19. A function f is **convex** if for all $x_1, x_2 \in D_f$ and $0 \le t \le 1$, we have

$$f(tx_1 + (1-t)x_2) \le tf(x_1) + (1-t)f(x_2).$$

Note that equality holds when $x_1 = x_2$.

Definition 2.3.20. A function f is **strictly convex** if for all $x_1, x_2 \in D_f$ with $x_1 \ne x_2$ and $0 < t < 1$, we have

$$f(tx_1 + (1-t)x_2) < tf(x_1) + (1-t)f(x_2).$$

Definition 2.3.21. A function f is **concave** if for all $x_1, x_2 \in D_f$ and $0 \le t \le 1$, we have

$$f(tx_1 + (1-t)x_2) \ge tf(x_1) + (1-t)f(x_2).$$

Note that equality holds when $x_1 = x_2$.

Definition 2.3.22. A function f is **strictly concave** if for all $x_1, x_2 \in D_f$ with $x_1 \ne x_2$ and $0 < t < 1$, we have

$$f(tx_1 + (1-t)x_2) > tf(x_1) + (1-t)f(x_2).$$

§2.3.7 Other Functions

Piecewise Functions

A function that has its domain divided into *separate partitions* and each partition of the domain given a different formula or rule is known as a **piecewise funtion**, i.e. a function defined "piece-wise".

- Absolute Value Function

  The **absolute function** of x, denoted by $|x|$, is the distance of x from 0 on the real number line. Distances are always non-negative, so for every $x \in \mathbb{R}$ we have

  $$|x| \ge 0.$$

  The most basic absolute value function is $f(x) = |x|$:

  $$|x| = \begin{cases} -x & x < 0 \\ x & x \ge 0 \end{cases}$$

- Floor Function

  We define the **floor function** $f(x) = \lfloor x \rfloor$ as the greatest integer smaller than or equal to x.

  For $x \in \mathbb{R}$ and $n \in \mathbb{Z}$,
  $$\lfloor x \rfloor = n \iff n \le x < n + 1.$$

- Ceiling Function

  The ceiling function $f(x) = \lceil x \rceil$ is the direct opposite of the floor function; it maps all real numbers in the domain to the smallest integer not smaller than it.
  $$\lceil x \rceil = \begin{cases} \lfloor x \rfloor + 1 & x \notin \mathbb{Z} \\ \lfloor x \rfloor & x \in \mathbb{Z} \end{cases}$$

Exercise 2.3.2

Prove that

(a) $\lfloor \sqrt{x} \rfloor = \left\lfloor \sqrt{\lfloor x \rfloor} \right\rfloor$

(b) $\lceil \sqrt{x} \rceil = \left\lceil \sqrt{\lceil x \rceil} \right\rceil$

Solution.

(a)

$$\begin{aligned} & \lfloor \sqrt{x} \rfloor = n \\ \iff & n \le \sqrt{x} < n + 1 \quad \text{[by definition of floor function]} \\ \iff & n^2 \le x < (n+1)^2 \quad \text{[square both sides]} \\ \iff & n^2 \le \lfloor x \rfloor \le x < (n+1)^2 \\ \iff & n \le \sqrt{\lfloor x \rfloor} < n + 1 \quad \text{[take square root throughout]} \\ \iff & \left\lfloor \sqrt{\lfloor x \rfloor} \right\rfloor = n \quad \text{[by definition of floor function]} \end{aligned}$$

(b)

$$\begin{aligned} & \lceil \sqrt{x} \rceil = n + 1 \\ \iff & n < \sqrt{x} \le n + 1 \quad \text{[by definition of ceiling function]} \\ \iff & n^2 < x \le (n+1)^2 \quad \text{[square both sides]} \\ \iff & n^2 < x \le \lceil x \rceil \le (n+1)^2 \\ \iff & n < \sqrt{\lceil x \rceil} \le n + 1 \quad \text{[take square root throughout]} \\ \iff & \left\lceil \sqrt{\lceil x \rceil} \right\rceil = n + 1 \quad \text{[by definition of ceiling function]} \end{aligned}$$

□

Symmetrical Functions

There are special functions with some form of geometric symmetry.

- Even Functions

  f is **even** if $f(-x) = f(x)$ for every $x \in D_f$.

  The graph of an even function is symmetric about the y-axis.

- Odd Functions

  f is **odd** if $f(-x) = -f(x)$ for every $x \in D_f$.

  The graph of an odd function is symmetric about the origin.

- Periodic Functions

  f is **periodic** if $f(x+p) = f(x)$ for every $x \in D_f$, where p is a positive constant. The smallest such p is known as the period.

Exercise 2.3.3

For a triangle ABC with corresponding angles a, b and c, show that

$$\sin a + \sin b + \sin c \le \frac{3\sqrt{3}}{2}$$

and determine when equality holds. (Hint: $y = \sin x$ is concave)

Solution. Since $f(x) = \sin x$ is strictly concave on $[0, \pi]$,

$$\begin{aligned}
&\frac{1}{3}f(a) + \frac{1}{3}f(b) + \frac{1}{3}f(c) \\
&= \frac{1}{3}f(a) + \frac{2}{3}\left(\frac{1}{2}f(b) + \frac{1}{2}f(c)\right) \\
&\le \frac{1}{3}f(a) + \frac{2}{3}\left(f\left(\frac{b}{2} + \frac{c}{2}\right)\right) \quad \text{[Concavity Inequality]} \\
&\le f\left(\frac{a}{3} + \frac{2}{3}\left(\frac{b+c}{2}\right)\right) \quad \text{[Concavity Inequality]} \\
&= f\left(\frac{a+b+c}{3}\right)
\end{aligned}$$

Hence

$$\sin a + \sin b + \sin c = f(a) + f(b) + f(c) \le 3f\left(\frac{a+b+c}{3}\right) = 3\sin\frac{\pi}{3} = \frac{3\sqrt{3}}{2}.$$

Equality holds when $a = b = c$, i.e. when ABC is an equilateral triangle. □

Exercises

Problem 29 (H3M Specimen N03)**.** Functions f and g are defined for $x \in \mathbb{R}$ by

$$f(x) = ax + b, \quad g(x) = cx + d$$

where a, b, c, d are constants with $a =\neq 0$. Given that $gf = f^{-1}g$, show that

- either g is a constant function, i.e. $g(x)$ is constant for all $x \in \mathbb{R}$,
- or f^2 is the identity function, i.e. $ff(x) = x$ for all $x \in \mathbb{R}$,
- or g^2 is the identity function.

[9]

Proof. Given that $gf = f^{-1}g$,

$$\begin{aligned} cf(x) + d &= f^{-1}(cx + d) \\ c(ax + b) + d &= \frac{(cx + d) - b}{a} \\ a^2cx + abc + ad &= cx + d - b \end{aligned}$$

Comparing coefficients,

$$\begin{cases} a^c = c \\ c(a - 1)(a + 1) = 0 \\ abc + ad = d - b \end{cases}$$

and we have three cases to work with. □

Problem 30. Let A be the set of all complex polynomials in n variables. Given a subset $T \subset A$, define the *zeros* of T as the set

$$Z(T) = \{P = (a_1, \dots, a_n) \mid f(P) = 0 \text{ for all } f \in T\}$$

A subset $Y \in \mathbb{C}^n$ is called an algebraic set if there exists a subset $T \subset A$ such that $Y = Z(T)$.

Prove that the union of two algebraic sets is an algebraic set.

Proof. We would like to consider $T = \{f_1, f_2, \dots\}$ expressed as indexed sets $T = \{f_i\}$. Then $Z(T)$ can also be expressed as $\{P \mid \forall i, f_i(P) = 0\}$.

Suppose that we have two algebraic sets X and Y. Let $X = Z(S)$, $Y = Z(T)$ where S, T are subsets of A (basically, they are certain sets of polynomials). Then

$$X = \{P \mid \forall f \in S, f(P) = 0\}$$

$$Y = \{P \mid \forall g \in T, g(P) = 0\}$$

We imagine that for $P \in X \cap Y$, we have $f(P) = 0$ or $g(P) = 0$. Hence we consider the set of polynomials

$$U = \{f \cdot g \mid f \in S, g \in T\}$$

For any $P \in X \cup Y$ and for any $fg \in U$ where $f \in S$ and $f \in g$, either $f(P) = 0$ or $g(P) = 0$, hence $fg(P) = 0$ and thus $P \in Z(U)$.

On the other hand if $P \in Z(U)$, suppose otherwise that P is not in $X \cup Y$, then P is neither in X nor in Y. This means that there exists $f \in S, g \in T$ such that $f(P) \neq 0$ and $g(P) \neq 0$, hence $fg(P) \neq 0$. This is a contradiction as $P \in Z(U)$ implies $fg(P) = 0$. Hence we have $X \cup Y = Z(U)$ and thus $X \cup Y$ is an algebraic set.

Now the other direction is simpler and can actually be generalised: The intersection of arbitrarily many algebraic sets is algebraic.

The basic result is that if $X = Z(S)$ and $Y = Z(T)$ then $X \cap Y = Z(S \cup T)$. □

Problem 31 (Modular Arithmetic)**.** Define the ring of integers modulo n:

$$\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}/\sim \text{ where } x \sim y \iff x - y \in n\mathbb{Z}.$$

The equivalence classes are called congruence classes modulo n.

(a) Define the sum of two congruence classes modulo n, $[x], [y] \in \mathbb{Z}/n\mathbb{Z}$, by

$$[x] + [y] = [x + y]$$

Show that the above definition is well-defined.

(b) Define the product of two congruence classes modulo n and show that such a definition is well-defined.

Solution.

(a) We often define such concepts by considering the **representatives** of the equivalence classes.

For example, here we define $[x] + [y] = [x + y]$ for two elements $[x]$ and $[y]$ in $\mathbb{Z}/n\mathbb{Z}$. So what we know here are the classes $[x]$ and $[y]$. But what exactly are x and y? They are just some element in the equivalence classes that was arbitrarily picked out. We then perform the sum $x + y$, and consequently, we used this to point towards the class $[x + y]$.

However, x and y are arbitrarily picked. We want to show that, regardless of which representatives are chosen from the equivalence classes $[x]$ and $[y]$, we will always obtain the same result.

In the definition itself, we have defined that, for the two representatives x and y we define $[x]+[y] = [x + y]$. So now, let's say that we take two other arbitrary representatives, $x' \in [x]$ and $y' \in [y]$. Then by definition, we have

$$[x] + [y] = [x' + y']$$

Thus, our goal is to show that $x' + y'] = [x + y]$. This expression means that the two sides of the equation are referring to the same equivalence class. Therefore, the expression above is completely equivalent to the condition.

$$x' + y' \sim x + y$$

We then check that this final expression is indeed true: Since $x' \in [x]$ and $y' \in [y]$, we have

$$\begin{aligned} & x' \sim x \text{ and } y' \sim y \\ \implies & x' - x, y' - y \in n\mathbb{Z} \\ \implies & (x' + y') - (x + y) = (x' - x) + (y' - y) \in n\mathbb{Z} \end{aligned}$$

(b) The product of two congruence classes is defined by

$$[x][y] = [xy]$$

For any other representatives x', y' we have

$$\begin{aligned} & x'y' - xy \\ & = x'y' - xy' + xy' - xy \\ & = (x' - x)y' + x(y' - y) \in n\mathbb{Z} \end{aligned}$$

Thus $[x'y'] = [xy]$ and the product is well-defined.

□

Problem 32. Let $A = \mathbb{R}$ and for any $x, y \in A$, $x \sim y$ if and only if $x - y \in \mathbb{Z}$. For any two equivalence classes $[x], [y] \in A/\sim$, define

$$[x] + [y] = [x + y] \text{ and } -[x] = [-x]$$

(a) Show that the above definitions are well-defined.

(b) Find a one-to-one correspondence $\phi : X \to Y$ between $X = A/\sim$ and $Y : |z| = 1$, i.e. the unit circle in $\mathbb{C}$, such that for any $[x_1], [x_2] \in X$ we have

$$\phi([x_1])\phi([x_2]) = \phi([x_1 + x_2])$$

(c) Show that for any $[x] \in X$,

$$\phi(-[x]) = \phi([x])^{-1}$$

Solution.

(a)

$$(x' + y') - (x + y) = (x' - x) + (y' - y) \in \mathbb{Z}$$

Thus $[x' + y'] = [x + y]$

$$(-x') - (-x) = -(x' - x) \in \mathbb{Z}$$

Thus $[-x'] = [-x]$.

(b) Complex numbers in the polar form: $z = re^{i\theta}$

Then the correspondence is given by $\phi([x]) = e^{2\pi i x}$

$$[x] = [y] \iff x - y \in \mathbb{Z} \iff e^{2\pi i(x-y)} = 1 \iff e^{2\pi i x} = e^{2\pi i y}$$

Hence this is a bijection.

Before that, we also need to show that ϕ is well-defined, which is almost the same as the above.

If we choose another representative x' then

$$\phi([x]) = e^{2\pi i x'} = e^{2\pi i x} \cdot e^{2\pi i(x'-x)} = e^{2\pi i x}$$

(c) You can either refer to the specific correspondence $\phi([x]) = e^{2\pi i x}$ or use its properties.

$$\phi(-[x])\phi([x]) = \phi([-x])\phi([x]) = \phi([-x + x]) = \phi([0]) = 1$$

□

Problem 33 (Set of Rational Numbers)**.** Let $\mathbb{Z}$ be the set of integers, and let $\mathbb{Z}^*$ be the set of nonzero integers. We define

$$\mathbb{Q} = \{(a,b) \mid a \in \mathbb{Z}, b \in \mathbb{Z}^*\}/\sim$$

where

$$(a,b) \sim (c,d) \iff ad = bc.$$

Let $\dfrac{a}{b}$ denote the equivalence class for (a,b). Such an equivalence class is called a rational number.

(a) For any two rational numbers $\dfrac{a}{b}$ and $\dfrac{c}{d}$, their sum is determined by

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

Show that the above definition is well-defined.

(b) Define the product of two rational numbers and show that such a definition is well-defined.

(c) Prove that for every equivalence class $\dfrac{a}{b} \in \mathbb{Q}$, there exists a unique integer pair (p,q) satisfying the following properties:

$$q > 0, (p,q) = 1 \text{ and } (p,q) \in \frac{a}{b}.$$

(d) Using the partial order of $\mathbb{Z}$, define the partial order of $\mathbb{Q}$.

Solution.

(a) For this problem, we are dealing with a "hidden" equivalence class.

The expressions $\frac{a}{b}$ and $\frac{c}{d}$ themselves are derived from their representatives (a,b) and (c,d).

So suppose that we choose other representatives (a',b') and (c',d'), then the sum would be

$$\frac{a'}{b'} + \frac{c'}{d'} = \frac{a'd'+b'c'}{b'd'}$$

We now have to show that $\frac{ad+bc}{bd} = \frac{a'd'+b'c'}{b'd'}$:

$$\begin{aligned}\frac{ad+bc}{bd} &\iff (ad+bc, bd) \sim (a'd'+b'c', b'd') \\ &\iff (ad+bc)b'd' = (a'd'+b'c')bd\end{aligned}$$

$$\begin{gathered}\frac{a}{b} = \frac{a'}{b'} \\ (a,b) \sim (a',b') \\ ab' = a'b\end{gathered}$$

Hence

$$\begin{aligned}(ad+bc)b'd' &= ab'dd' + bb'cd' \\ &= a'bdd' + bb'c'd \\ &= (a'd'+b'c')bd\end{aligned}$$

(b) The definition would be $\frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}$.

This is actually a lot simpler to check.

$$a'c'bd = (a'b)(c'd) = (ab')(cd') = acb'd'$$

Hence $\frac{a'c'}{b'd'} = \frac{ac}{bd}$.

(c) We basically try to do this step by step as we would in simplifying fractions.

First pick b to be positive, otherwise we swap a and b with $-a$ and $-b$.

Then simplify the common factors. For this one we let $(a, b) = d$, and $a = dp, b = dq$. Then (p, q) is the pair that we need

(d) In order to define the partial order we need to account for whether the denominators are negative.

$\frac{a}{b} \leq \frac{c}{d}$, and if $b, d > 0$ then we can safely draw a connection to the expression $ad \leq bc$

In order to show that this does in fact give a partial order we check that

(a) 1: $ab \leq ab$ and hence $\frac{a}{b} \leq \frac{a}{b}$

(b) 2: If $\frac{a}{b} \leq \frac{c}{d}$ and $\frac{c}{d} \leq \frac{a}{b}$, then $ad \leq bc$ and $bc \leq ad$, hence $ad = bc$ and thus $\frac{a}{b} = \frac{c}{d}$

(c) 3: This is trickier due to complications arising from inequalities and multiplication

If $\frac{a}{b} \leq \frac{c}{d}$ and $\frac{c}{d} \leq \frac{e}{f}$, note that $b, d, f > 0$ and so $ad \leq bc$ and $cf \leq de$.

i) $e < 0$, then $c < 0$ and $a < 0$, thus $-ad \geq -bc$, $-cf \geq -de$ and we have $acdf \geq bcde$

$af \leq be(c < 0, d > 0)$

Thus $\frac{a}{b} \leq \frac{e}{f}$

ii) $e \geq 0$ but $a < 0$, then $af < 0 \leq be$ and thus $\frac{a}{b} < \frac{e}{f}$

iii) $a \geq 0$, then $c \geq 0$ and $e \geq 0$, and we have the ordinary case.

Hence proven.

□

Problem 34 (Complex Numbers)**.** Let $\mathbb{R}[x]$ denote the set of real polynomials. Define

$$\mathbb{C} = \mathbb{R}[x]/(x^2+1)\mathbb{R}[x]$$

where

$$f(x) \sim g(x) \iff x^2+1 \text{ divides } f(x)-g(x).$$

The complex number $a+bi$ is defined to be the equivalence class of $a+bx$.

(a) Define the sum and product of two complex numbers and show that such definitions are well-defined.

(b) Define the reciprocal of a complex number.

Part II

Real Analysis

3 The Real Number System

§3.1 Rational numbers $\mathbb{Q}$

Rational numbers $\mathbb{Q}$ can be introduced following a general procedure called the construction of field of fractions. Every rational number can be written as

$$\frac{m}{n}, \quad m, n \in \mathbb{Z}, n \neq 0.$$

Moreover, every nonzero rational number can be uniquely written as

$$\frac{p}{q}, \quad \in \mathbb{Z}^{+}, q \in \mathbb{Z}, \gcd(p, q) = 1.$$

§3.1.1 $\mathbb{Q}$ is a field

We explain the meaning of a **field** using $\mathbb{Q}$ as an example.

Proposition 3.1.1. $(\mathbb{Q}, +, \cdot)$ is a field, which means

- $(\mathbb{Q}, +)$ is an abelian group with identity 0:
  - $\mathbb{Q}$ is closed under addition $+$ ($+$ is a binary operation over $\mathbb{Q}$).
  - Addition $+$ is associative: $(a + b) + c = a + (b + c)$.
  - Addition $+$ has identity element 0: $a + 0 = 0 = 0 + a = a$.
  - Any element has inverse element: $a + (-a) = (-a) + a = 0$.
  - Addition $+$ is commutative: $a + b = b + a$.
- $(\mathbb{Q}, +, \cdot)$ is a commutative, unital ring:
  - $\mathbb{Q}$ is closed under multiplication $\cdot$ ($\cdot$ is a binary operation over $\mathbb{Q}$).
  - Multiplication $\cdot$ is associative: $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.
  - Multiplication $\cdot$ has unity element 1: $a \cdot 1 = 1 \cdot a = a$.
  - Multiplication $\cdot$ is commutative: $a \cdot b = b \cdot a$.
  - Addition and multiplication satisfy distribution law: $(a + b) \cdot c = a \cdot c + b \cdot c$.
- Any nonzero element has a multiplicative inverse: $a \cdot \frac{1}{a} = \frac{1}{a} \cdot a = 1$ for any $a \in \mathbb{Q}, a \neq 0$.

Exercise 3.1.1

Prove that $(\mathbb{Z}, +, \cdot)$ is a commutative, unital ring, but it is not a field.

Proof.

- Check $(\mathbb{Z}, +, \cdot)$ is a commutative, unital ring.
- The number $2 \in \mathbb{Z}$ (in fact, every nonzero number except ± 1) has no multiplication inverse in $\mathbb{Z}$.

□

§3.1.2 $\mathbb{Q}$ is an ordered set

Let A be a set.

Definition 3.1.1. An **order** on A is a relation, denoted by $<$, with the following two properties:

(i) $\forall x, y \in A$, one and only one of the following statements is true: $x < y, x = y, y < x$.

(ii) $\forall x, y, z \in A$, if $x < y$ and $y < z$, then $x < z$.

Notation. The notation $x \leq y$ indicates that $x < y$ or $x = y$, without specifying which of these two is to hold. In other words, $x \leq y$ is the negation of $x > y$.

Definition 3.1.2. An **ordered set** is a set S in which an order is defined.

Example 3.1.1. $\mathbb{Q}$ is an ordered set if $r < s$ is defined to mean that $s - r$ is a positive rational number.

Let $A \subset \mathbb{R}$.

Definition 3.1.3. A is **bounded from above** if there exists an **upper bound** $M \in \mathbb{R}$ such that $x \leq M$ for all $x \in A$. A is **bounded from below** if there exists a **lower bound** $m \in \mathbb{R}$ such that $x \geq m$ for all $x \in A$. A is **bounded** in $\mathbb{R}$ if it is bounded above and below.

Definition 3.1.4. The **supremum** (or *least upper bound*) of A, denoted by $\sup A$, is the smallest real number M such that $x \leq M$ for all $x \in A$.

(i) M is an upper bound for A.

(ii) If N is an upper bound for A, then $M \leq N$.

Remark. If $M \in A$, then M is the *maximum value* of A.

The following proposition is convenient in working with suprema.

Proposition 3.1.2. Let A be a nonempty subset of $\mathbb{R}$ that is bounded above. Then $M = \sup A$ if an only if

(i) $x \leq M$ for all $x \in A$

(ii) For any $\varepsilon > 0$, there exists $a \in A$ such that $M - \varepsilon < a$.

Proof. Suppose first that $M = \sup A$. Then clearly (i) holds (since this is identical to condition (1) in the definition of supremum). Now let $\varepsilon > 0$. Since $M - \varepsilon < a$, condition (ii) in the definition of supremum implies that $M - \varepsilon$ is not an upper bound of A. Therefore, there must exist an element a in A such that $M - \varepsilon < a$, as desired. □

Definition 3.1.5. The **infimum** (or *greatest lower bound*) of A, denoted by $\inf A$, is defined as the largest real number m such that $x \geq m$ for all $x \in A$.

(i) m is a lower bound for A.

(ii) If n is a lower bound for A, then $m \geq n$.

Remark. If $m \in A$, then m is the *minimum value* of A.

Proposition 3.1.3 (Uniqueness of suprenum)**.** If a set $A \subset \mathbb{R}$ has a supremum, then it is unique.

Proof. Assume that M and N are suprema of a set A.

Since N is a supremum, it is an upper bound for A. Since M is a supremum, then it is the least upper bound and thus $M \leq N$.

Similarly, since M is a supremum, it is an upper bound for A; since N is a supremum, it is a least upper bound and thus $N \leq M$.

Since $N \leq M$ and $M \leq N$, thus $M = N$. Therefore, a supremum for a set is unique if it exists. □

Theorem 3.1.1 (Comparison Theorem)**.** Let $S, T \subset \mathbb{R}$ be non-empty sets such that $s \leq t$ for every $s \in S$ and $t \in T$. If T has a supremum, then so does S, and $\sup S \leq \sup T$.

Proof. Let $\tau = \sup T$. Since τ is a supremum for T, then $t \leq \tau$ for all $t \in T$. Let $s \in S$ and choose any $t \in T$. Then, since $s \leq t$ and $t \leq \tau$, then $s \leq t$. Thus, τ is an upper bound for S.

By the Completeness Axiom, S has a supremum, say $\sigma = \sup S$. We will show that $\sigma \leq \tau$. Notice that, by the above, τ is an upper bound for S. Since σ is the least upper bound for S, then $\sigma \leq \tau$. Therefore,

$$\sup S \leq \sup T.$$

□

Let's explore some useful properties of sup and inf.

Proposition 3.1.4. Let S, T be non-empty subsets of $\mathbb{R}$, with $S \subseteq T$ and with T bounded above. Then S is bounded above, and $\sup S \leq \sup T$.

Proof. Since T is bounded above, it has an upper bound, say b. Then $t \leq b$ for all $t \in T$, so certainly $t \leq b$ for all $t \in S$, so b is an upper bound for S.

Now S, T are non-empty and bounded above, so by completeness each has a supremum. Note that $\sup T$ is an upper bound for T and hence also for S, so $\sup T \geq \sup S$ (since $\sup S$ is the least upper bound for S). □

Proposition 3.1.5. Let $T \subseteq \mathbb{R}$ be non-empty and bounded below. Let $S = \{-t \mid t \in T\}$. Then S is non-empty and bounded above. Furthermore, $\inf T$ exists, and $\inf T = -\sup S$.

Proof. Since T is non-empty, so is S. Let b be a lower bound for T, so $t \geq b$ for all $t \in T$. Then $-t \leq -b$ for all $t \in T$, so $s \leq -b$ for all $s \in S$, so $-b$ is an upper bound for S.

Now S is non-empty and bounded above, so by completeness it has a supremum. Then $s \leq \sup S$ for all $s \in S$, so $t \geq -\sup S$ for all $t \in T$, so $-\sup S$ is a lower bound for T.

Also, we saw before that if b is a lower bound for T then $-b$ is an upper bound for S. Then $-b \geq \sup S$ (since $\sup S$ is the least upper bound), so $b \leq -\sup S$. So $-\sup S$ is the greatest lower bound.

So $\inf T$ exists and $\inf T = -\sup S$. □

Proposition 3.1.6 (Approximation Property)**.** Let $S \subseteq \mathbb{R}$ be non-empty and bounded above. For any $\varepsilon > 0$, there is $s_\varepsilon \in S$ such that $\sup S - \varepsilon < s_\varepsilon \leq \sup S$.

Proof. Take $\varepsilon > 0$.

Note that by definition of the supremum we have $s \leq \sup S$ for all $s \in S$. Suppose, for a contradiction, that $\sup S - \varepsilon \geq s$ for all $s \in S$.

Then $\sup S - \varepsilon$ is an upper bound for S, but $\sup S - \varepsilon < \sup S$, which is a contradiction.

Hence there is $s_\varepsilon \in S$ with $\sup S - \varepsilon < s_\varepsilon$. □

Theorem 3.1.2. Any set bounded from above/below must have a supremum/infimum.

Proof. We prove the above theorem for supremum, using Dedekind cuts.

Let S be a real number set. We consider the rational number set[1]

$$A = \{x \in \mathbb{Q} \mid \exists y \in S, x < y\}$$

Now we go through the definitions to check that $(A|B)$ is a Dedekind cut

1. Since $S \neq \varnothing$, pick $y \in S$, then $[y] - 1$ is a real number smaller than some element in S, hence $[y] - 1 \in A$ and thus $A \neq \varnothing$.

   Also, since we are given that S is bounded, there exists a positive integer M as an upper bound for S, thus $B \neq \varnothing$. (Note that an upper bound is simply a number that is bigger than anything from the set, and is not the supremum.)

2. We define B to be the complement of A in $\mathbb{Q}$ so condition 2 is trivial.

3. For any $x, y \in A$, if $x < y$ and $y \in A$, then there exists $z \in S$ such that $y < z$, hence $x < z$ and thus $x \in A$.

4. Suppose otherwise that $x \in A$ is the largest element in A, then there exists $y \in S$ such that $x < y$.

   We then pick a rational number z between x and y. Since we still have $z < y$, we have $z \in A$ but $z > x$, contradictory to z being the largest.

Now there is actually an issue with the proof for property 4 here: How exactly are we finding z?

First $x \in Q$. Then $y \in \mathbb{R}$ so we rewrite it as $y = (C|D)$ via definition.

$x < y$ translates to the fact that $x \in C$. Now, since y is real, by definition we know that C must not have a largest element. In particular, x is not largest and we can pick $z \in C$ such that $z > x$. This is in fact the z that we need.

Now that all the properties of a real number are validated, we may finally conclude that $\alpha = (A|B)$ is indeed a real number.

Now we need to show that α is in fact the supremum of S.

Let $x \in S$. If x is not the maximum value of S, i.e. $\exists y \in S$ such that $x < y$. Then $x \in A$ and thus $x < \alpha$.

If x is the maximum value of S, then for any rational number $y < x$ we have $y \in A$, and for any rational number $y \geq x$ we have $y \in B$. Thus $x = (A|B) = \alpha$.

In conclusion, $x \leq \alpha$ for all $x \in S$.

For any upper bound x of S, since $\forall y \in S, x \geq y$ we have $x \in B$ and thus $x \geq \alpha$.

Therefore, α is the smallest upper bound of S and thus $\sup S = \alpha$ exists. □

Definition 3.1.6. An ordered set S is said to have the **least-upper-bound property** if the following is true: If $E \subset S$, E is not empty, and E is bounded above, then $\sup E$ exists in S.

Example 3.2.1 shows that $\mathbb{Q}$ does not have the least-upper-bound property.

We shall now show that there is a close relation between greatest lower bounds and least upper bounds, and that every ordered set with the least-upper-bound property also has the greatest-lower-bound property.

Theorem 3.1.3. Suppose S is an ordered set with the least-upper-bound property, $B \subset S$, B is not empty, and B is bounded below. Let L be the set of all lower bounds of B. Then

$$\alpha = \sup L$$

exists in S, and $\alpha = \inf B$.

In particular, $\inf B$ exists in S.

[1] very important that you remember this for the definition of Dedekind cuts

Proof. Since B is bounded below, L is not empty. Since L consists of exactly those $y \in S$ which satisfy the inequality $y \leq x$ for every $x \in B$, we see that every $x \in B$ is an upper bound of L. Thus L is bounded above. Our hypothesis about S thus implies that L has a supremum in S; call it α.

If $\gamma < \alpha$ then γ is not an upper bound of L, hence $\gamma \notin B$. It follows that $\alpha \leq x$ for every $x \in B$. Thus $\alpha \in L$.

If $\alpha < \beta$ then $\beta \notin L$, since α is an upper bound of L.

We have shown that $\alpha \in L$ but $\beta \notin L$ if $\beta > \alpha$. In other words, α is a lower bound of B, but β is not if $\beta > \alpha$. This means that $\alpha = \inf B$. □

Problem 35. Consider the set $\{\frac{1}{n} \mid n \in \mathbb{Z}^+\}$.

(a) Show that $\max S = 1$.

(b) Show that if d is a lower bound for S, then $d \leq 0$.

(c) Use (b) to show that $0 = \inf S$.

Proof. □

If we are dealing with rational numbers, the sup/inf of a set may not exist. For example, a set of numbers in $\mathbb{Q}$, defined by $\{[\pi \cdot 10^n]/10^n\}$. 3,3.1,3.14,3.141,3.1415,3.14159,... But this set does not have an infimum in $\mathbb{Q}$.

By ZFC, we have the Completeness Axiom, which states that any non-empty set $A \subset \mathbb{R}$ that is bounded above has a supremum; in other words, if A is a non-empty set of real numbers that is bounded above, there exists a $M \in \mathbb{R}$ such that $M = \sup A$.

Problem 36. Find, with proof, the supremum and/or infimum of $\{\frac{1}{n}\}$.

Proof. For the suprenum,

$$\sup\left\{\frac{1}{n}\right\} = \max\left\{\frac{1}{n}\right\} = 1.$$

For the infinum, for all positive a we can pick $n = [\frac{1}{a}] + 1$, then $a > \frac{1}{n}$. Hence

$$\inf\left\{\frac{1}{n}\right\} = 0.$$

□

Problem 37. Find, with proof, the supremum and/or infimum of $\{\sin n\}$.

Proof. The answer is easy to guess: ±1

For the supremum, we need to show that 1 is the smallest we can pick, so for any $a = 1 - \varepsilon < 1$ we want to find an integer n close enough to $2k\pi + \dfrac{\pi}{2}$ so that $\sin n > a$.

Whenever we want to show the approximations between rational and irrational numbers we should think of the **pigeonhole principle**.

$$2k\pi + \frac{\pi}{2} = 6k + (2\pi - 6)k + \frac{\pi}{2}$$

Consider the set of fractional parts $\{(2\pi - 6)k\}$. Since this an infinite set, for any small number δ there is always two elements $\{(2\pi - 6)a\} < \{(2\pi - 6)b\}$ such that

$$|\{(2\pi - 6)b\} - \{(2\pi - 6)a\}| < \varepsilon$$

Then $\{(2\pi - 6)(b - a)\} < \delta$

We then multiply by some number m (basically adding one by one) so that

$$0 \leq \{(2\pi - 6) \cdot m(b - a)\} - \left(2 - \frac{\pi}{2}\right) < \delta$$

Picking $k = m(b - a)$ thus gives

$$\begin{aligned} 2k\pi + \frac{\pi}{2} &= 6k + (2\pi - 6)k + \frac{\pi}{2} \\ &= 6k + [(2\pi - 6)k] + 2 + (2\pi - 6)k - \left(2 - \frac{\pi}{2}\right) \end{aligned}$$

Thus $n = 6k + [(2\pi - 6)k] + 2$ satisfies $\left|2k\pi + \dfrac{\pi}{2} - n\right| < \delta$

Now we're not exactly done here because we still need to talk about how well $\sin n$ approximates to 1.

We need one trigonometric fact: $\sin x < x$ for $x > 0$. (This simply states that the area of a sector in the unit circle is larger than the triangle determined by its endpoints.)

$$\begin{aligned}\sin n &= sin\left(n - \left(2k\pi + \frac{\pi}{2}\right) + \left(2k\pi + \frac{\pi}{2}\right)\right)\\ &= \cos\left(n - \left(2k\pi + \frac{\pi}{2}\right)\right)\\ &= \cos\theta\end{aligned}$$

$$1 - \sin n = 2\sin^2\frac{\theta}{2} = 2\sin^2\left|\frac{\theta}{2}\right| \le \frac{\theta^2}{2} < \delta$$

Hence we simply pick $\delta = \varepsilon$ to ensure that $1 - \sin n < \varepsilon$, and we're done. □

Theorem 3.1.4. Archimedean Principle If $a, b \in \mathbb{R}$ with $a > 0$, then there exists $n \in \mathbb{N}$ such that $na > b$.

Proof. We prove by contradiction: suppose otherwise, that the Archimedean Property is false. Then there exists $a, b \in \mathbb{R}, a > 0$ such that $na \le b$ for all $n \in \mathbb{N}$.

For these particular a and b, we can say that b is an upper bound of $S \coloneqq \{na \mid n \in \mathbb{N}\}$. From the completeness axiom, $s_0 \coloneqq \sup S$ exists. Let $n \in \mathbb{N}$, we have $n + 1 \in \mathbb{N}$. So $s_0 \ge (n+1)a = na + a$.

Then we have $s_0 - a \ge na$. This is true for all $n \in \mathbb{N}$. So $s_0 - a$ is an upper bound of S. However, $s_0 - a < s_0$, which contradicts that s_0 is the least upper bound of S. This contradiction shows that the Archimedean Property is true. □

§3.2 Real Numbers $\mathbb{R}$

§3.2.1 Dedekind cuts

This book assumes familiarity with the rational numbers $\mathbb{Q}$, i.e. numbers of the form $\frac{m}{n}$, where m, n are integers and $n \neq 0$). $\mathbb{Q}$ contains gaps at irrational numbers such as $\sqrt{2}$ and π. In this section, we aim to construct $\mathbb{R}$ from $\mathbb{Q}$.

The rational number system is inadequate for many purposes, both as a field and as an ordered set (which we will discuss later). For instance, there is no rational p such that $p^2 = 2$. (We shall prove this presently.) This leads to the introduction of so-called "irrational numbers" which are often written as infinite decimal expansions and are considered to be "approximated" by the corresponding finite decimals. Thus the sequence

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

"tends to $\sqrt{2}$". But unless the irrational number $\sqrt{2}$ has been clearly defined, the question must arise: Just what is it that this sequence "tends to"?

This sort of question can be answered as soon as the so-called "real number system" is constructed.

Example 3.2.1. We now show that the equation

$$p^2 = 2$$

is not satisfied by any rational p. If there were such a p, we could write $p = \frac{m}{n}$ where m and n are integers that are not both even. Let us assume this is done. Then we have

$$m^2 = 2n^2.$$

This shows that m^2 is even. Hence m is even, and so m^2 is divisible by 4. It follows that the RHS is divisible by 4, so n^2 is even, which implies that n is even. Thus both m and n are even, a contradiction. Hence $p^2 = 2$ is impossible for rational p.

We now examine this situation a little more closely. Let A be the set of all positive rationals p such that $p^2 < 2$ and let B consist of all positive rationals p such that $p^2 > 2$. We shall show that A *contains no largest number and* B *contains no smallest.*

More explicitly, for every $p \in A$ we can find a rational $q \in A$ such that $p < q$, and for every $p \in B$ we can find a rational $q \in B$ such that $q < p$.

To do this, we associate with each rational $p > 0$ the number

$$q = p - \frac{p^2 - 2}{p + 2} = \frac{2p + 2}{p + 2}. \tag{1}$$

Then

$$q^2 - 2 = \frac{2(p^2 - 2)}{(p + 2)^2}. \tag{2}$$

If $p \in A$ then $p^2 - 2 < 0$, (1) shows that $q > p$, and (2) shows that $q^2 < 2$, thus $q \in A$; if $p \in B$ then $p^2 - 2 > 0$, (1) shows that $0 < q < p$, and (2) shows that $q^2 > 2$, thus $q \in B$.

In fact in 1872, German mathematician Richard Dedekind pointed out that a real number x can be determined by its lower set A and upper set B:

$$A \coloneqq \{a : \mathbb{Q} \mid a < x\}$$

$$B \coloneqq \{b : \mathbb{Q} \mid x < b\}$$

He defined a "real number" as a pair of sets of rational numbers, the lower and upper sets shown above. Such a pair of sets of rational numbers are known as a Dedekind cut.

- A is a **lower set**: $\forall a, b \in \mathbb{R}$, if $a < b$ where $b \in A$, then $a \in A$.

- B is an **upper set**: $\forall a, b \in \mathbb{R}$, if $a < b$ where $a \in B$, then $b \in B$.

Definition 3.2.1. Given that B is the complement of A in the reals, a non-empty subset $(A, B) \subset \mathbb{Q}$ is a **Dedekind cut** if:

(i) A is non-empty: $A \neq \varnothing$

(ii) A and B are disjoint: $A \cap B = \varnothing, A \cup B = \mathbb{Q}$

(iii) A is closed downwards: $\forall x, y \in \mathbb{Q}$ with $x < y, y \in A$, then $x \in A$

(iv) A does not contain a greatest element: $\forall x \in A, \exists y \in A$ such that $x < y$

Perhaps a not-so-intuitive fact here is that there are two possible things happening to B:

1. B contains a least element
2. B does not contain a least element

Case 1 and 2 are known as rational and irrational Dedekind cuts respectively.

Definition 3.2.2 (Real numbers)**.** The set of real numbers $\mathbb{R}$ is defined to be the set of all Dedekind cuts.

Remark. The way we think about this is that Dedekind cuts are real numbers, and real numbers are Dedekind cuts.

Order relations

Given real numbers α and β, let $\alpha = (A, B)$ and $\beta = (C, D)$. Then

$$\alpha < \beta \iff A \subset C$$

Remark. Since B is the complement of A, α is completely determined by A itself.

This ordering on the real numbers satisfies the following properties:

- $x < y$ and $y < z \implies x < z$
- Exactly one of $x < y$, $x = y$ or $x > y$ holds
- $x < y \implies x + z < y + z$

Proposition 3.2.1 (Ordering property)**.** For any two real numbers α and β, one of the following must hold:

$$\alpha < \beta \quad \alpha = \beta \quad \alpha > \beta$$

Proof. We prove by contradiction.

Note that $\alpha \le \beta \iff A \subseteq C$ ($A = C$ is possible).

Suppose otherwise, that all three of the above are false, then neither of the sets A and C can be a subset of the other.

We pick two rational numbers from each set: Pick p where $p \in A$, $p \notin C$, pick q where $q \in C$, $q \notin A$

- Obviously we cannot have $p = q$.
- If $p < q$, then since $q \in C$, according to property 3, we have $p \in C$, a contradiction.
- Similarly for $p > q$, we would find that $q \in A$, a contradiction.

Hence our assumption is false.

$\therefore$ One of the three cases $\alpha < \beta$, $\alpha = \beta$, $\alpha > \beta$ must hold. □

Addition

Property 3.2.1 (Addition)**.** Let $\alpha = (A, B)$, $\beta = (C, D)$, then $\alpha + \beta = (X, Y)$ where

$$X = \{a + c \mid a \in A, c \in C\}$$

Proof. To show that (X, Y) is a Dedekind cut, we simply need to check the conditions for Dedekind cuts.

- Property 1 is trivial.
- Property 2 is by definition.
- Property 3:
  Let $x, y \in X$ satisfy $x < y$, $y \in X$.
  Let $y = a + c$, $a \in A$, $c \in C$.
  Let $\varepsilon = y - x$.
  Let $a' = a - \frac{\varepsilon}{2}$, $c' = c - \frac{\varepsilon}{2}$.
  Then
  $$a' + c' = a + c - \varepsilon = x$$
  $a' < a, a \in A \implies a' \in A$. Similarly, $c' \in C$.
  $\therefore x = a' + c' \in X$.
- Property 4:
  $\forall a + c \in X, a \in A, c \in C$, $\exists a' \in A, c' \in C$ such that $a < a', c < c'$.
  $\therefore a' + c' \in X$ satisfies $a + c < a' + c'$.

