Undergraduate Mathematics

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/undergrad-maths.

Typeset using LaTeX.

Last updated January 28, 2025.

Preface

Part II covers **abstract algebra**, which follows [DF04; Art11]. Chapter 3 introduces groups; **??** introduces rings.

Part III covers **linear algebra**, which follows [Axl24]. Chapter 4 introduces vector spaces, subspaces, span, linear independence, bases and dimension. Chapter 5 concerns linear maps and related concepts.

Part IV covers **real analysis**, which follows [Rud76; Apo57; BS11]. The lecture series by Professor Francis Su is very helpful; [Alc14] is also a good read to get some intuition into some abstract notions. Chapter 8 introduces the real and complex number systems; Chapter 9 covers basic topology required for subsequent chapters; Chapter 10 and **??** cover numerical sequences and series, and sequences and series of functions respectively; **??** covers continuity of functions; **??** and **??** cover differentiation and Riemann–Stieljes integration respectively; **??** covers some special functions such as power series, exponential and logarithmic functions, trigonometric functions, fourier series and the gamma function.

The reader is not assumed to have any mathematical prerequisites, although some experience with proofs may be helpful. **Preliminary topics** such as logic and methods of proofs (Chapter 1), and basic set theory (Chapter 2) are covered in the appendix.

Note on Presentation

The following are some common mathematical terms used in this book. They are neither exhaustive nor rigorous, but they should give you a good idea of what is meant when these terms are used. The following terms are **bolded** when they are used, for readability.

- **Definition**: 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.
- **Theorem**: a mathematical statement that is proved using rigorous mathematical reasoning. It is often reserved for the most important results.
- **Lemma**: 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 (Zorn's lemma, Urysohn's lemma, Burnside's lemma, Sperner's lemma).
- **Corollary**: 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").
- **Proposition**: a proved and often interesting result, but generally less important than a theorem.
- **Conjecture**: a statement that is unproved, but is believed to be true (Collatz conjecture, Goldbach conjecture, twin prime conjecture).

- **Claim**: an assertion that is then proved. It is often used like an informal lemma.
- **Axiom/Postulate**: a statement that is assumed to be true without proof. These are the basic building blocks from which all theorems are proved (Euclid's five postulates, Zermelo–Fraenkel axioms, Peano axioms).
- **Identity**: a mathematical expression giving the equality of two (often variable) quantities (trigonometric identities, Euler's identity).
- **Paradox**: 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 (Russell's paradox). The term paradox is often used informally to describe a surprising or counterintuitive result that follows from a given set of rules (Banach–Tarski paradox, Alabama paradox, Gabriel's horn).

Important terms are ***coloured*** when they are first defined, and are included in the glossary at the end of the book. Less important terms are instead *italicised* when they are first defined, and are not included in the glossary.

Note on Problem Solving

Mathematics is about problem solving. In [Pól45], George Pólya outlined the following problem solving cycle.

1. **Understand the problem**

   Ask yourself the following questions:

   - Do you understand all the words used in stating the problem?
   - Is it possible to satisfy the condition? Is the condition sufficient to determine the unknown? Or is it insufficient? Or redundant? Or contradictory?
   - What are you asked to find or show? Can you restate the problem in your own words?
   - Draw a figure. Introduce suitable notation.
   - Is there enough information to enable you to find a solution?

2. **Devise a plan**

   A partial list of heuristics – good rules of thumb to solve problems – is included:

   - Guess and check
   - Look for a pattern
   - Make an orderly list
   - Draw a picture
   - Eliminate possibilities
   - Solve a simpler problem
   - Use symmetry
   - Use a model
   - Consider special cases
   - Work backwards
   - Use direct reasoning
   - Use a formula

- Solve an equation
- Be ingenious

3. **Execute the plan**

   This step is usually easier than devising the plan. In general, all you need is care and patience, given that you have the necessary skills. Persist with the plan that you have chosen. If it continues not to work discard it and choose another. Don't be misled, this is how mathematics is done, even by professionals.

   - Carrying out your plan of the solution, check each step. Can you see clearly that the step is correct? Can you prove that it is correct?

4. **Check and expand**

   Pólya mentions that much can be gained by taking the time to reflect and look back at what you have done, what worked, and what didn't. Doing this will enable you to predict what strategy to use to solve future problems.

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

   - Can you check the result? Can you check the argument?
   - Can you derive the solution differently? Can you see it at a glance?
   - Can you use the result, or the method, for some other problem?

Building on Pólya's problem solving strategy, Schoenfeld [Sch92] came up with the following framework for problem solving, consisting of four components:

1. **Cognitive resources**: the body of facts and procedures at one's disposal.
2. **Heuristics**: 'rules of thumb' for making progress in difficult situations.
3. **Control**: having to do with the efficiency with which individuals utilise the knowledge at their disposal. Sometimes, this is referred to as metacognition, which can be roughly translated as 'thinking about one's own thinking'.

   (a) These are questions to ask oneself to monitor one's thinking.
   - What (exactly) am I doing? [Describe it precisely.] Be clear what I am doing NOW. Why am I doing it? [Tell how it fits into the solution.]
   - Be clear what I am doing in the context of the BIG picture – the solution. Be clear what I am going to do NEXT.

   (b) Stop and reassess your options when you
   - cannot answer the questions satisfactorily [probably you are on the wrong track]; OR
   - are stuck in what you are doing [the track may not be right or it is right but it is at that moment too difficult for you].

   (c) Decide if you want to
   - carry on with the plan,

- abandon the plan, OR
- put on hold and try another plan.

4. **Belief system**: one's perspectives regarding the nature of a discipline and how one goes about working on it.

Contents

I

Preliminary Topics

1 Mathematical Reasoning and Logic

Learning Outcomes

In this chapter, we will

- introduce basic logic;
- introduce common methods of proof.

§1.1 Zeroth-order Logic

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 do some algebra on propositions, which include

(i) ***equivalence***, denoted by $P \iff Q$, 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.

Proposition 1.1.

(i) Double negation law: $P \iff \neg(\neg P)$.