□

Property 3.2.2 (Commutativity)**.** Addition is **commutative**:

$$\alpha + \beta = \beta + \alpha$$

Proof. The proof is trivial. □

Property 3.2.3 (Associativity)**.** Addition is **associative**:

$$\alpha + (\beta + \gamma) = (\alpha + \beta) + \gamma$$

Proof. Let $\alpha = (A, A')$, $\beta = (B, B')$, $\gamma = (C, C')$

$$\beta + \gamma = (B + C, (B + C)')$$

In this notation we only need to show that $A + (B + C) = (A + B) + C$.

$$\begin{aligned} x \in A + (B + C) &\iff \exists a \in A, p \in B + C \text{ s.t. } x = a + p \\ &\iff \exists a \in A, b \in B, c \in C \text{ s.t. } x = a + b + c \\ &\iff x \in (A + B) + C \end{aligned}$$

Hence proven. □

Exercise 3.2.1

Prove that

$$\alpha + 0 = \alpha = 0 + \alpha$$

Proof. Let $0 = (O, O')$ where $O = \{x \mid x < 0\}, O' = \{x \mid x \geq 0\}$.

Let $\alpha = (A, B)$, then $\alpha + 0 = (C, D)$ where

$$\begin{aligned} C &= \{a + \varepsilon \mid a \in A, \varepsilon < 0\} \\ &= \{a - \varepsilon \mid a \in A, \varepsilon > 0\} \end{aligned}$$

$$a - \varepsilon < a, a \in A \implies a - \varepsilon \in A \implies C \subseteq A$$

According to Property 4, $\forall a \in A, \exists a' \in A$ such that $a < a'$.

Let $\varepsilon = a' - a > 0$, then

$$a = a' - \varepsilon, a' \in A, \varepsilon > 0 \implies a \in C$$

So $A = C$.

$\therefore \alpha + 0 = \alpha$ □

Exercise 3.2.2

Express $-\alpha$ in terms of α; show

$$\alpha + (-\alpha) = 0 = (-\alpha) + \alpha$$

Proof. We split this into two cases.

Case 1: α is a rational number, then $\alpha = (A, B)$ where $A = \{x \mid x < \alpha\}$, $B = \{x \mid x \geq \alpha\}$.

Let $-\alpha = (A', B')$, where $A' = \{x \mid x < -\alpha\}$, $B' = \{x \mid x \geq -\alpha\}$. We see that $\alpha + (-\alpha) \leq 0$ is obvious.

On the other hand, since $0 = (O, O')$, for any $\varepsilon < 0$ we have

$$\varepsilon = \left(\alpha + \frac{\varepsilon}{2}\right) + \left(-\alpha + \frac{\varepsilon}{2}\right) \in A + A'$$

Hence $\alpha + (-\alpha) = 0$.

Case 2: α is irrational, let $\alpha = (A, B)$ where B does not have a lowest value. Then $-B = \{-x \mid x \in B\}$ does not have a highest value.

We wish to define $-\alpha = (-B, -A)$, but first we need to show that this is well-defined by checking through all the conditions.

- Property 1: This is trivial.
- Property 2: Prove that $-A$ and B are disjoint.

  Note that $\forall x \in \mathbb{R}$, if $x = -y$, then exactly one out of $y \in A$ and $y \in B$ is true $\implies$ exactly one out of $x \in -B$ and $x \in -A$ is true.
- Property 3: Prove $-B$ is closed downwards.

  Suppose otherwise, that $x < y, y \in -B$ but $x \notin -B$. Then $-y \in B$, $-x \notin B$. Since A is the complement of B, $-y \notin A$, $-x \in A$. But $-y < -x$, which is a contradiction.
- Property 4 is already guaranteed by the irrationality of α.

All of these properties imply that the real numbers form a commutative group by addition. □

Negation

Given any set $X \subset \mathbb{R}$, let $-X$ denote the set of the negatives of those rational numbers. That is $x \in X$ if and only if $-x \in -X$.

If (A, B) is a Dedekind cut, then $-(A, B)$ is defined to be $(-B, -A)$.

This is pretty clearly a Dedekind cut. - proof

Signs

A Dedekind cut (A, B) is **positive** if $0 \in A$ and **negative** if $0 \in B$. If (A, B) is neither positive nor negative, then (A, B) is the cut representing 0.

If (A, B) is positive, then $-(A, B)$ is negative. Likewise, if (A, B) is negative, then $-(A, B)$ is positive. The cut (A, B) is non-negative if it is either positive or 0.

Multiplication

Let $\alpha = (A, B)$ and $\beta = (C, D)$ where α, β are both non-negative.

We define $\alpha \times \beta$ to be the pair (X, Y) where

X is the set of all products ac where $a \in A, c \in C$ and at least one of the two numbers is non-negative. Y is the set of all products bd where $b \in B, d \in D$.

Intermediate Value Theorem

Bolzano-Weiersstrass Theorem

Connectedness of $\mathbb{R}$

§3.2.2 $\mathbb{R}$ is archimedian

Theorem 3.2.1 (Archimedian Principle)**.** For any $x \in \mathbb{R}^+$ and $y \in \mathbb{R}$, there exists some $n \in \mathbb{Z}^+$ such that

$$n \cdot x > y.$$

In particular, if we take $n = 1$ from this theorem, we immediately get the following statement.

Proposition 3.2.2. For any $y \in \mathbb{R}$, there exists some positive integer n such that $n > y$.

We now give a proof of Proposition 3.2.2 directly without using Theorem 3.2.1, and then we prove 3.2.1 from Proposition 3.2.2. This shows that these two statements are in fact equivalent, though Proposition 3.2.2 looks much simpler.

§3.2.3 $\mathbb{Q}$ is dense in $\mathbb{R}$

§3.2.4 $\mathbb{R}$ is closed under taking roots

§3.2.5 The extended real number system

§3.3 The Euclidean Plane $\mathbb{R}^2$

We consider the Cartesian product of $\mathbb{R}$ with $\mathbb{R}$; that is,

$$\mathbb{R}^2 \coloneqq \mathbb{R} \times \mathbb{R} \coloneqq \{(x_1, x_2) \mid x_1, x_2 \in \mathbb{R}\}.$$

Over $\mathbb{R}^2$, we can define operations

- Addition $+$: $(x_1, x_2) + (y_1, y_2) = (x_1 + y_1, x_2 + y_2)$;
- Scalar multiplication $\mathbb{R} \times \mathbb{R}^2 \to \mathbb{R}^2$: $c \cdot (x_1, x_2) = (c \cdot x_1, c \cdot x_2)$.

This two operations make $\mathbb{R}^2$ a 2-dimensional vector space (linear space) over the real field $\mathbb{R}$. We also say $\mathbb{R}^2$ is a $\mathbb{R}$-linear space of real dimension 2. For example, $\{(1, 0), (0, 1)\}$ form a basis of $\mathbb{R}^2$.

Moreover, over the linear space $\mathbb{R}^2$, one can define an inner product as

$$\langle (x_1, x_2), (y_1, y_2) \rangle = x_1 y_1 + x_2 y_2.$$

The inner product induces a norm

$$|(x_1, x_2)| = \sqrt{\langle (x_1, x_2), (x_1, x_2) \rangle} = \sqrt{x_1^2 + x_2^2}.$$

From now on, we use $\vec{x}$ to denote (x_1, x_2).

Proposition 3.3.1.

- $|\vec{x}| \geq 0$, where equality holds if and only if $\vec{x} = \vec{0}$.
- $|c \cdot \vec{x}| = |c||\vec{x}|$
- $|\vec{x} + \vec{y}| \leq |\vec{x}| + |\vec{y}|$
- $|\langle \vec{x}, \vec{y} \rangle| \leq |\vec{x}||\vec{y}|$

All constructions here can be easily generalised to any $\mathbb{R}^n$ with $n \in \mathbb{Z}^+$.

§3.4 The Complex Numbers $\mathbb{C}$

Over $\mathbb{R}^2$, we can define a multiplication $\cdot$ as

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

If we identity $\mathbb{R}^2$ with

$$\mathbb{C} \coloneqq \{x + yi \mid x, y \in \mathbb{R}\}$$

via $(x, y) \mapsto x + yi$, then all structures defined above are induced to $\mathbb{C}$. In particular, the multiplication is induced to $\mathbb{C}$ via requiring $i^2 = -1$. A nontrivial fact is that $(\mathbb{C}, +, \cdot)$ is a field. A element in $\mathbb{C}$ is called a complex number. Usually, people prefer to use $z = x + yi$, $x, y, \in \mathbb{R}$, to denote a complex number. Here x is called the real part of z and y is called the imaginary part of z. We use $|z|$ to denote its norm.

§3.5 Completeness

§3.5.1 Completeness axiom

Theorem 3.5.1 (Completeness axiom for the real numbers)**.** Let A be a non-empty subset of $\mathbb{R}$ that is bounded above. Then A has a supremum.

Any set in the reals bounded from above/below must have a supremum/infimum.

Proof. We prove this using Dedekind cuts.

Let S be a real number set. We consider the rational number set $A = \{x \in \mathbb{Q} \mid \exists y \in S\}$. Set B is defined to be the complement of A in $\mathbb{Q}$.

We go through the definitions to check that $(A|B)$ is a Dedekind cut.

1. Since $S \neq \varnothing$, pick $y \in S$, then $[y] - 1$ is a real number smaller than some element in S, hence $[y] - 1 \in A$ and thus $A \neq \varnothing$.

   Since we're given that S is bounded, $\exists M > 0$ as the upper bound for S, thus $B \neq \varnothing$.

   (Note that an upper bound is simply a number that is bigger than anything from the set, and is not the supremum

2. We defined B to be the complement of A in $\mathbb{Q}$, so this condition is trivial.

3. For any $x, y \in A$, if $x < y$ and $y \in A$, then $\exists z \in S$ such that $y < z \implies x < z \implies x \in A$.

4. Suppose otherwise that $x \in A$ is the largest element in A, then $\exists y \in S$ such that $x < y$ We then pick a rational number z between x and y. Since we still have $z < y$, we have $z \in A$ but $z > x$, contradictory to z being the largest.

   Now there's actually an issue with the proof for property 4 here How exactly are we finding z?

   First $x \in \mathbb{Q}$. Then $y \in \mathbb{R}$ so we rewrite it as $y = (C|D)$ via definition.

   $x < y$ translates to the fact that $x \in C$.

   Since y is real, by definition we know that C must not have a largest element.

   In particular, x is not largest and we can pick $z \in C$ such that $z > x$. This is in fact the z that we need

Now that all the properties of a real number are validated, we may finally conclude that $\alpha = (A|B)$ is indeed a real number.

Now we need to show that $\alpha = \sup S$.

Let $x \in S$. If x is not the maximum value of S, i.e. $\exists y \in S, x < y$, then $x \in A$ and thus $x < \alpha$.

If x is the maximum value of S, then for any rational number $y < x$ we have $y \in A$, and for any rational number $y \geq x$ we have $y \in B$. Thus $x = (A|B) = \alpha$.

In conclusion, $x \leq \alpha$ for all $x \in S$.

For any upper bound x of S, since $\forall y \in S, x \geq y$ we have $x \in B$ and thus $x \geq \alpha$.

$\therefore$ α is the smallest upper bound of S and thus $\sup S = \alpha$ exists. □

Theorem 3.5.2 (Archimedean property of $\mathbb{N}$)**.** $\mathbb{N}$ is not bounded above.

Proof. Suppose, for a contradiction, that $\mathbb{N}$ is bounded above. Then $\mathbb{N}$ is non-empty and bounded above, so by completeness (of $\mathbb{R}$) $\mathbb{N}$ has a supremum.

By the Approximation property with $\varepsilon = \frac{1}{2}$, there is a natural number $n \in \mathbb{N}$ such that $\sup \mathbb{N} - \frac{1}{2} < n \leq \sup \mathbb{N}$.

Now $n + 1 \in \mathbb{N}$ and $n + 1 > \sup \mathbb{N}$. This is a contradiction. □

4 Numerical Sequences and Series

§4.1 Sequences in $\mathbb{R}$

§4.1.1 Convergent Sequences

We first review some definitions and basic properties of sequences. For our current purpose, we state for $\mathbb{R}$ only, but they work for any metric space with corresponding modifications.

Definition 4.1.1. A sequence, which we denote by $\{x_n\}$, in $\mathbb{R}$ is a map from $\mathbb{Z}^+ \to \mathbb{R}$, which maps $n \in \mathbb{Z}^+$ to $x_n \in \mathbb{R}$. The range of the map is called the range of the sequence.

A **subsequence** of $\{x_n\}$ is defined via an injective map s from $\mathbb{Z}^+$ to a subset of $\mathbb{Z}^+$ satisfying

$$s(k_1) < s(k_2) \quad \forall k_1, k_2 \in \mathbb{Z}^+, k_1 < k_2,$$

and denoted as $\{x_{n_k}\}$ with $x_{n_k} = x_{s(k)}$.

A sequence $\{x_n\}$ in $\mathbb{R}$ is called **convergent**, if there exists some $L \in \mathbb{R}$ such that for any $\varepsilon > 0$, there exists some $N \in \mathbb{Z}^+$ so that for all $n > N$, $|x_n - L| < \varepsilon$. We denote it as $\lim_{n\to\infty} x_n = L$, and call L the **limit** of the sequence $\{x_n\}$.

A sequence $\{x_n\}$ in $\mathbb{R}$ is called **divergent**, if it has no limit in $\mathbb{R}$.

Remark. Take note of the use of logical statements:

- ε is independent, so it is literally for all $\varepsilon > 0$.
- N is dependent on ε; if ε is very small we would expect the sequence $\{a_n\}$ to get close enough to L further down the line.
- The order of the quantifiers matters.

Proposition 4.1.1.

- The limit of a convergent sequence in $\mathbb{R}$ is unique.
- The sequence $\{x_n\}$ converges to $L \in \mathbb{R}$ if and only if every open disk centred at L contains all but finitely many of terms in the sequence.
- The sequence $\{x_n\}$ converges to $L \in \mathbb{R}$ if and only if every subsequence of it converges to $L \in \mathbb{R}$.
- If a sequence $\{x_n\}$ in $\mathbb{R}$ is convergent, then it must be bounded.
- The set of all subsequential limits of a sequence $\{x_n\}$ in $\mathbb{R}$ is closed.

Exercise 4.1.1

What do we really mean by saying that $\frac{1}{n} \to 0$ as $n \to \infty$?
We mean that the sequence of numbers $\frac{1}{n}$ converges to 0, proven as follows:

Proof. $\forall \varepsilon > 0$, pick $N = \frac{1}{\varepsilon} + 1$. Then $\forall n > N$,

$$\frac{1}{n} < \frac{1}{N} < \frac{1}{\frac{1}{\varepsilon}} = \varepsilon.$$

□

We shall cover some characteristics of limits.

Proposition 4.1.2. Given a sequence of points $\{x_k\}$ and a point $x \in \mathbb{R}^n$, x_k converges to x if and only if all neighbourhoods of x "eventually" contain all x_k.

By "eventually" we mean something similar to the definition above: there exists some large N such that the property is satisfied for all $n > N$.

Proof. **Forward direction:**

If $\{x_k\}$ converges to x, we wish to prove: given any neighbourhood U of x, U eventually contains all x_k.

Since U is a neighbourhood of x, we pick a ball of radius ε centered at x, $B(x,\varepsilon)$, so that $B(x,\varepsilon)$ is contained in U.

Then since $B(x,\varepsilon)$ is precisely the set of points whose distance to x is no larger than ε, we then apply the fact that $\{x_k\}$ converges to x.

So for this particular ε, we take a natural number N so that $|x_k - x| < \varepsilon$, or $x_k \in B(x,\varepsilon)$, for all $k > N$.

Then simultaneously x_k are in U since $B(x,\varepsilon)$ is a subset of U, thus we've shown that U will contain all x_k after a certain point N.

Backward direction:

Suppose that all neighbourhoods of x will eventually contain all x_k, then in particular for any $\varepsilon > 0$, since $B(x,\varepsilon)$ is a neighbourhood of x, it will also eventually contain all x_k.

This then easily translates to the fact that $\{x_k\}$ converges to x. □

Proposition 4.1.3 (Uniqueness of the limit)**.** Suppose that $\{x_k\}$ converges to both x and x', then $x = x'$.

Proof. $\forall \varepsilon > 0$, we know that the terms in $\{x_k\}$ must be less than ε away from its limit after a certain point.

However, this certain point may not be the same for both limits; for the two limits x and x', we must first assume two separate numbers N and N' so that $|x_k - x| < \varepsilon$ when $k > N$, and $|x_k - x'| < \varepsilon$ when $k > N'$.

Now if you look at the book here, it says that we have a stronger requirement: $|x_k - x| < \frac{\varepsilon}{2}$ when $k > N$, $|x_k - x'| < \frac{\varepsilon}{2}$ when $k > N'$. This is simply because we want to prove certain statements strictly by definition

There is an important detail to take note, regarding $\max\{N, N'\}$.

We're taking the larger one of these, so it means that, after this certain point, we in fact have $|x_k - x| < \frac{\varepsilon}{2}$ and $|x_k - x'| < \frac{\varepsilon}{2}$ at the same time.

Therefore by triangle inequality,

$$|x - x'| \le |x_k - x| + |x_k - x| < \varepsilon$$

The choice of k actually vanished in the final statement; you can think of this as if picking this particular choice of k helps us to establish some kind of property for the original objects

Finally, since we've in fact proven that $|x - x'| < \varepsilon$ holds for any given positive $\varepsilon > 0$, we must have $|x - x'| = 0$ and therefore $x = x'$.

Strictly speaking, for the first part we need to explain why $a < \varepsilon$ for any positive ε implies that $a \leq 0$. This is very easy to prove (by contradiction) so let's not be too redundant The second part simply relies on the fact that |x-y| is the Euclidean metric and so by positive definiteness |x-y|=0 if and only if x=y. □

Proposition 4.1.4 (Boundedness of converging sequences)**.** If $\{x_k\}$ converges, then $\{x_k\}$ is bounded.

We simply take the limit x and note that the sequence is eventually contained in some ball centered at x, say $B(x, 1)$.

There are several outlying points prior to this, but since there are only a finite number of these, it doesn't change the fact that the sequence (viewed as a set) is bounded nevertheless.

This argument is precisely expressed by the construction of r given in the book: let $|x_k - x| < 1$ whenever $k > N$, then $\{x_k\}$ is in $B(x, r)$ where $r = \max\{1, |x_1 - x|, \dots, |x_N - x|\}$

1. We talk about the relationship between the limit of a sequence and the limit points of a set.

   Generally, limit points are a weaker construction.

   Suppose that $\{x_k\}$ converges to x If we view $\{x_k\}$ as a set, then x will be a limit point of this set

   The converse, however, is not true

   Exercise 1: Construct a sequence in R that is bounded and contains a single limit point but is divergent (not convergent)

   The thing about convergence of a series is that, unlike for limit points where we only require that there are other points that get arbitrarily close, but moreover we have to ensure that this pattern ensues for each and every term in the sequence

   Me:Suppose that $\{x_k\}$ converges to x If we view $\{x_k\}$ as a set, then x will be a limit point of this set" - - - - - - - - - - - - - - - Sorry I forgot something crucial about this : There is the strange possibility that the sequence $\{x_k\}$ is constant : (or at least eventually constant) : Then in fact x by definition is not a limit point of x_k because you can find a ball around x that only contains the element x itself, since that point is merely what the entire sequence $\{x_k\}$ amounts to : Anyways, we simply can't say that a sequence $\{x_k\}$ converges to x if we're only provided with the fact that x is a limit point of $\{x_k\}$

   However, we can say the following: (d) If x is a limit point of E, then there exists a sequence $\{x_n\}$ in $E \setminus x$ such that $\{x_n\}$ converges to x

   In fact this is correct in both ways so let's rewrite this as follows: (d) x is a limit point of E, if and only if there exists a sequence $\{x_n\}$ in $E \setminus x$ such that $\{x_n\}$ converges to x

   ($E \setminus x$ is important here, otherwise we simply pick the constant sequence $x_k = x$)

   $\rightarrow$: If x is a limit point, then for all $\varepsilon > 0$, $B_0(x, \varepsilon)$ contains points in E We then construct such a sequence $\{x_k\}$ in $E \setminus x$: pick any $x_k \in E$ so that x_k is contained in $B_0(x, 1/k)$

   Then it is easy to show that $\{x_k\}$ is a sequence in $E \setminus x$ which converges to x.

   $\leftarrow$: Suppose that there exists a sequence $\{x_n\}$ in $E \setminus x$ such that $\{x_n\}$ converges to x We wish to show that $B_0(x, \varepsilon)$ contains points in E for all $\varepsilon > 0$

   Since $\{x_n\}$ converges to x, for all $\varepsilon > 0$ the sequence is eventually contained in $B(x, \varepsilon)$ However because we have the precondition that $\{x_n\}$ has to be in $E \setminus x$, the sequence is in fact eventually contained in $B_0(x, \varepsilon)$.

Proposition 4.1.5. $\{x_k\}$ converges to x if and only if every subsequence of $\{x_k\}$ converges to x.

Proof. We only need to prove this in the forwards direction Every subsequence of $\{x_k\}$ can be written in the form $\{x_{k_i}\}$ where $k_1 < k_2 < \dots$ is a strictly increasing sequence of natural numbers

Intuitively, if every neighbourhood of x eventually contains all x_k, then since $\{x_{k_i}\}$ is just a subset of $\{x_k\}$ they should all be contained in the neighbourhood eventually as well. For every $\varepsilon > 0$, pick N such that for $k > N$, $|x_k - x| < \varepsilon$. Pick M such that $k_M > N$, then for all $i > M$ we have $|x_(k_i) - x| < \varepsilon$. □

Proposition 4.1.6. Subsequential limits of a sequence are precisely the limit points of the sequence (viewed as a set)

Proof. This is just part (d) of the previous section.

Again, to make this work, we need to assume that nothing funny is going on at subsequential limits If the limits appear due to eventually constant subsequences, then they need not be limit points of the original sequence when viewed as a set

3.6, 3.7 are precisely the statements we've prepared for last week □

Proposition 4.1.7. If $\{x_n\}$ is a sequence in a compact set (bounded closed set), then there exists a convergent subsequence of $\{x_n\}$

Proof. This is Weierstrass-Bolzano together with part (b)

Ah yes, regarding compact sets I need to emphasize this again, but the definition that we are currently using for compact sets is not the actual definition

I've sent a video before the lesson which talks about the real definition for compact sets Essentially, compact sets satisfies the property akin to the statement in Heine-Borel: Given a topological space (X, τ), a compact set K in X is a set satisfying that, given any open covering $\{U_i\}$ of X, there exists a finite open cover $\{U_1, \dots, U_n\}$ of X

This is difficult to process at this stage Since we're currently only working with Euclidean spaces it would be more beneficial if you consider the Heine-Borel Theorem as a property first It would be a lot easier to accept the definition after you're more accustomed to applying the theorem □

Proposition 4.1.8. (Rudin 3.7) Subsequential limits form a closed subset

Proof. Actually we've done this two weeks before, it is simply saying that A" is a subset of A'.

(A" is not always A'; consider the set in R^2 given by (1/n,1/m)|n,m in N Then (1,0),(0,1) are in A' but not in A" □

§4.1.2 Cauchy Sequences

This is a very very helpful way to determine whether a sequence is convergent or divergent, as it does not require the limit to be known. In the future you will see many instances where the convergence of all sorts of limits are compared with similar counterparts; generally we describe such properties as **Cauchy criteria**.

Cauchy sequences have to deal with the differences between terms within the sequence itself.

Definition 4.1.2. A sequence $\{x_k\}$ in $\mathbb{R}^n$ is **Cauchy**, if the distances between any two terms is sufficiently small after a certain point.

Strictly speaking, $\forall \varepsilon > 0$, there exists integer N such that

$$\forall k, l > N, |x_k - x_l| < \varepsilon.$$

It is easy to prove that a converging sequence is Cauchy using the triangle inequality. The idea is that, if all the points are becoming arbitrarily close to a given point p, then they are also becoming close to each other. The converse is not always true, however.

Proposition 4.1.9. A sequence $\{x_k\}$ in $\mathbb{R}^n$ is convergent if and only if it is Cauchy.

Proof. **Forward direction:**

Suppose that $\{x_k\}$ converges to x, then there exists N such that for $k > N$, $|x_k - x| < \frac{\varepsilon}{2}$ Then for $k, l > N$,

$$|x - k - x_l| \leq |x_k - x| + |x_l - x| < \varepsilon$$

Backward direction:

First, we show that $\{x_k\}$ must be bounded. Pick N such that for all $k, l > N$ we have $|x_k - x_l| < 1$. Centered at x_k, we show that $\{x_k\}$ is bounded; to do this we pick

$$r = \max\{1, |x_k - x_1|, \ldots, |x_k - x_N|\}$$

Then the sequence x_k is in $B(x_k, r)$ and thus is bounded.

Since $\{x_k\}$ is bounded, by the collolary of Bolzano-Weierstrass we know that $\{x_k\}$ contains a subsequence $\{x_{k_i}\}$ that converges to a limit x.

Then for all $\varepsilon > 0$, pick N_1 such that for all $k, l > N$, $|x_k - x_l| < \dfrac{\varepsilon}{2}$. Simultaneously, since $\{x_{k_i}\}$ converges to x, pick M such that for $i > M$, $|x_{k_i} - x| < \dfrac{\varepsilon}{2}$.

Now, since $k_1 < k_2 < \ldots$ is a sequence of strictly increasing natural numbers, we can pick $i > M$ such that $k_i > N$. Then for all $k > N$, by setting $l = k_i$ we obtain

$$|x_k - x_{k_i}| < \frac{\varepsilon}{2}, \quad |x_{k_i} - x| < \frac{\varepsilon}{2}$$

and hence

$$|x_k - x| \le |x_k - x_{k_i}| + |x_{k_i} - x| < \varepsilon$$

□

§4.1.3 Upper and Lower Limits

§4.1.4 Limits of Multiple Sequences

We shall cover some of the more basic aspects of limits in this section.

Inequalities

First let's consider two converging sequences $\{a_n\}$ and $\{b_n\}$

If $a_n \le b_n$, then $\lim a_n \le \lim b_n$.

Remark. One important thing to take note for limits is that, even if you have $a_n < b_n$, you cannot say that $\lim a_n < \lim b_n$; for example, $\frac{1}{n} > -\frac{1}{n}$ but their limits are both 0.

Proof. Let's say that $A = \lim a_n$ and $B = \lim b_n$. Suppose otherwise that $A > B$, then we try to cause some chaos with $\varepsilon = A - B > 0$.

Since $\frac{\varepsilon}{2} > 0$, then there exists N_1 such that for $n > N_1$ we have $|a_n - A| < \frac{\varepsilon}{2}$; and there exists N_2 such that for $n > N_2$ we have $|b_n - B| < \frac{\varepsilon}{2}$.

Let $N = \max\{N_1, N_2\}$, then for any $n > N$, the two inequalities above will hold simultaneously But then we would have

$$a_n > A - \frac{\varepsilon}{2}, b_n < B + \frac{\varepsilon}{2}$$

and thus

$$a_n - b_n > A - B - \varepsilon = 0,$$

so $a_n > b_n$, a contradiction □

A corollary is that limits essentially preserve signs, if you include 0 in your consideration

A converging sequence of nonnegative numbers will always be nonnegative, and same goes to nonpositive numbers : Now as we can see in the proof above, there is actually a place where the restrictions of limits overpower the statement itself : What I mean by that is, suppose that you want to form a proof by contradiction : What you need here is just one term $a_n > b_n$ But you actually have $a_n > b_n$ eventually for all terms in the sequence : In fact, a better exercise would have been to show that limsups and liminfs also preserves inequalities

I'll just use limsups for example If $a_n \le b_n$, let $A = \limsup a_n$, $B = \limsup b_n$. Suppose otherwise that $A > B$. Let $\varepsilon = A - B > 0$; since $\frac{\varepsilon}{2} > 0$, then for all N_1, there exists $n > N_1$ such that $a_n > A - \frac{\varepsilon}{2}$; and there exists N_2 such that for all $n > N_2$, $b_n < B + \frac{\varepsilon}{2}$.

Now we arrange our thoughts logically First, we pick $N_2 = N$ such that for all $n > N$, $b_n < B + \frac{\varepsilon}{2}$. Then we may fix $N_1 = N$.

Due to the first condition, we see that it is possible to pick $n_0 > N$ such that $a_{n_0} > A - \frac{\varepsilon}{2}$. Now due to the second condition, since $n_0 > N$, this exact same n_0 would satisfy $b_{n_0} < B + \frac{\varepsilon}{2}$.

Therefore, n_0 satisfies $a_{n_0} - b_{n_0} > A - B - \varepsilon = 0$ and we are done.

Sandwich Theorem

Theorem 4.1.1 (Sandwich Theorem)**.** Let $a_n \le c_n \le b_n$ where $\{a_n\}, \{b_n\}$ are converging sequences such that $\lim a_n = \lim b_n = L$, then $\{c_n\}$ is also a converging sequence and $\lim c_n = L$.

Now, one very very very important thing about this theorem

The purpose of this theorem is to investigate some difficult sequence $\{c_n\}$ with two simpler sequences $\{a_n\}$ and $\{b_n\}$ which bounds it from below and from above respectively If you look closely at the statement, you may realize that we're only working under the condition that $\{a_n\}$ and $\{b_n\}$ are converging sequences

In other words, at this point we don't know whether $\{c_n\}$ is convergent.

In fact, this is supposed to be the main implication

Of course, $\lim c_n = L$ is proven at the exact same time, so both implications constitute the two parts of the conclusion

What I want to say is that you cannot simply take lim over $a_n \le c_n \le b_n$ and say that lim preserves inequalities, because in order to apply this inequality-preserving property, you need to ensure that all sequences are converging before you can apply it; clearly, this does not work here since we have not shown that c_n is convergent, therefore this idea does not work.

There are two ways to circumvent this One is to use $\varepsilon - N$; basically, just do it

But if you're really lazy, then the second method is to use the idea above except you first take limsup and liminf

The advantage of these two is that you don't need the original sequences to be convergent in order to apply them, and that they preserve inequalities even if the original sequences show no signs of convergence

So basically,

$$\limsup a_n \le \limsup c_n \le \limsup b_n,$$

and

$$\liminf a_n \le \liminf c_n \le \liminf b_n.$$

Then since $\{a_n\}$ and $\{b_n\}$ actually converge to L, all the liminfs and limsups of a_n and b_n are L, so we obtain $\limsup c_n = L$ and $\liminf c_n = L$.

In particular, $\limsup c_n = \liminf c_n$, thus c_n is convergent and it follows that $\lim c_n = L$.

Arithmetic properties

Proposition 4.1.10. For converging $\{a_n\}$ and real constant k,

$$\lim k a_n = k \lim a_n.$$

Proof. The proof is left as an exercise. You will need to k into cases where it is positive, negative or 0. □

Remark. In multivariable calculus there's a similar property that is more interesting:

If T is a linear map on $\mathbb{R}^n$, and $\{x_n\}$ is a converging sequence of points, then $\{Tx_n\}$ is also converging; moreover if $x_n \to x_0$ then $Tx_n \to Tx_0$.

Proposition 4.1.11. If $\{a_n\}$ and $\{b_n\}$ are converging sequences of real numbers, then

$$\lim(a_n + b_n) = \lim a_n + \lim b_n.$$

Proof. Let $A = \lim a_n$ and $B = \lim b_n$, then for all $\varepsilon > 0$, there exists N_1 such that for all $n > N_1$, $|a_n - A| < \frac{\varepsilon}{2}$; there exists N_2 such that for all $n > N_2$, $|b_n - B| < \frac{\varepsilon}{2}$.

Let $N = \max\{N_1, N_2\}$, then for all $n > N$, by the triangle inequality we have

$$|(a_n + b_n) - (A + B)| \leq |a_n - A| + |b_n - B| < \varepsilon.$$

□

Remark. This proof is simple enough to generalise to any normed vector spaces.

The following corollary can be easily derived from the above.

Corollary 4.1.1. If $\{a_n\}$ and $\{b_n\}$ are converging sequences of real numbers, then

$$\lim(a_n - b_n) = \lim a_n - \lim b_n.$$

Proposition 4.1.12. If $\{a_n\}$ and $\{b_n\}$ are converging, then

$$\lim(a_n b_n) = \lim a_n \cdot \lim b_n.$$

Proof. Let $A = \lim a_n$ and $B = \lim b_n$.

Consider the limit $\lim(a_n b_n - AB)$, as it would be sufficient to prove that this is equal to 0.

Now we will use a common technique to deal with such products:

$$\lim(a_n b_n - AB) = \lim(a_n b_n - Ab_n + Ab_n - AB)$$

The idea is to show that this is equal to

$$\lim(a_n b_n - Ab_n) + \lim(Ab_n - AB)$$

(Note that we cannot write this yet because we have not shown that these two sequences are convergent)

So let's examine these two sequences. The second one is easier since we have proved proposition 4.1.11:

$$\lim b_n = B \implies \lim(b_n - B) = 0$$

Thus $\lim(Ab_n - AB) = A \lim(b_n - B) = 0$.

As for the first one, we want to show that $\lim(a_n - A)b_n = 0$. Since we know that b_n is itself a converging sequence, thus in particular b_n is bounded, so suppose that $M > 0$ is a bound of b_n, i.e. for all natural number n, $|b_n| \leq M$.

Since $\lim a_n = a$, for all $\varepsilon > 0$, there exists N such that for all $n > N$, $|a_n - a| < \frac{\varepsilon}{M}$.

Combining the two above, we then conclude that for all $\varepsilon > 0$, there exists N such that for all $n > N$,

$$|a_n b_n - Ab_n| = |(a_n - A)b_n| < \frac{\varepsilon}{M} \cdot M = \varepsilon.$$

Therefore, this implies that $\lim(a_n b_n - Ab_n) = 0$.

Since we have shown that the two parts are equal to 0, we can conclude that $\lim(a_n b_n - AB) = 0$. □

Proposition 4.1.13. If $\{a_n\}$ and $\{b_n\}$ are converging, b_n is never 0 and $\lim b_n \neq 0$, then

$$\lim \frac{a_n}{b_n} = \frac{\lim a_n}{\lim b_n}.$$

Proof. Since we already have third proposition, it is sufficient for us to show that $\lim \frac{1}{b_n} = \frac{1}{\lim b_n}$.

Let $b = \lim b_n$, then we consider the limit

$$\lim \left(\frac{1}{b_n} - \frac{1}{b} \right) = \lim \left(\frac{b - b_n}{b_n b} \right).$$

Again, the important term here is $b - b_n$, but there is an extra term of $\frac{1}{b_n b}$, so we'll need to control this.

Since we need this to be bounded, we actually cannot have b_n to be close to 0. The good thing here is that $b \neq 0$, so we can restrict b_n to be close enough to b so that it stays away from 0.

So we can first pick N_1 such that for all $n > N_1$,

$$|b_n - b| < \frac{|b|}{2}.$$

Then

$$|b_n b - b^2| < \frac{b^2}{2}$$
$$\frac{b^2}{2} < b_n b < \frac{3b^2}{2}$$

This show that if $n > N_1$, $b_n b$ would always be positive, and $\frac{1}{b_n b} < \frac{2}{b^2}$.

Let $M = \frac{2}{b^2}$, then we may refer back to the original statement

$$\left| \frac{b - b_n}{b_n b} \right| < M|b - b_n|$$

We pick N_2 such that for all $n > N_2$, $|b_n - b| < \frac{\varepsilon}{M}$.

Let $N = \max\{N_1, N_2\}$, then for all $n > N$,

$$\left| \frac{b - b_n}{b_n b} \right| < M \cdot \frac{\varepsilon}{M} = \varepsilon.$$

□

Now let's talk a little bit about the arithmetic properties of limsups and liminfs : There are quite a number of differences for this; essentially the arithmetical properties aren't as well-behaved as the more specific case of limits : (i) $\limsup ka_n = k \limsup a_n$ holds if $k > 0$ However, if $k < 0$, then $\limsup ka_n = k \liminf a_n$.

(ii) $\limsup(a_n + b_n)$ is in general not equal to $\limsup a_n + \limsup b_n$ However, we do have the following:

$$\limsup(a_n + b_n) \leq \limsup a_n + \limsup b_n$$

Moreover, $\limsup(a_n + b_n)$ may be bounded from below as follows:

$$\limsup(a_n + b_n) \geq \limsup a_n + \liminf b_n$$

Your homework for today is to write down the analogous properties for liminf, and to prove (i) and (ii)

Now you should try to prove (i) for liminf as well; as for (ii), try to explain why properties (i),(ii) for limsup and property (i) for liminf would imply property (ii) for $\liminf$

Problem 38. Let $\{x_n\}$ be a sequence of real numbers and let $\alpha \geq 2$ be a constant. Define the sequence $\{y_n\}$ as follows:

$$y_n = x_n + \alpha x_{n+1}, n = 1, 2, \ldots$$

Show that if $\{y_n\}$ is convergent, then $\{x_n\}$ is also convergent.

§4.2 Series in $\mathbb{R}$ ($\mathbb{C}$)

§4.2.1 Definition and basic properties

§4.2.2 Comparison test

§4.2.3 Root and ratio tests

§4.2.4 Addition and multiplication of series

§4.2.5 Rearrangement

5 Continuity

§5.1 Limit of Functions

Assume (X, d_x) is metric space and $E \subset X$ is a subset of X. Then the metric d_X induces a metric on E. We now consider another metric space (Y, d_Y). A map $f : E \to Y$ is also called a function over E with values in Y. In particular, if $Y = \mathbb{R}$, then f is called a real-valued function; and if $Y = \mathbb{C}$, f is called a complex-valued function.

Definition 5.1.1. Consider a limit point $p \in E$ and a point $q \in Y$. We say the **limit** of the funcion $f(x)$ at p is q, denoted as

$$\lim_{x \to p} f(x) = q$$

if for any $\varepsilon > 0$, there exists some $\delta > 0$ such that for any $x \in E$ with $0 < d_X(x, p) < \delta$, there is

$$d_Y\left(f(x), q\right) < \varepsilon.$$

We can recast this definition in terms of limits of sequences:

$$\lim_{n \to \infty} f(p_n) = q$$

for every sequence $(p_n) \in E$ so that $p_n \neq p$ and $\lim_{n \to \infty} p_n = p$.

By the same proofs as for sequences, limits are unique, and in $\mathbb{R}$ they add/multiply/divide as expected.

Definition 5.1.2. f is **continuous** at p if

$$\lim_{x \to p} f(x) = f(p).$$

In the case where p is not a limit point of the domain E, we say f is continuous at p. If f is continuous at all points of E, then we say f is continuous on E.

The sequential definition of continuity follows almost directly from the sequential definition of limits: f is continuous at p if for every sequence x_n converging to p, the sequence $f(x_n)$ converges to $f(p)$.

§5.2 Continuous Functions

§5.3 Continuity and Compactness

§5.4 Continuity and Connectedness

§5.5 Discontinuities

§5.6 Monotonic Functions

§5.7 Infinite Limits and Limits at Infinity

6 Differentiation

We focus on real valued functions defined on open or closed intervals.

§6.1 The Derivative of a Real Function

Definition 6.1.1. A function $f : [a,b] \to \mathbb{R}$ is called **differentiable** at $x_0 \in [a,b]$, if the limit of the function

$$\phi(t) := \frac{f(t) - f(x_0)}{t - x_0}, \quad a < t < b, t \neq x_0$$

exists as $t \to x_0$. For this case, we write

$$f'(x_0) = \lim_{t \to x_0} \phi(t) = \lim_{t \to x_0} \frac{f(t) - f(x_0)}{t - x_0}. \tag{6.1}$$

The function f is differentiable over $[a,b]$ if it is differentiable for each $x \in [a,b]$. It induces the function

$$\frac{\mathrm{d}f}{\mathrm{d}x} = f' : [a,b] \to \mathbb{R},$$

which is called the **derivative** of f.

Theorem 6.1.1. If $f : [a,b] \to \mathbb{R}$ is differentiable at $x_0 \in [a,b]$, then it must be continuous at x_0.

Proof. As $t \to x$,

$$f(t) - f(x) = \frac{f(t) - f(x)}{t - x} \cdot (t - x) \to f'(x) \cdot 0 = 0.$$

□

Remark. The converse of this theorem is not true. It is easy to construct continuous functions which fail to be differentiable at isolated points.

Notation. We use $C_1[a,b]$ to denote the set of differentiable functions over $[a,b]$ whose derivative is continuous. More generally, we use $C_k[a,b]$ to denote the set of functions whose k-th ordered derivative is continuous. In particular, $C_0[a,b]$ is the set of continuous functions over $[a,b]$.

Later on when we talk about properties of differentiation such as the intermediate value theorems, we usually have the following requirement on the function:

> f is a continuous function on $[a,b]$ which is differentiable in (a,b).

Theorem 6.1.2 (Differentiation rules)**.** Suppose $f, g : [a,b] \to \mathbb{R}$ are differentiable at $x_0 \in [a,b]$. Then $f \pm g$, fg and $\dfrac{f}{g}$ (when $g(x_0) \neq 0$) are differentiable at x_0. Moreover,

1. $(f \pm g)'(x_0) = f'(x_0) \pm g'(x_0)$;

2. $(fg)'(x_0) = f'(x_0)g(x_0) + f(x_0)g'(x_0)$;

3. $\left(\dfrac{f}{g}\right)'(x_0) = \dfrac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{g(x_0)^2}$

Proof. We take (2) as an example.

We calculate

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

□

Exercise 6.1.1

Prove the chain rule.

Proof. Now the proof for chain rule is actually quite nontrivial The proof of the chain rule that is usually taught in high school actually has some holes in it

First we list out the conditions: 1. f is continuous on some interval $[a,b]$ and differentiable at $x \in [a,b]$ (if x is either a or b, then the limit taken in the derivative is a right or left limit). 2. g is a function defined on the range of f (otherwise $g \circ f$ wouldn't exist) and differentiable at $f(x)$.