(ii) Commutative: $P \wedge Q \iff Q \wedge P$, $P \vee Q \iff Q \vee P$.

(iii) Conjunction is associative: $(P \wedge Q) \wedge R \iff P \wedge (Q \wedge R)$.

(iv) Disjunction is associative: $(P \vee Q) \vee R \iff P \vee (Q \vee R)$.

(v) Conjunction distributes over disjunction: $P \wedge (Q \vee R) \iff (P \wedge Q) \vee (P \wedge Q)$.

(vi) Disjunction distributes over conjunction: $P \vee (Q \wedge R) \iff (P \vee Q) \wedge (P \vee R)$.

Proposition 1.2 (de Morgan's laws)**.**

$$\neg(P \vee Q) \iff (\neg P \wedge \neg Q)$$

$$\neg(P \wedge Q) \iff (\neg P \vee \neg Q)$$

If, only if

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". $P \implies Q$ is known as a *conditional statement*, where P is known as the *hypothesis* and Q is known as the *conclusion*. The only case when $P \implies Q$ is false is when the hypothesis P is true and the conclusion Q is false.

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.

Given $P \implies Q$,

- its ***converse*** is $Q \implies P$; both are not logically equivalent;
- its ***inverse*** is $\neg P \implies \neg Q$, i.e. the hypothesis and conclusion of the statement are both negated; both are not logically equivalent;
- the ***contrapositive*** is $\neg Q \implies \neg P$; both are logically equivalent.

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).

If and only if, iff

Bidirectional implication is denoted by $P \iff Q$, which means both $P \implies Q$ and $Q \implies P$; $P \iff Q$ is known as a *biconditional statement*. 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".

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

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

§1.2 First-order Logic

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

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

Proposition 1.3 (de Morgan's laws)**.**

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

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

Exercise

Negate the statement

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

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

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

□

Exercise

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 $f(x) \neq y$.

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

□

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 $x \in X$.

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.

§1.3 Methods of 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.

A ***proof by contrapositive*** of $P \implies Q$ is to prove instead $\neg Q \implies \neg P$.

A ***disproof by counterexample*** is to providing a counterexample in order to refute or disprove a conjecture. The counterexample must make the hypothesis a true statement, and the conclusion a false statement. 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.

A ***proof by cases*** is to first dividing the situation into cases which exhaust all the possibilities, and then show that the statement follows in all cases.

Proof by Contradiction

A ***proof by contradiction*** of P involves first supposing P is false, i.e. $\neg P$; to prove $P \implies Q$ by contradiction, suppose $P \wedge \neg Q$. 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 (Irrationality of $\sqrt{2}$)

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

Solution. We prove by contradiction. Suppose otherwise, that $\sqrt{2}$ is rational. Then $\sqrt{2} = \dfrac{a}{b}$ for some $a, b \in \mathbf{Z}, b \neq 0$, a, b coprime.

Squaring both sides gives

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

Since RHS is even, LHS must also be even. Hence it follows that a is even. Let $a = 2k$ where $k \in \mathbf{Z}$. 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 contradicts the assumption that a and b coprime, so we are done. □

Exercise (Euclid)

Prove that there are infinitely many prime numbers.

Solution. Suppose otherwise, that only finitely many prime numbers exist. List them as $p_1, \ldots, p_n$. The number $N = p_1 p_2 \cdots p_n + 1$ is divisible by a prime p, yet is coprime to $p_1, \ldots, p_n$. Therefore, p does not belong to our list of all prime numbers, a contradiction. □

To ***prove uniqueness***, we can either assume $\exists x, y \in S$ such that $P(x) \wedge P(y)$ is true and show $x = y$, or argue by assuming that $\exists x, y \in S$ are distinct such that $P(x) \wedge P(y)$, then derive a contradiction. $\exists!$ denotes "there exists a unique". To prove uniqueness and existence, we also need to show that $\exists x \in S$ s.t. $P(x)$ is true.

Proof of Existence

To prove existential statements, we can adopt two approaches:

1. Constructive proof (direct 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

   (a) belongs to the domain X, and

   (b) satisfies the condition P.

2. Non-constructive proof (indirect 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**
>
> 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 $n, n+1, n+2, \ldots, n+99$, 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 $k = 1, 2, \ldots, 99$.

Hence the existential statement is proven. □

> **Exercise**
>
> 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 $\mathbf{Q}$ is closed under addition).

□

Exercise

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 $r = \dfrac{p+q}{2}$; 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 $c = \dfrac{q-p}{k}$ for some $k > 1$; to make c irrational, we take k to be irrational.

Claim. $r = p + \dfrac{q-p}{\sqrt{2}}$.

We shall show that (i) $p < r < q$, and (ii) r is irrational.

(i) Since $q - p > 0$, $\dfrac{q-p}{\sqrt{2}} > 0$ so $r = p + \dfrac{q-p}{\sqrt{2}} > p + 0 = p$.

$\dfrac{q-p}{\sqrt{2}} < q - p$ so $r < p + (q - p) = q$.

(ii) We prove by contradiction. Suppose r is rational. We have $\sqrt{2} = \dfrac{q-p}{r-p}$. 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, which is a contradiction.

□

Non-constructive proof:

Exercise

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.

Since there can only be a finite number of d_i's between 1 and n, this process must stop after finite steps. 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 on $\mathbf{N}$.

Exercise

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. □

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.

The *well-ordering principle* on $\mathbf{N}$ states the following: every non-empty subset $S \subset \mathbf{N}$ has a smallest element; that is, there exists $m \in S$ such that $m \leqslant k$ for all $k \in S$.

The *principle of induction* states the following: Let $S \subset \mathbf{N}$. If (i) $1 \in S$, and (ii) $k \in S \implies k+1 \in S$, then $S = \mathbf{N}$.

Lemma 1.4. The well-ordering principle is equivalent to the principle of induction.

Proof.

$\boxed{\implies}$ Suppose otherwise, for a contradiction, that S exists with the given properties in the principle of induction, but $S \neq \mathbf{N}$.