Again the notion of limits here must be taken over all 'relevant points'

For example let $g(x) = x$ be defined only on the non-negative real numbers, and let $f(x) = x^2$.

Then $g \circ f(x) = x^2$ is defined, and in fact differentiable over $\mathbb{R}$, even though there is a worrying issue at $x = 0$ (here $f(0) = 0$ so we will be dealing with an extremal input for g)

The saving grace here is that g is still differentiable at 0, even though here we need to interpret differentiability through a right limit: $\lim_{x\to 0^+} \frac{g(x)}{x} = 1$, but $\lim_{x\to 0^-} \frac{g(x)}{x}$ does not exist.

If g does not even have such differentiability at the endpoint, then we cannot expect $g \circ f$ to be differentiable, e.g. $g(x) = \sqrt{x}$ and $f(x) = x^2$, then $g(f(x)) = |x|$ is not differentiable at $x = 0$.

Now let's dive into the proof

We know that

$$f'(x) = \lim_{t\to x} \frac{f(t) - f(x)}{t - x},$$

so under the assumption that t stays within the domain of f, $\frac{f(t)-f(x)}{t-x}$ should be a good approximation to $f'(x)$.

To actually quantify this, let $u(t) = \frac{f(t)-f(x)}{t-x} - f'(x)$.

Then the differentiability of f tells us that $\lim_{t\to x} u(t) = 0$.

Similarly, let $v(s) = \frac{g(s)-g(y)}{s-y} - g'(y)$, then $\lim_{s\to y} v(s) = 0$, as long as s stays in the domain of g (in my opinion it is quite helpful to think of these limits as sequential ones)

What's nice here is that we can let $s = f(t)$, then by our assumption s always stays in the domain of g, so nothing fishy will happen

Ah I forgot a small detail here Additionally we also need to define u(x)=0 and v(y)=0

Now let $h(t) = g(f(t))$, then h is defined on $[a,b]$, and we deduce that

$$h(t) - h(x) = (t - x)[f'(x) + u(t)][g'(y) + v(s)]$$

We then check that

$$\lim_{t\to x} \frac{h(t) - h(x)}{t - x} = \lim_{t\to x}[f'(x) + u(t)][g'(y) + v(s)] = f'(x)g'(f(x))$$

and we are done. □

Example 6.1.1

One of the best (worst?) family of pathological examples in calculus are functions of the form

$$f(x) = x^p \sin\frac{1}{x}.$$

Here we're given the examples where p=1 and p=2 For p=1, the function is continuous and differentiable everywhere other than x=0 For p=2, the function is differentiable everywhere, but the derivative is discontinuous

Other more advanced pathological results (just for fun): 1. The graph for $y = \sin\frac{1}{x}1/x$ on (0,1], together with the interval [-1,1] on the y-axis, is a connected closed set that is not path-connected

2. For $0 < p < 1$, we obtain functions that are continuous and bounded, but the graphs are of infinite length (ps. I think that this is also true for p=1)

§6.2 Mean Value Theorems

We say that x is a **stationary point** of f if $f'(x) = 0$.

Local maximum/minimum are points which attains maximum/minimum in some neighbourhood of that particular point.

Theorem 6.2.1 (Fermat's Theorem (Interior Extremum Theorem))**.** If the differential exists, then by comparing the left and right limits it is easy to see that the differential for a local maximum/maximum can only be 0.

To summarize in four words: Local extrema are stationary

There are three mean value theorems, from specific to general:

1. Rolle's Theorem
2. (Lagrange's) Mean Value Theorem
3. Generalised (Cauchy's) Mean Value Theorem

Theorem 6.2.2 (Rolle's Theorem)**.** If f is continuous on $[a,b]$, differentiable in (a,b) and $f(a) = f(b)$, then there exists $c \in (a,b)$ such that

$$f'(c) = 0.$$

Proof. Let $h(x)$ be a function defined on $[a,b]$ where $h(a) = h(b)$.

The idea is to show that h has a local maximum/minimum, then by Fermat's Theorem this will then be the stationary point that we're trying to find.

First note that h is continuous on $[a,b]$, so h must have a maximum M and a minimum m.

If M and m were both equal to $h(a) = h(b)$, then h is just a constant function and so $h'(x) = 0$ everywhere.

Otherwise, h has a maximum/minimum that is not $h(a) = h(b)$, so this extremal point lies in (a,b).

In particular, this extremal point is also a local extremum. Since h is differentiable on (a,b), by Fermat's theorem this extremum point is stationary, thus Rolle's Theorem is proven. □

Theorem 6.2.3 (Mean Value Theorem)**.** If f is continuous on $[a,b]$ and differentiable in (a,b), then there exists $c \in (a,b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Exercise 2: Show that the Mean Value Theorem results directly from Rolle's Theorem (the other direction is trivial) : This isn't a very significant exercise because we're going to prove something more general

Theorem 6.2.4 (Generalised Mean Value Theorem)**.** If f and g are continuous on $[a, b]$ and differentiable in (a, b), then there exists $c \in (a, b)$ such that

$$\frac{f'(c)}{g'(c)} = \frac{f(b) - f(a)}{g(b) - g(a)}.$$

Now we return to the proof of the generalized MVT

We set the function $h(t) = [f(b) - f(a)]g(t) - [g(b) - g(a)]f(t)$, then h is continuous on $[a, b]$ and differentiable on (a, b)

Moreover, $h(a) = f(b)g(a) - f(a)g(b) = h(b)$, thus by Rolle's Theorem, there exists $c \in (a, b)$ such that $h'(c) = 0$, i.e. $[g(b) - g(a)]f'(c) = [f(b) - f(a)]g'(c)$

Corollary: If f and g are continuous on $[a, b]$ and differentiable in (a, b), and $g'(x) \neq 0$ for all $x \in (a, b)$, then there exists

$$c \in (a, b) \text{ s.t. } f'(c)/g'(c) = [f(b) - f(a)]/[g(b) - g(a)]$$

This form of the generalized MVT will be used to prove the most beloved rule of high school students

exercises for the Mean Value Theorem

Exercise 6.2.1

Let f and g be continuous on $[a, b]$ and differentiable on (a, b). If $f'(x) = g'(x)$, then $f(x) = g(x) + C$.

Exercise 6.2.2

Given that $f(x) = x^{\alpha}$ where $0 < \alpha < 1$. Prove that f is uniformly continuous on $[0, +\infty)$.

Exercise 6.2.3: Olympiad level

Let f be a function continuous on $[0, 1]$ and differentiable on $(0, 1)$ where $f(0) = f(1) = 0$. Prove that there exists $c \in (0, 1)$ such that

$$f(x) + f'(x) = 0.$$

§6.3 Applications of MVT

§6.3.1 Darboux's Theorem

Darboux's Theorem implies some sort of a 'intermediate value' property of derivatives that is similar to continuous functions

This is Theorem 5.12 in the book

Now first and foremost, the requirement for this statement is that f must be differentiable on [a,b], not just in (a,b) Otherwise f'(a) and f'(b) may not make sense : One common theme in many of these problems is to construct auxiliary functions Suppose that $f'(a) < \lambda < f'(b)$, then we construct the auxiliary function $g(x) = f(x) - \lambda x$: Then we only need to find a point $x \in (a, b)$ such that g'(x)=0 : This means that we only need to find a local maximum/minimum, which by Fermat's Theorem has to be a stationary point as well : Now we look at the values of g near a and b : Exercise 1: Using the fact that $g'(a) < 0$ and $g'(b) > 0$, show that a and b are local maxima of g

Here we regard g as simply a function on $[a, b]$, so we only need to show that a,b are maximum and corresponding semi-open neighbourhoods $[a, a + \varepsilon)$ and $(b - \varepsilon, b]$: Let m=g'(a)<0 be the slope of the tangent at a : Then lim(h→0+)[g(a+h)-g(a)]/h=m<0 : This means that there should exist $\delta > 0$ such that for $0 < h < \delta$, [g(a+h)-g(a)]/h<m/2<0 : Now we can rewrite the above as g(a+h)<g(a)+mh/2 : Since m<0 and h>0, we obtain $g(a + h) < g(a)$ for $0 < h < \delta$: Thus this proves that x=a is a local maximum of g A similar proof applies for x=b : Now since g is differentiable on [a,b], in particular it has to be continuous on [a,b] : Since [a,b] is compact, g([a,b]) is compact in R and thus g has both maximum and minimum values in [a,b] : Here we'll just focus on the minimum value : As we've shown, x=a is a 'strict' local maxima, in the sense that for any point $x \in (a, a+\varepsilon)$, we actually have the strict inequality $g(x) < g(a)$: This means that x=a cannot be a local minimum : Similarly, x=b cannot be a local minimum, and therefore g achieves its minimum strictly inside (a,b) : Only then we can say that this local minimum is stationary (This will not work otherwise; note that a and b are both local maxima but are not stationary points of g) : An interesting implication of Darboux's Theorem is that if f is differentiable on [a,b], then f' cannot have simple discontinuities (removable or jump discontinuities), simply because these discontinuities do not allow this 'intermediate value' property : However, we should recall certain pathological examples like f(x)=x² sin 1/x (f(0)=0) Here f'(0)=lim(h→0)[x² sin 1/x-0]/x=0, but f'(x)=2x sin 1/x - cos 1/x, so f' is discontinuous at x=0

§6.3.2 L'Hopital's Rule

First, a counterexamples : Let's say that we apply this rule to $\lim_{x\to\infty} \frac{\sin x}{x}$: Then we have

$$\lim_{x\to\infty} \frac{\sin x}{x} = \lim_{x\to\infty} \frac{\cos x}{1}$$

The limit on the RHS doesn't exist because cos x oscillates between -1 and 1 : However, the limit on the LHS does in fact exist and is equal to 0 : So this tells us that there are certain cases where we can apply L'Hopital, and other cases where we can't That being said, the case that we can apply the rule is actually the more useful case, so this situation does not jeopardize the effectiveness of L'Hopital : The entire statement is consequently rather long, so we'll split it into a few sections : 1. f and g are differentiable in (a,b) and $g'(x) \neq 0$ in (a,b) (or at least in a small neighbourhood of a) : 2. f(x)/g(x) is an indeterminate of the form 0/0 or $\frac{\infty}{\infty}$ (Now for the second one here we only really need $g(x) \to \infty$, but if f(x) does not approach infinity then the limit would simply be zero, so L'Hopital's Rule would not be required here) : 3. lim(x→a)f'(x)/g'(x) = A This is the most important one : From this we obtain lim(x→a)f(x)/g(x) = A : So for example, let's say that we want to calculate the following limit: lim(x→0) (sin x - x)/x³ : Repeated application of L'Hopital gives lim(x→0) (sin x - x)/x³ =lim(x→0) (cos x - 1)/3x² =lim(x→0) -sin x/6x =-1/6 : Now what we're really doing here is that, first we know that lim(x→0) sin x/x =1, so lim(x→0) -sin x/6x =-1/6 : Then by L'Hopital, lim(x→0) (cos x - 1)/3x²=-1/6 : Finally, again by L'Hopital, lim(x→0) (sin x - x)/x³=-1/6 : So, one very important thing to take note is that if you're calculating some complicated limit and you end up with the conclusion that it doesn't exist, you must make sure that you have not used L'Hopital during the process, because the rule never applies in such situations : As a side note, from the above calculation we see that as x→0, $\sin x \approx x - \frac{x^3}{6}$ This will later lead to the discussion of the Taylor series of sin x

Now the entire proof is quite tedious because there's actually eight main cases to think of 1. $\frac{0}{0}$ or $\frac{\infty}{\infty}$ 2. a is normal or $a = -\infty$ 3. A is normal or $A = \pm\infty$

We'll only prove the most basic one here: 0/0, a and A are normal This is the case which will be required for Taylor series

First we define f(a)=g(a)=0, so that f and g are continuous at $x = a$

Now let $x \in (a, b)$, then f and g are continuous on $[a, x]$ and differentiable in (a, x) : Thus by Cauchy's Mean Value Theorem, there exists $\xi \in (a, x)$ such that

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

For each x, we pick ξ which satisfies the above, so that ξ may be seen as a function of x satisfying $a < \xi(x) < x$

Then by squeezing we have $\lim_{x\to a^+} \xi(x) = a$.

Since $\frac{f'}{g'}$ is continuous near a, the theorem regarding the limit of composite functions give

$$\lim_{x\to a^+} \frac{f(x)}{g(x)} = \lim_{x\to a^+} \frac{f'(\xi)}{g'(\xi)} = \lim_{x\to a^+} \left(\frac{f'}{g'}\right)(\xi(x)) = A$$

Now the same reasoning can be used for b where we will use lim(x→b-) to replace all the $\lim_{x\to a^+}$, and ξ will be a function which maps to (x,b).

The ones with infinity are generally more complicated, but as we've mentioned we won't go through all the trouble : In particular, because you have the statement which works for both the left and right hand limits, L'Hopital works for ordinary limits as well

§6.3.3 Taylor Series

The main expression is as follows:

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots \tag{6.2}$$

So for example we have the following (we've used the ones for e^x and $\ln x$ for generating functions):

$$\begin{aligned} e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \\ \sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \\ \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots \\ \ln(1+x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \end{aligned}$$

There's a lot of things to say about these equations, for example the one for $\ln(1+x)$ only works for $|x| < 1$

Also, if you want the RHS of the expression to be an infinite power series, $f(x)$ has to be smooth (infinitely differentiable)

Even then, the power series may never converge to $f(x)$ at any interval, no matter how small The most common example given here is $f(x) = e^{\frac{-1}{x^2}}$ (f(0)=0); the Taylor series for $f(x)$ is just 0

Now sometimes we don't actually that nice of a property for f, we're often given that fact that f is only finitely differentiable

Then we will have something along the lines of

$$f(x) \approx f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n$$

where $f^{(n)}$ denotes the n-th differential.

There are two main forms of the statement regarding the error between the original function and the Taylor series estimate

The simpler form is what's known as the Peano form: Given that f is n times differentiable at a, then

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + o((x-a)^n)$$

To show this, we only need to show that we have the following limit:

$$\lim_{x\to a} \frac{f(x) - f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n}{(x-a)^n} = 0$$

The basic idea is to use the L'Hopital Rule n times. The numerator becomes $f^{(n)}(x) - f^{(n)}(a)$ which approaches 0, whereas the denominater is just $n!$, so the limit exists and is equal to 0.

However, we need to verify all the necessary conditions for L'Hopital : Here the main problem is that we don't know if we have the $0/0$ indeterminate at each step, so we'll need to check this for the k-th step where k=1,...,n

Fortunately, the k-th derivative of the numerator is $f^{(k)}(x) - f^{(k)}(a) - (x-a)F_k(x)$ where F_k is just a bunch of random stuff, so the numerator approaches 0 as $x \to a$ The k-th derivative of the denominator is $n(n-1)\cdots(n-k+1)(x-a)^{n-k}$ so it also approaches 0, and we're done

The other form is actually a family of similar statements which gives more precise values for the error The Peano form has a fundamental obstacle when used in approximation, we don't have any control on the size of the final term other than its asymptotic behaviour : We'll be talking about the one given in the book, known as the Lagrange form: : Given that f is n times differentiable on (a,b) such that $f^{(n-1)}$ is continuous on $[a,b]$, then

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n-1)}(a)}{(n-1)!}(x-a)^(n-1) + \frac{f^{(n)}(\xi)}{n!}(x-a)^n$$

Just like in L'Hopital, we intuitively think of (a,b) as just a very small interval at the right hand side of x=a : Here we are giving up on the second final term of Peano by combining it with the infinitesimal (small o) term to give an accurate description of the error

For the proof of this one we'll be using Cauchy's MVT

Fix any $x \in (a,b)$, then we construct the functions

$$F(t) = f(x) - \left(f(t) + \frac{f'(t)}{1!}(x-t) + \frac{f''(t)}{2!}(x-t)^2 + \cdots + \frac{f^{(n-1)}(t)}{(n-1)!}(x-t)^{n-1} \right)$$

$$G(t) = (x-t)^n$$

We calculate $F'(t)$ as follows:

$$-[f'(t) + \frac{f''(t)}{1!} - f'(t) + \frac{f'''(t)}{2!} - \frac{f''(t)}{1!} + \cdots + \frac{f^{(n)}(t)}{(n-1)!}(x-t)^{n-1} - \frac{f^{(n-1)}(t)}{(n-2)!}(x-t)^{n-2}] = -\frac{f^{(n)}(t)}{(n-1)!}(x-t)^{n-1}$$

$G'(t) = -n(x-t)^{n-1}$, so we have

$$\frac{F'(t)}{G'(t)} = \frac{f^{(n)}(t)}{n!}$$

The main reason for why we come up with the strange-looking F and G is that we specifically swap out a for t so that $F(x) = G(x) = 0$, in hopes of getting rid of x:

We apply Cauchy's MVT to F and G on $[a,x]$, so that we obtain $\xi \in (a,x)$ satisfying

$$\frac{F'(\xi)}{G'(\xi)} = \frac{F(x) - F(a)}{G(x) - G(a)} = \frac{F(a)}{G(a)}.$$

Thus the Lagrange form of the remainder is given by

$$F(a) = \frac{f^{(n)}(\xi)}{n!}G(a).$$

Theorem 5.19 is important, so do go through that proof as an exercise

7 The Riemann–Stieltjes Integral

§7.1 Definition of Riemann–Stieltjes Integral

Assume $[a,b]$ is a closed interval in $\mathbb{R}$. By a **partition** P, we mean a finite set of points

$$P : a = x_0 \le x_1 \le \cdots \le x_{n-1} \le x_n = b.$$

Assume f is a bounded real-valued function over $[a,b]$ and α is an increasing function over $[a,b]$. Denote by

$$M_i = \sup_{[x_{i-1},x_i]} f(x), \quad m_i = \inf_{[x_{i-1},x_i]} f(x)$$

and by

$$\Delta\alpha_i = \alpha(x_i) - \alpha(x_{i-1}).$$

Define the **upper sum** of f with respect to the partition P and α as

$$U(f,\alpha;P) = \sum_{i=1}^{n} M_i \Delta\alpha_i$$

and the **lower sum** of f with respect to the partition P and α as

$$L(f,\alpha;P) = \sum_{i=1}^{n} m_i \Delta\alpha_i.$$

Define the upper Riemann–Stieltjes integral as

$$\overline{\int_a^b} f(x)\,\mathrm{d}\alpha(x) \coloneqq \inf_P U(f,\alpha;P)$$

and the lower Riemann–Stieltjes integral as

$$\underline{\int_a^b} f(x)\,\mathrm{d}\alpha(x) \coloneqq \sup_P L(f,\alpha;P).$$

It is easy to see from definition that

$$\underline{\int_a^b} f(x)\,\mathrm{d}\alpha(x) \le \overline{\int_a^b} f(x)\,\mathrm{d}\alpha(x).$$

Definition 7.1.1. A function f is **Riemann–Stieltjes integrable** with respect to α over $[a,b]$, if

$$\underline{\int_a^b} f(x)\,\mathrm{d}\alpha(x) = \overline{\int_a^b} f(x)\,\mathrm{d}\alpha(x).$$

Notation. We use $\int_a^b f(x)\,\mathrm{d}\alpha(x)$ to denote the common value, and call it the Riemann–Stieltjes of f with respect to α over $[a,b]$.

Notation. We use the notation $R_\alpha[a,b]$ to denote the set of Riemann–Stieljes integrable functions with respect to α over $[a,b]$.

In particular, when $\alpha(x) = x$, we call the corresponding Riemann–Stieljes integration the **Riemann integration**, and use $R[a,b]$ to denote the set of Riemann integrable functions.

Now we move on to integrability conditions for f. The first one looks a lot like the $\varepsilon - N$ or $\varepsilon - \delta$ definition of limits:

Theorem 7.1.1. $f \in R_\alpha[a,b]$ if and only if for each $\varepsilon > 0$, there exists some partition P such that

$$U(f,\alpha;P) - L(f,\alpha;P) < \varepsilon.$$

Proof. If $f \in R_\alpha[a,b]$, then the two Darboux integrals would be equal and so

$$\inf_P U(f,P,\alpha) = \sup_P L(f,P,\alpha) = I$$

where I is the RS-integral of f wrt α.

Then for any $\varepsilon > 0$, we can find P_1 and P_2 such that

$$I \le U(f,P_1,\alpha) < I + \frac{\varepsilon}{2}$$

and

$$I - \frac{\varepsilon}{2} < L(f,P_2,\alpha) \le I.$$

Let P' be the common partition of P_1 and P_2, then

$$L(f,P_2,\alpha) \le L(f,P',\alpha) \le U(f,P',\alpha) \le U(f,P_1,\alpha)$$

And so from the above inequalities we have

$$I - \frac{\varepsilon}{2} < L(f,P',\alpha) \le U(f,P',\alpha) < I + \frac{\varepsilon}{2}$$

Therefore $U(f,P',\alpha) - L(f,P',\alpha) < \varepsilon$ (Sorry for the inconsistent notation) : On the other hand, for any given partition P we can always bound the Darboux integrals between the lower and upper Darboux sums over P.

So whenever we find a partition P such that

$$0 \le U(f,P,\alpha) - L(f,P,\alpha) < \varepsilon$$

We immediately obtain that

$$0 \le U(f,\alpha) - L(f,\alpha) < \varepsilon$$

(I'm not sure if this notation down here is universal so please stick to the original Darboux integral notations)

Since $\varepsilon > 0$ is arbitrary, this gives us that the two Darboux integrals must be equal. □

Example 7.1.1: Dirichlet function

The Dirichlet function is given by

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

We try to calculate the two on the interval $[0,1]$.
The Dirichlet function is pathological because for each subinterval $[x_{i-1}, x_i]$, the supremum is always 1 and the infimum is always 0.
So no matter what partition we use, $U(f,P)$ is always 1 whereas $L(f,P)$ is always 0. This means that $U(f) = 1$ and $L(f) = 0$, so there are two different values for "the integral of f".
This is like the case where we try to find the limit of the Dirichlet function where x is approaching any given real number r, there exists two sequences approaching r whose image approaches two different values.

Now, a very important and fun case about the more general RS-integral, which we'll discuss next week (do try the exercise yourself first)

Exercise 7.1.1

The Heaviside step function H is a real-valued function defined by the following:

$$H(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

For the purpose of this question we assume the convention $\infty \cdot 0 = 0$.

(a) Let f be a real-valued function over $\mathbb{R}$. Show that $f \in \mathbb{R}_H[a,b]$ iff f is continuous at 0, and find the RS-integral $\int_{-\infty}^{\infty} f \, \mathrm{d}H$.

(b) Suppose that the definition for H is changed for $x = 0$, say $H(0) = \frac{1}{2}$. Show that the above result still holds.

(c) Examine the RS-integral of f over $\mathbb{R} \setminus \{0\}$ wrt H, where f is a real-valued function over $\mathbb{R} \setminus \{0\}$ such that $\lim_{x\to 0} f(x) = \infty$ or $-\infty$.

(You may read up on more information regarding the Heaviside function, and the (in)famous Dirac delta function)

To to recap last time, we defined when a function is Riemann integrable over a closed interval [a,b], or more generally, when it is Riemann-Stieltjes integrable over [a,b] with respect to α

Now we'll need a key idea regarding partitions which will enable us to develop much deeper properties of integrals, namely refinements

This is quite simple: given that

$$P : a \leq x_0 \leq x_1 \leq \ldots \leq x_n = b$$

and

$$P' : a \leq y_0 \leq y_1 \leq \ldots \leq y_m = b,$$

we say that P' is a refinement of P if all points in P are in P', i.e. $\{x_i\}$ is a subset of $\{y_j\}$.

Furthermore, given two partitions P_1 and P_2, we can construct a refinement P' which collects all partition points of P_1 and P_2, giving what's called the **common partition** of P_1 and P_2.

Intuitively, a refinement will give a better estimation than the original partition, so the upper and lower sums of a refinement should be more restrictive

So there's Theorem 6.4 which states that, indeed, one has

$$L(P,f,\alpha) \leq L(P',f,\alpha) \leq U(P',f,\alpha) \leq U(P,f,\alpha)$$

This can be proven without induction: Suppose that

$$P : a \leq x_0 \leq x_1 \leq \ldots \leq x_n = b$$

and

$$P' : a \leq y_0 \leq y_1 \leq \ldots \leq y_m = b.$$

Then there exists a strictly increasing sequence of indices $j_0 = 0, j_1, \ldots, j_n = m$ such that $y_{j_k} = x_k$.

Now consider each closed interval $[x_{i-1}, x_i]$

Focusing on the upper sum, we have

$$\sup_{[x_{i-1},x_i]} f \geq \sup_{[y_{k-1},y_k]} f$$

for $k = j_{i-1} + 1, \ldots, j_i$. This is because $[y_{k-1}, y_k]$ is contained in $[x_{i-1}, x_i]$

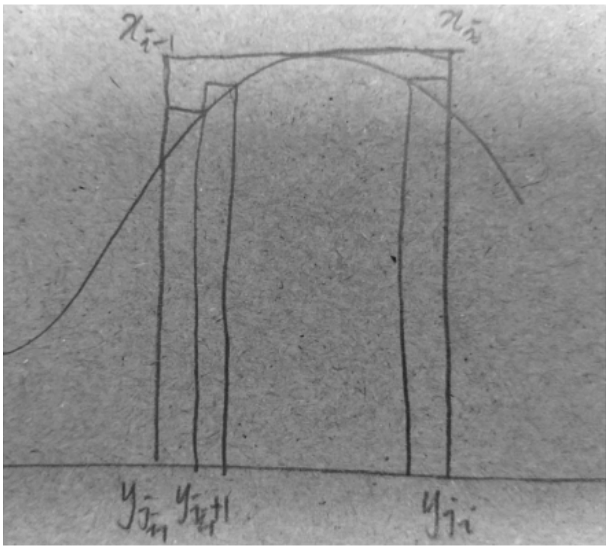


Figure 7.1: Partitions

Continuing from

$$\sup_{[x_{i-1},x_i]} f \geq \sup_{[y_{k-1},y_k]} f,$$

We then multiply by $\alpha(y_k) - \alpha(y_{k-1})$ on both sides and then take the sum from $k = j_{i-1} + 1$ to $k = j_i$: The RHS corresponds to the (weighted) sum of the thin rectangles that you see in the above picture : The LHS is actually a telescoping sum, and the sum would be

$$(\sup_{[x_{i-1},x_i]} f) \cdot [\alpha(y_{j_i}) - \alpha(y_{j_{i-1}})] = (\sup_{[x_{i-1},x_i]} f) \cdot [\alpha(x_i) - \alpha(x_{i-1})]$$

Finally, we take the sum from $i = 1$ to $i = n$ of the above inequality LHS ≥ RHS (sorry I don't know of a better way to put it) We then obtain $U(P, f, \alpha) \geq U(P', f, \alpha)$

(On the LHS we're collecting all the rectangles for the upper sum wrt P, but on the RHS we're collecting up collections of upper rectangles to obtain the entire collective of upper rectangles for the upper sum wrt P') : Lower sum is similar : Now, a lemma used to prove 6.5 Given any two partitions P_1 and P_2, we have

$$L(P_1, f, \alpha) \leq U(P_2, f, \alpha)$$

So a lower sum will always be no larger than any other upper sum : So this includes the cases where we have the most refined of P_1's and P_2's, with no information regarding the partition points whatsoever To be honest, the result seems to be both intuitive and unclear at the same time

The key here is to use common refinements as a link for both sums The idea is stated in the proof of 6.5 and I don't think I need to elaborate further

What's nice here is that now we have two completely independent partitions P_1 and P_2, so by fixing one partition, say P_2, and taking the 'limit' over the other (here we take the supremum over all possible P_1) we then obtain an inequality between a Darboux integral and a Darboux sum (here it's the lower integral and an upper sum)

Since the Darboux integral is just a number, we can then safely take the 'limit' over the other partition to obtain the inequality in 6.5

As we've already defined, f is Riemann-Stieltjes integrable wrt α if these two Darboux integrals coincide

Now we've been talking a lot about upper and lower sums because they're arguably the simplest way to define integrals, in the sense that there's not a whole lot of things that we could go wrong here By considering only upper and lower bound, we're essentially picking the most conservative route possible

It would be nice if we could just pick like one random point within each interval and consequently calculate the Riemann(-Stieltjes) sums

This method, of course, fails to be well defined for pathological functions like the Dirichlet function On the other hand, by using upper and lower sums, we could give a persuasive explanation as to why the Dirichlet function is not Riemann integrable

However, instead of throwing this idea away, there's actually a way for us to make this into a strict definition

When we were talking about the sequential definition for limits of functions, we noted that there are certain scenarios where the limit cannot exist because there may be two distinct sequences may give different limit Based on this observation, we then gave a reasonable condition as follows: "$\lim_{x\to a} f(x)$ exists and is equal to L iff for all sequences x_n converging but not containing a, $f(x_n)$ converges to L"

Well here, it's actually the same kind of scenario Given any partition P, we consider the Riemann sum $\sum f(\xi_i)\Delta x_i$ where ξ_i is any point where $x_{i-1} \le \xi_i \le x_i$

For the Dirichlet function over [0,1], given any partition P (here we may assume that the partition points are distinct), we will always be able to specifically pick $\xi_i, \eta_i \in [x_{i-1}, x_i]$ such that ξ_i is rational but η_i is irrational

Then $\sum f(\xi_i)\Delta x_i = 1$ but $\sum f(\eta_i)\Delta x_i = 0$

Now be very mindful that this alone cannot be evidence that f is non-integrable The key is that this somehow occured for all partitions P, no matter how refined they are; for every single partition P, there exists two sets of 'representing points' ξ_i, η_i such that the two Riemann sums are constantly far apart (1 and 0 in this case)

Let $\varepsilon_0 = 1$, then this ultimately translates to the following: The Dirichlet function cannot be Riemann integrable because There exists some $\varepsilon_0 > 0$, such that for any given partition P, there exists two sets of representing points ξ_i, η_i such that their corresponding Riemann sums satisfy that

$$|\sum f(\xi_i)\Delta x_i - \sum f(\eta_i)\Delta x_i| \ge \varepsilon_0.$$

Now if we always pick the representatives such that $\xi_i > \eta_i$ then we can neglect the absolute value

So now, let's take the converse A function f is said to be RS-integrable if For every $\varepsilon > 0$, There exists a partition P, such that For any two sets of representing points ξ_i, η_i, Their corresponding Riemann sums satisfy that

$$\sum[f(\xi_i) - f(\eta_i)]\Delta x_i < \varepsilon$$

(The last one should be $\Delta\alpha_i$ for RS-integrals, not Δx_i)

Unfortunately this is still not quite the correct definition according to Apostol, but we're pretty close The problem with this definition is that it is too weak if we're considering general α of bounded variation; if we were only talking about monotonically increasing α then this will actually be an equivalent definition

The official definition for the RS-integral wrt α of bounded variation is as follows:

Definition 7.1.2. For every $\varepsilon > 0$, there exists a partition P, such that [For any refinement P' of P, and] For any two sets of representing points ξ_i, η_i [of P'], their corresponding Riemann sums satisfy that

$$\sum[f(\xi_i) - f(\eta_i)]\Delta x_i < \varepsilon.$$

Now this definition is what mathematicians would refer to as a 'Cauchy' definition, since it defines a notion by comparing a pair of arbitrary values that are similar to one another, and if they agree in some sense then we say that that something satisfies some property.

The integral is then obtained as follows: If f were to satisfy the above Cauchy definition, then we may pick an arbitrary sequence of refinements

$$P_1 \subset P_2 \subset P_3 \subset \ldots;$$

and for each partition we pick a set of representatives to obtain a sequence RS-sum $I_1, I_2, I_3, \ldots :$ This sequence will be a Cauchy sequence of real numbers, and so will converge to a specific value I which we consider to be RS-integral of f : Now the reason why Apostol needed to strengthen the definition is that, otherwise this value I may not be unique : So if you look at the statement you see in 6.7(b)(c), then they correspond to the Cauchy definition and the 'value-based' definition respectively For monotonically increasing α, it is much easier to discuss them using upper and lower sums So your exercise today will be to read the statements and proofs in Theorem 6.7

§7.2 Properties of the Integral

Theorem 7.2.1.

1. Linearity of f:
2. Linearity of α:

§7.3 Fundamental Theorem of Calculus

Theorem 7.3.1.

8 Sequence and Series of Functions

§8.1 Uniform Convergence

Part III

Topology

9 n-dimensional Euclidean space

§9.1 What is $\mathbb{R}^n$?

$\mathbb{R}^n$, as a set, is defined as the set of vertical vectors with n coordinates in the real numbers. Algebraically, $\mathbb{R}^n$ is a n-dimensional vector space over $\mathbb{R}$; vectors in $\mathbb{R}^n$ are expressed as vertical vectors:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

We usually express the above vector compactly as follows:

$$x = (x_1, \dots, x_n)^T$$

Since $\mathbb{R}^n$ is a vector space (over $\mathbb{R}$), $\mathbb{R}^n$ has the following extra properties

- For any two vectors x, y we may perform addition

  $$x + y = (x_1 + y_1, \dots, x_n + y_n)^T$$

  Properties of addition:

  1. $x + y = y + x$
  2. $(x + y) + z = x + (y + z)$
  3. Zero vector $0 = (0, \dots, 0)$ satisfies $x + 0 = 0 + x = x$
  4. For any vector x, its negative $-x$ satisfies $x + (-x) = (-x) + x = 0$

- For any vector x and real number (scalar) k we may perform scalar multiplication

  $$kx = (kx_1, \dots, kx_n)^T$$

  Properties of scalar multiplication:

  1. $0 \cdot x = 0, 1 \cdot x = x$
  2. $(kl)x = k(lx) = l(kx)$
  3. $k(x + y) = kx + ky$
  4. $(k + l)x = kx + lx$

These properties make up the algebraic structure of $\mathbb{R}^n$, which may then be further expanded in linear algebra. However, in this section we shall focus on the analytical/topological aspects of Euclidean space.

So what's the difference? The Euclidean space is something that builds upon the vector space $\mathbb{R}^n$. Specifically speaking, it is $\mathbb{R}^n$ endowed with two extra notions:

- The **norm** of the Euclidean space $\|\cdot\|$ is a real-valued function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$. Given a vector $x = (x_1, \ldots, x_n)^T$ in $\mathbb{R}^n$, the norm of x is defined as

$$\|x\| \coloneqq \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{x_1^2 + \cdots + x_n^2}.$$

- The **metric** d of the Euclidean space is a real-valued function $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. Given two vectors $x = (x_1, \ldots, x_n)^T$ and $y = (y_1, \ldots, y_n)^T$, the distance between x and y is defined as

$$d(x,y) \coloneqq \|x - y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}.$$

As you can see here, the norm is something like the length of the vector itself (distant to the origin; absolute value in the case of $\mathbb{R}^1$). The **metric** on the other hand refers to the distance function which measures the length between two points in $\mathbb{R}^n$ (determined by their positional vectors x and y). Now these two notions may seem similar to each other, but in fact they are pretty asymmetrical. Essentially, the metric is a much more general notion than the norm: the norm can only be defined on vector spaces; the metric can literally be defined on any set.

Before explaining this, we will look at the fundamental properties of the norm and the metric. In fact, these properties are precisely the definition of a norm over some random vector space, and of a metric over some random set.

Remark. Note that the definitions for norm and metric depends on the space; in the above definitions the algebraic expression defining these two are only applicable to Euclidean spaces, and in fact they define what an Euclidean space should be, lengths intuitive to our common perception.

Norms are required to satisfy the following:

1. **Positive Definiteness**: For any vector x, $\|x\| \geq 0$, and $\|x\| = 0$ if and only if $x = 0$.
2. **Absolute Homogeneity**: For any vector x and scalar a, $\|ax\| = |a| \cdot \|x\|$.
3. **Subadditivity (Triangle Inequality)**: For any two vectors x and y, $\|x + y\| \leq \|x\| + \|y\|$.

Metrics are required to satisfy the following:

1. **Positive Definiteness**: For any two elements x and y, $d(x,y) \geq 0$, and $d(x,y) = 0$ if and only if $x = y$.
2. **Symmetry**: For any two elements x and y, $d(x,y) = d(y,x)$.
3. **Triangle Inequality**: For any three elements x,y and z, $d(x,z) \leq d(x,y) + d(y,z)$.

Generally, if there is a norm $\|\cdot\|$ on some random vector space, then this norm naturally determines a metric $d(x,y) = \|x - y\|$, which is precisely the case for Euclidean spaces.

§9.2 Concepts in Euclidean Space

§9.2.1 Bounded Sets

Definition 9.2.1 (Bounded set)**.** A set E in $\mathbb{R}^n$ is called a *bounded set* if there exists $M > 0$ such that $\|x\| \leq M$ for all x in E.

Exercise 9.2.1

Given E and F in $\mathbb{R}^n$ and real number k, define

$$kE = \{kx \mid x \in E\}$$

$$E + F = \{x + y \mid x \in E, y \in F\}$$

(a) Show that if E is bounded, then kE is bounded;

(b) Show that if E and F are bounded, then $E + F$ is bounded.

§9.2.2 Diameter

Definition 9.2.2 (Diameter)**.** Given a set $E \subset \mathbb{R}^n$, the *diameter* of E is defined as

$$\operatorname{diam} E = \sup_{x,y \in E} d(x,y).$$

Exercise 9.2.2

Find the diameter of the open unit ball in $\mathbb{R}^n$ given by

$$B = \{x \in \mathbb{R}^n \mid \|x\| < 1\}.$$

Solution. First note that

$$d(x,y) = \|x - y\| \le \|x\| + \|-y\| = \|x\| + \|y\| < 1 + 1 = 2.$$

On the other hand, for any $\varepsilon > 0$, we pick

$$x = \left(1 - \frac{\varepsilon}{4}, 0, \ldots, 0\right), \quad y = \left(-\left(1 - \frac{\varepsilon}{4}\right), 0, \ldots, 0\right).$$

Then $d(x,y) = 2 - \dfrac{\varepsilon}{2} > 2 - \varepsilon$.

Therefore $\operatorname{diam} B = 2$. □

Exercise 9.2.3

Given a set E in $\mathbb{R}^n$, show that E is bounded iff $\operatorname{diam} E < +\infty$.

Proof. **Forward direction:**

If E is bounded, then there exists $M > 0$ such that $\|x\| \le M$ for all $x \in E$.

Thus for any $x, y \in E$,

$$d(x,y) = \|x - y\| \le \|x\| + \|y\| \le 2M.$$

Thus $\operatorname{diam} E = \sup d(x,y) \le 2M < +\infty$.

Backward direction:

Suppose that $\operatorname{diam} E = r$. Pick a random point $x \in E$, suppose that $\|x\| = R$.

Then for any other $y \in E$,

$$\|y\| = \|x + (y - x)\| \le \|x\| + \|y - x\| \le R + r.$$

Thus, by picking $M = R + r$, we obtain $\|y\| \le M$ for all $y \in E$, and we are done.

Remark. Basically you use x to confine E within a ball, which is then confined within an even bigger ball centered at the origin.

□

§9.2.3 Distance Between Sets

Definition 9.2.3 (Distance between sets)**.** Given two sets $E, F \subset \mathbb{R}^n$, the *distance between sets* E and F is defined as

$$d(E,F) = \inf_{x \in E, y \in F} \|x - y\| .$$

Obviously $d(E,F) > 0$ implies that E and F are disjoint, but E and F may still be disjoint even if $d(E,F) = 0$. For example, the closed intervals $E = (-1,0)$ and $F = (0,1)$.

Exercise 9.2.4

Suppose that E and F are sets in $\mathbb{R}^n$ where E and F is finite. Prove that E and F are disjoint iff $d(E,F) > 0$.

§9.3 Topology in Euclidean Space

Before we move on, we need to talk about how we think about topology. The concept first begins with an attempt to say that two points are close to one another. Of course, we did define the metric earlier, But as it turns out, this particular notion can be made extremely abstract. Specifically speaking, we could theoretically define closeness simply with set theory.

Definition 9.3.1 (Neighbourhood basis)**.** Given a set X, we define a family of subsets in X, denoted by $\mathcal{B}$, to describe points close to each other; points that belong to the same set U in $\mathcal{B}$ are considered to be close to each other w.r.t. U.

Definition 9.3.2 (Neighbourhood)**.** Given a point $x \in X$, we use the term neighbourhood to describe a particular construction for x; N is said to be a neighbourhood of x, if there exists U in $\mathcal{B}$ containing x such that U is a subset of N.

Definition 9.3.3 (Neighbourhood system)**.** Given a point $x \in X$, the neighbourhood system of x, denoted by $\mathcal{N}(x)$, is the set of all neighbourhoods of x.

$$N \in \mathcal{N}(x) \iff \exists U \in \mathcal{B} \text{ s.t. } x \in U \subset N$$

There are still some problems regarding the above definitions.

- If M and N are neighbourhoods of x, is $M \cap N$ a neighbourhood of x?

  Supposedly the answer should be yes, because M and N should contain all the points that are close to x.

- If y is close to x w.r.t. N, then supposedly the points in N should contain points close to Y as well.

  There are some of the more basic requirements, such as x must have a way to define closeness so there must be a neighbourhood containing x (even if it's just one set $\{x\}$, which sort of implies that every other element is far away from x).

People realised that the above requirements can be formalised simply with the neighbourhood systems themselves

These are the axioms for the neighbourhood systems:

1. $\mathcal{N}(x)$ is nonempty, and for all $U \in \mathcal{N}(x)$, $x \in U$.
2. If $U, V \in \mathcal{N}(x)$, then there exists $W \in \mathcal{N}(x)$ such that W is a subset of $U \cap V$.
3. If $U \in \mathcal{N}(x)$ and $y \in U$, then there exists $V \in \mathcal{N}(y)$ such that V is a subset of U.