Consider the set $\mathbf{N} \setminus S$. Then $\mathbf{N} \setminus S$ is not empty. By the well-ordering principle, $\mathbf{N} \setminus S$ has a least element p. Since $1 \in S$, $1 \notin \mathbf{N} \setminus S$ so $p \neq 1$, thus we must have $p > 1$.

Now consider $p-1$. Since p is the least element of $\mathbf{N} \setminus S$, $p-1 \notin \mathbf{N} \setminus S$ so $p-1 \in S$. But by (ii) of the principle of induction, $p-1 \in S$ implies $p \in S$, which contradicts the fact that $p \in \mathbf{N} \setminus S$.

$\boxed{\impliedby}$ Suppose the principle of induction is true. Then this implies that Theorem 1.5 is true, which in turn implies that Theorem 1.7 is true. In order to prove the well-ordering of $\mathbf{N}$, we prove the following statement $P(n)$ by strong induction on n: If $S \subset \mathbf{N}$ and $n \in S$, then S has a least element.

The basis step is true, because if $1 \in S$ then 1 is the smallest element of S, since there are no smaller elements of $\mathbf{N}$.

Now suppose that $P(k)$ is true for $k = 1, \ldots, n$. To show that $P(n+1)$ is true, let $S \subset \mathbf{N}$ contain $n+1$. If $n+1$ is the smallest element of S, then we are done. Otherwise, S has a smaller element k, and $P(k)$ is true by the inductive hypothesis, so again S has a smallest element.

Hence by strong induction, $P(n)$ is true for all $n \in \mathbf{N}$. This implies the well-ordering of $\mathbf{N}$, because if S is a non-empty subset of $\mathbf{N}$, then pick $n \in S$. Since $n \in \mathbf{N}$, $P(n)$ is true, and therefore S has a smallest element. □

Theorem 1.5 (Principle of mathematical induction)**.** Let $P(n)$ be a family of statements indexed by $\mathbf{N}$. Suppose that

(i) $P(1)$ is true;

(ii) for all $k \in \mathbf{N}$, $P(k) \implies P(k+1)$.

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

(i) is known as the *base case*; (ii) is known as the *inductive step*, where we assume $P(k)$ to be true—this is called the *inductive hypothesis*—and show that $P(k+1)$ is true.

Proof. Apply the principle of induction to the set $S = \{n \in \mathbf{N} \mid P(n) \text{ is true}\}$. □

Exercise

Prove that for any $n \in \mathbf{N}$,

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

Proof. We induct on n. Let $P(n) : \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$.

Clearly $P(1)$ holds. Now suppose $P(k)$ holds for some $k \in \mathbf{N}$, $k \geqslant 1$; that is,

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

Adding $k+1$ to both sides,

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

thus $P(k+1)$ is true. Hence by induction, the result holds. □

Exercise (Bernoulli's inequality)

Let $x \in \mathbf{R}$, $x > -1$. Then for all $n \in \mathbf{N}$,

$$(1+x)^n \geqslant 1 + nx.$$

Proof. We prove by induction on n. Fix $x > -1$. Let $P(n) : (1+x)^n \geqslant 1 + nx$.

The base case $P(1)$ is clear. Suppose that $P(k)$ is true for some $k \in \mathbf{Z}^+$, $k \geqslant 1$. That is, $(1+x)^k \geqslant 1+kx$. Note that $1 + x > 0$, and $kx^2 \geqslant 0$ (since $k > 0$ and $x^2 \geqslant 0$). Then

$$\begin{aligned}(1+x)^{k+1} &= (1+x)(1+x)^k \\ &\geqslant (1+x)(1+kx) \quad \text{[induction hypothesis]} \\ &= 1 + (k+1)x + kx^2 \\ &\geqslant 1 + (k+1)x \quad [\because kx^2 \geqslant 0]\end{aligned}$$

so $P(k+1)$ is true. Hence by induction, the result holds. □

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

Corollary 1.6. Let $P(n)$ be a family of statements indexed by integers $n \geqslant N$ for $N \in \mathbf{Z}$. Suppose that

(i) $P(N)$ is true;

(ii) for all $k \geqslant N$, $P(k) \implies P(k+1)$.

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

Proof. Apply Theorem 1.5 to the statement $Q(n) = P(n+N)$ for $n \in \mathbf{N}$. □

Another variant on induction is when the inductive step relies on some earlier case(s) but not necessarily the immediately previous case.

Theorem 1.7 (Strong induction)**.** Let $P(n)$ be a family of statements indexed by $\mathbf{N}$. Suppose that

(i) $P(1)$ is true;

(ii) for all $k \in \mathbf{N}$, $P(1) \wedge \cdots \wedge P(k) \implies P(k+1)$.

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

Proof. Let $Q(n)$ be the statement "$P(k)$ holds for $k = 1, \ldots, n$". Then the conditions for the strong form are equivalent to (i) $Q(1)$ holds and (ii) for $n \in \mathbf{N}$, $Q(n) \implies Q(n+1)$. By Theorem 1.5, $Q(n)$ holds for all $n \in \mathbf{N}$, and hence $P(n)$ holds for all n. □

Exercise (Fundamental theorem of arithmetic)

Prove that 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 $n \geqslant 2$ be a natural number and suppose that $P(k)$ holds for all $k < n$.

- 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.

In both cases, $P(n)$ holds. Hence by strong induction, $P(n)$ is true for all $n \in \mathbf{N}$. □

The following is also another variant on induction.

Theorem 1.8 (Cauchy induction)**.** Let $P(n)$ be a family of statements indexed by $\mathbf{N}_{\geqslant 2}$. Suppose that

(i) $P(2)$ is true;

(ii) for all $k \in \mathbf{N}$, $P(k) \implies P(2k)$ and $P(k) \implies (k-1)$.

Then $P(n)$ is true for all $n \in \mathbf{N}_{\geqslant 2}$.

Exercise (AM–GM inequality)

Given $n \in \mathbf{N}$, prove that for positive reals $a_1, a_2, \ldots, a_n$,

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

Proof. Let $P(n) : \dfrac{a_1 + a_2 + \cdots + a_n}{n} \geqslant \sqrt[n]{a_1 a_2 \cdots a_n}$.