As for the Euclidean plane, we have a natural way of defining the neighbourhood systems. First we pick the neighbourhood basis to be

$$\mathcal{B} = \{B(x, \varepsilon) \mid x \in \mathbb{R}^n, \varepsilon > 0\}$$

Then we say that N is a neighbourhood of x if there exists $\varepsilon > 0$ such that $B(x, \varepsilon)$ is in N.

$B(x, \varepsilon)$ represents the points close to x, whereas a neighbourhood N of x should contain all the points close to x, at least from the perspective of $B(x, \varepsilon)$.

Once we have neighbourhood systems, we can then define the two most important kinds of sets in topology, open and closed sets.

(In topology, we actually define open and closed sets with axioms first and then define neighbourhoods from there, but this definition is very opaque and will only be revised in the far future.)

- Basic constructions

  A **ball** in $\mathbb{R}^n$ is determined by its center $x \in \mathbb{R}^n$ and its radius $r > 0$, denoted by $B(x, r)$.

  A **punctured ball** in $\mathbb{R}^n$ is a ball excluding its center, denoted by $B_0(x, r)$.

- Neighbourhood, interior and open sets

  A set $A \subset \mathbb{R}^n$ containing x is called a **neighbourhood** of x if $B(x, \varepsilon) \subset A$ for some $\varepsilon > 0$.

  An element $x \in A$ is called an **interior point** if A is a neighbourhood of x.

  The **interior** of a set A, denoted by A°, is the set of all interior points in A.

  A set $A \subset \mathbb{R}^n$ is called an **open set** if $A^\circ = A$, i.e. all points in A are interior points.

- Limit points, closure and closed sets

  An element $x \in A$ is called a limit point of A if $B_0(x, \varepsilon) \cap A \neq \varnothing$ for all $\varepsilon > 0$.

  The induced set of a set, denoted by A', is the set of all limit points of A.

  The closure of a set A, denoted by $\bar{A}$, is the union set $A \cup A'$.

  A set $A \subset RR^n$ is called a closed set if $\bar{A} = A$, i.e. all limit points of A are contained in A

- Further topological constructions of points

  An element $x \in A$ is called an isolated point of A if it is not a limit point of A.

  The boundary of a set A, denoted by ∂A, is the set difference $\bar{A} \setminus A^\circ$.

  An element $x \in \mathbb{R}^n$ is called a boundary point of A if it is in ∂A.

  An element $x \in \mathbb{R}^n$ is called an exterior point of A if it is an interior point of A^c.

- Further topological constructions of sets

  A set $A \subset \mathbb{R}^n$ is compact if it is a bounded closed set.

  A subset $B \subset A$ is called a dense subset of A if $\bar{B} = A$.

  A set $A \subset \mathbb{R}^n$ is nowhere dense its closure has no interior, i.e. $(\bar{A})^\circ = \varnothing$.

In general topology, these are the axioms used to define open and closed sets. At the moment we only consider them to be certain properties regarding open and closed sets in $\mathbb{R}^n$.

P1 A is open if and only if A^c is closed.

Proof. **Forward direction**: Let A be open, we consider the punctured balls of $x \notin A$ (if $x \notin A$, we consider the punctured balls centered at x).

Our goal is to show that $B_0(x, r)$ always intersects with A^c

So suppose otherwise that $B_0(x, \varepsilon)$ is a subset of A for some $\varepsilon > 0$

Ah no sorry, we consider x not in A^c

The thing is we want to show that A^c is closed, i.e. all limit points of A^c are in A^c

So suppose otherwise that x is a limit point of A^c that is not in A^c

x is a limit point of A^c, hence for all $\varepsilon > 0$, $B_0(x, \varepsilon)$ always intersects with A^c

This is equivalent to saying that $B_0(x, \varepsilon)$ is never a subset of $(A^c)^c = A$

However, x is not in $x \notin A^c$, so $x \in A$.

But if A is open, then there exists $\varepsilon > 0$ such that $B(x, \varepsilon)$ is a subset of A, a contradiction

Backward direction: Let A^c be closed. Suppose otherwise that A is not open, i.e. there is a point $x \in A$ such that $B(x, \varepsilon)$ is never a subset of A; that is to say, $B(x, \varepsilon)$ always intersects with A^c

Since $x \in A$, then $B(x, \varepsilon) \cap A^c = B_0(x, \varepsilon) \cap A^c$

But this means that $B_0(x, \varepsilon) \cap A^c$ is never empty, hence x is a limit point of A^c.

However, $x \in A$, contradictory to A^c being closed and thus should contain all of its limit points □

P2 An arbitrary union of open sets is open; a finite intersection of open sets is open.

Proof. Let A be an arbitrary union of open sets $\{U_i\}_{i \in I}$.

Then for any $x \in A$, suppose that $x \in U_i$, then since U_i is open we can pick $B(x, \varepsilon)$ subset of U_i subset of A

On the other hand, let U and V be open sets and let $x \in U \cap V$. Since U and V are open, we can pick ε_1 and ε_2 such that $B(x, \varepsilon_1)$ is in U whereas $B(x, \varepsilon_2)$ is in V. Then we simply pick $\varepsilon = \min\{\varepsilon_1, \varepsilon_2\}$ so that $B(x, \varepsilon)$ is in $U \cap V$. □

P3 An arbitrary intersection of closed sets is closed; a finite union of closed sets is closed.

Proof. This follows from de Morgan's Law on P1 and P2. □

Problem 39. Compare the sizes of the following pairs of sets, i.e. determine if they are equal, or if one set may be a subset of the other.

1. $(A \cup B)^\circ$, $A^\circ \cup B^\circ$
2. $(A \cap B)^\circ$, $A^\circ \cap B^\circ$
3. $\overline{A \cup B}$, $\bar{A} \cup \bar{B}$
4. $\overline{A \cap B}$, $\bar{A} \cap \bar{B}$

Proof.

1. $(A \cup B)^\circ$ may be bigger

   In $\mathbb{R}$ we consider the intervals $A = (-1, 0]$ and $B = [0, 1)$, then

   $$A^\circ \cup B^\circ = (-1, 0) \cup (0, 1), \quad (A \cup B)^\circ = (-1, 1)$$

   For $x \in A^\circ \cup B^\circ$, we have either $x \in A^\circ$ or $x \in B^\circ$, so there is some ball centered at x that is contained in either A or B and thus must be contained in $A \cup B$ as well.

2. Equal

   If $x \in (A \cap B)^\circ$, then there exists a ball U centered at x such that U is in both A and B, so x is in both A° and B°.

   On the other hand, $A^\circ \cap B^\circ$ is a subset of $A \cap B$; taking the interior of both sides, then since the intersection between two open sets is open we find that $A^\circ \cap B^\circ$ is a subset of $(A \cap B)^\circ$.

3. Equal

4. $\bar{A} \cap \bar{B}$ may be bigger

□

Problem 40. Prove that the set of exterior points, $(A^c)^\circ$ is the same as $(\bar{A})^c$.

Proof.

$$\begin{aligned} x \in (A^c)^\circ &\iff \exists \varepsilon > 0 \text{ such that } B(x,\varepsilon) \subset A^c \\ &\iff B(x,\varepsilon) \cap A = \varnothing \\ &\iff x \notin A \text{ and } B_0(x,\varepsilon) \cap A = \varnothing \\ &\iff x \notin A \cup A' = \bar{A} \\ &\iff x \in (\bar{A}^c) \end{aligned}$$

□

Problem 41. Regarding alternative descriptions:

1. A is a neighbourhood of x if and only if there exists an open set U such that x is in U, U is subset of A (trivial except you'll actually need to prove that balls are open sets).
2. If x is a limit point of A, then in fact for any $\varepsilon > 0$, $B(x,\varepsilon)$ contains infinitely many elements of A (you don't need to mention the punctured ball here because of obvious reasons; converse is trivial but a good and intuitive description).
3. x is a boundary point of A if and only if for all $\varepsilon > 0$, $B(x,\varepsilon)$ intersects with both A and A^c.

Proof.

1. We show that $B(x,\varepsilon)$ is open:
   $\forall y \in B(x,\varepsilon)$,
   $$|y - x| < \varepsilon$$
   $\forall z \in B(y, \varepsilon - |y - x|)$,
   $$|z - x| \le |z - y| + |y - x| < \varepsilon - |y - x| + |y - x| = \varepsilon$$
   $\therefore B(y, \varepsilon - |y - x|) \subset B(x,\varepsilon)$
2. We construct a sequence $\{x_n\}$ recursively as follows:
   - Pick $x_1 \in B_0(x,\varepsilon) \cap A$
   - Pick $x_{n+1} \in B_0(x, |x_n - x|) \cap A$

   It is easy to see that the balls above are getting smaller so all x_n are both mutually distinct and all contained in $B(x,\varepsilon)$.
3. x is a boundary point if and only if $x \in \bar{A} \smallsetminus A^\circ$

   Forward direction:

   We consider two cases

   - $x \in A$, then all $B(x,\varepsilon)$ intersects with A at x, but since x is not in A° they must always intersect with A^c as well.
   - $x \notin A$, then all $B(x,\varepsilon)$ intersect with A^c at x, but since $x \in \bar{A}$, x is a limit point of A and thus $B(x,\varepsilon)$ always intersects with A.

   Backward direction:

   We consider two cases

   - $x \in A$, then since $B(x,\varepsilon)$ always intersects with A^c, x cannot be in A°.
   - $x \notin A$, then since $B(x,\varepsilon)$ always intersects with A, x must be in $\bar{A}$.

   In fact we can describe the closure without referring to punctured balls and induced sets: $x \in \bar{A}$ if and only if $B(x,\varepsilon)$ always intersects with A

   Also as a side note, $A \circ \cup \partial A \cup (A^c) \circ = \mathbb{R}^n$

□

Problem 42. Regarding closures (The following properties are relatively nontrivial compared to its 'open-set' counterparts):

(a) A' is closed.

(b) $\bar{A}$ is closed, i.e. bar(barA)=barA

Proof.

(a) In order to show that A' is closed, we need to show that if x is a limit point of A', then $x \in A'$, i.e. x is a limit point of A.

So we need to show that limit points of A' are always limit points of A: Let x be a limit point of A', then for all $\varepsilon > 0$, $B_0(x, \varepsilon/2)$ intersects with A' and we may pick $y \in B_0(x, \varepsilon/2) \cap A'$

Now here's the tricky part Since $y \in A'$, y is a limit point of A, hence $B_0(y, |y - x|)$ intersects with A and thus we may pick $z \in B_0(y, |y - x|) \cap A$.

We show that $z \in B_0(x, \varepsilon)$:

$$|z - x| \le |z - y| + |y - x| < 2|y - x| < \varepsilon,$$

hence $z \in B(x, \varepsilon)$.

$$|z - y| < |x - y|,$$

hence $z \neq x$

$\therefore z \in B_0(x, \varepsilon)$

(b) As for (b), it is just (a) and problem 39 item 3.

□

For homework, you'll work out some properties regarding dense sets

1. A is a dense set in X if and only if A intersects with all open sets in X. 2. If A is dense in X and B is dense in A, then B is dense in X 3. If A and B are dense in X where A is open, then $A \cap B$ is dense in X

§9.4 Important Theorems

Theorem 9.4.1: Cantor's Intersection Theorem

Given a decreasing sequence of compact sets $A_1 \supset A_2 \supset \cdots$, there exists a point $x \in \mathbb{R}^n$ such that x belongs to all A_i. In other words, $\bigcap_{i=1}^{\infty} A_i \neq \varnothing$. Moreover, if for all $i \in \mathbb{N}$ we have $\operatorname{diam} A_{i+1} \le c \cdot \operatorname{diam} A_k$ for some constant $c < 1$, then such a point must be unique, i.e. $\bigcap_{i=1}^{\infty} A_k = \{x\}$ for some $x \in \mathbb{R}^n$.

Theorem 9.4.2: Heine–Borel Theorem

A set $A \subset \mathbb{R}^n$ is compact if and only if every open covering has a finite subcover, i.e. for any family of open sets $\mathscr{U} = \{U_i\}_{i \in I}$ satisfying $A \subset \bigcup_{i \in I} U_i$, there exists $\{U_1, \ldots, U_n\} \subset \mathscr{U}$ such that $A \subset \bigcup_{i=1}^{n} U_i$.

Theorem 9.4.3: Bolzano–Weierstrass Theorem

Infinite bounded sets in $\mathbb{R}^n$ must contain limit points.

We will follow a very specific sequence of steps to prove these three theorems:

(a) Cantor Intersection for $n = 1$

(b) Bolzano–Weierstrass for $n = 1$

(c) Bolzano–Weierstrass for general n

(d) Cantor Intersection for general n

(e) Heine–Borel for general n

Proof.

(a) Suppose that there is a decreasing sequence of compact sets $A_1, A_2, \dots$ in the real numbers

Since A_k are bounded, we may let $a_k = \inf A_k$ Also since A_k are closed, $a_k \in A_k$

Note that since A_k is a decreasing sequence of sets we have $a_1 \leq a_2 \leq \dots$

Also, whenever we have $n > k$, we have $a_n \in A_n$, but $A_n \subset A_k$ and thus $a_n \in A_k$.

Let $b_1 = \sup A_1$, then $a_k \in A_1$ and thus $a_k \leq b_1$ for all k.

This tells us that the sequence $\{a_k\}$ is bounded above, and thus we may let $a = \sup a_k$.

Our goal is to show that the number a appears in all A_k, thus showing that the entire intersection $\bigcap A_k$ contains a and thus must be non-empty.

Now we split this in two cases, which asks whether a is simply made from isolated points, or if it is actually some nontrivial point obtained from the boundaries of A_k

Case 1: $a_k = a$ for some k In this case we see that $a_k \leq a_n \leq a$ for all $n > k$ and thus $a_n = a$ in this case, therefore a is an element in A_n for all n

In this case you can imagine that there is a possibility where a is an isolated minimum point of A_n which stays there forever in the decreasing sequence of sets

Case 2: $a_k < a$ for all k; in this case we see that a is the limit point of the increasing sequence $\{a_k\}$

Exercise 1: Show that a is a limit point of each A_k.

Note that a_n is in A_k for each $n > k$, and since $a = \sup\{a_k\}$ where a_k is increasing, we can actually show that a is a limit point of $\{a_n \mid n \leq k\}$: For every $\varepsilon > 0$, we pick n_0 such that $0 < a - a_{n_0} < \varepsilon$ Pick $n\prime > \max\{k, n_0\}$, then $a_{n'} \geq a_{n_0}$ and so

$$0 < a - a_{n\prime} \leq a_{n_0} < \varepsilon$$

This shows that there exists a_n' in $B_0(a, \varepsilon) \cap \{a_n \mid n > k\}$ for all ε, and so a is a limit point of $\{a_n \mid n > k\}$.

Now since $\{a_n | n \geq k\}$ is a subset of A_k we also see that a is a limit point of A_k Finally, since A_k is closed, we conclude that a is in A_k for all k, and we are done

Wait hold on, I forgot about the second part

Now we consider a decreasing sequence of compact sets $A_1, A_2, \dots$ such that $\operatorname{diam} A_{k+1} \leq c \operatorname{diam} A_k$ for $c < 1$.

Suppose otherwise that there exists x, y in $\bigcap A_k$

You can imagine that this will form a fixed distance between two points, and thus there is a constant positive lower bound for the diameters:

$$\operatorname{diam} A_k \geq |x - y| > 0 \forall k$$

But this cannot be true because $\operatorname{diam} A_{k+1} \leq c \operatorname{diam} A_k$ and so the diameter is controlled by a decreasing geometric sequence:

$$\operatorname{diam} A_{k+1} \leq c^k \operatorname{diam} A_1$$

So we can simply pick a natural number k such that

$$k > \log_c \frac{|x - y|}{\operatorname{diam} A_1}$$

(b) We consider an infinite bounded set A in the real numbers. Since A is bounded, we can pick a closed interval $[a_1, b_1]$ containing A.

We then perform a series of binary cuts: Consider the two halves of $[a_1, b_1]$. We know that at least one of these two must contain infinitely many elements in A, otherwise A cannot be infinite. We pick this half of the interval and denote it by $[a_2, b_2]$. We continue this to pick a decreasing sequence of closed intervals $[a_n, b_n]$.

Now $\operatorname{diam}[a_{n+1}, b_{n+1}] = \frac{1}{2}\operatorname{diam}[a_n, b_n]$, so by the Cantor Intersection Theorem, there exists a unique real number c in the intersection $\bigcap[a_n, b_n]$.

We show that this c is in fact a limit point of A.

For any $\varepsilon > 0$, we need to show that $B_0(c, \varepsilon) \cap A \neq \varnothing$, i.e. we need to find an element $x \neq c$ in A that is less than ε apart from c.

We then realize that we can simply exploit the decreasing sequence $[a_n, b_n]$ Since $\operatorname{diam}[a_n, b_n]$ is controlled by a decreasing sequence:

$$\operatorname{diam}[a_{n+1}, b_{n+1}] \le 1/2^n \operatorname{diam}[a_1, b_1]$$

We take a sufficiently large n so that $b_n - a_n < \varepsilon$ Since c is in $[a_n, b_n]$, for all x in $[a_n, b_n]$ we have $|x - c| \le b_n - a_n < \varepsilon$ and therefore $[a_n, b_n]$ is within $B(c, \varepsilon)$.

Here's the funny part: $[a_n, b_n]$ contains infinitely many elements of A, so it must contain at least one element in A that is not c.

Therefore this element $x \neq c$ is in $B_0(c, \varepsilon)$.

(c) Now we have an infinte bounded set A in $\mathbb{R}^n$

The idea here is to consecutively come up with better and better sequences of points in A. We denote x_i to be the i-th coordinate in $\mathbb{R}^n$.

Our first wish is to pick some elements in A so that they sort of converge at x_1.

Because such considerations of 'restricting to a single coordinate' is important here, we define the projection map to the i-th coordinate by

$$f_i(x_1, \dots, x_n) = x_i$$

So, we look at $f_i(A)$ and try to apply BW for the case where $n = 1$.

However, the problem is that $f_i(A)$ need not be infinite. For example, the set $\{(0,0), (0,1), (0,2), \dots\}$ projected onto the first coordinate is simply $\{0\}$.

This forces us to consider two cases

Exercise 2: Show that $f_i(A)$ is bounded This is simple 1. $f_1(A)$ is infinite, then we can apply BW(n=1) to find a real number c_1 which is a limit point in $f_1(A)$

Here we can construct a sequence of points

$$\{x^{(1),1}, x^{(1),2}, ...\}$$

so that their first coordinates satisfy

$$|x_1^{(1),n} - c_1| < 1/n$$

for all natural number n (I know this notation is cumbersome but the problem is that we need multiple sequences for this proof)

2. $f_1(A)$ is finite, then by the Pigeonhole Principle there exists a real number c_1 such that its preimage $f_1^{-1}(c_1)$ in A is infinite

In this case we can randomly pick a sequence $\{x^{(1),1}, x^{(1),2}, \dots\}$ in A so that their first coordinate is equal to c_1

I forgot to mention something that is implied, but we actually do have the need to emphasize that the sequence $\{x^{(1),1}, x^{(1),2}, \dots\}$ can be chosen to contain mutually distinct entries

Now that we have a sequence that behaves nice on the first coordinate, we may then move on to the second coordinate

Let $A_1 = \{x^{(1),1}, x^{(1),2}, \dots\}$ We again consider $f_2(A_1)$ in two cases, infinite or finite

In any case, we are able to find a subsequence $\{x^{(2),1}, x^{(2),2}, \dots\}$, where $x^{(2),k} = x^{(1),n_k}$ for some strictly increasing sequence of natural numbers n_k

So that, for the limit point/point with infinite preimage c_2, this sequence satisfies

$$|f_2(x^{(2),n}) - c_2| < \frac{1}{n}$$

Note that the property we have for the second case (we in fact have $f_2(x^{(2),n}) = c_2$) is just a better version of this.

Now, take note that picking this subsequence does no harm whatsoever towards the first coordinate (if anything it would turn out to be better) since

$$|f_1(x^{(2),k}) - c_1| = |f_1(x^{(1),n_k} - c_1| < \frac{1}{n_k} \leq \frac{1}{k}$$

($n_1 < \cdots < n_k$ is a strictly increasing sequence of natural numbers so $n_k \geq k$)

This continues on until we obtain a sequence of points $\{x^{(n),1}, x^{(n),2}, \dots\}$ in A so that

$$|f_i(x^{(n),k} - c_i| < \frac{1}{k} \quad \forall i, k$$

As we can see, the point $c = (c_1, \dots, c_n)$ is in fact a limit point of A as we can always choose a big enough k so that $x^{(n),k}$ is in $B(c, \varepsilon) \cap A$.

Since $\{x^{(n),k}\}$ was always chosen to be a sequence of distinct entries, there is no danger for this sequence to always be c, and so c must be a limit point of A.

(d) We may now return to the general case of Cantor.

Suppose that there is a sequence of decreasing compact sets $A_1, A_2, \dots$ in $\mathbb{R}^n$. Note that every point is contained in A_1, so boundedness will never be an issue here.

Since A_k are all nonempty, we can simply pick any element a_k from A_k.

For the uncannily specific case that there are only finitely many $\{a_k\}$ chosen, we simply note that, again by Pigeonhole Principle, one of the a_k appears infinitely often; thus for each A_n we simply pick $n_k > n$ so that A_{n_k} contains a_k, then a_k is in A_{n_k} which is a subset of A_n.

Otherwise, we can then note that $\{a_k\}$ is an infinite bounded set of points, so there must exist a limit point a of $\{a_k\}$.

We can now see that a is always an element of A_k: Using the same technique as Exercise 1, we see that a is a limit point of $\{a_n \mid n > k\}$ and so is a limit point of A_k, therefore a is in A_k as A_k is closed.

This proves the first part of the statement The second part is completely identical to the second part of the $n = 1$ case so we don't need to waste our time there either

(e) We now consider a compact set A with some open covering $\mathscr{U}$.

This theorem is proved by contradiction: Suppose otherwise that set A cannot be covered by any finite collection of open sets in $\mathscr{U}$

Since A is compact, we may enclose it in a closed cube Q_1 (whose edges are parallel to the axes)

Now, for each step, we partition Q into 2^n cubes by cutting it in half from each direction.

Then, starting from Q_1, there must exist one of these smaller cubes, denoted by Q_2, such that $A \cap Q_2$ cannot be covered by a finite collection of open sets in $\mathscr{U}$. Otherwise, if each $A \cap Q$ has a finite cover, then we simply collect all of these open sets together to form a finite cover of A, which violates our assumption.

We continue on to partition Q_n and pick Q_{n+1} so that A_{n+1} has no finite cover (denote $A_n = A \cap Q_n$).

Note that A and Q_n are both compact, so A_n is compact Also we see that there is a decreasing sequence $A_1, A_2, \dots$ (we can't exactly obtain a relation between $\operatorname{diam} A_n$ and $\operatorname{diam} A_{n+1}$ here)

By Cantor Intersection Theorem we can always find a point x in A located in the intersection $\bigcap A_k$.

Now, since $\mathscr{U}$ is an open covering of A, there exists an open set U in $\mathscr{U}$ such that $x \in U$.

The final key step is to exploit the sequence of decreasing cubes Q_n. So even though there isn't a clear cut way to control the sizes of $\operatorname{diam} A_n$, we do in fact have the property that $\operatorname{diam} Q_{n+1} = \frac{1}{2^n} \operatorname{diam} Q_1$.

Therefore, by picking a sufficiently large n, we can obtain Q_n that is contained in U.

But this is a contradiction. This is because we've specifically chosen the sequence A_n to be sets that do not possess any finite cover $\{U_1, ..., U_n\}$ in $\mathscr{U}$. But here A_n simply would have a one-element cover $\{U\}$.

This completes our proof.

□

10 Metric Spaces

§10.1 Metric Spaces

§10.1.1 The definition of a metric space

One of the key definitions of Analysis was that of the continuity of a function. Recall that if $f : \mathbb{R} \to \mathbb{R}$ is a function, we say that f is continuous at $a \in \mathbb{R}$ if, for any $\varepsilon > 0$, we can find a $\delta > 0$ such that if $|x - a| < \delta$ then $|f(x) - f(a)| < \varepsilon$.

Stated somewhat more informally, this means that no matter how small an ε we are given, we can ensure $f(x)$ is within distance ε of $f(a)$ provided we demand x is sufficiently close to – that is, within distance δ of – the point a.

Now consider what it is about real numbers that we need in order for this definition to make sense: Really we just need, for any pair of real numbers x_1 and x_2, a measure of the distance between them. (Note that we needed this notion of distance in the above definition of continuity for both the pairs (x, a) and $(f(x), f(a))$.) Thus we should be able to talk about continuous functions $f : X \to X$ on any set X provided it is equipped with a notion of distance. Even more generally, provided we have prescribed a notion of distance on two sets X and Y, we should be able to say what it means for a function $f : X \to Y$ to be continuous. In order to make this precise, we will therefore need to give a mathematically rigorous definition of what a "notion of distance" on a set X should be. This is the concept of a metric space.

Definition 10.1.1. Let X be a set. Then a **distance function** on X is a function $d : X \times X \to \mathbb{R}$ with the following properties:

(i) (positivity) $d(x, y) > 0$ and $d(x, y) = 0$ if and only if $x = y$;

(ii) (symmetry) $d(x, y) = d(y, x)$;

(iii) (triangle inequality) if $x, y, z \in X$ then we have $d(x, z) \leq d(x, y) + d(y, z)$.

The pair (X, d) consisting of a set X together with a distance function d on it is called a **metric space**.

Notation. We usually abbreviate (X, d) as just X. Occasionally, we will be more formal, for instance when we have two metric spaces (X, d_X) and (Y, d_Y) and wish to make it clear which distance we are talking about.

§10.1.2 Some examples of metric spaces

In this section we look at some examples of metric spaces. A very basic example is that of the real numbers.

Example 10.1.1. Take $X = \mathbb{R}$ and $d(x, y) = |x - y|$.

Example 10.1.2. Take $X = \mathbb{R}^n$. Then each of the following functions define metrics on X.

$$d_1(v,w) = \sum_{i=1}^{n} |v_i - w_i|$$

$$d_2(v,w) = \left(\sum_{i=1}^{n} (v_i - w_i)^2\right)^{\frac{1}{2}}$$

$$d_\infty(v,w) = \max_{i \in \{1,2,\ldots,n\}} |v_i - w_i|$$

These are called the ℓ^1 ("ell one"), ℓ^2 (or Euclidean), and ℓ^∞-distances respectively. Of course, the Euclidean distance is the most familiar one.

The proof that each of d_1, d_2, d_∞ defines a distance is mostly very routine, with the exception of proving that d_2, the Euclidean distance, satisfies the triangle inequality. To establish this, recall that the Euclidean norm $kvk2$ of a vector $\mathbf{v} = (v1, ..., vn) \in Rni$

Example 10.1.1: Metric spaces of $\mathbb{R}^2$

(a) We can make $\mathbb{R}^2$ into a metric space by imposing the Euclidean distance function

$$d((x_1,y_1),(x_2,y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

(b) Just like with the first example, any subset $S \subseteq \mathbb{R}^2$ can be made into a metric space, such as the unit disk, unit circle, and the unit square $[0,1]^2$.

Example 10.1.2: Metric spaces of $\mathbb{R}^n$

To generalise the above examples, for positive integer n,

(a) Let $\mathbb{R}^n$ be the metric space whose points are points in n-dimensional Euclidean space, and whose metric is the Euclidean metric

$$d((a_1,\ldots,a_n),(b_1,\ldots,b_n)) = \sqrt{(a_1 - b_1)^2 + \cdots + (a_n - b_n)^2}$$

This is the n-dimensional **Euclidean space**.

(b) The open unit ball B^n is the subset of $\mathbb{R}^n$ consisting of the points $(x_1,\ldots,x_n)$ such that $x_1^2 + \cdots + x_n^2 < 1$.

Notation. We will refer to $\mathbb{R}^n$ with the Euclidean metric by just $\mathbb{R}^n$; if we wish to take the metric space for a subset $S \subseteq \mathbb{R}^n$ with the inherited metric, we will just write S.

§10.1.3 Norms

§10.1.4 New metric spaces from old ones

§10.1.5 Balls and boundedness

§10.2 Convergence

Since we can talk about the distance between two points, we can talk about what it means for a sequence of points to converge. This is the same as the typical epsilon–delta definition, with absolute values replaced by the distance function.

Definition 10.2.1: Convergence

Let $(x_n)_{n\geq 1}$ be a sequence of points in a metric space X. We say that x_n **converges** to x if the following condition holds: for all $\varepsilon > 0$, there exists an integer N (depending on ε) such that $d(x_n, x) < \varepsilon$ for each $n \geq N$. This is written as

$$\lim_{n\to\infty} x_n = x.$$

We say that a sequence converges in X if it converges to a point in X.

§10.3 Limits and Continuity

Chapter 2. Limits and continuity 13 2.1. Basic definitions and properties 13 2.2. Continuity of linear functions in normed spaces 14 2.3. Function spaces 15

From calculus, the ε–δ definition of a continuous function is

> A function $f : \mathbb{R} \to \mathbb{R}$ is continuous at a point $p \in \mathbb{R}$ if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $|x - p| < \delta \implies |f(x) - f(p)| < \varepsilon$.

For the definition in metric space, all we have do is replace the absolute values with the more general distance functions: this gives us a definition of continuity for any function $M \to N$.

Definition 10.3.1: Continuity

For metric spaces $X = (X, d_X)$ and $Y = (Y, d_Y)$, a function $f : X \to Y$ is **continuous** at a point $p \in X$ if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$d_X(x, p) < \delta \implies d_Y(f(x), f(p)) < \varepsilon.$$

Moreover, the function f is continuous if it is continuous at every point $p \in X$.

Here is an equivalent condition for sequences.

Theorem 10.3.1: Sequential continuity

A function $f : X \to Y$ of metric spaces is **continuous** at a point $p \in X$ if and only if the following property holds: if $x_1, x_2, \ldots$ is a sequence in X converging to p, then the sequence $f(x_1), f(x_2), \ldots$ in Y converges to $f(p)$.

Proposition 10.3.1 (Composition of continuous functions is continuous)**.** Let $f : X \to Y$ and $g : Y \to Z$ be continuous maps of metric spaces. Then their composition $g \circ f$ is continuous.

§10.4 Isometries, homeomorphisms and equivalence

§10.5 Open and closed sets

§10.6 Interiors, closures, limit points

§10.7 Structures

to incorporate into other sections

Let X be a metric space. All points and sets mentioned below are understood to be elements and subsets of X respectively.

- A *ball* in $\mathbb{R}^n$ is determined by its center $x \in \mathbb{R}^n$ and its radius $r > 0$, and is denoted by $B(x, r)$.

  $$B(x, r) = \{p \in X \mid d(x, p) < r\}$$

  A *punctured ball* in $\mathbb{R}^n$ is a ball excluding its center, and is denoted by $B_0(x, r)$.

- A containing x is a *neighborhood* of x if $B(x, \varepsilon) \subset A$ for some $\varepsilon > 0$.

- The *complement* of A, denoted by A^c, is the set of all points $x \in X$ such that $x \notin A$.

- A point $x \in A$ is an *interior point* of A if A is a neighbourhood of x.

  The *interior* of A, denoted by A°, is the set of all interior points in A.

  A is *open* if every point of A is an interior point of A, i.e. $A^\circ = A$.

- A point $x \in A$ is a *limit point* of A if every neighborhood of x contains a point $y \neq x$ such that $y \in A$.

  This means $B_0(x, \varepsilon) \cap A \neq \varnothing$ for all $\varepsilon > 0$.

  The *induced set* of A, denoted by A', is the set of all limit points of A.

  The *closure* of A, denoted by $\bar{A}$, is the union set $A \cup A'$. A is *closed* if all limit points of A are contained in A, i.e. $\bar{A} = A$.

- A point $x \in A$ is an *isolated point* of A if x is not a limit point of A.

- The *boundary* of a set A, denoted by ∂A, is the set difference $\bar{A} \setminus A^\circ$.

  A point x is a *boundary point* of A if $x \in \partial A$.

- A point x is an *exterior point* of A if it is an interior point of A^c.

- A is *perfect* if A is closed and if every point of A is a limit point of A.

- A is *bounded* if there is a real number M and a point $p \in X$ such that $d(x, p) < M$ for all $x \in A$.

- A is *compact* if it is a bounded closed set.

- A is *dense* in X if every point of X is a limit point of A, or a point of A (or both).

- A subset $B \subset A$ is a *dense subset* of A if $\bar{B} = A$.

- A is *nowhere dense* if its closure has no interior, i.e. $(\bar{A})^\circ = \varnothing$.

Theorem 10.7.1

Every neighborhood is an open set.

Proof. Consider a neighborhood $E = N_r(p)$, and let q be any point of E. Then there is a positive real number h such that

$$d(p, q) = r - h.$$

For all points s such that $d(q, s) < h$, we have then

$$d(p, s) \leq d(p, q) + d(q, s) < r - h + h = r,$$

so $s \in E$. Hence q is an interior point of E. □

§10.8 Open sets

to incorporate into other sections

Definition 10.8.1: Neighbourhood

For metric space X and point $p \in X$, an *r-neighborhood* of p, denoted by $N_r(p)$, is the set of all q with $d(p,q) < r$ for some radius $r > 0$.

$$N_r(p) = \{q \in X \mid d(p,q) < r\}$$

Remark. Others define a neighborhood as any set that contains one of these neighborhoods, which are instead called “the open ball of radius r about p".

Such an open ball is sometimes referred to as the open neighborhood of p of radius r.

Open balls are instances of open sets.

Definition 10.8.2: Open set

A subset $U \subset X$ is open if, for every point $x \in U$, there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subset U$.

The idea is that, in a open set, there exists a “safety margin" around every point. Given a point p, one can *move around in the set a certain distance and remain* in the sense.

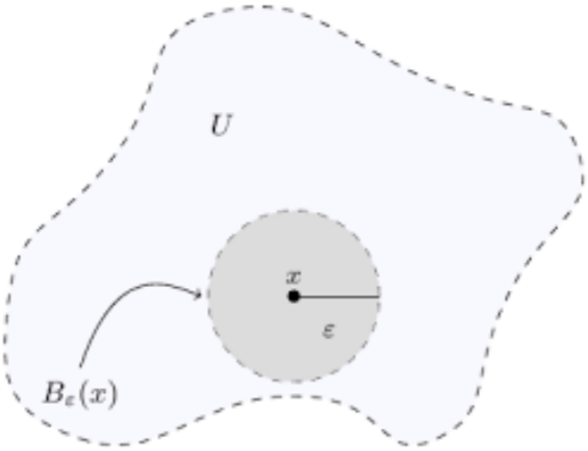


Figure 10.1: Open set

§10.9 Completeness

6.1. Basic definitions and examples 35 6.2. First properties of complete metric spaces 36 6.3. Completeness of function spaces 37 6.4. The contraction mapping theorem 38 6.5. *Completions 40

§10.10 Connectedness and Path-connectedness

7.1. Connectedness 43 7.2. *Connected subsets of R 46 7.3. Path-connectedness 47 7.4. Connectedness and path-connectedness 48

§10.11 Sequential Compactness

8.1. Definitions 51 8.2. Closure and boundedness properties 52 8.3. Continuous functions on sequentially compact spaces 53 8.4. Product spaces 54 8.5. Sequentially compact equals complete and totally bounded

55 8.6. The Arzel'a-Ascoli theorem 57

§10.12 Compactness

9.1. Open covers and the definition of compactness 61 9.2. Compactness implies sequential compactness 62 9.3. The Heine-Borel theorem 63 9.4. Sequential compactness implies compactness 63

Definition 10.12.1: Open cover

By an **open cover** of a set A in a metric space X we mean a collection $\{G_\alpha\}$ of open subsets of X such that $A \subset \bigcup_\alpha G_\alpha$.

Definition 10.12.2: Compact set

A subset K of a topological (or metric) space is compact if every open cover of K has a *finite* subcover.

An open cover of A is a collection of open sets that collectively cover A.

A subcover is a subcollection of these open sets that still collectively cover A.

This means that any infinite collection of open sets that together cover a compact set always “overcovers" it.

The simplest kind of compact set is just a finite set: a collection of finitely many points.

§10.13 Structures on Euclidean Space

to remove

Definition 10.13.1: Limit and isolated point

A point p is a limit point of E if every neighborhood of p contains a point $q \neq p \in E$.
If p is not a limit point but is in E, then p is an isolated point.

Definition 10.13.2: Closed set

E is closed if every limit point of E is in E. Intuitively, this means E “contains all its edges".
The closure $\bar{E}$ of E is the union of E and the set of its limit points.

Definition 10.13.3: Interior point

A point p is an interior point of E if there is a neighborhood N of p such that $N \subset E$. Note that interior points must be in E itself, while limit points need not be.

Definition 10.13.4: Open set

E is open if every point of E is an interior point of E. Intuitively, E “doesn’t have edges”.

Definition 10.13.5: Dense set

E is dense in X if every point of X is a limit point of E or a point of E, or both.

Definition 10.13.6: Interior

The interior E^0 of E is the set of all interior points of E, or equivalently the union of all open sets contained in E.

11 Knot Theory

Readings:

- Knot Theory by Stanford University
- The Knot Book by Colin C. Adams

§11.1 Knot and Knot Types

Part IV

Linear Algebra

12 Vectors

§12.1 Coordinate Space and the Algebra of Vectors

Definition 12.1.1. By a **vector** we will mean a list of n real numbers $x_1, x_2, \ldots, x_n$ where n is a positive integer. Mostly this list will be written as a row vector:

$$(x_1, x_2, \ldots, x_n)$$

Sometimes the numbers will be arranged as a coluumn vector:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Often we will denote such a vector by a single letter in bold, say $\mathbf{x}$, and refer to x_i as the i-th coordinate of $\mathbf{x}$.

Definition 12.1.2. For a given n, we denote the set of all vectors with n coordinates as $\mathbb{R}^n$, and often refer to $\mathbb{R}^n$ as n-dimensional coordinate space or simply as n-dimensional space. If $n = 2$ then we commonly use x and y as coordinates and refer to $\mathbb{R}^2 = \{(x, y) \mid x, y \in \mathbb{R}\}$ as the xy-plane. If $n = 3$ then we commonly use x, y and z as coordinates and refer to $\mathbb{R}^3 = \{(x, y, z) \mid x, y, z \in \mathbb{R}\}$ as xyz-space.

Definition 12.1.3. There is a special vector $(0, 0, \ldots, 0)$ in $\mathbb{R}^n$ which we denote as $\mathbf{0}$ and refer to as the **zero vector**.

A vector is an object that has both magnitude and direction. In simple terms, especially when we are thinking of $\mathbb{R}^2$ or $\mathbb{R}^3$, a vector is an arrow.

Definition 12.1.4. The points $(0, 0, \ldots, 0, x_i, 0, \ldots, 0)$ in $\mathbb{R}^n$, where x_i is a real number, comprise the x_i-axis, with the origin lying at the intersection of all the axes.

Given two vectors $\mathbf{u} = (u_1, u_2, \ldots, u_n)$ and $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ in $\mathbb{R}^n$, we can add and subtract them much as you would expect, by separately adding the corresponding coordinates; that is,

$$\mathbf{u} + \mathbf{v} = (u_1 + v_1, u_2 + v_2, \ldots, u_n + v_n); \quad \mathbf{u} - \mathbf{v} = (u_1 - v_1, u_2 - v_2, \ldots, u_n - v_n).$$

Remark. Note that two vectors may be added if and only if they have the same number of coordinates. No immediate sense can be made of adding a vector in $\mathbb{R}^2$ to one from $\mathbb{R}^3$, for example.

Given a vector $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ and a real number k then the scalar multiple $k\mathbf{v}$ is defined as

$$k\mathbf{v} = (kv_1, kv_2, \ldots, kv_n).$$

Definition 12.1.5. The n vectors

$$(1, 0, \ldots, 0), \quad (0, 1, 0, \ldots, 0), \quad \ldots, \quad (0, \ldots, 0, 1, 0), \quad (0, \ldots, 0, 1)$$

in $\mathbb{R}^n$ are known as the **standard** (or canonical) basis for $\mathbb{R}^n$. We will denote these, respectively, as $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$.

Notation. When $n = 2$, the vectors $(1,0)$ and $(0,1)$ form the standard basis for $\mathbb{R}^2$. These are also commonly denoted by the symbols $\mathbf{i}$ and $\mathbf{j}$ respectively. Note that any vector $\mathbf{v} = (x,y)$ can be written uniquely as a linear combination of $\mathbf{i}$ and $\mathbf{j}$: that is $(x,y) = x\mathbf{i}+y\mathbf{j}$ and this is the only way to write (x,y) as a sum of scalar multiples of $\mathbf{i}$ and $\mathbf{j}$.

When $n = 3$, the vectors $(1,0,0)$, $(0,1,0)$, $(0,0,1)$ form the standard basis for $\mathbb{R}^3$ being respectively denoted $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$.

§12.2 The Geometry of Vectors

Definition 12.2.1. The **length** (or **magnitude**) of a vector $\mathbf{v} = (v_1, v_2, \ldots, v_n)$, which is written $|\mathbf{v}|$, is defined by

$$|\mathbf{v}| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}.$$

We say a vector $\mathbf{v}$ is a **unit vector** if it has length 1.

This formula formalises our intuitive idea of a vector as an arrow having a length; the length of the arrow is exactly what you'd expect it to be from Pythagoras' Theorem. We see this is the distance of the point $\mathbf{v}$ from the origin, or equivalently the distance a point moves when it is translated by $\mathbf{v}$. So if $\mathbf{p}$ and $\mathbf{q}$ are points in $\mathbb{R}^n$, then the vector that will translate $\mathbf{p}$ to $\mathbf{q}$ is $\mathbf{q} - \mathbf{p}$, and hence we define:

Definition 12.2.2. The distance between two points $\mathbf{p}$ and $\mathbf{q}$ in Rn is $|\mathbf{q} - \mathbf{p}|$ (or equally $|\mathbf{p} - \mathbf{q}|$). In terms of their coordinates p_i and q_i we have

$$\text{distance between } \mathbf{p} \text{ and } \mathbf{q} = \sqrt{\sum_{i=0}^{n} (p_i - q_i)^2}.$$

Note that $|\mathbf{v}| > 0$ and that $|\mathbf{v}| = 0$ if and only if $\mathbf{v} = \mathbf{0}$.