Base case $P(2)$ is true because

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

Next we show that $P(n) \implies P(2n)$

$$\frac{a_1 + a_2 + \cdots + a_{2n}}{2n} = \frac{\frac{a_1+a_2+\cdots+a_n}{n} + \frac{a_{n+1}+a_{n+2}+\cdots+a_{2n}}{n}}{2}$$

$$\frac{\frac{a_1+a_2+\cdots+a_n}{n} + \frac{a_{n+1}+a_{n+2}+\cdots+a_{2n}}{n}}{2} \geqslant \frac{\sqrt[n]{a_1 a_2 \cdots a_n} + \sqrt[n]{a_{n+1} a_{n+2} \cdots a_{2n}}}{2}$$

$$\frac{\sqrt[n]{a_1 a_2 \cdots a_n} + \sqrt[n]{a_{n+1} a_{n+2} \cdots a_{2n}}}{2} \geqslant \sqrt{\sqrt[n]{a_1 a_2 \cdots a_n}\, \sqrt[n]{a_{n+1} a_{n+2} \cdots a_{2n}}}$$

$$\sqrt{\sqrt[n]{a_1 a_2 \cdots a_n}\, \sqrt[n]{a_{n+1} a_{n+2} \cdots a_{2n}}} = \sqrt[2n]{a_1 a_2 \cdots a_{2n}}$$

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 $P(n) \implies P(n-1)$. By n-variable AM–GM, $\frac{a_1+a_2+\cdots+a_n}{n} \geqslant \sqrt[n]{a_1 a_2 \cdots a_n}$ 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,

$$\frac{a_1 + a_2 + \cdots + a_{n-1}}{n-1} \geqslant \sqrt[n]{a_1 a_2 \cdots a_{n-1} \cdot \frac{a_1 + a_2 + \cdots + a_{n-1}}{n-1}}$$

$$\Rightarrow \left(\frac{a_1 + a_2 + \cdots + a_{n-1}}{n-1}\right)^n \geqslant a_1 a_2 \cdots a_{n-1} \cdot \frac{a_1 + a_2 + \cdots + a_{n-1}}{n-1}$$

$$\Rightarrow \left(\frac{a_1 + a_2 + \cdots + a_{n-1}}{n-1}\right)^{n-1} \geqslant a_1 a_2 \cdots a_{n-1}$$

$$\Rightarrow \frac{a_1 + a_2 + \cdots + a_{n-1}}{n-1} \geqslant \sqrt[n-1]{a_1 a_2 \cdots a_{n-1}}$$

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

Pigeonhole Principle

Theorem 1.9 (Pigeonhole principle)**.** If $kn+1$ objects are distributed among n boxes, one of the boxes will contain at least $k+1$ objects.

Exercise (IMO 1972)

Prove that every set of 10 two-digit integer numbers has two disjoint subsets with the same sum of elements.

Solution. Let S be the set of 10 numbers. It has $2^{10} - 2 = 1022$ subsets that differ from both S and the empty set. They are the "pigeons".

If $A \subset S$, the sum of elements of A cannot exceed $91 + 92 + \cdots + 99 = 855$. The numbers between 1 and 855, which are all possible sums, are the "holes".

Because the number of "pigeons" exceeds the number of "holes", there will be two "pigeons" in the same "hole". Specifically, there will be two subsets with the same sum of elements. Deleting the common elements, we obtain two disjoint sets with the same sum of elements. □

Exercise (Putnam 2006)

Prove that for every set $X = \{x_1, x_2, \ldots, x_n\}$ of n real numbers, there exists a nonempty subset S of X and an integer m such that

$$\left| m + \sum_{x \in S} s \right| \leqslant \frac{1}{n+1}.$$

Solution. Recall that the fractional part of a real number x is $x - \lfloor x \rfloor$. Let us look at the fractional parts of the numbers $x_1, x_1 + x_2, \ldots, x_1 + x_2 + \cdots + x_n$. If any of them is either in the interval $\left[0, \frac{1}{n+1}\right]$ or $\left[\frac{n}{n+1}, 1\right]$, then we are done. If not, we consider these n numbers as the "pigeons" and the $n-1$ intervals $\left[\frac{1}{n+1}, \frac{2}{n+1}\right], \left[\frac{2}{n+1}, \frac{3}{n+1}\right], \ldots, \left[\frac{n-1}{n+1}, \frac{n}{n+1}\right]$ as the "holes". By the pigeonhole principle, two of these sums, say $x_1 + x_2 + \cdots + x_k$ and $x_1 + x_2 + \cdots + x_{k+m}$, belong to the same interval. But then their difference $x_{k+1} + \cdots + x_{k+m}$ lies within a distance of $\frac{1}{n+1}$ of an integer, and we are done. □

Exercises

Problem 1.1. 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 $a, b \in \mathbf{Z}$. 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. 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. □

Problem 1.2. 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: $p < \dfrac{p+q}{2} < q$.

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

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

Problem 1.3. Given n real numbers $a_1, a_2, \ldots, a_n$. Show that there exists an a_i $(1 \leqslant i \leqslant n)$ 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 1.4. 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 $\dfrac{\sqrt{2}}{a}$.

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

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

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

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

Problem 1.5. 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 $1 \leqslant i \leqslant m$. 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 1.6. Prove that, for any positive integer n, there is a perfect square m^2 (m is an integer) such that $n \leqslant m^2 \leqslant 2n$.

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

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

which simplifies to

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

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

Problem 1.7. Prove that for every positive integer $n \geqslant 4$,

$$n! > 2^n.$$

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

For the base case $P(4)$, the LHS equals to $4! = 4 \times 3 \times 2 \times 1 = 24$, and the RHS equals $2^4 = 16 < 24$. Since LHS equals RHS, $P(4)$ is true.

Now suppose that $P(k)$ is true for some $k \in \mathbf{N}_{\geqslant 4}$. Then

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

thus $P(k+1)$ is true, so we have shown $P(k) \implies P(k+1)$ for all $k \in \mathbf{N}_{\geqslant 4}$.

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

Problem 1.8. Prove by mathematical induction, for $n \geqslant 2$,

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

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

For the base case, when $n = 2$, $\sqrt{2} = 1.41\cdots < 2 - \dfrac{1}{2} = 1.5$ which is true. Hence $P(2)$ is true.

Now assume $P(k)$ is true for $k \geqslant 2, k \in \mathbf{N}$; that is,

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

We want to prove that $P(k+1)$ is true; that is,

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

Since $k > 2$, we have

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

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 1.9. Prove that for all integers $n \geqslant 3$,

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

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

Assume that $P(k)$ is true for some $k \in \mathbf{N}_{\geqslant 3}$; that 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 $1 \leqslant i \leqslant n$, 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 1.10. Let (a_n) be a sequence of integers defined recursively by the 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 $n \in \mathbf{N}$, prove that

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

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

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 $k \in \mathbf{N}$.

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

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

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

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

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

By inductive hypothesis, for $k \in \mathbf{N}$ we have $a_k \leqslant 2^k, a_{k+1} \leqslant 2^{k+1}, a_{k+2} \leqslant 2^{k+2}$.

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

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

Problem 1.11. For $m, n \in \mathbf{N}$, prove that

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

Proof. We induct on n. Let $P(n) : F_{n+m+1} = F_n F_m + F_{n+1} F_{m+1}$ for all $m \in \mathbf{N}$ in the cases $k = n$ and $k = n + 1$.

To show that $P(0)$ is true, 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 $F_1 = F_2 = 1$.

Now assume $P(n)$ is true; that is, for all $m \in \mathbf{N}$,

$$\begin{aligned} F_{n+m+1} &= F_n F_m + F_{n+1} F_{m+1}, \\ F_{n+m+2} &= F_{n+1} F_m + F_{n+2} F_{m+1}. \end{aligned}$$

Then

$$\begin{aligned} F_{n+m+3} &= F_{n+m+2} + F_{n+m+1} \\ &= F_n F_m + F_{n+1} F_{m+1} + F_{n+1} F_m + F_{n+2} F_{m+1} \\ &= (F_n + F_{n+1}) F_m + (F_{n+1} + F_{n+2}) F_{m+1} \\ &= F_{n+2} F_m + F_{n+3} F_{m+1} \end{aligned}$$

thus $P(n+1)$ is true, for all $m \in \mathbf{N}$. □

2 Set Theory

Learning Outcomes

In this chapter, we will

- recap basic definitions relating to sets (excluding detailed axiomatic discussions);
- define relations and related concepts including binary relation, partial order, total order, well order, equivalence relations, equivalence relations, equivalence class, quotient set, partition of a set;
- define functions, injectivity, surjectivity, bijectivity, composition, invertibility, monotonicity;

§2.1 Basics

Definitions and Notations

A ***set*** S can be loosely defined as a collection of objects[1]. For a set S, we write $x \in S$ to mean that x is an ***element*** of S, and $x \notin S$ if otherwise.

To describe a set, one can list its elements explicitly. A set can also 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:

- $\mathbf{N} = \{1, 2, 3, \dots\}$ denotes the natural numbers (non-negative integers).
- $\mathbf{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$ denotes the integers.
- $\mathbf{Q} = \left\{ \frac{p}{q} \,\middle|\, p, q \in \mathbf{Z}, q \neq 0 \right\}$ denotes the rational numbers.

[1] *Russell's paradox*, after the mathematician and philosopher Bertrand Russell (1872–1970), provides a warning as to the looseness of our definition of a set. Suppose H is the collection of sets that are not elements of themselves; that is,

$$H = \{S \mid S \notin S\}.$$

The problem arises when we ask the question of whether or not H is itself in H? On one hand, if $H \notin H$ then H meets the precise criterion for being in H and so $H \in H$, a contradiction. On the other hand, if $H \in H$ then by the property required for this to be the case, $H \notin H$, another contradiction. Thus we have a paradox: H is neither in H, nor not in H.

The modern resolution of Russell's paradox is that we have taken too naive an understanding of "collection", and that Russell's "set" H is in fact not a set. It does not fit within axiomatic set theory (which relies on the so-called ZF axioms), and so the question of whether or not H is in H simply doesn't make sense.

- $\mathbf{R}$ denotes the real numbers (the construction of which using Dedekind cuts will be discussed in Chapter 8).
- $\mathbf{C} = \{x + yi \mid x, y \in \mathbf{R}\}$ denotes the complex numbers.

We have that $\mathbf{N} \subset \mathbf{Z} \subset \mathbf{Q} \subset \mathbf{R} \subset \mathbf{C}$.

The ***empty set*** is the set with no elements, denoted by $\emptyset$.

A is a ***subset*** of B if every element of A is in B, denoted by $A \subset B$:

$$A \subset B \iff (\forall x)(x \in A \implies x \in B)$$

We denote $A \subsetneq B$ to explicitly mean that $A \subset B$ and $A \neq B$; we call A a *proper subset* of B.

Proposition 2.1 ($\subset$ is transitive)**.** If $A \subset B$ and $B \subset C$, then $A \subset C$.

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

A and B are ***equal*** if and only if they contain the same elements, denoted by $A = B$.

Proposition 2.2 (Double inclusion)**.** Let $A \subset S$ and $B \subset S$. Then

$$A = B \iff (A \subset B) \wedge (B \subset A)$$

Proof. We have

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

□

Remark. Double inclusion is a useful tool to prove that two sets are equal.