Also $|\lambda\mathbf{v}| = |\lambda||\mathbf{v}|$ for any real number λ.

Proposition 12.2.1 (Triangle Inequality)**.** Let $\mathbf{u}$ and $\mathbf{v}$ be vectors in $\mathbb{R}^n$. Then

$$|\mathbf{u} + \mathbf{v}| \le |\mathbf{u}| + |\mathbf{v}|. \tag{12.1}$$

If $\mathbf{v} \ne \mathbf{0}$ then equality holds if and only if $\mathbf{u} = \lambda\mathbf{v}$ for some $\lambda \ge 0$.

Geometrically, this is intuitively obvious.

Proof. Let $\mathbf{u} = (u_1, u_2, \ldots, u_n)$ and $\mathbf{v} = (v_1, v_2, \ldots, v_n)$. The inequality is trivial if $\mathbf{v} = 0$, so suppose $\mathbf{v} \ne \mathbf{0}$. Note that for any real number t,

$$0 \le |\mathbf{u} + t\mathbf{v}|^2 = \sum_{i=1}^{n} (u_i + tv_i)^2 = |\mathbf{u}|^2 + 2t\sum_{i=1}^{n} u_i v_i + t^2|\mathbf{v}|^2.$$

As $|\mathbf{v}| \ne 0$, the RHS of the above inequality is a quadratic in t which is always non-negative, and thus has non-positive discriminant ($b^2 \le 4ac$). Hence

$$4\left(\sum_{i=0}^{n} u_i v_i\right)^2 \le 4|\mathbf{u}|^2|\mathbf{v}|^2 \quad \text{giving} \quad \left|\sum_{i=1}^{n} u_i v_i\right| \le |\mathbf{u}||\mathbf{v}|.$$

Finally

□

13 Linear Systems and Matrices

§13.1 Systems of linear equations

Definition 13.1.1. By a **linear system**, or **linear system of equations**, we will mean a set of m simultaneous equations in n real variables $x_1, x_2, \dots, x_n$ which are of the form

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n & = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n & = b_2 \\ & \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n & = b_m \end{cases} \tag{13.1}$$

where a_{ij} and b_i are real constants.

Any vector $(x_1, x_2, \dots, x_n)$ which satisfies eq. (13.1) is said to be a **solution**; if the linear system has one or more solutions then it is said to be **consistent**. The **general solution** to the system is any description of all the solutions of the system. We will see later that such linear systems can have zero, one or infinitely many solutions.

We will often write the linear system eq. (13.1) as the **augmented matrix** $(A \mid \mathbf{b})$ where

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}.$$

For now, we won't consider a matrix (such as A) or vector (such as $\mathbf{b}$) to be anything more than an array of numbers.

To solve systems of linear equations efficiently, we introduce such a process called **row-reduction**. It relies on three types of operation, called elementary row operations (EROs), which importantly do not affect the set of solutions of a linear system as we apply them.

Definition 13.1.2. Given a linear system of equations, an **elementary row operation** (ERO) is an operation of one of the following three kinds.

(a) Swap two equations.

(b) Multiply an equation by a non-zero constant.

(c) Add a multiple of one equation to another equation.

Notation.

(a) Let S_{ij} denote the ERO which swaps rows i and j (or equivalently the i-th and j-th equations).

(b) Let $M_i(\lambda)$ denote the ERO which multiplies row i by $\lambda \neq 0$ (or equivalently both sides of the ith equation).

(c) For $i \neq j$, let $A_{ij}(\lambda)$ denote the ERO which adds λ times row i to row j (or does the same to the equations).

Note this is not standard notation in any way, but I have introduced it here for convenience.

§13.2 Matrices and matrix algebra

At its simplest, a matrix is just a two-dimensional array of numbers.

Definition 13.2.1. Let m and n be positive integers. An $m \times n$ **matrix** is an array of real numbers arranged into m rows and n columns.

The numbers in a matrix are its **entries**. Given an $m \times n$ matrix A, we will write a_{ij} for the entry in the i-th row and j-th column. Note that i can vary between 1 and m, and that j can vary between 1 and n. So

$$i\text{-th row} = (a_{i1}, \dots, a_{in}) \quad \text{and} \quad j\text{-th column} = \begin{pmatrix} a_{ij} \\ \vdots \\ a_{mj} \end{pmatrix}$$

Notation. We shall denote the set of real $m \times n$ matrices as M_{mn}. Note that $M_{1n} = \mathbb{R}^n$ and that $M_{n1} = \mathbb{R}^n_{\text{col}}$.

There are three important operations that can be performed with matrices: matrix addition, scalar multiplication and matrix multiplication. As with vectors, not all pairs of matrices can be meaningfully added or multiplied.

Addition: Let $A = (a_{ij})$ be an $m \times n$ matrix (recall: m rows and n columns) and $B = (b_{ij})$ be a $p \times q$ matrix. As with vectors, matrices are added by adding their corresponding entries. So, as with vectors, to add two matrices they have to be the same size – that is, to add A and B, we must have $m = p$ and $n = q$. If we write $C = A + B = (c_{ij})$ then $c_{ij} = a_{ij} + b_{ij}$ for $1 \leq i \leq m$ and $1 \leq j \leq n$.

In general, matrix addition is commutative as for matrices M and N of the same size we have

$$M + N = N + M.$$

Addition of matrices is also associative as

$$L + (M + N) = (L + M) + N$$

for any matrices of the same size.

Definition 13.2.2. The $m \times n$ **zero matrix** is the matrix with m rows and n columns whose every entry is 0. This matrix is simply denoted as 0 unless we need to specify its size, in which case it is written 0_{mn}.

A simple check shows that $A + 0_{mn} = A = 0_{mn} + A$ for any $m \times n$ matrix A.

Scalar multiplication: Let $A = (a_{ij})$ be an $m \times n$ matrix and k be a real number (a scalar). Then the matrix kA is defined to be the $m \times n$ matrix with (i, j)-th entry equal to ka_{ij}.

More generally the following identities hold. Let A, B, C be $m \times n$ matrices and λ, μ be real numbers.

- $A + 0_{mn} = A$
- $A + B = B + A$
- $0A = 0_{mn}$
- $A + (-A) = 0_{mn}$
- $(A + B) + C = A + (B + C)$
- $1A = A$

- $(\lambda + \mu)A = \lambda A + \mu A$
- $\lambda(A + B) = \lambda A + \lambda B$
- $\lambda(\mu A) = (\lambda\mu)A$

These are readily verified and show that M_{mn} is a real vector space.

Based on how we added matrices then you might think that we multiply matrices in a similar fashion, namely multiplying corresponding entries, but we do not. At first glance the rule for multiplying matrices is going to seem rather odd but, in due course, we will see why matrix multiplication is done as follows and that this is natural in the context of matrices representing linear maps.

Matrix multiplication: We can multiply an $m \times n$ matrix $A = (a_{ij})$ with an $p \times q$ matrix $B = (b_{ij})$ if $n = p$. That is, A must have as many columns as B has rows. If this is the case then the product $C = AB$ is the $m \times q$ matrix with entries

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} \tag{13.2}$$

for $1 \leq i \leq m$ and $1 \leq j \leq q$.

It may help to write the rows of A as $\mathbf{r}_1, \ldots, \mathbf{r}_m$ and the columns of B as $\mathbf{c}_1, \ldots, \mathbf{c}_q$. Then the above equation is equivalent to

$$\text{the } (i,j)\text{-th entry of } AB = \mathbf{r}_i \cdot \mathbf{c}_j$$

for $1 \leq i \leq m$ and $1 \leq j \leq q$.

Definition 13.2.3. The $n \times n$ **identity matrix** I_n is the $n \times n$ matrix with entries

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}.$$

The identity matrix will be simply denoted as I unless we need to specify its size. The (i,j)-th entry of I is denoted as δ_{ij} which is referred to as the **Kronecker delta**.

remove below

Matrices: linear transformations, kernels and images; inner products, inner product spaces, orthonormal sets, and the Gram-Schmidt process; eigenvectors and eigenvalues; matrix diagonalisation and its applications; symmetric and Hermitian matrices; quandratic forms and bilinear forms; Jordan normal form and other canonical forms.

• determinant of a square matrix and inverse of a non-singular matrix (2 × 2 and 3 × 3 matrices only)
• use of matrices to solve a set of linear equations (including row reduction and echelon forms, and geometrical interpretation of the solution)

Here are some special matrices:

- **Square matrix** of order n is a matrix with n rows and n columns, i.e. # rows = # columns

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix}$$

- **Diagonal matrix**

$$\mathbf{A} = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{mm} \end{bmatrix}$$

- **Symmetric matrix**

$$\mathbf{A} = \mathbf{A}^T$$

- **Row matrix**: matrix with only one row (sometimes used to represent a vector)

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \end{bmatrix}$$

- **Column matrix**: matrix with only one column (sometimes used to represent a vector)

$$\mathbf{A} = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix}$$

Conjugate matrix

§13.2.1 Identity Matrix, Determinant and Inverse of a Matrix

Identity Matrix

The identity matrix has the property that when multiplied with another matrix it leaves the other matrix unchanged:

$$\mathbf{AI} = \mathbf{A} = \mathbf{IA} \tag{13.3}$$

Transpose of Matrix

The **transpose** of an $m \times n$ matrix $\mathbf{A}$ is the $n \times m$ matrix $\mathbf{A}^T$ formed by turning rows into columns and vice versa:

$$(\mathbf{A}^T)_{i,j} = \mathbf{A}_{j,i}$$

For example:

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & -6 & 7 \end{bmatrix}^T = \begin{bmatrix} 1 & 0 \\ 2 & -6 \\ 3 & 7 \end{bmatrix}$$

Determinant of Matrix

The **determinant** of a 2×2 matrix $\mathbf{A}$, denoted by $|\mathbf{A}|$ or $\det \mathbf{A}$, is

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

and the determinant of a 3×3 matrix is

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} + a_{12} \begin{vmatrix} a_{23} & a_{21} \\ a_{33} & a_{31} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

More generally, for a $n \times n$ matrix $\mathbf{A}$, the formal method is as follows. First we will require some definitions:

- The (i,j)-**minor** of $\mathbf{A}$ is the determinant of the submatrix obtained by deleting the i-th row and j-th column of $\mathbf{A}$. We denote this submatrix as $M_{ij}(\mathbf{A})$.
- The (i,j)-**cofactor** of $\mathbf{A}$ is the matrix $C_{ij}(\mathbf{A}) = (-1)^{i+j} M_{ij}(\mathbf{A})$.

Now, in order to calculate the determinant of an $n \times n$ matrix $\mathbf{A}$, we calculate

$$|\mathbf{A}| = \sum_{i=1}^{n} a_{1n} C_{1n}(\mathbf{A}) = a_{11} C_{11}(\mathbf{A}) + a_{12} C_{12}(\mathbf{A}) + a_{13} C_{13}(\mathbf{A}) + \dots + a_{1n} C_{1n}(\mathbf{A}) \tag{13.4}$$

A matrix whose determinant is zero, i.e. $|\mathbf{A}| = 0$, is said to be **singular**; a matrix whose determinant is non-zero, i.e. $|\mathbf{A}| \neq 0$, is said to be **non-singular**.

Inverse of Matrix

Only non-singular matrices have an inverse matrix. The **inverse** of a matrix $\mathbf{A}$ is denoted $\mathbf{A}^{-1}$ and has the following property:

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \tag{13.5}$$

To find the inverse of a 2 × 2 matrix $\mathbf{A}$ given by

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

we have the following formula:

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \tag{13.6}$$

where $|\mathbf{A}| = ad - bc \neq 0$.

Remark. There exists more complicated methods for finding the inverses of 3×3 matrices and square matrices of larger size, which we will not discuss here.

§13.2.2 Rank of Matrix

§13.2.3 Orthogonal Matrix

A square matrix A is called **orthogonal** if

$$\mathbf{A}\mathbf{A}^T = \mathbf{I} \text{ and } \mathbf{A}^T\mathbf{A} = \mathbf{I}.$$

Show that if $\mathbf{A}$ and $\mathbf{B}$ are orthogonal matrices, then $\mathbf{AB}$ is an orthogonal matrix.

§13.2.4 System of Linear Equations

Example 13.2.1

Solve the following linear system by performing elementary row operations:

$$\begin{aligned} x - 3y &= 2 \\ -x + y + 5z &= 2 \\ 2x - 5y + z &= 0 \end{aligned}$$

Solution. The augmented matrix of the linear system is

$$\begin{bmatrix} 1 & -3 & 0 & 2 \\ -1 & 1 & 5 & 2 \\ 2 & -5 & 1 & 0 \end{bmatrix}$$

Hence

$$\begin{pmatrix} 1 & -3 & 0 & 2 \\ -1 & 1 & 5 & 2 \\ 2 & -5 & 1 & 0 \end{pmatrix} \xrightarrow{R2+R1} \begin{pmatrix} 1 & -3 & 0 & 2 \\ 0 & -2 & 5 & 4 \\ 2 & -5 & 1 & 0 \end{pmatrix} \xrightarrow{R3+(-2)R1} \begin{pmatrix} 1 & -3 & 0 & 2 \\ 0 & -2 & 5 & 4 \\ 0 & 1 & 1 & -4 \end{pmatrix} \xrightarrow{R2\leftrightarrow R3} \begin{pmatrix} 1 & -3 & 0 & 2 \\ 0 & 1 & 1 & -4 \\ 0 & -2 & 5 & 4 \end{pmatrix}$$

$$\xrightarrow{R3+2R2} \begin{pmatrix} 1 & -3 & 0 & 2 \\ 0 & 1 & 1 & -4 \\ 0 & 0 & 7 & -4 \end{pmatrix} \xrightarrow{\frac{R3}{7}} \begin{pmatrix} 1 & -3 & 0 & 2 \\ 0 & 1 & 1 & -4 \\ 0 & 0 & 1 & -\frac{4}{7} \end{pmatrix}$$

By backward substitution, we obtain the solution of the linear system:

$$x = -\frac{58}{7}, \quad y = -\frac{24}{7}, \quad z = -\frac{4}{7}.$$

□

Consider the following two linear systems:

$$\begin{aligned} x + 2y - z + 5w &= -1 \\ y + 3z - w &= 2 \\ z + 2w &= 3 \\ w &= 1 \end{aligned} \tag{1}$$

and

$$\begin{aligned} x &= 3 \\ y &= 1 \\ z &= 2 \\ w &= 5 \end{aligned} \tag{2}$$

The solution to (1) can be obtained by backward substitution, while the solution to (2) is immediate. The augmented matrices of the linear systems (1) and (2) are respectively

$$\begin{bmatrix} 1 & 2 & -1 & 5 & -1 \\ 0 & 1 & 3 & -1 & 2 \\ 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 0 & 0 & 0 & 3 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 1 & 5 \end{bmatrix}$$

The first matrix is an example of a matrix in **row-echelon form**, while the second matrix is an example of a matrix in **reduced row-echelon form**.

Definition 13.2.1: Row-echelon form

A matrix is said to be in **row-echelon form** if it satisfies all the following properties:

1. If there are any rows that consist entirely of zeros, then they are grouped together at the bottom of the matrix.
2. If a row does not consist of entirely of zeros, then the first nonzero number in the row is a 1. We call this a leading 1.
3. In any two successive rows that do not consists entirely of zeros, the leading 1 in the lower row occurs further to the right than the leading 1 in the higher row.

The matrix is said to be in **reduced row-echelon form** if, in addition to the above three properties, the following property is satisfied:

4. Each column that contains a leading 1 has zeros everywhere else in that column.

Example 13.2.2: Linear system with a unique solution

The augmented matrix of a linear system in (x, y, z) has been reduced to the given row-echelon form:

$$\begin{bmatrix} 1 & 2 & -1 & 2 \\ 0 & 1 & 3 & -1 \\ 0 & 0 & 1 & 4 \end{bmatrix}$$

Solve the linear system.

Solution. The corresponding linear system is

$$\begin{aligned} x + 2y - z &= 2 \\ y + 3z &= -1 \\ z &= 4 \end{aligned}$$

By backward substitution, we obtain the solution $x = 32$, $y = -13$ and $z = 4$. □

Example 13.2.3: Linear system with infinitely many solutions

Write down all the solutions of

$$x + 2y - z = 3.$$

Solution. Let $y = s$ and $z = t$, then $x = 3 - 2s + t$.

Thus all the solutions are $x = 3 - 2s + t$, $y = s$ and $z = t$, where $s, t \in \mathbb{R}$. □

Remark. Note that s and t are called **parameters**, and the set of all solutions expressed in terms of the parameters is called the **general solution** of the linear system.

Example 13.2.4

The augmented matrix of a linear system in (x, y, z, w) has been reduced to the reduced-row echelon form:

$$\begin{bmatrix} 1 & 0 & 0 & 2 & -7 \\ 0 & 1 & 0 & 1 & 5 \\ 0 & 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Solve the linear system.

Solution. The corresponding linear system is

$$\begin{aligned} x + 2w &= -7 \\ y + w &= 5 \\ z + 3w &= 1 \end{aligned}$$

The variables (unknowns) that corresponding to the leading 1's, namely x, y and z, are called **leading variables**. The non-leading variables (w in this case) are called **free variables**.

Solving for leading variables in terms of variables, we can assign any arbitrary value to the free variable w, say t, which then determines the values of the leading variable. Thus this linear system has *infinitely many solutions* given by

$$x = -7 - 2t, \quad y = 5 - t, \quad z = 1 - 3t, \quad w = t \quad \text{where } t \in \mathbb{R}$$

□

Definition 13.2.2: Gaussian elimination

The method of solving a linear system by reducing the corresponding augmented matrix to row-echelon form (respectively reduced row-echelon form) is unknown as **Gaussian elimination** (respectively **Gauss-Jordan elimination**).

Example 13.2.5

Without using a calculator, solve the linear system

$$\begin{aligned} 3x + 4y - 2z + 13w &= 9 \\ x + 2y - 2z + 7w &= 5 \\ 2x + y + 4z + 6w &= -3 \end{aligned}$$

Solution. We write down the augmented matrix of the linear system and then perform elementary row operations to reduce it to row-echelon form or reduced row-echelon form:

$$\begin{pmatrix} 3 & 4 & -2 & 13 & 9 \\ 1 & 2 & -2 & 7 & 5 \\ 2 & 1 & 4 & 6 & -3 \end{pmatrix} \xrightarrow{R1\leftrightarrow R2} \begin{pmatrix} 1 & 2 & -2 & 7 & 5 \\ 3 & 4 & -2 & 13 & 9 \\ 2 & 1 & 4 & 6 & -3 \end{pmatrix} \xrightarrow[R3-R1\times 2]{R2-R1\times 3} \begin{pmatrix} 1 & 2 & -2 & 7 & 5 \\ 0 & -2 & 4 & -8 & -6 \\ 0 & -3 & 8 & -8 & -13 \end{pmatrix}$$

$$\xrightarrow{R2\times\left(-\frac{1}{2}\right)} \begin{pmatrix} 1 & 2 & -2 & 7 & 5 \\ 0 & 1 & -2 & 4 & 3 \\ 0 & -3 & 8 & -8 & -13 \end{pmatrix} \xrightarrow{R3+R2\times 3} \begin{pmatrix} 1 & 2 & -2 & 7 & 5 \\ 0 & 1 & -2 & 4 & 3 \\ 0 & 0 & 2 & 4 & -4 \end{pmatrix} \xrightarrow{R3\times\frac{1}{2}} \begin{pmatrix} 1 & 2 & -2 & 7 & 5 \\ 0 & 1 & -2 & 4 & 3 \\ 0 & 0 & 1 & 2 & -2 \end{pmatrix}$$

The linear system corresponding to the row-echelon form is

$$\begin{aligned} x + 2y - 2z + 7w &= 5 \\ y - 2z + 4w &= 3 \\ z + 2w &= -2 \end{aligned}$$

which has the same set of solutions as the given linear system. Now x, y and z are the leading variables, and w is the free variable. Let $w = t$ where $t \in \mathbb{R}$ is an arbitrary number. By backward substitution, $z = -2 - 2t, y = -1 - 8t, x = 3 + 5t$. Thus the general solution of the given linear system is

$$x = 3 + 5t, \quad y = -1 - 8t, \quad z = -2 - 2t, \quad w = t \quad \text{where } t \in \mathbb{R}$$

Alternatively, we can further reduce the row-echelon form to reduced row-echelon form, then assign $w = t$ to obtain the same general solution. □

Example 13.2.6: Geometrical interpretation

The general solution of the system of linear equations

$$\begin{aligned} x + y &= -1 \\ 2x + y + z &= 3 \\ x + z &= 4 \end{aligned}$$

is given by $x = 4 - t, y = -5 + t, z = t$. What is the geometrical interpretation of the solution?

Solution. The three planes $x + y = -1$, $2x + y + z = 3$ and $x + z = 4$ intersect in a common line, with vector equation

$$r = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 4 - t \\ -5 + t \\ t \end{pmatrix} = \begin{pmatrix} 4 \\ -5 \\ 0 \end{pmatrix} + t \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}, \quad t \in \mathbb{R}$$

□

Homogenous Linear Systems

Definition 13.2.3: Homogenous linear system

A linear system of the form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= 0 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= 0 \end{aligned}$$

is known as a **homogeneous linear system**.

Every homogeneous linear system is consistent, since $x_1 = x_2 = \cdots = x_n = 0$ is a solution; this solution is called the **trivial solution**; if there are other solutions, then they are called **non-trivial solutions**, i.e. a solution $x_1 = s_1, x_2 = s_2, \ldots, x_n = s_n$ is a non-trivial solution if *at least one* of $s_i \neq 0$.

Theorem 13.2.1

Every homogeneous system of linear equations with more unknowns than equations has infinity many solutions.

Example 13.2.7

Determine whether the homogeneous linear system has non-trivial solution.

$$\begin{aligned} x + y + 3z &= 0 \\ -x + 2y + 6z &= 0 \\ 2x - y - 3z &= 0 \end{aligned}$$

Solution. The augmented matrix is

$$\begin{bmatrix} 1 & 1 & 3 & 0 \\ -1 & 2 & 6 & 0 \\ 2 & -1 & -3 & 0 \end{bmatrix}$$

Performing elementary row operations on the augmented matrix gives us:

$$\begin{bmatrix} 1 & 1 & 3 & 0 \\ 0 & 3 & 9 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The corresponding homogeneous system

$$\begin{aligned} x + y + 3z &= 0 \\ 3y + 9z &= 0 \end{aligned}$$

has 3 unknowns and 2 equations.

Hence the homogeneous linear system has non-trivial solution. Since it is equivalent to the given homogeneous system, it also has non-trivial solution. □

§13.2.5 Linear Transformations

• linear spaces and subspaces, and the axioms (restricted to spaces of finite dimension over the field of real numbers only) • linear independence and span • basis and dimension (in simple cases), including use of terms such as 'column space', 'row space', 'range space' and 'null space' • rank of a square matrix and relation between rank, dimension of null space and order of the matrix • linear transformations and matrices from $\mathbb{R}^n$ to $\mathbb{R}^m$ • eigenvalues and eigenvectors of square matrices (2 × 2 and 3 × 3 matrices, restricted to cases where the eigenvalues are real and distinct) • diagonalisation of a square matrix M by expressing the matrix in the form QDQ–1, where D is a diagonal matrix of eigenvalues and Q is a matrix whose columns are eigenvectors, and use of this expression such as to find the powers of M

§13.2.6 Eigenvalues and Eigenvectors

Bases: Spans and Spanning Sets, Linear Independence

Dimension

Linear Transformations

Linear Maps and Matrices

Inner Product Spaces

14 Vectors

§14.1 Basic Properties

§14.1.1 Coordinate Space and the Algebra of Vectors

Definition 14.1.1: Vector

By a *vector* we will mean a list of n real numbers $x_1, x_2, x_3, \ldots, x_n$ where n is a positive integer. Mostly this list will be treated as a row vector and written as

$$(x_1, x_2, \ldots, x_n).$$

Sometimes (for reasons that will become apparent) the numbers will be arranged as a column vector

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

Often we will denote such a vector by a single letter in bold, say $\mathbf{x}$, and refer to x_i as the i-th coordinate of $\mathbf{x}$.

Definition 14.1.2: Coordinate space

For a given n, we denote the set of all vectors with n coordinates as $\mathbb{R}^n$, and often refer to $\mathbb{R}^n$ as *n-dimensional coordinate space* or simply as *n-dimensional space.*

If $n = 2$ then we commonly use x and y as coordinates and refer to $\mathbb{R}^2 = \{(x, y) \mid x, y \in \mathbb{R}\}$ as the xy-plane.

If $n = 3$ then we commonly use x, y and z as coordinates and refer to $\mathbb{R}^3 = \{(x, y, z) \mid x, y, z \in \mathbb{R}\}$ as xyz-space.

Remark. Note that the order of the coordinates matters; so, for example, $(2, 3)$ and $(3, 2)$ are different vectors in $\mathbb{R}^2$.

There is a special vector $(0, 0, \ldots, 0)$ in $\mathbb{R}^n$ which we denote as $\mathbf{0}$ and refer to as the *zero vector.*

A vector is an object that has both *magnitude* and *direction.* In simple terms, especially when we are thinking of $\mathbb{R}^2$ or $\mathbb{R}^3$, a vector is an arrow. A vector can be used in different ways. Consider the case of vectors in $\mathbb{R}^3$, the xyz-space: we can use a vector to represent a point that has coordinates x, y and z. We call this vector the *position vector* of that point.

The points $(0, 0, \ldots, 0, x_i, 0, \ldots, 0)$ in $\mathbb{R}^n$, where x_i is a real number, comprise the x_i-axis, with the origin lying at the intersection of all the axes.

Similarly in three (and likewise higher) dimensions, the triple (x, y, z) can be thought of as the point in $\mathbb{R}^3$ which is x units along the x-axis from the origin, y units parallel to the y-axis and z units parallel to the z-axis, or it can represent the translation which would take the origin to that point.

Definition 14.1.3: Vector addition

Given two vectors $\mathbf{u} = (u_1, u_2, \ldots, u_n)$ and $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ in $\mathbb{R}^n$, we can add and subtract them much as you would expect, by separately adding the corresponding coordinates. That is

$$\mathbf{u} + \mathbf{v} = (u_1 + v1, u_2 + v2, \ldots, u_n + v_n); \quad \mathbf{u} - \mathbf{v} = (u_1 - v_1, u_2 - v_2, \ldots, u_n - v_n).$$

Geometrically, the vector $\mathbf{u} + \mathbf{v}$ is constructed by moving the start of the $\mathbf{v}$ arrow to the end of the $\mathbf{u}$ arrow: $\mathbf{u} + \mathbf{v}$ is then the arrow from the start of $\mathbf{u}$ to the end of $\mathbf{v}$.

Definition 14.1.4: Scalar multiple

Given a vector $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ and a real number k then the scalar multiple $k\mathbf{v}$ is defined as

$$k\mathbf{v} = (kv_1, kv_2, \ldots, kv_n).$$

We write $-\mathbf{v}$ for $(-1)\mathbf{v} = (-v_1, -v_2, \ldots, -v_n)$.

Definition 14.1.5: Standard basis

The n vectors

$$(1, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1, 0), (0, \ldots, 0, 1)$$

in $\mathbb{R}^n$ are known as the *standard (or canonical) basis* for $\mathbb{R}^n$. We will denote these, respectively, as $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$.

When $n = 2$, the vectors $(1, 0)$ and $(0, 1)$ form the standard basis for $\mathbb{R}^2$. These are also commonly denoted by the symbols $\mathbf{i}$ and $\mathbf{j}$ respectively. Note that any vector $\mathbf{v} = (x, y)$ can be written uniquely as a linear combination of $\mathbf{i}$ and $\mathbf{j}$: that is $(x, y) = x\mathbf{i} + y\mathbf{j}$ and this is the only way to write (x, y) as a sum of scalar multiples of $\mathbf{i}$ and $\mathbf{j}$.

When $n = 3$, the vectors $(1, 0, 0), (0, 1, 0), (0, 0, 1)$ form the standard basis for $\mathbb{R}^3$ being respectively denoted $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$.

§14.1.2 Geometry of Vectors. Some Geometric Theory

Definition 14.1.6: Magnitude

The *length (or magnitude)* of a vector $\mathbf{v} = (v_1, v_2, \ldots, v_n)$, denoted by $|\mathbf{v}|$, is defined by

$$|\mathbf{v}| = \sqrt{{v_1}^2 + {v_2}^2 + \cdots + {v_n}^2}.$$

We say a vector $\mathbf{v}$ is a *unit vector* if it has length 1.

This formula formalises our intuitive idea of a vector as an arrow having a length; the length of the arrow is exactly what you would expect it to be from Pythagoras' Theorem. We see this is the distance of the point $\mathbf{v}$ from the origin, or equivalently the distance a point moves when it is translated by $\mathbf{v}$.

So if $\mathbf{p}$ and $\mathbf{q}$ are points in $\mathbb{R}^n$, then the vector that will translate $\mathbf{p}$ to $\mathbf{q}$ is $\mathbf{q} - \mathbf{p}$, and hence we define:

Definition 14.1.7: Distance

The distance between two points $\mathbf{p}$ and $\mathbf{q}$ in $\mathbb{R}^n$ is $|\mathbf{q} - \mathbf{p}|$ (or equally $|\mathbf{p} - \mathbf{q}|$). In terms of their coordinates p_i and q_i we have

$$|\mathbf{q} - \mathbf{p}| = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Remark. Note that $|\mathbf{v}| \geq 0$ and that $|\mathbf{v}| = 0$ if and only if $\mathbf{v} = \mathbf{0}$. Also $|\lambda\mathbf{v}| = |\lambda|\,|\mathbf{v}|$ for any real number λ.

Theorem 14.1.1: Triangle Inequality

Let $\mathbf{u}$ and $\mathbf{v}$ be vectors in $\mathbb{R}^n$. Then

$$|\mathbf{u}+\mathbf{v}| \le |\mathbf{u}| + |\mathbf{v}| \tag{14.1}$$

If $\mathbf{v} \neq \mathbf{0}$ then there is equality in eq. (14.1) if and only if $\mathbf{u} = \lambda\mathbf{v}$ for some $\lambda > 0$.

Proof. Let $\mathbf{u} = (u_1, u_2, \dots, u_n)$ and $\mathbf{v} = (v_1, v_2, \dots, v_n)$. The inequality eq. (14.1) is trivial if $\mathbf{v} = \mathbf{0}$, so suppose $\mathbf{v} \neq \mathbf{0}$. Note that for any real number t,

$$0 \le |\mathbf{u}+t\mathbf{v}|^2 = \sum_{i=1}^{n}(u_i + tv_i)^2 = |\mathbf{u}|^2 + 2t\sum_{i=1}^{n}u_iv_i + t^2|\mathbf{v}|^2.$$

As $|\mathbf{v}| \neq 0$, the RHS of the above inequality is a quadratic in t which is always non-negative, and so has non-positive discriminant ($b^2 \le 4ac$). Hence □

§14.1.3 Equations of lines and planes

§14.1.4 The Question Of Consistency

§14.2 Vector Product and Vector Algebra

§14.2.1 Vector Product

§14.2.2 Scalar and vector triple product

§14.2.3 Cross product equation of a line

§14.2.4 Properties of Determinants

§14.3 Vectors

§14.3.1 Linear Combinations

Linear combinations of vectors $\mathbf{u}$ and $\mathbf{v}$ are given by

$$\lambda\mathbf{u} + \mu\mathbf{v}$$

where $\lambda, \mu \in \mathbb{R}$.

For $a_1, a_2, a_3 \in \mathbb{R}$,

- the combinations $a_1\mathbf{u}$ fill a **line** through the origin;
- the combinations $a_1\mathbf{u} + a_2\mathbf{v}$ fill a **plane** through the origin;
- the combinations $a_1\mathbf{u} + a_2\mathbf{v} + a_3\mathbf{w}$ fill the **three-dimensional space**. (Provided $\mathbf{w}$ does not lie in the plane of $\mathbf{u}$ and $\mathbf{v}$.)

The **Euclidean space** $\mathbb{R}^n$, as a set, is defined as the set of vertical vectors with n coordinates in the real numbers. Algebraically, $\mathbb{R}^n$ is an n-dimensional vector space over $\mathbb{R}$. Vectors in $\mathbb{R}^n$ are expressed as vertical vectors

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

To save space, we usually express the above vector compactly as follows:

$$\mathbf{x} = (x_1, \dots, x_n)$$

§14.3.2 Length and Dot Product

Definition 14.3.1: Dot product

The dot product (or inner product) of $\mathbf{v} = (v_1, \dots, v_n)$ and $\mathbf{w} = (w_1, \dots, w_n)$ is given by

$$\mathbf{v} \cdot \mathbf{w} = \sum_{i=1}^{n} v_i w_i = v_1 w_1 + \cdots + v_n w_n \tag{14.2}$$

It is easy to verify that the dot product is commutative; that is, $\mathbf{v} \cdot \mathbf{w} = \mathbf{w} \cdot \mathbf{v}$.

For perpendicular vectors, the dot product is zero.

An important case is the dot product of a vector *with itself.* In this case $\mathbf{v}$ equals $\mathbf{w}$. The dot product $\mathbf{v} \cdot \mathbf{v}$ gives the **length of v squared**.

Definition 14.3.2: Length

The length $\|\mathbf{v}\|$ of a vector $\mathbf{v} = (v_1, \dots, v_n)$ is the square root of $\mathbf{v} \cdot \mathbf{v}$, given by

$$\|\mathbf{v}\| = \sqrt{\mathbf{v} \cdot \mathbf{v}} = \sqrt{\sum_{i=1}^{n} v_i^2} \tag{14.3}$$

Proof. This simply follows from Pythagoras' theorem. □

The word "unit" indicates that some measurement equals "one". Hence we can define the **unit vector** as follows.

Definition 14.3.3: Unit vector

A unit vector of vector $\mathbf{v}$, denoted by $\hat{\mathbf{v}}$, is a vector whose length equals one; that is, $\hat{\mathbf{v}} \cdot \hat{\mathbf{v}} = 1$.

The standard unit vectors along the x- and y-axes are written $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ respectively. In the xy-plane, the unit vector that makes an angle θ with the x-axis is $(\cos\theta, \sin\theta)$.

$$\hat{\mathbf{i}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \hat{\mathbf{j}} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \hat{\mathbf{u}} = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$$

To get the unit vector, divide any non-zero vector $\mathbf{v}$ by its length $\|v\|$.

$$\hat{\mathbf{v}} = \frac{\mathbf{v}}{\|\mathbf{v}\|} \tag{14.4}$$

is a unit vector in the same direction as $\mathbf{v}$.

Cosine formula If $\mathbf{v}$ and $\mathbf{w}$ are non-zero vectors then

$$\frac{\mathbf{v} \cdot \mathbf{w}}{\|\mathbf{v}\| \|\mathbf{w}\|} = \cos\theta \tag{14.5}$$

where θ is the angle between the two vectors.

Since $|\cos\theta|$ never exceeds 1, the cosine formula gives two great inequalities:

Theorem 14.3.1: Schwarz inequality

$$|\mathbf{v} \cdot \mathbf{w}| \le \|\mathbf{v}\| \|\mathbf{w}\| \tag{14.6}$$

Theorem 14.3.2: Triangle inequality

$$\|\mathbf{v} + \mathbf{w}\| \le \|\mathbf{v}\| + \|\mathbf{w}\| \tag{14.7}$$

§14.4 Solving Linear Equations

15 Vector Spaces

§15.1 Real and Complex Numbers

This text assumes that the reader should be familiar with the sets of real and complex numbers, denoted by $\mathbb{R}$ and $\mathbb{C}$ respectively.

Euclidean spaces, linear combinations and linear span, subspaces, linear independence, bases and dimension, rank of a matrix, inner products, eigenvalues and eigenvectors, diagonalisation, linear transformations between Euclidean spaces

§15.2 Definition

The motivation for the definition of a vector space comes from properties of addition and scalar multiplication in $\mathbb{F}^n$: Addition is commutative, associative, and has an identity. Every element has an additive inverse. Scalar multiplication is associative. Scalar multiplication by 1 acts as expected. Addition and scalar multiplication are connected by distributive properties.

We will define a vector space to be a set V with an addition and a scalar multiplication on V that satisfy the properties in the paragraph above.

Definition 15.2.1: Addition, scalar multiplication

n **addition** on V is a function that assigns an element $u + v \in V$ to each pair of elements $u, v \in V$.
A **scalar multiplication** on V is a function that assigns an element $\lambda v \in V$ to each $\lambda \in \mathbb{F}$ and each $v \in V$.

Now we are ready to give the formal definition of a vector space.

Definition 15.2.1 (Vector space)**.** A *vector space* is a set V along with an addition on V and a scalar multiplication on V such that the following properties hold:

(i) Commutativity: $\forall u, v \in V, u + v = v + u$

(ii) Associativity: $\forall u, v, w \in V, u + (v + w) = (u + v) + w$

(iii) Existence of additive identity: there exists $0 \in V$ such that $\forall v \in V, v + 0 = v = 0 + v$

(iv) Existence of additive inverse: $\forall v \in V$ there exists $w \in V$ such that $v + w = 0_V = w + v$

(v) Existence of multiplicative identity: $\forall v \in V, 1v = v$

(vi) Distributivity of scalar multiplication over vector addition: $\forall u, v \in V, \lambda \in \mathbb{F}, \lambda(u + v) = \lambda u + \lambda v$

(vii) Distributivity of scalar multiplication over field addition: $\forall v \in V, \lambda, \mu \in \mathbb{F}, (\lambda + \mu)v = \lambda v + \mu v$

(viii) Scalar multiplication interacts well with field multiplication: $\forall v \in V, \lambda, \mu \in \mathbb{F}, (\lambda\mu)v = \lambda(\mu v)$

Elements of a vector space are called *vectors* or *points*.

The scalar multiplication in a vector space depends on $\mathbb{F}$. Thus when we need to be precise, we will say that V is a vector space over $\mathbb{F}$ instead of saying simply that V is a vector space.

Example 15.2.1 ($\mathbb{R}^n$ and $\mathbb{C}^n$)**.** $\mathbb{R}^n$ is a vector space over $\mathbb{R}$, and $\mathbb{C}^n$ is a vector space over $\mathbb{C}$.

A vector space over $\mathbb{R}$ is called a *real vector space*; a vector space over $\mathbb{C}$ is called a *complex vector space*.

Proposition 15.2.1 (Uniqueness of additive identity)**.** A vector space has a unique additive identity.

Proof. Suppose 0 and $0'$ are both additive identities for some vector space V.

Then

$$0' = 0' + 0 = 0 + 0' = 0$$

where the first equality holds because 0 is an additive identity, the second equality comes from commutativity, and the third equality holds because $0'$ is an additive identity.

Thus $0' = 0$, proving that V has only one additive identity. □

Proposition 15.2.2 (Uniqueness of additive inverse)**.** Every element in a vector space has a unique additive inverse.

Proof. Suppose V is a vector space. Let $v \in V$. Suppose w and w' are additive inverses of v. Then

$$w = w + 0 = w + (v + w') = (w + v) + w' = 0 + w' = w'$$

Thus $w = w'$, as desired. □

Because additive inverses are unique, the following notation now makes sense.

Notation. Let $v, w \in V$. Then $-v$ denotes the additive inverse of v; $w - v$ is defined to be $w + (-v)$.

Notation. For the rest of the book, V denotes a vector space over $\mathbb{F}$.

§15.3 Subspaces

Definition 15.3.1 (Subspace)**.** A subset $U \subset V$ is called a subspace of V if U is also a vector space (with the same addition and scalar multiplication as on V).

A subset U of V is a subspace of V if and only if U satisfies the following three conditions:

1. Existence of additive identity: $0 \in U$
2. Closed under addition: $u + w \in U \implies u + w \in U$
3. Closed under scalar multiplication: $a \in F$ and $u \in U$ implies $au \in U$.

Proof. If U is a subspace of V, then U satisfies the three conditions above by the definition of vector space.

Conversely, suppose U satisfies the three conditions above. The first condition above ensures that the additive identity of V is in U.

The second condition above ensures that addition makes sense on U. The third condition ensures that scalar multiplication makes sense on U. □

Part V

Abstract Algebra

16 Group Theory

§16.1 Group Axioms

§16.1.1 Binary Operations

Definition 16.1.1. A **binary operation** $*$ on a set S is a map $* : S \times S \to S$. We write $a * b$ for the image of (a, b) under $*$.

Example 16.1.1. The following are examples of binary operations.

- $+, -, \times$ on $\mathbb{R}$; $\div$ is not a binary operation on $\mathbb{R}$ as, for example $1 \div 0$ is undefined;
- $\wedge$, the cross product, on $\mathbb{R}^3$;
- min and max on $\mathbb{N}$;
- $\circ$, composition, on the set Sym(S) of bijections of a set S to itself.

A binary operation $*$ on a set S is said to be **associative** if, for any $a, b, c \in S$,

$$(a * b) * c = a * (b * c).$$

In particular, this means an expression such as $a_1 * a_2 * \cdots *$ an always yields the same result, irrespective of how the individual parts of the calculation are performed.

A binary operation $*$ on a set S is said to be **commutative** if, for any $a, b \in S$,

$$a * b = b * a.$$

An element $e \in S$ is said to be an **identity element** (or simply an identity) for an operation $*$ on S if, for any $a \in S$,

$$e * a = a = a * e.$$

Proposition 16.1.1 (Uniqueness of identity)**.** Let $*$ be a binary operation on a set S and let $a \in S$. If an identity e exists then it is unique.