Some frequently occurring subsets of $\mathbf{R}$ are known as ***intervals***, which can be visualised as sections of the real line. We define *bounded intervals*

$$\begin{aligned} (a, b) &= \{x \in \mathbf{R} \mid a < x < b\}, \\ [a, b] &= \{x \in \mathbf{R} \mid a \leqslant x \leqslant b\}, \\ [a, b) &= \{x \in \mathbf{R} \mid a \leqslant x < b\}, \\ (a, b] &= \{x \in \mathbf{R} \mid a < x \leqslant b\}, \end{aligned}$$

and *unbounded intervals*

$$\begin{aligned}(a,\infty) &= \{x \in \mathbf{R} \mid a < x\},\\ [a,\infty) &= \{x \in \mathbf{R} \mid a \leqslant x\},\\ (-\infty,a) &= \{x \in \mathbf{R} \mid x < a\},\\ (\infty,a] &= \{x \in \mathbf{R} \mid x \leqslant a\}.\end{aligned}$$

An interval of the first type (a,b) is called an *open interval*; an interval of the second type $[a,b]$ is called a *closed interval*. Note that if $a = b$, then $[a,b] = \{a\}$, while $(a,b) = [a,b) = (a,b] = \emptyset$.

The ***power set*** $\mathcal{P}(A)$ of A is the set of all subsets of A (including the set itself and the empty set):

$$\mathcal{P}(A) = \{S \mid S \subset A\}.$$

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:

$$A \times B \coloneqq \{(a,b) \mid a \in A, b \in B\}.$$

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

Example

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

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

Algebra of Sets

We now disuss the algebra of sets. Given $A \subset S$ and $B \subset S$,

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

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

(ii) 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: $A \cap B = \emptyset$.

More generally, we can take unions and intersections of arbitrary numbers of sets (could be finitely or infinitely many). Given a family of sets $\{A_i \mid i \in I\}$ 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

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

(iii) The ***complement*** of A, denoted by A^c, is the set containing elements that are not in A:

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

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

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

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

Proposition 2.3 (Distributive laws)**.** Let $A, B, C \subset S$. Then

(i) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$;

(ii) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Proof.

(i) Suppose $x \in A \cup (B \cap C)$. Then

$$\begin{aligned} x \in A \cup (B \cap C) &\iff x \in A \quad \vee \quad x \in B \cap C \\ &\iff x \in A \quad \vee \quad (x \in B) \wedge (x \in C) \\ &\iff (x \in A) \vee (x \in B) \quad \wedge \quad (x \in A) \vee (x \in C) \\ &\iff x \in A \cup B \quad \wedge \quad x \in A \cup C \\ &\iff x \in (A \cup B) \cap (A \cup C). \end{aligned}$$

Thus $A \cup (B \cap C) \subset (A \cup B) \cap (A \cup C)$.

Conversely suppose that $x \in (A \cup B) \cap (A \cup C)$. Then go in the reverse direction of the above steps to show that $(A \cup B) \cap (A \cup C) \subset A \cup (B \cap C)$.

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

(ii) Similar.

□

Proposition 2.4 (de Morgan's laws)**.** Let $A, B \subset S$. Then

(i) $(A \cup B)^c = A^c \cap B^c$;

(ii) $(A \cap B)^c = A^c \cup B^c$.

Proof.

(i)

$$\begin{aligned} x \in (A \cup B)^c &\iff x \notin A \cup B \\ &\iff x \notin A \quad \wedge \quad x \notin B \\ &\iff x \in A^c \quad \wedge \quad x \in B^c \\ &\iff x \in A^c \cap B^c \end{aligned}$$

(ii) Similar.

□

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

$$\left(\bigcap_{i \in I} A_i\right)^c = \bigcup_{i \in I} {A_i}^c,$$
$$\left(\bigcup_{i \in I} A_i\right)^c = \bigcap_{i \in I} {A_i}^c.$$

Exercise

Prove the following:

(i) $\left(\bigcup_{i \in I} A_i\right) \cup B = \bigcup_{i \in I} (A_i \cup B)$

(ii) $\left(\bigcap_{i \in I} A_i\right) \cup B = \bigcap_{i \in I} (A_i \cup B)$

(iii) $\left(\bigcup_{i \in I} A_i\right) \cup \left(\bigcup_{j \in J} B_j\right) = \bigcup_{(i,j) \in I \times J} (A_i \cup B_j)$

(iv) $\left(\bigcap_{i \in I} A_i\right) \cup \left(\bigcap_{j \in J} B_j\right) = \bigcap_{(i,j) \in I \times J} (A_i \cup B_j)$

Exercise

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.)

§2.2 Relations

Definition and Examples

Definition 2.5 (Relation)**.** R is a ***relation*** between A and B if $R \subset A \times B$; $a \in A$ and $b \in B$ are said to be *related* if $(a, b) \in R$, denoted 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.

Example

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

- The "less than or equal to" relation $\leqslant$ on the set of real numbers is
$$\{(x, y) \in \mathbf{R}^2 \mid x \leqslant y\} \subset \mathbf{R}^2;$$
we write $x \leqslant y$ if (x, y) is in this set.
- The "divides" relation $\mid$ on $\mathbf{N}$ is
$$\{(m, n) \in \mathbf{N}^2 \mid m \text{ divides } n\} \subset \mathbf{N}^2;$$
we write $m \mid n$ if (m, n) is in this set.
- For a set S, the "subset" relation $\subset$ on $\mathcal{P}(S)$ is
$$\{(A, B) \in \mathcal{P}(S)^2 \mid A \subset B\} \subset \mathcal{P}(S)^2;$$
we write $A \subset B$ if (A, B) is in this set.

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

Example

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

Definition 2.6 (Binary relation)**.** A ***binary relation*** in A is a relation between A and itself; that is, $R \subset A \times A$.

Properties of Relations

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

(i) R is ***reflexive*** if xRx for all $x \in A$;

(ii) R is ***symmetric*** if $xRy \implies yRx$;

(iii) R is ***anti-symmetric*** if xRy and $yRx \implies x = y$;

(iv) R is ***transitive*** if xRy and $yRz \implies xRz$.

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

Definition 2.7. A ***partial order*** on a non-empty set A is a relation on A satisfying reflexivity, anti-symmetry and transitivity.

A ***total order*** on A is a partial order on A such that if for every $x, y \in A$, either xRy or yRx.

A ***well order*** on A is a total order on A such that every non-empty subset of A has a minimal element; that is, for each non-empty $B \subset A$ there exists $s \in B$ such that $s \leqslant b$ for all $b \in B$.

Example
- Less than: the relation $<$ on R is not reflexive, symmetric, or anti-symmetric, but it is transitive.
- Not equal to: the relation $\neq$ on R is not reflexive, anti-symmetric or transitive, but it is symmetric.

Equivalence Relations

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 2.8 (Equivalence relation)**.** A relation $\sim$ on a set A is an ***equivalence relation*** if it is reflexive, symmetric and transitive.

Notation. We denote $a \sim b$ for $(a, b) \in R$.

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

Definition 2.9 (Equivalence class)**.** Given an equivalence relation $\sim$ on a set A, and given $x \in A$, the ***equivalence class*** of x is

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

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

Definition 2.10 (Partition)**.** A ***partition*** of a set A is a collection of subsets $\{A_i \subset A \mid i \in I\}$, where I is an indexing set, with the property that

(i) $A_i \neq \emptyset$ for all $i \in I$ (all the subsets are non-empty)

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

(iii) $A_i \cap A_j = \emptyset$ for every $i \neq j$ (the subsets are disjoint)

The subsets are called the *parts* of the partition.

Proposition 2.11. Let $\sim$ be an equivalence relation on a non-empty set X. Then the equivalence classes under $\sim$ are a partition of X.

To prove this, we need to show that

(i) every equivalence class is non-empty;

(ii) every element of X is an element of an equivalence class;

(iii) every element of X lies in exactly one equivalence class.

Proof.

(i) An equivalence class $[x]$ contains x as $x \sim x$, by reflexivity of the relation. Thus $[x] \neq \emptyset$.

(ii) From (i), note that every $x \in X$ is in the equivalence class $[x]$, so every element of X is an element of at least one equivalence class.

(iii) Suppose otherwise, for a contradiction, that some element of X lies in more than one equivalence class. Let $x \in X$ such that $x \in [y]$ and $x \in [z]$; we want to show that $[y] = [z]$ (using double inclusion).

Let $a \in [y]$, so $a \sim y$. ALso $x \in [y]$ so $x \sim y$. By symmetry, $y \sim x$. By transitivity, $a \sim x$. Now $x \in [z]$ so $x \sim z$ and similarly $a \sim z$ thus $a \in [z]$. Hence $[y] \subset [z]$.

By the same argument, $[z] \subset [y]$. Hence $[y] = [z]$.

□

Definition 2.12 (Quotient set)**.** The ***quotient set*** is the set of all equivalence classes, denoted by $A/\sim$.

Example (Modular arithmetic)

Let n be a fixed positive integer. Define a relation on $\mathbf{Z}$ by

$$a \sim b \iff n \mid (b - a).$$

Proposition 2.13. $a \sim b$ is a equivalence relation.

Proof.

(i) $a \sim a$ so $\sim$ is reflexive.

(ii) $a \sim b \implies b \sim a$ for any integers a and b, so $\sim$ is symmetric.

(iii) If $a \sim b$ and $b \sim c$ then $n \mid (a-b)$ and $n \mid (b-c)$, so $n \mid (a-b)+(b-c) = (a-c)$, so $a \sim c$ and $\sim$ is transitive.

□

Notation. We write $a \equiv b \pmod{n}$ if $a \sim b$.

Notation. For any $k \in \mathbf{Z}$ we denote the equivalence class of a by $[a]$, called the *congruence class* (or *residue class*) of a mod n, which consists of the integers which differ from a by an integral multiple of n; that is,

$$[a] = \{a + kn \mid k \in \mathbf{Z}\}.$$

There are precisely n distinct congruence classes mod n, namely

$$[0], [1], \ldots, [n-1],$$

determined by the possible remainders after division by n; and these residue classes partition the integers $\mathbf{Z}$. The set of equivalence classes under this equivalence relation is denoted by $\mathbf{Z}/n\mathbf{Z}$, and called the *integers modulo* n.

Define addition and multiplication on $\mathbf{Z}/n\mathbf{Z}$ as follows: for $[a], [b] \in \mathbf{Z}/n\mathbf{Z}$,

$$[a] + [b] = [a+b]$$
$$[a][b] = [ab].$$

This means that to compute the sum / product of two elements $[a], [b] \in \mathbf{Z}/n\mathbf{Z}$, take any *representative* $a \in [a]$, $b \in [b]$, and add / multiply integers a and b as usual in $\mathbf{Z}$, then take the congruence class containing the result.

Proposition 2.14. Addition and mulltiplication on $\mathbf{Z}/n\mathbf{Z}$ are well-defined; that is, they do not depend on the choices of representatives for the classes involved. More precisely, if $a_1, a_2 \in \mathbf{Z}$ and $b_1, b_2 \in \mathbf{Z}$ with $\overline{a_1} = \overline{b_1}$ and $\overline{a_2} = \overline{b_2}$, then $\overline{a_1 + a_2} = \overline{b_1 + b_2}$ and $\overline{a_1 a_2} = \overline{b_1 b_2}$, i.e., If

$$a_1 \equiv b_1 \pmod{n}, \quad a_2 \equiv b_2 \pmod{n}$$

then

$$a_1 + a_2 \equiv b_1 + b_2 \pmod{n}, \quad a_1 a_2 \equiv b_1 b_2 \pmod{n}.$$

Proof. Suppose $a_1 \equiv b_1 \pmod{n}$, i.e., $n \mid (a_1 - b_1)$. Then $a_1 = b_1 + sn$ for some integer s. Similarly, $a_2 \equiv b_2 \pmod{n}$ means $a_2 = b_2 + tn$ for some integer t.

Then $a_1 + a_2 = (b_1 + b_2) + (s + t)n$ so that $a_1 + a_2 \equiv b_1 + b_2 \pmod{n}$, which shows that the sum of the residue classes is independent of the representatives chosen.