Proof. Suppose that e_1 and e_2 are two identities for $*$. Then

$$e_1 * e_2 = e_1 \quad \text{as } e_2 \text{ is an identity;}$$

$$e_1 * e_2 = e_2 \quad \text{as } e_1 \text{ is an identity.}$$

Hence $e_1 = e_2$. □

If an operation $*$ on a set S has an identity e and $a \in S$, then we say that $b \in S$ is an **inverse** of a if

$$a * b = e = b * a.$$

Proposition 16.1.2 (Uniqueness of inverse)**.** Let $*$ be an associative binary operation on a set S with an identity e and let $a \in S$. Then an inverse of a, if it exists, is unique.

Proof. Suppose that b_1 and b_2 are inverses of a. Then

$$b_1 * (a * b_2) = b_1 * e = b_1;$$

$$(b_1 * a) * b_2 = e * b_2 = b_2.$$

By associativity $b_1 = b_2$. □

Notation. If $*$ is an associative binary operation on a set S with identity e, then the inverse of a (if it exists) is written a^{-1}.

Example 16.1.2. The following are examples of binary operations.

- $+$ on $\mathbb{R}$ is associative, commutative, has identity 0 and $x^{-1} \coloneqq -x$ for any x; $-$ on $\mathbb{R}$ is not associative or commutative and has no identity; $\times$ on $\mathbb{R}$ is associative, commutative, has identity 1 and $x^{-1} \coloneqq \frac{1}{x}$ for any non-zero x.
- min on $\mathbb{N}$ is both associative and commutative but has no identity; max on $\mathbb{N}$ is both associative and commutative and has identity 0 (being the least element of $\mathbb{N}$) though no positive integer has an inverse;
- $\circ$ is associative, but not commutative, with the identity map $x \to x$ being the identity element and as permutations are bijections they each have inverses.

§16.1.2 Group Axioms

A group is an algebraic structure that captures the idea of symmetry without an object.

Definition 16.1.2. A **group** is a pair $(G,*)$, where G is a set and $*$ is a binary operation on G satisfying the following group axioms:

1. **(associativity)** for all $a,b,c \in G$, $a*(b*c)=(a*b)*c$.
2. **(identity)** there exists an identity element $e \in G$ such that for all $a \in G$, $a*e=e*a=a$.
3. **(invertibility)** for all $a \in G$, there exists a unique inverse $a^{-1} \in G$ such that $a*a^{-1}=a^{-1}*a=e$.
4. **(closure)** for all $a,b,c \in G$, $a*b \in G$.

Notation. A group $(G,*)$ is usually simply denoted by G.

Notation. We abbreviate $a*b$ to just ab. Also, since the operation $*$ is associative, we can omit unnecessary parentheses: $(ab)c=a(bc)=abc$.

Notation. For any $g \in G$ and $n \in \mathbb{N}$ we abbreviate $g^n = \underbrace{g*\cdots*g}_{n\text{ times}}$.

Example 16.1.3 (Additive integers)**.** The pair $(\mathbb{Z},+)$ is a group. Note that

- The element $0 \in \mathbb{Z}$ is an identity: $a+0=0+a=a$ for any a.
- Every element $a \in \mathbb{Z}$ has an additive inverse: $a+(-a)=(-a)+a=0$.

Example 16.1.4 (Addition mod n)**.** Let $n>1$ be an integer, and consider the residues (remainders) modulo n. These form a group under addition. We call this the cyclic group of order n, and denote it as $\mathbb{Z}/n\mathbb{Z}$, with elements $0,1,\dots,n-1$. The identity is 0.

Proposition 16.1.3. Cancellation laws hold in groups.

Proof. By invertibility axiom,

$$ab=ac \implies b=c, \quad ba=ca \implies b=c$$

by multiplying a^{-1} on LHS or RHS. □

Proposition 16.1.4 (Inverse of products)**.** For $a,b \in G$, $(ab)^{-1}=b^{-1}a^{-1}$.

Proof. Direct computation. We have

$$(ab)(b^{-1}a^{-1})=a(bb^{-1})a^{-1}=aa^{-1}=e.$$

Similarly,

$$(b^{-1}a^{-1})(ab)=e.$$

Hence equating both gives us $(ab)^{-1}=b^{-1}a^{-1}$. □

Proposition 16.1.5 (Left multiplication is a bijection)**.** For a group G, pick a $g \in G$. Then the map $G \to G$ given by $x \mapsto gx$ is a bijection.

Proof. Check this by showing injectivity and surjectivity directly. □

G is **abelian**[1] if the operation is commutative; it is **non-abelian** if otherwise.

Example 16.1.5. The sets $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ form abelian groups under $+$ with $e=0$ and $x^{-1} \coloneqq -x$ in each case.

Example 16.1.6. The sets $\mathbb{Q}\setminus\{0\}$, $\mathbb{R}\setminus\{0\}$ and $\mathbb{C}\setminus\{0\}$ form abelian groups under $\times$ with $e=1$ and $x^{-1} \coloneqq \dfrac{1}{x}$ in each case. These groups are respectively denoted as $\mathbb{Q}^\times$, $\mathbb{R}^\times$ and $\mathbb{C}^\times$.

[1] after the Norwegian mathematician Niels Abel (1802–1829)

An important (if rather elementary) family of groups is the cyclic groups.

Definition 16.1.3 (Cyclic group)**.** A group G is called **cyclic** if there exists $g \in G$ such that

$$G = \{g^k \mid k \in \mathbb{Z}\}.$$

Such a g is called a **generator**.

As $g^i g^j = g^{i+j} = g^j g^i$ then cyclic groups are abelian.

Example 16.1.7. $\mathbb{Z}$ is cyclic and has generators 1 and -1.

Example 16.1.8. Let $n \geq 1$. The n-th cyclic group C_n is the group with elements

$$e, g, g_2, \ldots, g^{n-1}$$

which satisfy $g^n = e$. So given two elements in C_n we define

$$g_i g_j = \begin{cases} g^{i+j} & \text{if } 0 \leq i+j < n, \\ g^{i+j-n} & \text{if } n \leq i+j \leq 2n-2. \end{cases}$$

Another important family of groups is the dihedral groups.

Definition 16.1.4 (Dihedral group)**.** Let $n \geq 3$ be an integer and consider a regular n-sided polygon P in the plane. We then write D_{2n} for the set of isometries of the plane which map the polygon back to itself.

Remark. Here “D” stands for “dihedral”, meaning two-sided.

It is clear that D_{2n} forms a group under composition as

(i) the identity map is in D_{2n},

(ii) the product of two isometries taking P to P is another such isometry,

(iii) the inverse of such an isometry is another such isometry,

(iv) composition is associative.

Given two groups G and H, there is a natural way to make their Cartesian product $G \times H$ into a group. Recall that as a set

$$G \times H = (g, h) \mid g \in G, h \in H.$$

We then define the product group $G \times H$ as follows.

Definition 16.1.5 (Product group)**.** Let $(G, *_G)$ and $(H, *_H)$ be groups. Then the operation $*$ defined on $G \times H$ by

$$(g_1, h_1) * (g_2, h_2) = (g_1 *_G g_2, h_1 *_H h_2)$$

is a group operation. $(G \times H, *)$ is called the **product group** or the product of G and H.

Proof. As $*_G$ and $*_H$ are both associative binary operations then it follows easily from the definition to see that $*$ is also an associative binary operation on $G \times H$. We also note

$$e_{G \times H} = (e_G, e_H) \quad \text{and} \quad (g, h)^{-1} = (g^{-1}, h^{-1})$$

as for any $g \in G$, $h \in H$,

$$(e_G, e_H) * (g, h) = (g, h) = (g, h) * (e_G, e_H);$$
$$(g^{-1}, h^{-1}) * (g, h) = (e_G, e_H) = (g, h) * (g^{-1}, h^{-1}).$$

□

Definition 16.1.6 (Order (of a group))**.** The cardinality $|G|$ of a group G is called the **order** of G. We say that a group G is **finite** if $|G|$ is finite.

One way to represent a finite group is by means of the group table or Cayley table[2].

Definition 16.1.7 (Cayley table)**.** Let $G = \{e, g_2, g_3, \dots, g_n\}$ be a finite group. The **Cayley table** (or group table) of G is a square grid which contains all the possible products of two elements from G. The product $g_i g_j$ appears in the i-th row and j-th column of the Cayley table.

Remark. Note that a group is abelian if and only if its Cayley table is symmetric about the main (top-left to bottom-right) diagonal.

Definition 16.1.8 (Subgroup)**.** Let G be a group. We say that a subset $H \subseteq G$ is a **subgroup** of G if the group operation $*$ restricts to make a group of H. That is H is a subgroup of G if:

(i) $e \in H$;

(ii) whenever $g_1, g_2 \in H$ then $g_1 g_2 \in H$.

(iii) whenever $g \in H$ then $g^{-1} \in H$.

Remark. Note that there is no need to require that associativity holds for products of elements in H as this follows from the associativity of products in G.

Example 16.1.9. The set of even integers is a subgroup of $\mathbb{Z}$; the set of odd integers is not a subgroup of $\mathbb{Z}$ because it does not even form a group, since it does not satisfy the closure axiom.

Definition 16.1.9 (Order (of a group element))**.** Let G be a group and $g \in G$. The **order** of g, written $o(g)$, is the least positive integer k such that $g^k = e$. If no such integer exists then we say that g has infinite order.

Remark. Note, now, that there are unfortunately two different uses of the word order: the order of a group is the number of elements it contains; the order of a group element is the least positive power of that element which is the identity.

[2] after the English mathematician Arthur Cayley (1821–1895)

§16.1.3 Isomorphism

Definition 16.1.10 (Isomorphism)**.** An **isomorphism** $\phi : G \to H$ between two groups $(G, *_G)$ and $(H, *_H)$ is a bijection such that for any $g_1, g_2 \in G$ we have

$$\phi(g_1 *_G g_2) = \phi(g_1) *_H \phi(g_2).$$

Two groups are said to be **isomorphic** if there is an isomorphism between them, denoted by $G \cong H$.

Example 16.1.10 ($\mathbb{Z} \cong 10\mathbb{Z}$)**.** Consider the two groups

$$\mathbb{Z} = (\{\dots, -2, -1, 0, 1, 2, \dots\}, +)$$

and

$$10\mathbb{Z} = (\{\dots, -20, -10, 0, 10, 20, \dots\}, +).$$

These groups are "different", but only superficially so — you might even say they only differ in the names of the elements.

Formally, the map

$$\phi : \mathbb{Z} \to 10\mathbb{Z} \text{ by } x \mapsto 10x$$

is a bijection of the underlying sets which respects the group operation. In symbols,

$$\phi(x + y) = \phi(x) + \phi(y).$$

In other words, ϕ is a way of re-assigning names of the elements without changing the structure of the group.

§16.2 Permutation Groups

§16.3 More on Subgroups & Cyclic Groups

§16.4 Lagrange's Theorem

Definition 16.4.1 (Coset)**.** Let H be a subgroup of G.

Then the **left cosets** of H (or left H-cosets) are the sets

$$gH = \{gh \mid h \in H\}.$$

The **right cosets** of H (or right H-cosets) are the sets

$$Hg = \{hg \mid h \in H\}.$$

Two (left) cosets aH and bH are either disjoint or equal.

Since multiplication is injective, the cosets of H are the same size as H, and thus H partitions G into equal-sized parts.

Notation. We write G/H for the set of (left) cosets of H in G. The cardinality of G/H is called the **index** of H in G.

An important result relating the order of a group with the orders of its subgroups is Lagrange's theorem.

Theorem 16.4.1 (Lagrange's theorem)**.** If G is a finite group and H is a subgroup of G, then $|H|$ divides $|G|$.

Groups of small order (up to order 8). Quaternions. Fermat–Euler theorem from the group-theoretic point of view.

Theorem 16.4.2 (Fermat's Little Theorem)**.** For every finite group G, for all $a \in G$, $a^{|G|} = e$.

Proof. Consider the subgroup H generated by a: $H = \{a^i \mid i \in \mathbb{Z}\}$. Since G is finite, the infinite sequence $a^0 = e, a^1, a^2, a^3, \ldots$ must repeat, say $a^i = a^j, i < j$. Let $k = j - i$. Multiplying both sides by $a^{-i} = (a^{-1})^i$, we get $a^{j-i} = a^k = e$. Suppose k is the least positive integer for which this holds. Then $H = \{a_0, a_1, a_2, \ldots, a^{k-1}\}$, and thus $|H| = k$. By Lagrange's Theorem, k divides $|G|$, so $a^{|G|} = (a^k)^{\frac{|G|}{k}} = e$. □

17 Ring Theory

Readings:

- Ring Theory by Brilliant
- Ring Theory (Math 113) by UC Berkeley

§17.1 Definition

A ring is just a set where you can add, subtract, and multiply. In some rings you can divide, and in others you can't. There are many familiar examples of rings, the main ones falling into two camps: "number systems" and "functions".

Definition 17.1.1. A **ring** is a set R endowed with two binary operations, addition and multiplication, denoted $+$ and $\times$, with elements $0, 1 \in R$, which maps $+ : R \times R \to R$ and $\times : R \times R \to R$, subject to three axioms:

1. $(R, +)$ is an abelian group with identity 0.
2. $(R, \times)$ is a commutative semigroup, i.e. $a \times (b \times c) = (a \times b) \times c$, $a \times 1 = 1 \times a = a$, and $a \times b = b \times a$ for all $a, b, c \in R$.
3. Distributivity: $a \times (b + c) = a \times b + a \times c$ for all $a, b, c \in R$.

Example 17.1.1. Examples of rings:

- $\mathbb{Z}$: the integers $\dots, -2, -1, 0, 1, 2, \dots$ with usual addition and multiplication, form a ring. Note that we cannot always divide, since $1/2$ is no longer an integer.
- $2\mathbb{Z}$: the even integers $\dots, -4, -2, 0, 2, 4, \dots$
- $\mathbb{Z}[x]$: this is the set of polynomials whose coefficients are integers.

  It is an extension of $\mathbb{Z}$, in the sense that we allow all the integers, plus an "extra symbol" x, which we are allowed to multiply and add, giving rise to x^2, x^3, etc., as well as $2x$, $3x$, etc. Adding up various combinations of these gives all the possible integer polynomials.
- $\mathbb{Z}[x, y, z]$: polynomials in three variables with integer coefficients.

  This is an extension of the previous ring. In fact you can continue adding variables to get larger and larger rings.
- $\mathbb{Z}/n\mathbb{Z}$: integers mod n.

  These are equivalence classes of the integers under the equivalence relation "congruence mod n". If we just think about addition (and subtraction), this is exactly the cyclic group of order n. However, when we call it a ring, it means we are also using the operation of multiplication.
- $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$

Ideals, homomorphisms, quotient rings, isomorphism theorems. Prime and maximal ideals. Fields. The characteristic of a field. Field of fractions of an integral domain. Factorization in rings; units, primes and irreducibles. Unique factorization in principal ideal domains, and in polynomial rings. Gauss' Lemma and Eisenstein's irreducibility criterion. Rings $\mathbb{Z}[\alpha]$ of algebraic integers as subsets of $\mathbb{C}$ and quotients of $\mathbb{Z}[x]$. Examples of Euclidean domains and uniqueness and non-uniqueness of factorization. Factorization in the ring of Gaussian integers; representation of integers as sums of two squares. Ideals in polynomial rings. Hilbert basis theorem

18 Field Theory

§18.1 Field Axioms

Definition 18.1.1. A **field** is a ring R that satisfies the following extra properties:

- $0 \neq 1$,
- every non-zero element of R has a multiplicative inverse (or reciprocal): if $r \in R$ and $r \neq 0$, then there exists $s \in R$ such that $rs = 1$; in other words: $R \setminus \{0\}$ is a group under × with identity 1.

Example 18.1.1. Examples and non-examples of fields:

- $\mathbb{Z}^+$ is not a field because, for example, 0 is not a positive integer, for no positive integer n is $-n$ a positive integer, for no positive integer n except 1 is n^{-1} a positive integer.
- $\mathbb{Z}$ is not a field because for an integer n, n^{-1} is not an integer unless $n = 1$ or $n = -1$.
- $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are fields.

Proposition 18.1.1. Suppose K is a field and $X \subseteq K$ is a subset of K, with the following properties:

- $0, 1 \in X$,
- if $x, y \in X$, then $x + y, x - y, x \times y \in X$; and if $y \neq 0$, then $\frac{x}{y} \in X$.

Then X is a field.

Proof. By assumption, X is closed under addition and multiplication. Moreover, X is clearly a ring, because X inherits all the axioms from K. Finally, $0 \neq 1$, and if $0 \neq x \in X$, then $x^{-1} \in X$ by assumption. Therefore, X is a field. □

We call X a **subfield** of K.

19 Galois Theory

Readings:

- Notes by Tom Leinster

20 Category Theory

Readings:

- Basic Category Theory, by Tom Leinster

Part VI

Calculus (Single Variable)

21 Limits

§21.1 Precise Definition of a Limit

The reader is assumed to know how a limit is defined intuitively. However, this definition is imprecise: phrases such as "x is close to a" and "$f(x)$ gets closer and closer to L" are vague. In order to be able to prove limits conclusively, we must make the definition of a limit precise.

Definition 21.1.1. Let $f(x)$ be a function defined on an open interval around x_0. We say that the **limit** of $f(x)$ as x approaches x_0 is L, that is $\lim_{x \to x_0} f(x) = L$, if for every $\varepsilon > 0$ there exists $\delta > 0$ such that $\forall x \in \mathbb{R}$,

$$0 < |x - x_0| < \delta \implies \left|f(x) - L\right| < \varepsilon$$

Visualising this graphically, as ε becomes smaller and smaller, there always exists a δ that satisfies the property that for any x in the open interval $(x_0 - \delta, x_0 + \delta)$, the value of $f(x)$ lies in the interval $(L - \varepsilon, L + \varepsilon)$.

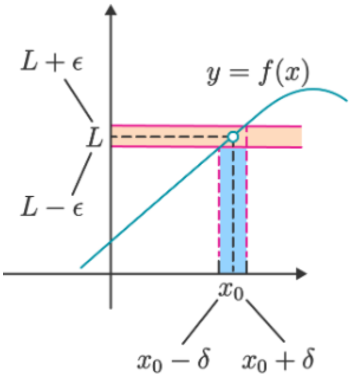


Figure 21.1: Epsilon–delta definition

Exercise 21.1.1

Prove that

$$\lim_{x \to 3} 2x + 4 = 10.$$

Before the proof, we work backwards to find the value of δ in terms of ε and x_0, which we then declare in our proof.

$\forall \varepsilon > 0$, $\exists \delta > 0$, $\forall x \in \mathbb{R}$,

$$|x - 3| < \delta \implies |f(x) - 10| < \varepsilon$$

Let $\varepsilon > 0$ be given.

$$|f(x) - 10| = |2x + 4 - 10| = |2x - 6| = 2|x - 3| < \varepsilon$$

Notice $|x-3| < \frac{\varepsilon}{2}$. We can thus define $\delta \coloneqq \frac{\varepsilon}{2}$. We now write our proof.

Proof. Let $\varepsilon > 0$ be given. Choose $\delta = \frac{\varepsilon}{2}$.

Then $\forall x \in \mathbb{R}$,

$$\begin{aligned}
|x-3| &< \delta = \frac{\varepsilon}{2} \\
2|x-3| &< \varepsilon \\
|2x-6| &< \varepsilon \\
|2x+4-10| &< \varepsilon \\
|f(x)-10| &< \varepsilon
\end{aligned}$$

□

Exercise 21.1.2

Use the formal definition of the limit to verify that

$$\lim_{x\to 3} \sqrt{2x+3} = 3.$$

We must prove that $\forall \varepsilon > 0$, $\exists \delta > 0$ such that $\sqrt{2x+3} - 3 < \varepsilon$ whenever $|x-3| < \delta$.

$$\begin{aligned}
\sqrt{2x+3} - 3 &= \left|\frac{(2x+3)-3^2}{\sqrt{2x+3}+3}\right| = \left|\frac{2x-6}{\sqrt{2x+3}+3}\right| \\
&\le \left|\frac{2(x-3)}{3}\right| \\
&= \frac{2}{3}|x-3| < \frac{2}{3}\delta
\end{aligned}$$

Hence, we can define

$$\varepsilon \coloneqq \frac{2}{3}\delta$$

which we can use in our proof.

Proposition 21.1.1 (Sum Law)**.** If $\lim_{x\to a} f(x) = L$ and $\lim_{x\to a} g(x) = M$ both exist, then

$$\lim_{x\to a}[f(x)+g(x)] = L + M.$$

Proof. Let $\varepsilon > 0$ be given. We must find $\delta > 0$ such that

$$0 < |x-a| < \delta \implies |f(x)+g(x)-(L+M)| < \varepsilon.$$

Using the Triangle Inequality we can write

$$\begin{aligned}
\big|f(x)+g(x)-(L+M)\big| &= \Big|\big(f(x)-L\big)+\big(g(x)-M\big)\Big| \\
&\le |f(x)-L| + |g(x)-M|
\end{aligned}$$

We make $\big|f(x)+g(x)-(L+M)\big|$ less than ε by making each of the terms $|f(x)-L|$ and $|g(x)-M|$ less than $\frac{\varepsilon}{2}$.

Since $\frac{\varepsilon}{2} > 0$ and $\lim_{x\to a} f(x) = L$, there exists $\delta_1 > 0$ such that

$$0 < |x-a| < \delta_1 \implies |f(x)-L| < \frac{\varepsilon}{2}.$$

Similarly, since $\lim_{x\to a} g(x) = M$, there exists $\delta_2 > 0$ such that

$$0 < |x-a| < \delta_2 \implies |g(x)-M| < \frac{\varepsilon}{2}.$$

Let $\delta = \min\{\delta_1, \delta_2\}$, the smaller of the numbers δ_1 and δ_2. Notice that

$$0 < |x-a| < \delta \implies 0 < |x-a| < \delta_1 \text{ and } 0 < |x-a| < \delta_2$$

and so

$$|f(x)-L| < \frac{\varepsilon}{2} \text{ and } |g(x)-M| < \frac{\varepsilon}{2}.$$

Therefore

$$\begin{aligned} \big|f(x)+g(x)-(L+M)\big| &\leq |f(x)-L| + |g(x)-M| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

To summarise,

$$0 < |x-a| < \delta \implies \big|f(x)+g(x)-(L+M)\big| < \varepsilon.$$

Thus by the definition of a limit,

$$\lim_{x\to a}[f(x)+g(x)] = L+M.$$

□

§21.2 Limit Laws

Let $f(x)$ and $g(x)$ be defined for all $x \neq a$ over some open interval containing a. Assume that L and M are real numbers such that $\lim_{x\to a} f(x) = L$ and $\lim_{x\to a} g(x) = M$, c is a constant. Then each of the following statements holds.

- **Sum law**: The limit of a sum is the sum of the limits.
$$\lim_{x\to a}(f(x) + g(x)) = \lim_{x\to a} f(x) + \lim_{x\to a} g(x) = L + M$$
- **Difference law**: The limit of a difference is the difference of the limits.
$$\lim_{x\to a}(f(x) - g(x)) = \lim_{x\to a} f(x) - \lim_{x\to a} g(x) = L - M$$
- **Constant multiple law**: The limit of a constant times a function is the constant times the limit of the function.
$$\lim_{x\to a} cf(x) = c\lim_{x\to a} f(x) = cL$$
- **Product law**: The limit of a product is the product of the limits.
$$\lim_{x\to a}(f(x) \cdot g(x)) = \lim_{x\to a} f(x) \cdot \lim_{x\to a} g(x) = L \cdot M$$
- **Quotient law**: The limit of a quotient is the quotient of the limits.
$$\lim_{x\to a} \frac{f(x)}{g(x)} = \frac{\lim_{x\to a} f(x)}{\lim_{x\to a} g(x)} = \frac{L}{M}$$
for $M \neq 0$.
- **Power law**
$$\lim_{x\to a}(f(x))^n = (\lim_{x\to a} f(x))^n = L^n$$
for every positive integer n.
- **Root law**
$$\lim_{x\to a} \sqrt[n]{f(x)} = \sqrt[n]{\lim_{x\to a} f(x)} = \sqrt[n]{L}$$
for all L if n is odd, and for $L \geq 0$ if n is even.

§21.3 Evaluating Limits

Indeterminate forms of a limit include:

$$\frac{0}{0} \quad \frac{\infty}{\infty} \quad 0 \times \infty \quad \infty - \infty \quad 0^0 \quad 1^\infty \quad \infty^0$$

As long as limits are in indeterminate forms, they can still be evaluated.

Methods:

- Direct substitution

  If f is a polynomial or a rational function and a is in the domain of f, then

  $$\lim_{x \to a} f(x) = f(a)$$

- Cancel common factors
- Multiply by the conjugate of the numerator or denominator

Exercise 21.3.1

Evaluate the following limit:

$$\lim_{x \to 0} x^2 \sin\left(\frac{1}{x}\right)$$

Solution. If we plot the graph of the function out, we see that we can try to find two functions to apply Squeeze Theorem.

Notice that

$$-1 \le \sin\left(\frac{1}{x}\right) \le 1$$

and hence

$$-x^2 \le x^2 \sin\left(\frac{1}{x}\right) \le x^2$$

thus x^2 and $-x^2$ are the two functions that “sandwich” the given function.

Since $\lim_{x \to 0} x^2 = 0$ and $\lim_{x \to 0} -x^2 = 0$, applying Squeeze Theorem gives us

$$\lim_{x \to 0} x^2 \sin\left(\frac{1}{x}\right) = 0$$

□

Exercise 21.3.2

Evaluate the following limit:

$$\lim_{x \to 3} \frac{-4x}{x-3}$$

Solution. Approaching from the left side,

$$\lim_{x \to 3^-} \frac{-4x}{x-3} = +\infty$$

Approaching from the right side,

$$\lim_{x \to 3^+} \frac{-4x}{x-3} = -\infty$$

Since $\lim_{x \to 3^-} \dfrac{-4x}{x-3} \ne \lim_{x \to 3^+} \dfrac{-4x}{x-3}$, the limit does not exist. □

Theorem 21.3.1. If $f(x) \leq g(x)$ when x is near a (except possibly at a) and the limits of f and g both exist as x approaches a, then

$$\lim_{x \to a} f(x) \leq \lim_{x \to a} g(x).$$

Theorem 21.3.2 (Squeeze theorem)**.** Suppose that $g(x) \geq f(x) \geq h(x)$ for all x in some open interval containing c except possibly at c itself. If $\lim_{x \to c} g(x) = L = \lim_{x \to c} h(x)$, then $\lim_{x \to c} f(x) = L$.

Proof. This can be proven using the epsilon–delta definition of limits.

Let $\varepsilon > 0$ be given. We are done if we find a $\delta > 0$ such that $|f(x) - L| < \varepsilon$ whenever $0 < |x - c| < \delta$.

Since $\lim_{x \to c} g(x) = L$, by definition of limits, there exists some $\delta_1 > 0$ such that for all $0 < |x - c| < \delta_1$, $|g(x) - L| < \varepsilon$. Thus,

$$-\varepsilon < g(x) - L < \varepsilon \quad \text{for all } 0 < |x - c| < \delta_1$$

so

$$L - \varepsilon < g(x) < L + \varepsilon \quad \text{for all } 0 < |x - c| < \delta_1 \tag{1}$$

Similarly, since $\lim_{x \to c} h(x) = L$, by definition of limits, there exists some $\delta_2 > 0$ such that

$$L - \varepsilon < h(x) < L + \varepsilon \quad \text{for all } 0 < |x - c| < \delta_2 \tag{2}$$

Additionally, since $g(x) \leq f(x) \leq h(x)$ for all x in some open interval containing c, there exists some $\delta_3 > 0$ such that for

$$g(x) \leq f(x) \leq h(x) \quad \text{for all } 0 < |x - c| < \delta_3 \tag{3}$$

Now, we choose $\delta = \min(\delta_1, \delta_2, \delta_3)$. Then by (1), (3), and (2), we have

$$L - \varepsilon < g(x) \leq f(x) \leq h(x) < L + \varepsilon \quad \text{for all } 0 < |x - c| < \delta.$$

Therefore, $-\varepsilon < f(x) - L < \varepsilon$ for all $0 < |x - c| < \delta$, so

$$|f(x) - L| < \varepsilon \quad \text{for all } 0 < |x - c| < \delta.$$

Hence, by definition of limits, $\lim_{x \to c} f(x) = L$. □

Theorem 21.3.3 (L'Hôpital's Rule)**.** Let $f(x)$ and $g(x)$ be differentiable on an interval I containing c, and that $g'(c) \neq 0$ on I for $x \neq c$. Suppose that $\lim_{x \to c} \frac{f(x)}{g(x)}$ is in an indeterminate form.

Then as long as the limits exist, we have

$$\lim_{x \to c} \frac{f(x)}{g(x)} = \lim_{x \to c} \frac{f'(x)}{g'(x)}.$$

Proof. □

Exercise 21.3.3

Let f be a differentiable function on $(0, \infty)$ and suppose that $\lim_{x \to \infty} \big(f(x) + f'(x)\big) = L$.

By considering $f(x) = \dfrac{e^x f(x)}{e^x}$, show that $\lim_{x \to \infty} f(x) = L$ and $\lim_{x \to \infty} f'(x) = 0$.

Solution. We may apply L'Hôpital's Rule as we encounter $\dfrac{\infty}{\infty}$ in $\lim\limits_{x \to \infty} \dfrac{e^x f(x)}{e^x}$.

$$\lim_{x \to \infty} f(x) = \lim_{x \to \infty} \frac{e^x f(x)}{e^x} = \lim_{x \to \infty} \frac{\big(e^x f(x)\big)'}{\left(e^x\right)'} = \lim_{x \to \infty} \frac{e^x f(x) + e^x f'(x)}{e^x} = L$$

Hence $\lim_{x \to \infty} f(x) = L$ and $\lim_{x \to \infty} f'(x) = 0$. □

§21.4 Important Limits

$$\lim_{x\to 0} \frac{\sin x}{x} = 1 \tag{21.1}$$

Proof. This can be proven using the squeeze theorem, which will be discussed later. □

$$\lim_{x\to 0} \frac{1 - \cos x}{x} = 0 \tag{21.2}$$

Proof. This can be proven using the squeeze theorem, which will be discussed later. □

$$\lim_{x\to 0} \frac{\arcsin x}{x} = 1 \tag{21.3}$$

$$\lim_{x\to \pm\infty} \left(1 + \frac{1}{x}\right)^x = e \tag{21.4}$$

§21.5 Continuity

Definition 21.5.1. A function $f(x)$ is **continuous** at $x = a$ if

$$\lim_{x\to a} f(x) = f(a).$$

A function is said to be continuous on the interval $[a, b]$ if it is continuous at each point in the interval.

Note that this definition is also implicitly assuming that both $f(a)$ and $\lim_{x\to a} f(x)$ exist. If either of these do not exist the function will not be continuous at $x = a$.

This definition can be turned around into the following fact.

Corollary 21.5.1. If $f(x)$ is continuous at $x = a$ then

$$\lim_{x\to a} f(x) = f(a) \qquad \lim_{x\to a^-} f(x) = f(a) \qquad \lim_{x\to a^+} f(x) = f(a)$$

A nice consequence of continuity is the following fact.

Corollary 21.5.2. If $f(x)$ is continuous at $x = b$ and $\lim_{x\to a} g(x) = b$ then

$$\lim_{x\to a} f(g(x)) = f\left(\lim_{x\to a} g(x)\right)$$

Exercise 21.5.1

Evaluate the following limit:

$$\lim_{x\to 0} e^{\sin x}$$

Solution. Since we know that exponentials are continuous everywhere we can use the fact above.

$$\lim_{x\to 0} e^{\sin x} = e^{\lim_{x\to 0}\sin x} = e^0 = \boxed{1}$$

□

Another very nice consequence of continuity is the Intermediate Value Theorem.

Theorem 21.5.1 (Intermediate Value Theorem)**.** Suppose that $f(x)$ is continuous on $[a, b]$ and let M be any number between $f(a)$ and $f(b)$. Then there exists $c \in (a, b)$ such that $f(c) = M$.

All the Intermediate Value Theorem is really saying is that a continuous function will take on all values between $f(a)$ and $f(b)$.

22 Derivative

§22.1 Definitions

Definition 22.1.1. The **derivative** of $f(x)$ with respect to x, denoted by $f'(x)$, is defined as

$$f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}. \tag{22.1}$$

The right-hand derivative of $f(x)$ at $x = x_0$ is defined as

$$f'_+(x_0) = \lim_{h\to 0^+} \frac{f(x_0+h)-f(x_0)}{h}.$$

Similarly, the left-hand derivative of $f(x)$ at $x = x_0$ is defined as

$$f'_-(x_0) = \lim_{h\to 0^-} \frac{f(x_0+h)-f(x_0)}{h}.$$

A function f has a derivative at $x = x_0$ if and only if $f'_+(x_0) = f'_-(x_0)$.

Definition 22.1.2. $f(x)$ is **differentiable** at x_0 if $f'(x_0)$ exists; $f(x)$ is differentiable on an interval I if the derivative exists for every $x \in I$.

Theorem 22.1.1. $f(x)$ is continuous at x_0 if $f(x)$ is differentiable at $x = x_0$.

Proof. To prove that f is continuous at x_0, we have to show that $\lim_{x\to x_0} f(x) = f(x_0)$. We will do this by showing that the difference $f(x) - f(x_0)$ approaches 0.

Given that f is differentiable at x_0,

$$f'(x_0) = \lim_{x\to x_0} \frac{f(x)-f(x_0)}{x-x_0}$$

exists. We can rewrite

$$f(x)-f(x_0) = \frac{f(x)-f(x_0)}{x-x_0}(x-x_0).$$

Then

$$\begin{aligned}
\lim_{x\to x_0}[f(x)-f(x_0)] &= \lim_{x\to x_0} \frac{f(x)-f(x_0)}{x-x_0}(x-x_0) \\
&= \lim_{x\to x_0} \frac{f(x)-f(x_0)}{x-x_0} \cdot \lim_{x\to x_0}(x-x_0) \\
&= f'(x_0)\cdot 0 = 0
\end{aligned}$$

To use what we have just proved,

$$\begin{aligned}
\lim_{x\to x_0} f(x) &= \lim_{x\to x_0}\left[f(x_0) + \big(f(x)-f(x_0)\big)\right] \\
&= \lim_{x\to x_0} f(x_0) + \lim_{x\to x_0}[f(x)-f(x_0)] \\
&= f(x_0) + 0 = f(x_0)
\end{aligned}$$

Therefore f is continuous at x_0. □

Remark. This means differentiability implies continuity. However the converse is not necessarily true, one notable example being the Weierstrass function.

§22.2 Theorems

Definition 22.2.1. Let c be a number in the domain D_f of a function f. Then $f(c)$ is the

- **absolute maximum** value of f on D_f if $f(c) \geq f(x)$ for all $x \in D_f$.
- **absolute minimum** value of f on D_f if $f(c) \leq f(x)$ for all $x \in D_f$.

An absolute maximum or minimum is sometimes called a *global* maximum or minimum. The maximum and minimum values of f are called **extreme values** of f.

Definition 22.2.2. $f(c)$ is a

- **local maximum** value of f if $f(c) \geq f(x)$ when x is near c.
- **local minimum** value of f if $f(c) \leq f(x)$ when x is near c.

We have seen that some functions have extreme values, whereas others do not. The following theorem gives conditions under which a function is guaranteed to possess extreme values.

Theorem 22.2.1 (Extreme Value Theorem)**.** For a function f continuous on $[a,b]$, it attains its maximum and minimum values on $[a,b]$.

Proof. We prove the case that f attains its maximum value on $[a,b]$.

Since f is continuous on $[a,b]$, we know it must be bounded on $[a,b]$ by the Boundedness Theorem. Let $M = \sup f$.

If there is some $c \in [a,b]$ where $f(c) = M$ there is nothing more to show – f attains its maximum on $[a,b]$.

Suppose otherwise, that there is no such c. Then $f(x) < M$ for all $x \in [a,b]$.

We define a new function g by

$$g(x) = \frac{1}{M - f(x)}.$$

Note that $g(x) > 0$ for every $x \in [a,b]$ and g is continuous on $[a,b]$, and thus also bounded on this interval, again by the Boundedness theorem.

Given that g is bounded on $[a,b]$, there must exist some $K > 0$ such that $g(x) \leq K$ for every $x \in [a,b]$.

Consequently,

$$\frac{1}{M - f(x)} \leq K \implies f(x) \leq M - \frac{1}{K}$$

for every $x \in [a,b]$. This contradicts the assumption that M is the least upper bound.

That leaves as the only possibility that there is some c in $[a,b]$ where $f(c) = M$. That is to say, f attains its maximum on $[a,b]$.

The proof that f attains its minimum on the same interval is argued similarly and is left as an exercise for the reader. □

Theorem 22.2.2 (Fermat's Theorem)**.** If f has a local maximum or minimum at c, and if $f'(c)$ exists, then $f'(c) = 0$.

Proof. Suppose, for the sake of definiteness, that f has a local maximum at c. Then $f(c) \geq f(x)$ if x is sufficiently close to c. This implies that if h is sufficiently close to 0, with h being positive or negative, then

$$f(c) \geq f(c+h)$$

and thus

$$f(c+h) - f(c) \leq 0.$$

We can divide both sides of an inequality by a positive number. Thus, if $h > 0$ and h is sufficiently small, we have

$$\frac{f(c+h) - f(c)}{h} \leq 0.$$

Taking the right-hand limit of both sides of the inequality,

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

but since $f'(c)$ exists, we have

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

and so we have show that $f'(c) \leq 0$.

On the other hand, if $h < 0$, then the direction of the inequality is reversed when we divide by h:

$$\frac{f(c+h) - f(c)}{h} \geq 0$$

so taking the left-hand limit, we have

$$f'(c) = \lim_{h \to 0} \frac{f(c+h) - f(c)}{h} = \lim_{h \to 0-} \frac{f(c+h) - f(c)}{h} \geq 0$$

thus $f'(c) \geq 0$.

Since $f'(c) \geq 0$ and $f'(c) \leq 0$, we have $f'(c) = 0$.

We have proved Fermat's Theorem for the case of a local maximum. The case of a local minimum can be proved in a similar manner. □

Theorem 22.2.3 (Rolle's Theorem)**.** Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b), and $f(a) = f(b)$. Then there exists $c \in (a, b)$ such that

$$f'(c) = 0.$$

Remark. Rolle's Theorem is simply a special case of the Mean Value Theorem, where $f(a) = f(b)$.

Proof. **Case 1:** $f(x) = k$ where k is a constant

Then $f'(x) = 0$, so the number c can be taken to be any number in (a, b).

Case 2: $f(x) > f(a)$ for some $x \in (a, b)$

By the Extreme Value Theorem, f has a maximum value somewhere in $[a, b]$. Since $f(a) = f(b)$, it must attain this maximum value at a number c in the open interval (a, b). Then f has a local maximum at c, thus f is differentiable at c. Therefore $f'(c) = 0$ by Fermat's Theorem.

Case 3: $f(x) < f(a)$ for some $x \in (a, b)$

By the Extreme Value Theorem, f has a minimum value in $[a, b]$ and, since $f(a) = f(b)$, it attains this minimum value at a number $c \in (a, b)$. Again $f'(c) = 0$ by Fermat's Theorem. □

Theorem 22.2.4 (Mean Value Theorem)**.** Let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). Then there exists $c \in (a, b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

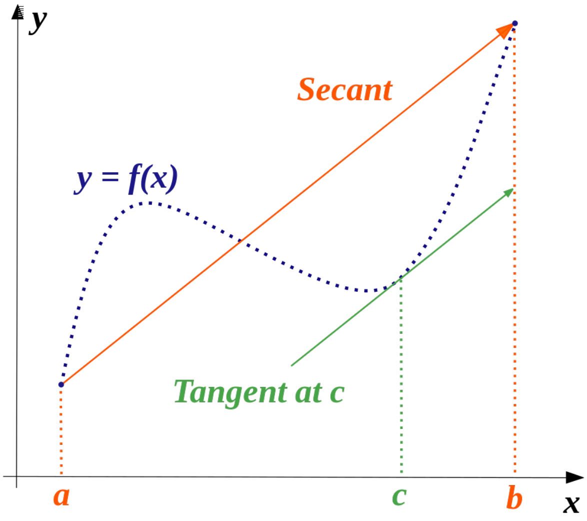


Figure 22.1: Mean value theorem

Theorem 22.2.5 (Cauchy's Generalised Theorem of the Mean)**.** Let f and g be continuous on $[a,b]$ and differentiable on (a,b), and where $g(x) \neq 0$ in (a,b). Then there exists $c \in (a,b)$ such that

$$\frac{f'(c)}{g'(c)} = \frac{f(b) - f(a)}{g(b) - g(a)}.$$

In layman's term, the theorem states that given two differentiable functions, there is some point c where the ratio of their average rates of change agrees with the ratio of their instantaneous rates of change.

Remark. When $g(x) = x$, it reduces to the Mean Value Theorem.

Exercise 22.2.1

Let $f(x) = \dfrac{x^e}{e^x}$, where $x > 0$. Find the maximum value of $f(x)$ and hence prove that $e^\pi > \pi^e$.

Proof.

$$f'(x) = x^{e-1}e^{-x}(e - x)$$

Since $x^{e-1}e^{-x} > 0$ for all $x > 0$, we have

$$f'(x) = \begin{cases} > 0 & x < e \\ 0 & x = e \\ < 0 & x > e \end{cases}.$$

Hence the maximum value of f occurs at $x = e$ since f is a continuous function, with maximum value $f(e) = \dfrac{e^e}{e^e} = 1$.

Therefore $\dfrac{x^e}{e^x} < 1$ for all $x \in \mathbb{R}^+ \setminus \{e\}$, i.e. $\dfrac{\pi^e}{e^\pi} < 1$ and thus $e^\pi > \pi^e$. □

Exercise 22.2.2

By applying Rolle's Theorem on the function $f(x) = e^{-x} - \sin x$, show that there is at least one real root of $e^x \cos x = -1$ between any two real roots of $e^x \sin x = 1$.

Proof. Let the two real roots involved of $e^x \sin x = 1$ be a and b with $a < b$.

Then $f(a) = f(b) = 0$. By Rolle's Theorem, there exist a value c with $a < c < b$ such that $f'(c) = 0$, i.e. $-e^{-c} - \cos c = 0$ which reduces to $e^c \cos c = -1$. □

Exercise 22.2.3

By using the Theorem of the Mean, show that

$$\frac{\pi}{6} + \frac{\sqrt{3}}{15} < \sin^{-1} 0.6 < \frac{\pi}{6} + \frac{1}{8}.$$

Proof. Let $f(x) = \sin^{-1} x$, we have $f'(x) = \dfrac{1}{\sqrt{1-x^2}}$.

By the Theorem of the Mean, we have, for $a < c < b$,

$$\frac{1}{\sqrt{1-a^2}} < \frac{1}{\sqrt{1-c^2}} = \frac{\sin^{-1}(b) - \sin^{-1}(a)}{b-a} < \frac{1}{\sqrt{1-b^2}}.$$

i.e. Interval min grad / Interval average grad/ Interval max grad

Using $a = 0.5$ and $b = 0.6$,

$$\begin{aligned}
&\frac{1}{\sqrt{1-0.5^2}} < \frac{\sin^{-1} 0.6 - \sin^{-1} 0.5}{0.6-0.5} < \frac{1}{\sqrt{1-0.6^2}} \\
&\frac{2}{\sqrt{3}} < \frac{\sin^{-1} 0.6 - \frac{\pi}{6}}{0.1} < \frac{1}{0.8} \\
&\frac{\sqrt{3}}{15} < \sin^{-1} 0.6 - \frac{\pi}{6} < \frac{1}{8} \\
&\frac{\pi}{6} + \frac{\sqrt{3}}{15} < \sin^{-1} 0.6 < \frac{\pi}{6} + \frac{1}{8}
\end{aligned}$$

□

§22.3 Differentiation Rules

For a constant $c \in \mathbb{R}$ and functions f and g of x, the following rules hold.

- **Scalar multiplication**
$$(cf)' = cf'$$

- **Addition rule**
$$(f+g)' = f' + g'$$

  Proof.
$$\begin{aligned}(f+g)'(x) &= \lim_{h\to 0}\frac{(f+g)(x+h)-(f+g)(x)}{h}\\ &= \lim_{h\to 0}\frac{f(x+h)+g(x+h)-f(x)-g(x)}{h}\\ &= \lim_{h\to 0}\left[\frac{f(x+h)-f(x)}{h}+\frac{g(x+h)-g(x)}{h}\right]\\ &= \lim_{h\to 0}\frac{f(x+h)-f(x)}{h}+\lim_{h\to 0}\frac{g(x+h)-g(x)}{h}\\ &= f'(x)+g'(x)\end{aligned}$$
□

- **Power rule**
$$\frac{\mathrm{d}}{\mathrm{d}x}x^n = nx^{n-1}$$

  Proof. Using implicit differentiation,
$$\begin{aligned}y &= x^n\\ \ln y &= \ln x^n\\ \ln y &= n\ln x\\ \frac{y'}{y} &= n\frac{1}{x}\\ y' = y\frac{n}{x} &= x^n\left(\frac{n}{x}\right) = nx^{n-1}\end{aligned}$$
□

- **Product rule**
$$(fg)' = f'g + fg'$$

  Proof.
$$\begin{aligned}(fg)'(x) &= \lim_{h\to 0}\frac{(fg)(x+h)-(fg)(x)}{h}\\ &= \lim_{h\to 0}\frac{f(x+h)g(x+h)-f(x)g(x)}{h}\\ &= \lim_{h\to 0}\frac{f(x+h)g(x)-f(x)g(x)+f(x+h)g(x+h)-f(x+h)g(x)}{h}\\ &= \lim_{h\to 0}\frac{f(x+h)g(x)-f(x)g(x)}{h}+\lim_{h\to 0}\frac{f(x+h)g(x+h)-f(x+h)g(x)}{h}\\ &= \lim_{h\to 0}\frac{f(x+h)-f(x)}{h}g(x)+\lim_{h\to 0}\frac{g(x+h)-g(x)}{h}f(x+h)\\ &= \left[\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}\right]g(x)+\left[\lim_{h\to 0}\frac{g(x+h)-g(x)}{h}\right]f(x)\\ &= f'(x)g(x)+f(x)g'(x)\end{aligned}$$
□

- **Quotient rule**

$$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$$

Proof.

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

□

- **Chain rule**

Theorem 22.3.1 (Chain rule)**.** If f and g are both differentiable functions and we define $F(x) = (f \circ g)(x)$, then the derivative of $F(x)$ is

$$F'(x) = f'(g(x))g'(x) \tag{22.2}$$

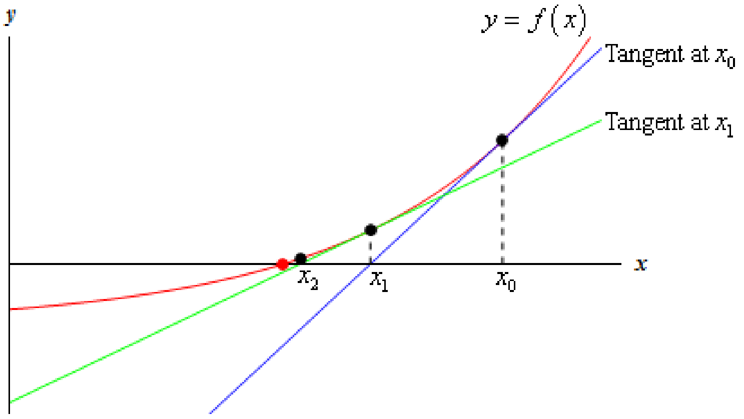
§22.4 Applications of Differentiation

§22.4.1 Implicit differentiation

Implicit differentiation simply means differentiating both sides of the equation with respect to a variable.

§22.4.2 Newton's Method

In this section we are going to look at a method for approximating solutions to equations.



Suppose that we want to approximate the solution to $f(x) = 0$. Suppose that we have somehow found an initial rough approximation to the solution: $x = x_0$. The tangent line to $f(x)$ at $x = x_0$ is

$$y = f(x_0) + f'(x_0)(x - x_0)$$

This tangent line crosses the x-axis much closer to the actual solution to the equation than x_0 does. Let the tangent at x_0 intersect x-axis at x_1. We use this point as our new approximation to the solution. x_1 is given by:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

Repeat the process; form up the tangent line at x_1 and use its root x_2 as a new approximation:

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

Here is the general Newton's Method:

Theorem 22.4.1: Newton's Method

If x_n is an approximation of a solution of $f(x) = 0$ and if $f'(x_n) \neq 0$, then the next approximation is given by

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

23 Integral

§23.1 Definition

We use the **Riemann** definition of an integral:

Definition 23.1.1. An **integral** is defined as an infinite sum over an interval.

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{n\to\infty}\sum_{i=1}^{n} f(x_i)\Delta x \tag{23.1}$$

A Riemann sum is an approximation of an integral by a finite sum.

Let f be defined on the closed interval $[a,b]$ and let Δx be a partition of $[a,b]$, with

$$a = x_1 < x_2 < \cdots < x_n < x_{n+1} = b.$$

Let Δx_i denote the length of the ith subinterval $[x_i, x_{i+1}]$ and let c_i denote any value in the ith subinterval.

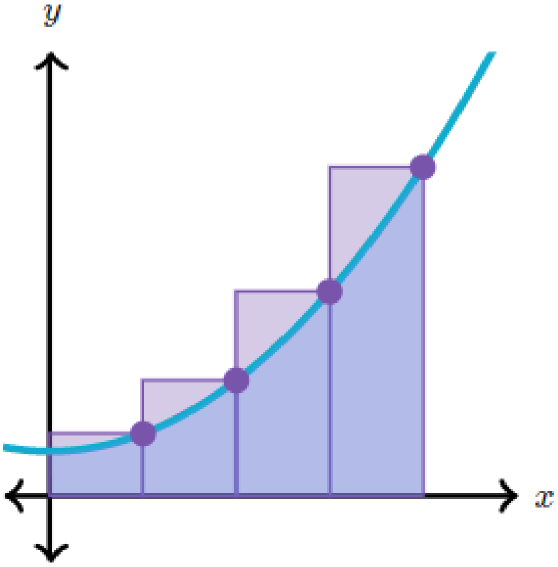
The sum

$$\sum_{i=1}^{n} f(c_i)\Delta x_i$$

is a Riemann sum of f on $[a,b]$.

As the subinterval becomes infinitesimally small,

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{\Delta x\to 0}\sum_{i=1}^{n} f(x_i)\Delta x_i$$



Theorem 23.1.1: Fundamental Theorem of Calculus

The fundamental theorem of (single variable) calculus states that if f' is continuous on $[a,b]$, then the integral of the derivative across the bounds is equal to the original function at the bounds:

$$\int_a^b f'(x)\,\mathrm{d}x = f(b) - f(a) \tag{23.2}$$

or equivalently,

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_a^x f(s)\,\mathrm{d}s = f(x) \tag{23.3}$$

Using the definition of the derivative, we differentiate the following integral:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\int_a^x f(s)\,\mathrm{d}s &= \lim_{h\to 0}\frac{\int_a^{x+h} f(s)\,\mathrm{d}s - \int_a^x f(s)\,\mathrm{d}s}{h}\\ &= \lim_{h\to 0}\frac{\int_x^{x+h} f(s)\,\mathrm{d}s}{h}\\ &= \lim_{h\to 0}\frac{hf(x)}{h}\\ &= f(x)\end{aligned}$$

§23.2 Integration Rules

For constant $k \in \mathbb{R}$ and functions $f(x)$ and $g(x)$, the following rules hold.

- **Sum and difference rule**
$$\int f(x) \pm g(x)\,\mathrm{d}x = \int f(x)\,\mathrm{d}x \pm \int g(x)\,\mathrm{d}x$$
- **Scalar multiplication**
$$\int kf(x)\,\mathrm{d}x = k\int f(x)\,\mathrm{d}x$$
- **Power rule**
$$\int x^n\,\mathrm{d}x = \frac{x^{n+1}}{n+1} + C$$
- **Constant rule**
$$\int a\,\mathrm{d}x = ax + C$$

§23.3 Integration Techniques

Integrals of powers and of trigonometric functions

Reciprocal rules:

$$\int \frac{1}{x}\,\mathrm{d}x = \ln|x| + C$$

$$\int \frac{1}{ax+b}\,\mathrm{d}x = \frac{1}{a}\ln(ax+b) + C$$

Exponential functions:

$$\int e^x\,\mathrm{d}x = e^x + C$$

$$\int a^x\,\mathrm{d}x = \frac{a^x}{\ln a} + C$$

Natural log rule:

$$\int \ln x\,\mathrm{d}x = x\ln x - x + C$$

Trigonometric functions:

$$\int \sin x\,\mathrm{d}x = -\cos x + C$$

$$\int \cos x\,\mathrm{d}x = \sin x + C$$

$$\int \tan x\,\mathrm{d}x = \ln|\sec x| + C$$

$$\int \operatorname{cosec} x\,\mathrm{d}x = \ln|\operatorname{cosec} x - \cot x| + C$$

$$\int \operatorname{cosec}^2 x\,\mathrm{d}x = -\cot x + C$$

$$\operatorname{cosec} x \cot x\,\mathrm{d}x = -\operatorname{cosec} x + C$$

$$\sec x\,\mathrm{d}x = \ln|\sec x + \tan x| + C$$

$$\int \sec^2 x\,\mathrm{d}x = \tan x + C$$

$$\int \sec x \tan x\,\mathrm{d}x = \sec x + C$$

$$\int \cot x\,\mathrm{d}x = \ln|\sin x| + C$$

Inverse trigonometric functions:

$$\int \frac{1}{\sqrt{1-x^2}}\,\mathrm{d}x = \sin^{-1} x + C$$

$$\int -\frac{1}{\sqrt{1-x^2}}\,\mathrm{d}x = \cos^{-1} x + C$$

$$\int \frac{x}{1+x^2}\,\mathrm{d}x = \tan^{-1} x + C$$

Splitting the Numerator

Exercise 23.3.1

Evaluate exactly

$$\int_0^1 \frac{x+2}{\sqrt{x+1}}\,\mathrm{d}x\,.$$

Solution.

$$\begin{aligned}\int_0^1 \frac{x+2}{\sqrt{x+1}}\,\mathrm{d}x &= \int_0^1 \left(\sqrt{x+1} + \frac{1}{\sqrt{x+1}}\right)\mathrm{d}x \\ &= \left[\frac{2}{3}(x+1)^{\frac{3}{2}} + 2(x+1)^{\frac{1}{2}}\right]_0^1 \\ &= \boxed{\frac{2}{3}\left(5\sqrt{3}-4\right)}\end{aligned}$$

□

Substitution

$$\int f(g(x))g'(x)\,\mathrm{d}x = \int f(u)\,\mathrm{d}u \tag{23.4}$$

where $u = g(x)$.

The most common way of doing a integral by substitution, and the only way for indefinite integrals, is as follows:

1. Change variables from x to u (hence the common name “u-substitution”)
2. Keep track of the relation between $\mathrm{d}x$ and $\mathrm{d}u$
3. If you chose correctly you can now do the u-integral
4. When you are done, substitute back for x

Exercise 23.3.2

Compute $\int \sin^n x \cos x\,\mathrm{d}x$.

Solution. Substitute $u = \sin x$ and $\mathrm{d}u = \cos x\,\mathrm{d}x$. This turns the integral into $\int u^n\,\mathrm{d}u$ which is easily valuated as $u^{n+1}/(n+1)$. Now plug back in $u = \sin x$ and you get the answer

$$\frac{\sin^{n+1} x}{n+1}.$$

□

Exercise 23.3.3

Compute $\displaystyle\int_1^2 \frac{x}{x^2+1}\,\mathrm{d}x$.

Solution. Let $u = x^2+1$ then $\mathrm{d}u = 2x\,\mathrm{d}x$, so the integrand becomes $(1/2)\,\mathrm{d}u/u$. If x goes from 1 to 2 then u goes from 2 to 5, thus the integral becomes

$$\int_2^5 \frac{1}{2}\frac{\mathrm{d}u}{u} = \frac{1}{2}(\ln 5 - \ln 2).$$

□

Exercise 23.3.4

Compute $\int xe^{x^2}\,\mathrm{d}x$.

Solution. To do this integral we'll use the following substitution.

$$u = x^2 \implies \mathrm{d}u = 2x\,\mathrm{d}x \implies x\,\mathrm{d}x = \frac{1}{2}\,\mathrm{d}u$$

$$\int xe^{x^2}\,\mathrm{d}x = \frac{1}{2}\int e^u\,\mathrm{d}u = \frac{1}{2}e^u + c = \frac{1}{2}e^{x^2} + c$$

□

Exercise 23.3.5

Prove the following result for $a > 0$:

$$\int_0^a f(x)\,\mathrm{d}x = \int_0^a f(a-x)\,\mathrm{d}x \qquad (*)$$

Hence evaluate

(a) $\displaystyle\int_0^{\frac{\pi}{2}} \frac{\cos^n x}{\sin^n x + \cos^n x}\,\mathrm{d}x$, where n is a real constant,

(b) $\displaystyle\int_0^{a} \frac{x^4}{x^4 + (x-a)^4}\,\mathrm{d}x$, where a is a positive real constant.

Solution. Substitute $x = a - u \implies \frac{\mathrm{d}x}{\mathrm{d}u} = -1$.

For the lower and upper limits, $x = 0 \implies u = a$ and $x = a \implies u = 0$ respectively.

Using the substitution, we have

$$\int_0^a f(x)\,\mathrm{d}x = \int_a^0 f(a-u)(-1)\,\mathrm{d}u = \int_0^a f(a-u)\,\mathrm{d}u = \int_0^a f(a-x)\,\mathrm{d}x\,.$$

(a) For $f(x) = \dfrac{\cos^n x}{\sin^n x + \cos^n x}$ with $a = \frac{\pi}{2}$,

$$f\left(\frac{\pi}{2} - x\right) = \frac{\cos^n\left(\frac{\pi}{2} - x\right)}{\sin^n\left(\frac{\pi}{2} - x\right) + \cos^n\left(\frac{\pi}{2} - x\right)} = \frac{\sin^n x}{\cos^n x + \sin^n x}.$$

By (*),

$$\int_0^{\frac{\pi}{2}} f(x)\,\mathrm{d}x = \int_0^{\frac{\pi}{2}} f\left(\frac{\pi}{2} - x\right)\mathrm{d}x$$

So we have

$$\int_0^{\frac{\pi}{2}} \frac{\cos^n x}{\sin^n x + \cos^n x}\,\mathrm{d}x = \int_0^{\frac{\pi}{2}} \frac{\sin^n x}{\sin^n x + \cos^n x}\,\mathrm{d}x$$

Let $I = \displaystyle\int_0^{\frac{\pi}{2}} \frac{\cos^n x}{\sin^n x + \cos^n x}\,\mathrm{d}x = \int_0^{\frac{\pi}{2}} \frac{\sin^n x}{\sin^n x + \cos^n x}\,\mathrm{d}x.$

Then

$$2I = I + I = \int_0^{\frac{\pi}{2}} \frac{\sin^n x + \cos^n x}{\sin^n x + \cos^n x}\,\mathrm{d}x = \int_0^{\frac{\pi}{2}} 1\,\mathrm{d}x = \frac{\pi}{2}$$

Hence $I = \displaystyle\int_0^{\frac{\pi}{2}} \frac{\cos^n x}{\sin^n x + \cos^n x}\,\mathrm{d}x = \boxed{\frac{\pi}{4}}$

(b) Let $f(x) = \dfrac{x^4}{x^4 + (x-a)^4}$.

Then $f(a-x) = \frac{(a-x)^4}{(a-x)^4 + (-x)^4} = \frac{(x-a)^4}{x^4 + (x-a)^4}$.

By (*),

$$I = \int_0^a \frac{x^4}{x^4 + (x-a)^4}\,\mathrm{d}x = \int_0^a \frac{(x-a)^4}{x^4 + (x-a)^4}\,\mathrm{d}x$$

Similarly,

$$2I = \int_0^a \frac{x^4 + (x-a)^4}{x^4 + (x-a)^4}\,\mathrm{d}x = \int_0^a 1\,\mathrm{d}x = a$$

Hence $I = \boxed{\frac{a}{2}}$

□

Integration by parts

Recall that the product rule for differentiation is given by

$$(fg)' = fg' + f'g.$$

Integrating both sides and rearranging gives us

$$\int fg'\,\mathrm{d}x = fg - \int f'g\,\mathrm{d}x \tag{23.5}$$

Alternatively, we can rewrite this as

$$\int u\,\mathrm{d}v = uv - \int v\,\mathrm{d}u \tag{23.6}$$

DI method

Note that after applying integration by parts multiple times, we may obtain a multiple of the original integral. A simple rearrangement of the integral gives the required solution.

Exercise 23.3.6

Evaluate $\int e^x \cos x\,\mathrm{d}x$.

Solution. Integrating by parts twice,

$$\begin{aligned}\int e^x \cos x\,\mathrm{d}x &= e^x \cos x - \int (e^x)(-\sin x)\,\mathrm{d}x = e^x \cos x + \int e^x \sin x\,\mathrm{d}x \\ &= e^x \cos x + \left(e^x \sin x - \int e^x \cos x\,\mathrm{d}x\right)\end{aligned}$$

Rearranging terms gives us

$$2\int e^x \cos x\,\mathrm{d}x = e^x \cos x + e^x \sin x + c$$

Hence $\int e^x \cos x\,\mathrm{d}x = \boxed{\frac{1}{2}(e^x \cos x + e^x \sin x) + C}$ where c and C are arbitrary constants. □

Finally in this section we look at an example of a **reduction formula**.

Exercise 23.3.7

Consider $I_n = \int \cos^n x\,\mathrm{d}x$ where n is a non-negative integer. Find a reduction formula for I_n and then use this formula to evaluate $\int \cos^7 x\,\mathrm{d}x$.

Solution. The aim here is to write I_n in terms of other I_k where $k < n$, so that eventually we are reduced to calculating I_0 or I_1, say, both of which are easily found. Using integration by parts we have:

$$\begin{aligned}
I_n &= \int \cos^{n-1} x \times \cos x \, \mathrm{d}x \\
&= \cos^{n-1} x \sin x - \int (n-1) \cos^{n-2} x (-\sin x) \sin x \, \mathrm{d}x \\
&= \cos^{n-1} x \sin x + (n-1) \int \cos^{n-2} x (1 - \cos^2 x) \, \mathrm{d}x \\
&= \cos^{n-1} x \sin x + (n-1)(I_{n-2} - I_n)
\end{aligned}$$

Rearranging this to make I_n the subject we obtain

$$I_n = \frac{1}{n} \cos^{n-1} x \sin x + \frac{n-1}{n} I_{n-2}, \quad n \geq 2.$$

With this reduction formula, I_n can be rewritten in terms of simpler and simpler integrals until we are left only needing to calculate I_0 if n is even, or I_1 if n is odd.

Therefore, I_7 can be found as follows:

$$\begin{aligned}
I_7 &= \frac{1}{7} \cos^6 x \sin x + \frac{6}{7} I_5 \\
&= \frac{1}{7} \cos^6 x \sin x + \frac{6}{7} \left(\frac{1}{5} \cos^4 x \sin x + \frac{4}{5} I_3 \right) \\
&= \frac{1}{7} \cos^6 x \sin x + \frac{6}{35} \cos^4 x \sin x + \frac{24}{35} \left(\frac{1}{3} \cos^2 x \sin x + \frac{2}{3} I_1 \right) \\
&= \frac{1}{7} \cos^6 x \sin x + \frac{6}{35} \cos^4 x \sin x + \frac{24}{105} \cos^2 x \sin x + \frac{48}{105} \sin x + c
\end{aligned}$$

□

Partial fraction decomposition

Trigonometric substitutions

- Pythagorean identity: $\sin^2 x + \cos^2 x = 1$
- Double-angle formulae
  These can be used in the integrals of $\sin^2 x$ and $\cos^2 x$.
- Product-to-sum identities

Integrals of powers of trigonometric functions

Integrals of hyperbolic functions

Completing the square

Elimination of radicals by substitution

Weierstrass substitution

Substituting the tangent of a half-angle: $t = \tan \dfrac{\theta}{2}$

Through trigonometric identities and manipulation, we have

$$\sin \theta = \frac{2t}{1+t^2} \qquad \cos \theta = \frac{1-t^2}{1+t^2} \qquad \mathrm{d}\theta = \frac{2\,\mathrm{d}t}{1+t^2}$$

Some examples here:

Useful info here: More info to be found on Youtube.

More problems here:

Odd and even functions

An odd function $f(x)$ satisfies $f(x) = -f(-x)$ for all x. Hence for any finite a,

$$\int_{-a}^{a} f(x)\,\mathrm{d}x = 0$$

An even function $f(x)$ satisfies $f(x) = f(-x)$ for all x. Hence for any finite a,

$$\int_{-a}^{a} f(x)\,\mathrm{d}x = 2\int_{0}^{a} f(x)\,\mathrm{d}x$$

Exercise 23.3.8

Evaluate $\displaystyle\int_{-\frac{\pi}{4}}^{\frac{\pi}{4}} \frac{x\cos x - 2\sin x + 1}{\cos^2 x}\,\mathrm{d}x.$

Solution. Breaking up into individual terms, we have

$$\int_{-\frac{\pi}{4}}^{\frac{\pi}{4}} \frac{x\cos x}{\cos^2 x}\,\mathrm{d}x - \int_{-\frac{\pi}{4}}^{\frac{\pi}{4}} \frac{2\sin x}{\cos^2 x}\,\mathrm{d}x + \int_{-\frac{\pi}{4}}^{\frac{\pi}{4}} \frac{1}{\cos^2 x}\,\mathrm{d}x$$

Notice that $\dfrac{x\cos x}{\cos^2 x}$ and $\dfrac{2\sin x}{\cos^2 x}$ are off functions, whose definite integrals equal to 0.
Hence we are left with

$$\int_{-\frac{\pi}{4}}^{\frac{\pi}{4}} \frac{1}{\cos^2 x}\,\mathrm{d}x = [\tan x]_{-\frac{\pi}{4}}^{\frac{\pi}{4}} = \boxed{2}$$

□

Exercise 23.3.9

Integrate a suitable Maclaurin series to obtain the value of

$$\frac{1}{\sqrt{2\pi}}\int_{-1}^{1} e^{-\frac{1}{2}x^2}\,\mathrm{d}x\,.$$

Solution. □

Reflections

This is known as **King's property**, which states that we can revere the interval of integration: to "integrate backwards".

$$\int_{a}^{b} f(x)\,\mathrm{d}x = \int_{a}^{b} f(a+b-x)\,\mathrm{d}x \tag{23.7}$$

Instead of the function being centred at 0, the function is now centred at $\frac{a+b}{2}$. Then

$$\int_{a}^{b} f(x)\,\mathrm{d}x = \frac{1}{2}\int_{a}^{b} f(x) + f(a+b-x)\,\mathrm{d}x$$

Inversions

Suppose the function f has bounded anti-derivative on $[0,\infty]$. Then via the u-substitution $x \to \frac{1}{x}$,

$$\int_{0}^{\infty} f(x)\,\mathrm{d}x = \frac{1}{2}\int_{0}^{\infty} f(x) + \frac{f(\frac{1}{x})}{x^2}\,\mathrm{d}x$$

Inverse functions

Suppose the function f is one-to-one and increasing. Then a geometric equivalence may be established:

Feynman's integration trick

DIfferentiating under the integral sign

§23.4 Approximation of Integral

§23.4.1 Trapezium Rule

We can sample the integrand at regular integrals and carry out an estimate based on this. One way of doing that is to approximate the function by a sequence of straight line segments. The area between each segment and the x-axis is a *trapezium*, meaning that if the width of the interval is h, and the y-values at each end of the interval are y_i and y_{i+1}, then the area of the trapezium is

$$\frac{h}{2}(y_i + y_{i+1}).$$

The entire area between the curve and the x-axis, which is to say the integral, can be approximated by adding together several such trapezia. If there are n trapezia, and n+1 y-values (ordinates) running from y_0 to y_n, then the integral is approximately

$$T_n = \frac{h}{2}\left(y_0 + 2\,y_2 + 2\,y_2 + \cdots + 2\,y_{n-2} + 2\,y_{n-1} + y_n\right) \tag{23.8}$$

§23.4.2 Simpson's Rule

Simpson's Rule is based on the fact that given any three points, you can find the *equation of a quadratic* through those points.

This fact inspired Simpson to approximate integrals using quadratics, as follows.

If you want to integrate $f(x)$ over the interval $[a, b]$:

1. Find $f(a)$, $f(b)$ and $f(m)$ where $m = \dfrac{a+b}{2}$.
2. Find a quadratic $P(x)$ that goes through these three points.

Since quadratics are easy to integrate, you simply need to integrate the quadratic over the interval. It turns out that the integral of the quadratic over the interval $[a, b]$ always comes out to

$$\frac{b-a}{6}[f(a) + 4f(m) + f(b)] \tag{23.9}$$

For even n subdivisions,

$$\int_a^b f(x)x' \approx \frac{\Delta x}{3}(f(x_0) + 4f(x_1) + 2f(x_2) + \cdots + 4f(x_{n-1}) + f(x_n)) \tag{23.10}$$

where $\Delta x = \dfrac{b-a}{n}$, $x_i = a + i\Delta x$.

§23.5 Applications of integrals

§23.5.1 Arc Length

The **arc length** of a function $y = f(x)$ in the interval $[a, b]$ is given by

$$L = \int \mathrm{d}s = \int_a^b \sqrt{1 + \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2}\,\mathrm{d}x \tag{23.11}$$

Similarly, the arc length of a function $x = g(y)$ in the interval $[c, d]$ is given by

$$L = \int \mathrm{d}s = \int_c^d \sqrt{1 + \left(\frac{\mathrm{d}x}{\mathrm{d}y}\right)^2}\,\mathrm{d}y$$

Proof. These formulae can be proven easily using Pythagoras' theorem.

Considering one small section of the arc,

$$\mathrm{d}s = \sqrt{(\mathrm{d}x)^2 + (\mathrm{d}y)^2}$$

Hence arc length is given by

$$L = \int \mathrm{d}s = \int \sqrt{(\mathrm{d}x)^2 + (\mathrm{d}y)^2} = \int \sqrt{1 + \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2}\,\mathrm{d}x$$

□

§23.5.2 Surface Area

A solid of revolution is a solid obtained by rotating a region bounded by two curves about a vertical or horizontal axis.

The surface area of the solid of revolution formed by rotating $y = f(x)$ in the interval $[a, b]$ by 2π about the x-axis is given by

$$S = \int 2\pi y\,\mathrm{d}s = \int_a^b 2\pi y \sqrt{1 + \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2}\,\mathrm{d}x \tag{23.12}$$

24 Sequences and Series

§24.1 Sequences

Definition 24.1.1. A **sequence** can be thought of as a list of numbers written in a definite order:

$$a_1, a_2, a_3, a_4, \ldots, a_n, \ldots$$

The number is called the first term, is the second term, and in general is the n-th term. We will deal exclusively with infinite sequences and so each term a_n will have a successor a_{n+1}.

Notice that for every positive integer n there is a corresponding number a_n and so a sequence can be defined as a function whose domain is the set of positive integers. But we usually write a_n instead of the function notation $f(n)$ for the value of the function at the number n.

Notation. The sequence $\{a_1, a_2, a_3, \ldots\}$ is also denoted by

$$\{a_n\} \text{ or } \{a_n\}_{n=1}^{\infty}$$

Example 24.1.1. The Fibonacci sequence $\{f_n\}$ is defined recursively by the conditions

$$f_1 = 1, f_2 = 1, f_n = f_{n-1} + f_{n-2} \text{ where } n \leq 3.$$

Using the ideas that we developed for limits of functions we can write down the following (informal) working definition for limits of sequences.

- We say that $\lim_{n\to\infty} a_n = L$ if we can make an as close to L as we want for all sufficiently large n. In other words, the value of the a_n's approach L as n approaches infinity.
- We say that $\lim_{n\to\infty} a_n = \infty$ if we can make an as large as we want for all sufficiently large n. Again, in other words, the value of the a_n's get larger and larger without bound as n approaches infinity.
- We say that $\lim_{n\to\infty} a_n = -\infty$ if we can make an as large and negative as we want for all sufficiently large n. Again, in other words, the value of the a_n's are negative and get larger and larger without bound as n approaches infinity.

Formally, we have the following precise definition.

Definition 24.1.1: Limit of sequence

We say that $\lim_{n\to\infty} a_n = L$ if for every number $\varepsilon > 0$ there exists an integer N such that

$$|a_n - L| < \varepsilon \quad \forall n > N.$$

Similarly, we say that $\lim_{n\to\infty} a_n = \infty$ if for every $M > 0$ there exists integer N such that $a_n > M$ whenever $n > N$; $\lim_{n\to\infty} a_n = -\infty$ if for every $M < 0$ there exists integer N such that $a_n < M$ whenever $n > N$.

Note that both definitions tell us that in order for a limit to exist and have a finite value all the sequence terms must be getting closer and closer to that finite value as n increases.

If $\lim_{n\to\infty} a_n$ exists and is finite we say that the sequence is **convergent**. If $\lim_{n\to\infty} a_n$ does not exist or is infinite we say the sequence **diverges**.

So just how do we find the limits of sequences? Most limits of most sequences can be found using one of the following theorems.

Theorem 24.1.1

Given the sequence $\{a_n\}$ if we have a function $f(x)$ such that $f(n) = a_n$ and $\lim_{x\to\infty} f(x) = L$ then $\lim_{n\to\infty} a_n = L$.

So, now that we know that taking the limit of a sequence is nearly identical to taking the limit of a function we also know that all the properties from the limits of functions will also hold. If $\{a_n\}$ and $\{b_n\}$ are both convergent sequences then

- $\lim_{n\to\infty}(a_n \pm b_n) = \lim_{n\to\infty} a_n \pm \lim_{n\to\infty} b_n$
- $\lim_{n\to\infty} ca_n = c\lim_{n\to\infty} a_n$
- $\lim_{n\to\infty}(a_n b_n) = (\lim_{n\to\infty} a_n)(\lim_{n\to\infty} b_n)$
- $\lim_{n\to\infty} \dfrac{a_n}{b_n} = \dfrac{\lim_{n\to\infty} a_n}{\lim_{n\to\infty} b_n}$, provided $\lim_{n\to\infty} b_n \neq 0$
- $\lim_{n\to\infty} {a_n}^p = (\lim_{n\to\infty} a_n)^p$, provided $a_n \geq 0$.

Next, just as we had a Squeeze Theorem for function limits we also have one for sequences and it is pretty much identical to the function limit version.

Theorem 24.1.2: Squeeze Theorem for sequences

If $a_n \leq c_n \leq b_n$ for all $n > N$ for some N and $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = L$ then $\lim_{n\to\infty} c_n = L$.

As we'll see not all sequences can be written as functions that we can actually take the limit of. This will be especially true for sequences that alternate in signs. While we can always write these sequence terms as a function we simply don't know how to take the limit of a function like that. The following theorem will help with some of these sequences.

Theorem 2 If $\lim_{n\to\infty} |a_n| = 0$ then $\lim_{n\to\infty} a_n = 0$.

Note that in order for this theorem to hold the limit MUST be zero and it won't work for a sequence whose limit is not zero. This theorem is easy enough to prove so let's do that.

§24.2 Series

Definition 24.2.1. A **series** is formed when the terms of a sequence are added, i.e. $a_1 + a_2 + \cdots + a_n = \sum_{i=1}^{n} a_i$ for a finite series, and $\sum_{i=1}^{\infty} a_i$ for an infinite series.

An **arithmetic progression** (AP) is a sequence in which each term differs from the preceding term by a constant called the **common difference** d.

A **geometric progression** (GP) is a sequence in which each term other than the first is obtained from the preceding one by multiplying by a non-zero constant called the common ratio r.

§24.2.1 Special series

In this section we are going to take a brief look at three special series.

- **Geometric series**
$$\sum_{n=0}^{\infty} ar^n$$
- **Telescoping series**
- **Harmonic series**
$$\sum_{n=1}^{\infty} \frac{1}{n}$$
The harmonic series is divergent, which we will prove in the subsequent sections.

§24.2.2 Convergence Tests

There are several convergence tests to determine if a series converges or diverges.

- **Divergence Test**

  If $\lim_{n\to\infty} a_n \neq 0$, then $\sum a_n$ will diverge.

- **Comparison Test**

  Suppose that we have two series $\sum a_n$ and $\sum b_n$ with $a_n, b_n \geq 0$ for all n and $a_n \leq b_n$ for all n. Then

  – if $\sum_{n=1}^{\infty} b_n$ converges, so does $\sum_{n=1}^{\infty} a_n$;

  – if $\sum_{n=1}^{\infty} a_n$ diverges, so does $\sum_{n=1}^{\infty} b_n$.

  Limit Comparison Test

  Suppose that we have two series $\sum a_n$ and $\sum b_n$ with $a_n, b_n \geq 0$ for all n. Define
$$c = \lim_{n\to\infty} \frac{a_n}{b_n}.$$
  If c is positive (i.e. $c > 0$) and is finite (i.e. $c < \infty$), then either both series converge or both series diverge.

- **Monotone Convergence Theorem**

  If a sequence of real numbers is increasing and bounded above, then its supremum is the limit.

  If a sequence of real numbers is decreasing and bounded below, then its infimum is the limit.

- **Absolute Convergence Test**

  A series $\sum a_n$ is **absolutely convergent** if $\sum |a_n|$ is convergent. If $\sum a_n$ is convergent and $\sum |a_n|$ is divergent we call the series **conditionally convergent**.

  If $\sum a_n$ is absolutely convergent then it is also convergent.

- **Alternating Series Test**

  Suppose that we have a series $\sum a_n$ and either $a_n = (-1)^n b_n$ or $a_n = (-1)^{n+1} b_n$ where $b_n \geq 0$ for all n. Then if

  - $\lim_{n\to\infty} b_n = 0$ and
  - $\{b_n\}$ is a decreasing sequence,

  the series $\sum a_n$ is convergent.

- **Integral Test**

  Suppose that $f(x)$ is a continuous, positive and decreasing function on the interval $[k, \infty)$ and that $f(n) = a_n$ then

  - If $\int_k^\infty f(x)\,\mathrm{d}x$ is convergent so is $\sum_{n=k}^\infty a_n$.
  - If $\int_k^\infty f(x)\,\mathrm{d}x$ is divergent so is $\sum_{n=k}^\infty a_n$.

- **Ratio Test**

  Suppose we have the series $\sum a_n$. Define

  $$L = \lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right|.$$

  Then

  - if $L < 1$, the series is absolutely convergent (and hence convergent);
  - if $L > 1$, the series is divergent;
  - if $L = 1$, the series may be divergent, conditionally convergent, or absolutely convergent.

- **Root Test**

  Suppose that we have the series $\sum a_n$. Define

  $$L = \lim_{n\to\infty} \sqrt[n]{|a_n|} = \lim_{n\to\infty} |a_n|^{\frac{1}{n}}.$$

  Then

  - if $L < 1$, the series is absolutely convergent (and hence convergent);
  - if $L > 1$, the series is divergent;
  - if $L = 1$, the series may be divergent, conditionally convergent, or absolutely convergent.

Exercise 24.2.1

Determine if the following series is convergent or divergent:

$$\sum_{n=0}^{\infty} \frac{4n^2 - n^3}{10 + 2n^3}$$

Solution. The first thing that we should do is take a look at the series terms and see if they go to zero or not. If it's clear that the terms don't go to zero use the Divergence Test and be done with the problem.

$$\lim_{n\to\infty} \frac{4n^2 - n^3}{10 + 2n^3} = -\frac{1}{2} \neq 0$$

The limit of the series terms is not zero and so by the Divergence Test the series diverges. □

Exercise 24.2.2

Determine if the following sequence converges or diverges:

$$u_n = \frac{n^3}{3^n}$$

Solution. Consider $f(x) = \dfrac{x^3}{3^x}$. Then

$$f'(x) = \frac{(3x^2)(3^x) - (x^3)(\ln 3)(3^x)}{3^{2x}} = \frac{x^2(3 - x\ln 3)}{3^x}.$$

Observe that for $x > 3$, we have $x^2 > 0$, $3^x > 0$ and $3 - x\ln 3 > 0$. Thus $f'(x) < 0$ for $x > 3$.

Also, note that $f(x) \geq 0$ for $x > 3$.

Thus $f(x)$ converges to a certain value as $x \to \infty$, i.e. u_n converges as $n \to \infty$.

(Bounded Convergence, provides the conclusion that the sequence converges but not providing the actual limit, which can be obtained from L'Hopital) □

Exercise 24.2.3

Show that the following series is divergent:

$$5 + \sqrt{5} + \sqrt[3]{5} + \sqrt[4]{5} + \cdots$$

Solution. The n-th term of the series is $u_n = \sqrt[n]{5} = 5^{\frac{1}{n}} \to 5^0 = 1$ as $n \to \infty$.

Since $u_n \not\to 0$ as $n \to \infty$, the series diverges. □

§24.2.3 Power Series

Definition 24.2.2. A **power series** about a is any series that can be written in the form

$$\sum_{n=0}^{\infty} c_n (x - a)^n$$

where a_i and c_i are constants. c_i are known as the coefficients of the series.

As we will see later, we will be able to show that there exists a number R, known as the **radius of convergence**, such that the power series converges for $|x - a| < R$ and diverges for $|x - a| > R$.

Secondly, the interval of all x's, including the endpoints if need be, for which the power series converges is called the **interval of convergence** of the series.

These two concepts are fairly closely tied together. If we know that the radius of convergence of a power series is R then we have the following:

- the power series converges if $a - R < x < a + R$, and
- the power series diverges if $x > a + R$.

Power Series And Functions

Taylor Series

A function f can be represented as a Taylor series about position a if

- it is continuous near a and
- all of its derivatives are continuous near a

Using the notation $\Delta x = x - a$:

$$f(x) = f(a) + \Delta x f'(a) + \frac{\Delta x}{2!} f''(a) + \frac{\Delta x}{3!} f^{(3)}(a) + \cdots + \frac{\Delta x}{n!} f^{(n)}(a) + \cdots$$

If infinitely many terms are used, this approximation is exact near a.

If all terms of order n and above are discarded then the error is approximately proportional to Δx^n (assuming that Δx is small). Then the approximation is said to be n-th order accurate. For example, a third order accurate approximation for $f(x)$ has error proportional to Δx^3. We say that the error is of order Δx^3 or $O(\Delta x^3)$.

$$f(x) = f(a) + \Delta x f'(a) + \frac{\Delta x}{2!} f''(a) + O(\Delta x^3)$$

Maclaurin series, determine radius and interval of convergence of a power series

Taylor Series – In this section we will discuss how to find the Taylor/Maclaurin Series for a function. This will work for a much wider variety of function than the method discussed in the previous section at the expense of some often unpleasant work.

Binomial Series

Binomial Series – In this section we will give the Binomial Theorem and illustrate how it can be used to quickly expand terms in the form (a + b) n when n is an integer. In addition, when n is not an integer an extension to the Binomial Theorem can be used to give a power series representation of the term.

25 Ordinary Differential Equations

Definition 25.0.1. An **ordinary differential equation** (ODE) is an equation relating a variable, say x, a function, say y, of the variable x, and finitely many of the derivatives of y with respect to x.

That is, an ODE can be written in the form

$$f\left(x, y, \frac{\mathrm{d}y}{\mathrm{d}x}, \frac{\mathrm{d}^2y}{\mathrm{d}x^2}, \cdots, \frac{\mathrm{d}^ky}{\mathrm{d}x^k}\right) = 0$$

for some function f and some natural number k. Here x is the **independent variable** and the ODE governs how the **dependent variable** y varies with x.

Remark. The equation may have no, one or many functions $y(x)$ which satisfy it; the problem is usually to find the most general solution $y(x)$, a function which satisfies the differential equation.

The derivative $\frac{\mathrm{d}^ky}{\mathrm{d}x^k}$ is said to be of order k. We say that an ODE has **order** k if it involves derivatives of order k and less. Hence, a first-order differential equation involves up to the first derivative $\frac{\mathrm{d}y}{\mathrm{d}x}$, whereas a second-order differential equation involves up to the second derivative $\frac{\mathrm{d}^2y}{\mathrm{d}x^2}$.

§25.1 First-order ODEs

First-order differential equations take the form

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

There are several standard methods for solving first order ODEs and we look at some of these now.

§25.1.1 Direct integration

If the ODE takes the form

$$\frac{\mathrm{d}y}{\mathrm{d}x} = f(x)$$

in other words the derivative is a function of x only, then we can integrate directly.

Exercise 25.1.1

Find the general solution of

$$\frac{\mathrm{d}y}{\mathrm{d}x} = x^2 \sin x$$

Solution. By direct integration,

$$y = \int x^2 \sin x \,\mathrm{d}x = (2 - x^2)\cos x + 2x \sin x + c$$

which is done using integration by parts. □

§25.1.2 Separation of variables

This method is applicable when the first order ODE takes the form

$$\frac{\mathrm{d}y}{\mathrm{d}x} = a(x)b(y)$$

where a is a function of x and b is a function of y.

Such an equation is called **separable**. These equations can be rearranged and solved as follows. First

$$\frac{1}{b(y)}\frac{\mathrm{d}y}{\mathrm{d}x} = a(x)$$

and then integrating with respect to x we find

$$\int \frac{1}{b(y)}\,\mathrm{d}y = \int a(x)\,\mathrm{d}x$$

Here we have assumed that $b(y) \neq 0$; if $b(y) = 0$ then the solution is $y = c$ where c is a constant.

Exercise 25.1.2

Find the general solution to the separable differential equation

$$x(y^2 - 1) + y(x^2 - 1)\frac{\mathrm{d}y}{\mathrm{d}x} = 0$$

where $0 < x < 1$.

Solution. We rearrange to obtain

$$\frac{y}{y^2 - 1}\frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{x}{x^2 - 1}$$

After integration we obtain

$$\frac{1}{2}\ln|y^2 - 1| = -\frac{1}{2}\ln|x^2 - 1| + c$$

where c is a constant. This can be arranged to give

$$(x^2 - 1)(y^2 - 1) = c.$$

Note that the constant functions $y = 1$ and $y = -1$ are also solutions of the differential equation, but are already included in the given general solution, for $c = 0$. □

§25.1.3 Reduction to separable form by substitution

Some first order differential equations can be transformed by a suitable substitution into separable form.

Exercise 25.1.3

Find the general solution of

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \sin(x + y + 1)$$

Solution. Let $u(x) = x + y(x) + 1$ so that $\frac{\mathrm{d}u}{\mathrm{d}x} = 1 + \frac{\mathrm{d}y}{\mathrm{d}x}$. Then the original equation can be written as $\frac{\mathrm{d}u}{\mathrm{d}x} = 1 + \sin u$, which is separable. We have

$$\frac{1}{1 + \sin u}\frac{\mathrm{d}u}{\mathrm{d}x} = 1$$

which integrates to

$$\int \frac{1}{1 + \sin u}\,\mathrm{d}u = x + c$$

Let us evaluate the integral on the left hand side:

$$\begin{aligned}\int \frac{1}{1+\sin u}\,\mathrm{d}u &= \int \frac{1-\sin u}{(1+\sin u)(1-\sin u)}\,\mathrm{d}u \\ &= \int \frac{1-\sin u}{1-\sin^2 u}\,\mathrm{d}u = \int \frac{1-\sin u}{\cos^2 u}\,\mathrm{d}u \\ &= \int \frac{1}{\cos^2 u}\,\mathrm{d}u - \int \frac{\sin u}{\cos^2 u}\,\mathrm{d}u \\ &= \tan u - \frac{1}{\cos u} + c\end{aligned}$$

Therefore

$$\tan u - \frac{1}{\cos u} = x + c$$

In terms of x and y, the solution is given by

$$\tan(x+y+1) - \frac{1}{\cos(x+y+1)} = x + c$$

or

$$\sin(x+y+1) - 1 = (x+c)\cos(x+y+1).$$

This solution, where we have not found y in terms of x, is called an **implicit solution**. □

A special group of first order differential equations is those of the form

$$\frac{\mathrm{d}y}{\mathrm{d}x} = f\left(\frac{y}{x}\right)$$

These differential equations are called **homogeneous** and they can be solved with a substitution of the form

$$y(x) = xv(x)$$

to get a new equation in terms of x and the new dependent variable v. This new equation will be separable:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = v + x\frac{\mathrm{d}v}{\mathrm{d}x}$$

which becomes

$$x\frac{\mathrm{d}v}{\mathrm{d}x} = f(v) - v$$

§25.1.4 Exact solutions

LHS is an exact derivative.

Exercise 25.1.4

Solve

$$\sin x\frac{\mathrm{d}y}{\mathrm{d}x} + y\cos x = 3x^2.$$

Solution. Notice that the LHS is an exact derivative.

$$\begin{aligned}\sin x\frac{\mathrm{d}y}{\mathrm{d}x} + y\cos x &= \frac{\mathrm{d}}{\mathrm{d}x}y\sin x \\ \frac{\mathrm{d}}{\mathrm{d}x}y\sin x &= 3x^2 \\ y\sin x &= \int 3x^2\,\mathrm{d}x = x^3 + c \\ y &= \frac{x^3+c}{\sin x}\end{aligned}$$

□

§25.2 First-order linear ODEs

In general, a k-th order **inhomogeneous linear ODE** takes the form

$$a_k(x)\frac{\mathrm{d}^k y}{\mathrm{d}x^k} + a_{k-1}(x)\frac{\mathrm{d}^{k-1} y}{\mathrm{d}x^{k-1}} + \cdots + a_1(x)\frac{\mathrm{d}y}{\mathrm{d}x} + a_0(x)y = f(x)$$

where $a_k(x) \neq 0$. The equation is **homogeneous** if $f(x) = 0$.

§25.2.1 Integrating Factor

Looking specifically at first order linear ODEs, which take the general form

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

we see that the homogeneous form, that is when $Q(x) = 0$, is separable. On the other hand, the inhomogeneous form can be solved using an **integrating factor** $I(x)$ given by

$$I(x) = e^{\int P(x)\mathrm{d}x}$$

Proof. Simply multiply the general equation for first-order linear ODEs through by the integrating factor to obtain

$$e^{\int P(x)\mathrm{d}x}\frac{\mathrm{d}y}{\mathrm{d}x} + P(x)e^{\int P(x)\mathrm{d}x}y = e^{\int P(x)\mathrm{d}x}Q(x)$$

Using the product rule for derivatives, we see that this gives

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(e^{\int P(x)\mathrm{d}x}y\right) = e^{\int P(x)\mathrm{d}x}Q(x)$$

and we can now integrate directly and rearrange, to obtain

$$y = e^{-\int P(x)\mathrm{d}x}\left[\int e^{\int P(x)\mathrm{d}x}Q(x)\,\mathrm{d}x + c\right].$$

□

Exercise 25.2.1

Solve the linear differential equation

$$\frac{\mathrm{d}y}{\mathrm{d}x} + 2xy = 2xe^{-x^2}.$$

Solution. We can easily see that the given differential equation is in the form of a first-order linear ODE.

First we find the integrating factor:

$$I(x) = e^{\int 2x\mathrm{d}x} = e^{x^2}$$

Multiplying the given differential equation through by the integrating factor this gives

$$e^{x^2}\frac{\mathrm{d}y}{\mathrm{d}x} + 2xe^{x^2}y = 2x$$

that is

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(e^{x^2}y\right) = 2x$$

Integrating this gives us

$$e^{x^2}y = x^2 + c$$

so the general solution is $y = (x^2 + c)e^{-x^2}$. □

§25.3 Second-order homogeneous linear ODEs

This section introduces a method for finding a second solution to a second order homogeneous linear ODE, when one solution has already been found. Suppose $z(x) \neq 0$ is a non-trivial solution to the second order homogeneous linear differential equation

$$P(x)\frac{\mathrm{d}^2y}{\mathrm{d}x^2} + Q(x)\frac{\mathrm{d}y}{\mathrm{d}x} + R(x)y = 0.$$

We can make the substitution $y(x) = u(x)z(x)$, so that

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}u}{\mathrm{d}x}z + u\frac{\mathrm{d}z}{\mathrm{d}x} \quad \text{and} \quad \frac{\mathrm{d}^2y}{\mathrm{d}x^2} = \frac{\mathrm{d}^2u}{\mathrm{d}x^2}z + 2\frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}z}{\mathrm{d}x} + u\frac{\mathrm{d}^2z}{\mathrm{d}x^2}.$$

Substituting these into the above equation and using the prime to denote the derivative with respect to x we obtain

$$P(x)\left(u''z + 2u'z' + uz''\right) + Q(x)\left(u'z + uz'\right) + R(x)uz = 0.$$

If we now rearrange the above equation and use the fact that z is a solution to the given equation then we get

$$P(x)zu'' + \left(2P(x)z' + Q(x)z\right)u' = 0,$$

which is a homogeneous differential equation of first order for u'. In theory this can be solved, to obtain the general solution to the given equation. The following example illustrates this technique.

Exercise 25.3.1

Verify that $z(x) = \frac{1}{x}$ is a solution to

$$x\frac{\mathrm{d}^2y}{\mathrm{d}x^2} + 2(1-x)\frac{\mathrm{d}y}{\mathrm{d}x} - 2y = 0,$$

and hence find its general solution.

Solution. □

To edit

§25.3.1 Linear with constant coefficients

For equations in the form

$$a\frac{\mathrm{d}^2y}{\mathrm{d}x^2} + b\frac{\mathrm{d}y}{\mathrm{d}x} + cy = 0$$

Auxiliary equation:

$$am^2 + bm + c = 0$$

which has solutions

$$m_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Roots of auxiliary equation	General solution
real and distinct	$y = Ae^{m_1x} + Be^{m_2x}$
real and repeated	$y = e^{mx}(A + Bx)$
imaginary: $m = \alpha \pm i\beta$	$y = e^{\alpha x}(A\cos\beta x + B\sin\beta x)$

§25.3.2 Non-linear

Equation

$$a\frac{\mathrm{d}^2y}{\mathrm{d}x^2}+b\frac{\mathrm{d}y}{\mathrm{d}x}+cy=f(x)$$

where $a\neq 0$, $f(x)\neq 0$

General solution y=Complementary Function (CF)+Particular Integral (PI) Complementary function: solution of the homogeneous equation

Particular integral (PI) find by seeing different cases of f(x) and then y,dy/dx,d2y/dx2 of general PI into DE to find unknown constants of PI If f(x) is a polynomial of degree n, $y=a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0$ When DE is a*d2y/dx2+b*dy/dx=f(x), method 1: find CF and let PI be

$$y=x(a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0)$$

method 2: integrate both sides wrt x and then solve accordingly

If $f(x)=qe^{kx}$ where k and q are constants,

If $f(x)=k\cos ax$ or $f(x)=k\sin ax$ or $f(x)=k\cos ax+k\sin ax$

If $f(x)$ is the sum of various functions

Example 25.3.1 (Vibrating springs)**.** Restoring force is given by $F=-kx$. By Newton's 2nd Law, we have $m\frac{\mathrm{d}^2x}{\mathrm{d}t^2}=-kx$, or

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2}+kx=0.$$

Hence we have auxilliary equation $mr^2+k=0$ with roots $r=\pm\omega i$ where angular frequency $\omega=\sqrt{\frac{k}{m}}$. Thus the general solution is

$$x(t)=c_1\cos\omega t+c_1\sin\omega t.$$

Using R-formula,

$$x(t)=R\cos(\omega t+\delta)$$

where amplitude $R=\sqrt{c_1^2+c_2^2}$, $\cos\delta=\frac{c_1}{R}$, $\sin\delta=-\frac{c_2}{R}$, δ is known as phase angle. Period $T=\frac{2\pi}{\omega}=2\pi\sqrt{\frac{m}{k}}$.

Example 25.3.2 (Damping vibrations)**.** Damping force is given by $F=-c\frac{\mathrm{d}x}{\mathrm{d}t}$. By Newton's 2nd law, $m\frac{\mathrm{d}^2x}{\mathrm{d}t^2}=-kx-c\frac{\mathrm{d}x}{\mathrm{d}t}$, or

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2}+c\frac{\mathrm{d}x}{\mathrm{d}t}+kx=0$$

which has auxilliary equation $mr^2+cr+k=0$.

- Case 1: $c^2-4mk>0$ (over-damping)

  Solution is
  $$x=c_1e^{r_1t}+c_2e^{r_2t}.$$

- Case 2: $c^2-4mk=0$ (critical damping)

  Solution is
  $$x=(c_1+c_2)e^{rt}.$$

- Case 3: $c^2-4mk<0$ (under-damping)

  Roots of auxilliary equation are $r=-\frac{c}{2m}\pm\omega i$ where $\omega=\frac{\sqrt{4mk-c^2}}{2m}$.

  Solution is
  $$x=e^{-\frac{c}{2m}t}(c_1\cos\omega t+c_2\sin\omega t).$$

  Amplitude follows the graph
  $$x=\pm Ae^{-\frac{c}{2m}t}$$
  where $A=\sqrt{c_1^2+c_2^2}$.

For damped forced vibrations,

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2}+c\frac{\mathrm{d}x}{\mathrm{d}t}+kx=F(t)$$

where $F(t)$ is external force.

Example 25.3.3 (Electric circuits)**.**

$$L\frac{\mathrm{d}I}{\mathrm{d}t}+RI+\frac{Q}{C}=E$$

where L is inductance, R is resistance, C is capacitance, Q is charge, $I=\frac{\mathrm{d}Q}{\mathrm{d}t}$ is current, $E(t)$ is e.m.f. Rewriting and taking time derivative,

$$L\frac{\mathrm{d}^2I}{\mathrm{d}t^2}+R\frac{\mathrm{d}I}{\mathrm{d}t}+\frac{1}{C}I=E'(t).$$

Population Dynamics and Population Growth Models • exponential growth model • logistic growth model, equilibrium points and their stability, and harvesting

To look for info from F Maths notes

Part VII

Calculus (Multivariable)

26 Partial differentiation

§26.1 Computation of partial derivatives

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function of n variables $x_1, x_2, \ldots, x_n$. Then the **partial derivative**

$$\frac{\partial f}{\partial x_i}(p_1, \ldots, p_n)$$

is the rate of change of f, at $(p_1, \ldots, p_n)$, when we vary only the variable x_i about p_i and keep all of the other variables constant. Precisely, we have

$$\frac{\partial f}{\partial x_i}(p_1, \ldots, p_n) = \lim_{h \to 0} \frac{f(p_1, \ldots, p_{i-1}, p_i + h, p_{i+1}, \ldots, p_n) - f(p_1, \ldots, p_n)}{h}. \tag{26.1}$$

Notation. We shall occasionally write f_x for $\delta f / \delta x$, etc.

Derivatives such as eq. (26.1), where f has been differentiated once, are called **first order** partial derivatives. We will look at higher orders in a moment.

Exercise 26.1.1

Find all the first order derivatives of

$$f(x, y, z) = x^2 + ye^{2x} + \frac{z}{y}.$$

Solution. Keep in mind that we only need to find the derivative of functions with respect to one variable by keeping the rest of the variables constant.

Thus we have

$$\frac{\partial f}{\partial x} = 2x + 2ye^{2x}, \quad \frac{\partial f}{\partial y} = e^{2x} - \frac{z}{y^2}, \quad \frac{\partial f}{\partial z} = \frac{1}{y}.$$

□

We define second and higher order partial derivatives in a similar manner to how we define them for full derivatives. So, in the case of second order partial derivatives of a function $f(x, y)$ we have

$$\begin{aligned}
f_{xx} &= \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial x^2} \\
f_{yy} &= \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial y^2} \\
f_{xy} &= \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial x\, \partial y} \\
f_{yx} &= \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial y\, \partial x}
\end{aligned}$$

Observe that

$$\frac{\partial^2 f}{\partial y\,\partial x} = \frac{\partial^2 f}{\partial x\,\partial y}, \quad \frac{\partial^2 f}{\partial z\,\partial x} = \frac{\partial^2 f}{\partial x\,\partial z}, \quad \frac{\partial^2 f}{\partial z\,\partial y} = \frac{\partial^2 f}{\partial y\,\partial z}.$$

This will typically be the case in the examples we will see in this course, but it is not guaranteed unless the derivatives in question are continuous.

we might have a function $f(u,v)$ of two variables u and v, each of which is itself a function of x and y. We can make the composition

$$F(x,y) = f(u(x,y), v(x,y))$$

which is a function of x and y, and we might then want to calculate the partial derivatives

$$\frac{\partial F}{\partial x} \text{ and } \frac{\partial F}{\partial y}.$$

Theorem 26.1.1 (Chain rule)**.** Let $F(t) = f(u(t), v(t))$ with u and v differentiable and f being continuously differentiable in each variable. Then

$$\frac{\mathrm{d}F}{\mathrm{d}t} = \frac{\partial f}{\partial u}\frac{\partial u}{\partial t} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial t} \tag{26.2}$$

Corollary 26.1.1. Let $F(x,y) = f(u(x,y), v(x,y))$ with u and v differentiable in each variable and f being continuously differentiable in each variable. Then

$$\frac{\partial F}{\partial x} = \frac{\partial f}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial x}, \quad \frac{\partial F}{\partial y} = \frac{\partial f}{\partial u}\frac{\partial u}{\partial y} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial y}.$$

Exercise 26.1.2

A particle P moves in three dimensional space on a helix so that at time t,

$$x(t) = \cos t, \quad y(t) = \sin t, \quad z(t) = t.$$

The temperature T at (x,y,z) equals $xy + yz + zx$. Use the chain rule to calculate $\frac{\mathrm{d}T}{\mathrm{d}t}$.

Solution. The chain rule in this case says that

$$\begin{aligned}\frac{\mathrm{d}T}{\mathrm{d}t} &= \frac{\partial T}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial T}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t} + \frac{\partial T}{\partial z}\frac{\mathrm{d}z}{\mathrm{d}t}\\ &= (y+z)(-\sin t) + (x+z)\cos t + (y+x)(1)\\ &= (\sin t + t)(-\sin t) + (\cos t + t)\cos t + (\cos t + \sin t)\end{aligned}$$

$$= -\sin^2 t + \cos^2 t + \sin t + \cos t + t\cos t - t\sin t.$$

□

Exercise 26.1.3

Let $z = f(xy)$, where f is an arbitrary differentiable function in one variable. Show that

$$x\frac{\partial z}{\partial x} - y\frac{\partial z}{\partial y} = 0.$$

Solution. By the chain rule,

$$\frac{\partial z}{\partial x} = yf'(xy) \quad \text{and} \quad \frac{\partial z}{\partial y} = xf'(xy),$$

where the prime denotes the derivative with respect to xy. Hence we have

$$x\frac{\partial z}{\partial x} - y\frac{\partial z}{\partial y} = xyf'(xy) - yxf'(xy) = 0.$$

□

§26.2 Directional Derivatives

To this point we've only looked at the two partial derivatives $f_x(x,y)$ and $f_y(x,y)$. Recall that these derivatives represent the rate of change of f as we vary x (holding y fixed) and as we vary y (holding x fixed) respectively.

We now discuss how to find the rate of change of f if we allow both x and y to change simultaneously. The problem here is that there are many ways to allow both x and y to change. For instance, one could be changing faster than the other and then there is also the issue of whether or not each is increasing or decreasing. So, before we get into finding the rate of change we need to get a couple of preliminary ideas taken care of first. The main idea that we need to look at is just how are we going to define the changing of x and/or y.

Let's start off by supposing that we wanted the rate of change of f at a particular point, say (x_0, y_0). Let's also suppose that both x and y are increasing and that, in this case, x is increasing twice as fast as y is increasing. So as y increases one unit of measure, x increases two units of measure.

Let's suppose that a particle is sitting at (x_0, y_0) and the particle will move in the direction given by the changing x and y. At this point, the particle can be said to be moving in the direction

$$\vec{v} = \langle 2, 1\rangle$$

There is still a small problem with this however. There are many vectors that point in the same direction. For instance, all of the following vectors point in the same direction as $\vec{v} = \langle 2, 1\rangle$:

$$\vec{v} = \left\langle \frac{1}{5}, \frac{1}{10} \right\rangle \quad \vec{v} = \langle 6, 3\rangle \quad \vec{v} = \left\langle \frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}} \right\rangle$$

We need a way to consistently find the rate of change of a function in a given direction. We will do this by insisting that the vector that defines the direction of change be a unit vector. This means that for the example that we started off thinking about we would want to use

$$\vec{v} = \left\langle \frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}} \right\rangle$$

Definition 26.2.1: Directional derivative

Rate of change of $f(x,y)$ in the direction of the unit vector $\vec{u} = \langle a, b\rangle$ is called the directional derivative and is denoted by $D_{\vec{u}} f(x,y)$.
The definition of the directional derivative is

$$D_{\vec{u}} f(x,y) = \lim_{h\to 0} \frac{f(x+ah, y+bh) - f(x,y)}{h} \tag{26.3}$$

To derive an equivalent formula for taking directional derivatives, we define a new function of a single variable

$$g(z) = f(x_0 + az, y_0 + bz)$$

where x_0, y_0, a, b are some fixed numbers. Note that this really is a function of a single variable z.

Then by the definition of the derivative for functions of a single variable we have

$$g'(z) = \lim_{h\to 0} \frac{g(z+h) - g(z)}{h}$$

and the derivative at $z = 0$ is given by

$$g'(0) = \lim_{h\to 0} \frac{g(h) - g(0)}{h}$$

If we now substitute in for $g(z)$ we get

$$g'(0) = \lim_{h\to 0} \frac{g(h) - g(0)}{h} = \lim_{h\to 0} \frac{f(x_0 + ah, y_0 + bh) - f(x_0, y_0)}{h} = D_{\vec{u}} f(x_0, y_0)$$

This gives us

$$g'(0) = D_{\vec{u}}f(x_0, y_0) \tag{1}$$

Now, let's look at this from another perspective. Let's rewrite $g(z)$ as $g(z) = f(x, y)$ where $x = x_0 + az$ and $y = y_0 + bz$. Applying chain rule,

$$g'(z) = \frac{\mathrm{d}g}{\mathrm{d}z} = \frac{\partial f}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}z} + \frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}z} = f_x(x, y)a + f_y(x, y)b$$

This gives us

$$g'(z) = f_x(x, y)a + f_y(x, y)b$$

If we take $z = 0$ we get $x = x_0$ and $y = y_0$. Plugging these into the above equation gives

$$g'(0) = f_x(x_0, y_0)a + f_y(x_0, y_0)b \tag{2}$$

Equating (1) and (2) gives

$$D_{\vec{u}}f(x_0, y_0) = f_x(x_0, y_0)a + f_y(x_0, y_0)b$$

Allowing x and y to be any number we get the following formula for computing directional derivatives:

$$D_{\vec{u}}f(x, y) = f_x(x, y)a + f_y(x, y)b$$

For three variables, directional derivative of $f(x, y, z)$ in the direction of the unit vector $\vec{u} = \langle a, b, c\rangle$ is given by

$$D_{\vec{u}}f(x, y, z) = f_x(x, y, z)a + f_y(x, y, z)b + f_z(x, y, z)c \tag{26.4}$$

We can write the directional derivative as a **dot product** and notice that the second vector is nothing more than the unit vector $\vec{u}$ that gives the direction of change.

$$D_{\vec{u}}f(x, y, z) = \langle f_x, f_y, f_z\rangle \cdot \langle a, b, c\rangle \tag{26.5}$$

Now let's give a name and notation to the first vector in the dot product since this vector will show up fairly regularly.

Definition 26.2.2: Gradient vector

he gradient vector of f is defined to be

$$\nabla f = \langle f_x, f_y, f_z\rangle \tag{26.6}$$

With the definition of the gradient we can now say that the directional derivative is given by

$$D_{\vec{u}}f = \nabla f \cdot \vec{u}$$

Theorem 26.2.1

Maximum value of $D_{\vec{u}}f(\vec{x})$ (and hence then the maximum rate of change of the function $f(\vec{x})$) is given by $\|\nabla f(\vec{x})\|$ and will occur in the direction given by $\nabla f(\vec{x})$.

Proof. □

27 Partial differential equations

§27.1 Definitions and Terminology

Definition 27.1.1. A **partial differential equation** is an equation involving a function and/or its partial derivatives.

We can classify PDEs based on:

- **Order.**

  The order is the number corresponding to the order of the highest partial derivative in the equation.

  For instance, the order of the following PDE is 2.

  $$\frac{\partial^2 f}{\partial x^2} = \frac{\partial f}{\partial t}$$

  This also applies to mixed partial derivatives. For instance, the order of the following PDE is 3.

  $$\frac{\partial^3 f}{\partial x^2\, \partial y} = \frac{\partial f}{\partial t}$$

- **Number of independent variables.**

  An independent variable is what we differentiate with respect to.

- **Linearity.**

  A linear PDE is one in which the *dependent* variable (the one being differentiated) appears only in a linear fashion.

  For instance, the two PDEs above are linear as the partial derivatives are not being raised to a power or multiplied with each other.

  The following PDE is non-linear.

  $$f\frac{\partial^2 f}{\partial x^2} = \frac{\partial f}{\partial t}$$

- **Homogeneity.**

  A homogenous PDE is one in which every term only involves the dependent variable and/or its derivatives.

  The first two PDEs above are homogenous as every term contains f or its derivatives.

  The following PDE is non-homogenous as there are two terms that do not contain f.

  $$\frac{\partial^2 f}{\partial x^2} = \frac{\partial f}{\partial t} + x^2 + \tan t$$

- **Coefficient type.**

  The coefficient here refers to the coefficient of the term involving the dependent variable and its derivatives. It can be either constant or variable.

  For instance, the coefficients of the terms in the first two examples are 1. We say that these two PDEs have constant coefficients.

  The following PDE has variable coefficients.

  $$\tan x \frac{\partial^2 f}{\partial x^2} = \frac{\partial f}{\partial t}$$

- **Parabolic, Hyperbolic, or Elliptic.**

  We can do this classification for linear 2nd order PDEs which take the form of

  $$A\frac{\partial^2 f}{\partial x^2} + B\frac{\partial^2 f}{\partial x\,\partial y} + C\frac{\partial^2 f}{\partial y^2} + D\frac{\partial f}{\partial x} + E\frac{\partial f}{\partial y} + Ff = G$$

  where the coefficients are generally functions of x or y.

  For a **hyperbolic** PDE, $B^2 - 4AC > 0$. Using variable substitutions to change x and y to η and ε respectively, we can reduce the PDE to

  $$\frac{\partial^2 f}{\partial \eta^2} - \frac{\partial^2 f}{\partial \varepsilon^2} + g = 0$$

  where g denotes the first and lower order terms. This is similar to the equation of a hyperbola: $x^2 - y^2 = 1$.

  For a **parabolic** PDE, $B^2 - 4AC = 0$. Using variable substitutions, we can reduce the PDE to

  $$\frac{\partial^2 f}{\partial \eta^2} + g = 0.$$

  This is similar to the equation of a parabola: $x^2 + y = 0$.

  For an **elliptic** PDE, $B^2 - 4AC < 0$. Using variable substitutions, we can reduce the PDE to

  $$\frac{\partial^2 f}{\partial \eta^2} + \frac{\partial^2 f}{\partial \varepsilon^2} + g = 0.$$

  This is similar to the equation of an ellipse: $x^2 + y^2 = 1$.

  Note that if the coefficients are constants, the PDE can be hyperbolic, parabolic or elliptic. However, if the coefficients are variables, then it is possible for the PDE to be hyperbolic in some regions, and elliptic or parabolic in some regions.

§27.2 Solutions and Auxiliary Conditions

Exercise 27.2.1

Show that $f(x,y) = \tan^{-1}\frac{y}{x}$ satisfies Laplace's equation in the plane:

$$\frac{\partial[order=2]}{\partial f}x + \frac{\partial[order=2]}{\partial f}y = 0.$$

Solution. The first order partial derivatives are

$$\frac{\partial f}{\partial x} = \frac{1}{1+\left(\frac{y}{x}\right)^2}\left(-\frac{y}{x^2}\right) = -\frac{y}{x^2+y^2}$$

and

$$\frac{\partial f}{\partial y} = \frac{1}{1+\left(\frac{y}{x}\right)^2}\left(\frac{1}{x}\right) = \frac{x}{x^2+y^2}.$$

Hence we have
$$\frac{\partial[order = 2]}{\partial f}x = \frac{2xy}{x^2+y^2} \quad \text{and} \quad \frac{\partial[order = 2]}{\partial f}y = -\frac{2xy}{(x^2+y^2)^2},$$
from which we see that
$$\frac{\partial[order = 2]}{\partial f}x + \frac{\partial[order = 2]}{\partial f}y = 0$$
as required. □

In the above example we verified that a given solution satisfied a given PDE. We now look at how to find solutions to simple PDEs.

Exercise 27.2.2

Find all the solutions of the form $f(x,y)$ of the PDEs

(a) $\frac{\partial^2 f}{\partial y\,\partial x} = 0$,

(b) $\frac{\partial[order=2]}{\partial f}x = 0$.

Solution.

(a) We have
$$\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = 0.$$
Those functions $g(x,y)$ which satisfy $\delta g/\delta y = 0$ are functions which solely depend on x. So we have
$$\frac{\partial f}{\partial x} = p(x),$$
where p is an arbitrary function of x. We can now integrate again, but this time with respect to x rather than y. Now, $\delta/\delta x$ sends to zero any function which solely depends on y. The solution to the PDE is therefore
$$f(x,y) = P(x) + Q(y),$$
where $Q(y)$ is an arbitrary function of y and $P(x)$ is an anti-derivative of $p(x)$, i.e. $P'(x) = p(x)$.

□

If a PDE involves derivatives with respect to one variable only, we can treat it like an ODE in that variable, holding all other variables constant. The difference, as noted above, is that our arbitrary "constants" will now be arbitrary functions of the variables that we have held constant.

Exercise 27.2.3

Find solutions $u(x,y)$ of the PDE
$$u_{xx} - u = 0.$$

Solution. Since there are no derivatives with respect to y, we can solve the associated ODE
$$\frac{\mathrm{d}^2u}{\mathrm{d}x^2} - u = 0,$$
where u is treated as being a function of x only. This ODE has solution $u(x) = C_1e^x + C_2e^{-x}$, where C_1 and C_2 are constants, and so the solution to the original PDE is
$$u(x,y) = A(y)e^x + B(y)e^{-x},$$
where A and B are arbitrary functions of y only. □

we look at one specific method for solving PDEs, that of **separating the variables**.

Exercise 27.2.4

Find all solutions of the form $T(x,t) = A(x)B(t)$ to the one-dimensional heat/diffusion equation

$$\frac{\partial T}{\partial t} = \kappa \frac{\partial[order = 2]}{\partial T} x,$$

where κ is a positive constant, called the thermal diffusivity.

Solutions of the form $A(x)B(t)$ are known as separable solutions

There are a lot of solutions to a given PDE, hence it is important for us to know the auxiliary conditions, i.e. boundary and initial conditions, which dictate which technique we use to solve the PDE.

- A boundary condition expresses the behavior of a function on the boundary (border) of its area of definition. An initial condition is like a boundary condition, but then for the time-direction.

28 Multiple Integrals

§28.1 Double Integrals

To motivate the idea of double integrals, we give an example of calculating areas.

We want to integrate a function of two variables, $f(x,y)$. With functions of one variable we integrated over an *interval* (i.e. a one-dimensional space) and so it makes some sense then that when integrating a function of two variables we will integrate over a *region* of $\mathbb{R}^2$ (two-dimensional space).

Exercise 28.1.1

Calculate the area of the disc $x^2 + y^2 \le a^2$.

Solution. we know the answer, namely πa^2. If we wish to capture all of the disc's area then we let x vary from $-a$ to a and, at each x we let y vary from $-\sqrt{a^2 - x^2}$ to $\sqrt{a^2 - x^2}$.

Thus we have

$$\begin{aligned} A &= \int_{x=-a}^{x=a} \int_{y=-\sqrt{a^2-x^2}}^{y=\sqrt{a^2-x^2}} \mathrm{d}y \, \mathrm{d}x \\ &= \int_{x=-a}^{x=a} 2\sqrt{a^2 - x^2} \, \mathrm{d}x \\ &= \int_{\theta=-\frac{\pi}{2}}^{\theta=\frac{\pi}{2}} 2\sqrt{a^2 - a^2 \sin^2\theta} a \cos\theta \, \mathrm{d}\theta \qquad [x = a\sin\theta] \\ &= a^2 \int_{\theta=-\frac{\pi}{2}}^{\theta=\frac{\pi}{2}} 2\cos^2\theta \, \mathrm{d}\theta = \pi a^2 \end{aligned}$$

□

Definition 28.1.1. Let $R \subset \mathbb{R}^2$. Then we define the **area** of R to be

$$A(R) = \iint_{(x,y)\in R} \mathrm{d}x \, \mathrm{d}y \, .$$

Recall that given a function $f(x)$, the definite integral $\int_a^b f(x)\,\mathrm{d}x$ evaluates the area under the curve $y = f(x)$ between $x = a$ and $x = b$. Similarly given a function $f(x,y)$, the volume of the solid below the graph $z = f(x,y)$ and above the region R in the xy-plane is given by

$$V = \iint_R f(x,y) \, \mathrm{d}A \, . \tag{28.1}$$

Remark. You can think of $\iint$ as a sum of heights $f(x,y)$ and areas $\mathrm{d}A$, which evaluates to give a volume.

Exercise 28.1.2

Let R be the triangle whose vertices are $(0,0,0)$, $(0,1,0)$ and $(1,0,0)$. Evaluate $\iint_R 1 \, \mathrm{d}A$.

Solution. $\iint_R 1\,\mathrm{d}A$ is the volume of a prism with height 1 and base of area R. Hence

$$\iint_R 1\,\mathrm{d}A = \text{Area of } R \times 1 = \frac{1}{2} \times 1 \times 1 \times 1 = \boxed{\frac{1}{2}}$$

□

To evaluate double integrals we need a notion called iterated integrals, which are basically two integrals with one nested inside the other.

Exercise 28.1.3

Evaluate

(a) $\int_0^1 \int_1^2 x + 2y\,\mathrm{d}x\,\mathrm{d}y$

(b) $\int_{-1}^1 \int_0^x 3x^2 + 2y\,\mathrm{d}y\,\mathrm{d}x$

Solution. (a)

$$\begin{aligned}\int_0^1 \int_1^2 x + 2y\,\mathrm{d}x\,\mathrm{d}y &= \int_0^1 \left(\int_1^2 x + 2y\,\mathrm{d}x\right)\mathrm{d}y \\ &= \int_0^1 \left[\frac{x^2}{2} + 2yx\right]_{x=1}^{x=2} \mathrm{d}y \\ &= \int_0^1 2y + \frac{3}{2}\,\mathrm{d}y \\ &= \left[y^2 + \frac{3}{2}y\right]_{y=0}^{y=1} = \boxed{\frac{5}{2}}\end{aligned}$$

(b)

$$\begin{aligned}\int_{-1}^1 \int_0^x 3x^2 + 2y\,\mathrm{d}y\,\mathrm{d}x &= \int_{-1}^1 \left(\int_0^x 3x^2 + 2y\,\mathrm{d}y\right)\mathrm{d}x \\ &= \int_{-1}^1 \left[3x^2y + y^2\right]_{y=0}^{y=x} \mathrm{d}x \\ &= \int_{-1}^1 3x^3 + x^2\,\mathrm{d}x \\ &= \left[\frac{3x^4}{4} + \frac{x^3}{3}\right]_{x=-1}^{x=1} = \boxed{\frac{2}{3}}\end{aligned}$$

□

§28.2 Iterated Integrals

Now we are going to discuss the relation between double integrals and iterated integrals.

R is called simple if R is a rectangle given by $a \le x \le b$ and $c \le y \le d$, in which case

$$\iint_R f(x,y)\,\mathrm{d}A = \int_a^b \int_c^d f(x,y)\,\mathrm{d}y\,\mathrm{d}x = \int_c^d \int_a^b f(x,y)\,\mathrm{d}x\,\mathrm{d}y\,.$$

This is known as Fubini's Theorem.

R is called vertically simple if R is given by $a \le x \le b$, $g(x) \le y \le h(x)$, in which case

$$\iint_R f(x,y)\,\mathrm{d}A = \int_a^b \int_{g(x)}^{h(x)} f(x,y)\,\mathrm{d}y\,\mathrm{d}x\,.$$

R is called horizontally simple if R is given by $c \le y \le d$, $g(y) \le x \le h(y)$, in which case

$$\iint_R f(x,y)\,\mathrm{d}A = \int_c^d \int_{g(y)}^{h(y)} f(x,y)\,\mathrm{d}x\,\mathrm{d}y\,.$$

29 Line Integrals

§29.1 Vector fields

A vector field is basically what you get when associating each point in space with a vector.

Definition 29.1.1: Vector field

A **vector field** in $\mathbb{R}^n$ is a function $F : \mathbb{R}^n \to \mathbb{R}^n$ that assigns to each $x \in \mathbb{R}^n$ a vector $F(x)$. A vector field in $\mathbb{R}^n$ with domain $U \subset \mathbb{R}^n$ is called a vector field on U.

The standard notation for the function $\vec{F}$ is:

$$\vec{F}(x,y) = P(x,y)\hat{i} + Q(x,y)\hat{j}$$
$$\vec{F}(x,y,z) = P(x,y,z)\hat{i} + Q(x,y,z)\hat{j} + R(x,y,z)\hat{k}$$

depending on whether or not we're in two or three dimensions. The functions P, Q, R are called **scalar functions**.

Exercise 29.1.1

Sketch the following vector field:

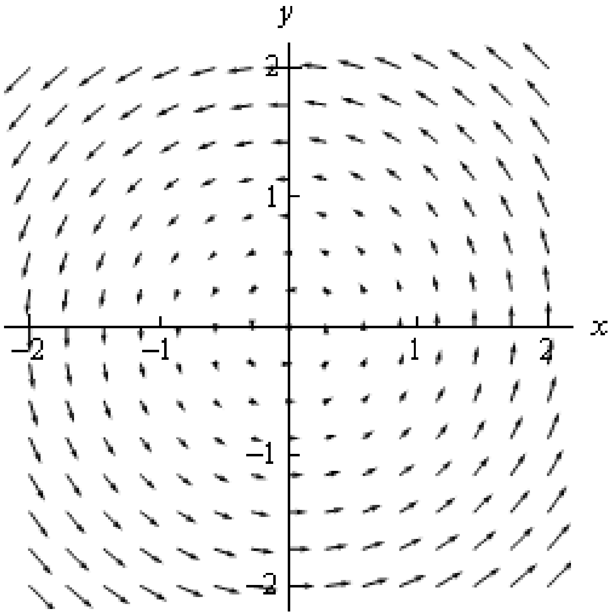
$$\vec{F}(x,y) = -y\hat{i} + x\hat{j}$$

Solution. To graph the vector field we need to get some "values" of the function. This means plugging in some points into the function. Here are a couple of evaluations:

$$\vec{F}\left(\frac{1}{2},\frac{1}{2}\right) = -\frac{1}{2}\hat{i} + \frac{1}{2}\hat{j}$$
$$\vec{F}\left(\frac{1}{2},-\frac{1}{2}\right) = -\left(-\frac{1}{2}\right)\hat{i} + \frac{1}{2}\hat{j} = \frac{1}{2}\hat{i} + \frac{1}{2}\hat{j}$$
$$\vec{F}\left(\frac{3}{2},\frac{1}{4}\right) = -\frac{1}{4}\hat{i} + \frac{3}{2}\hat{j}$$

So what do these evaluations tell us? The first one tells us that at the point $\left(\frac{1}{2},\frac{1}{2}\right)$ we plot the vector $-\frac{1}{2}\hat{i} + \frac{1}{2}\hat{j}$.

Plotting points gives us the following sketch of the vector field:



□

Definition 29.1.2: Gradient vector

Given a function $f(x, y, z)$, the gradient vector is defined by

$$\nabla f = \langle f_x, f_y, f_z \rangle \tag{29.1}$$

This is a vector field and is often called a gradient vector field.

§29.2 Types of line integrals

§29.3 Fundamental Theorem for Line Integrals

Theorem 29.3.1: Fundamental Theorem of Line Integrals

Suppose that C is a smooth curve from points A to B parameterised by $\mathbf{r}(t)$ for $t \in [a, b]$. Let f be a differentiable function whose domain includes C and whose gradient vector ∇f is continuous on C. Then

$$\int_C \nabla f \, \mathrm{d}\mathbf{r} = f(\mathbf{r}(b)) - f(\mathbf{r}(a)) = f(B) - f(A) \tag{29.2}$$

Remark. Similar to the fundamental theorem of calculus, the primary change is that gradient ∇f takes the place of the derivative f'.

§29.4 Conservative Vector Fields

§29.5 Green's Theorem

to compute arc lengths, areas of curves

applications of integrals to find area and volume

30 Surface Integrals

Part VIII

Differential Geometry

Readings:

- Introduction to Differentiable Manifolds and Riemannian Geometry
- Differential Geometry of Curves and Surfaces