Similarly, $a_1 a_2 = (b_1 + sn)(b_2 + tn) = b_1 b_2 + (b_1 t + b_2 s + stn)n$ shows that $a_1 a_2 \equiv b_1 b_2 \pmod{n}$ and so the product of the residue classes is also independent of the representatives chosen. □

An important subset of $\mathbf{Z}/n\mathbf{Z}$ consists of the collection of congruence classes which have a multiplicative inverse in $\mathbf{Z}/n\mathbf{Z}$:

$$(\mathbf{Z}/n\mathbf{Z})^\times := \{[a] \in \mathbf{Z}/n\mathbf{Z} \mid \exists [c] \in \mathbf{Z}/n\mathbf{Z}, [a][c] = [1]\}.$$

Proposition 2.15. $(\mathbf{Z}/n\mathbf{Z})^\times$ is also the collection of congruence classes whose representatives are relatively prime to n:

$$(\mathbf{Z}/n\mathbf{Z})^\times = \{[a] \in \mathbf{Z}/n\mathbf{Z} \mid (a, n) = 1\}.$$

Axiom of Choice and Its Equivalences

Definition 2.16. Let $(P, \leqslant)$ be a partially ordered set. Suppose $A \subset P$.

(i) $u \in P$ is an ***upper bound*** for A if $x \leqslant u$ for all $x \in A$.

(ii) $m \in P$ is a ***maximal element*** of P if $x \in P$ and $m \leqslant x$ implies $m = x$.

(iii) Similarly we define ***lower bound*** and ***minimal element***.

(iv) $C \subset P$ is called a ***chain*** if either $x \leqslant y$ or $y \leqslant x$ for all $x, y \in C$.

This terminology of partially ordered sets will often be applied to an arbitrary family of sets. When this is done, it should be understood that the family is being regarded as a partially ordered set under the relation $\subsetneq$. Thus a maximal member of $\mathscr{A}$ is a set $M \in \mathscr{A}$ such that M is a proper subset of no other member of $\mathscr{A}$; a chain of sets is a family $\mathscr{C}$ of sets such that $A \subsetneq B$ or $B \subsetneq A$ for all $A, B \in \mathscr{C}$.

Definition 2.17. Let $\mathscr{F}$ be a family of sets. Then $\mathscr{F}$ is said to be a *family of finite character* if for each set A, we have $A \in \mathscr{F}$ if and only if each finite subset of A is in $\mathscr{F}$.

We shall need the following technical fact.

Lemma 2.18. Let $\mathscr{F}$ be a family of finite character, and let $\mathscr{C}$ be a chain in $\mathscr{F}$. Then $\bigcup \mathscr{C} \in \mathscr{F}$.

Proof. It suffices to show that each finite subset of $\bigcup \mathscr{C}$ is in $\mathscr{F}$. Let $F = \{x_1, \ldots, x_n\} \subset \bigcup \mathscr{C}$. Then there exist sets $C_1, \ldots, C_n \in \mathscr{C}$ such that $x_i \in C_i$ $(i = 1, \ldots, n)$. Since $\mathscr{C}$ is a chain, there exists $i_0 \in \{1, \ldots, n\}$ such that $C_i \subsetneq C_{i_0}$ for $i = 1, \ldots, n$. Then $F \subset C_{i_0} \in \mathscr{F}$. But $\mathscr{F}$ is of finite character, and so $F \in \mathscr{F}$. □

Theorem 2.19. The following are equivalent:

(i) *Axiom of choice*: The Cartesian product of any non-empty collection of non-empty sets is non-empty.

(ii) *Tukey's lemma*: Every non-empty family of finite character has a maximal member.

(iii) *Hausdorff maximality principle*: Every non-empty partially ordered set contains a maximal chain.

(iv) *Zorn's lemma*: Every non-empty partially ordered set in which every chain has an upper bound has a maximal element.

(v) *Well-ordering principle*: Every non-empty set has a well-ordering.

Proof. We direct the reader to Section 3 of [HS65] for the complete proof. □

Remark. It is a non-trivial result that Zorn's lemma is independent of the usual (Zermelo–Fraenkel) axioms of set theory in the sense that if the axioms of set theory are consistent, then so are these axioms together with Zorn's lemma; and if the axioms of set theory are consistent, then so are these axioms together with the negation of Zorn's lemma.

§2.3 Functions

Definitions and Examples

Definition 2.20 (Function)**.** A ***function*** $f : X \to Y$ is a mapping of every element of X to some element of 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 $f : \mathbf{R} \to \mathbf{R}$, 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 2.21. Given a function $f : X \to Y$, the ***image*** (or *range*) of f is

$$f(X) \coloneqq \{f(x) \mid x \in X\} \subset Y.$$

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

$$f(A) \coloneqq \{f(x) \mid x \in A\} \subset Y.$$

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

$$f^{-1}(B) \coloneqq \{x \mid f(x) \in B\} \subset X.$$

Remark. Note the distinction between "codomain" and "range".

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

Definition 2.22 (Restriction)**.** Given a function $f : X \to Y$ and a subset $A \subset X$, the ***restriction*** of f to A is the map $f|_A : A \to Y$.

Remark. The restriction is almost the same function as the original function—just the domain has changed.

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

Definition 2.23 (Identity map)**.** Given a set X, the ***identity*** $\mathrm{id}_X : X \to X$ is defined by

$$\mathrm{id}_X(x) = x \quad (\forall x \in X)$$

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

Injectivity, Surjectivity, Bijectivity

Definition 2.24. Let $f : X \to Y$ be a function.

(i) f is ***injective*** (or *one-to-one*) if each element of Y has at most one element of X that maps to it:

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

(ii) f is ***surjective*** (or *onto*) if every element of Y is mapped to at least one element of X:

$$\forall y \in Y, \quad \exists x \in X, \quad f(x) = y$$

Equivalently, f is surjective if $f(X) = Y$.

(iii) f is ***bijective*** if it is both injective and surjective; a bijective function is termed a *bijection*.

Notation. We write $X \sim Y$ if there exists a bijection $f : X \to Y$.

Composition

Definition 2.25 (Composition)**.** Given $f : X \to Y$ 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 2.26 (Associativity of composition)**.** Let $f : X \to Y$, $g : Y \to Z$, $h : Z \to W$. Then

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

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

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

□

Proposition 2.27 (Composition preserves injectivity and surjectivity)**.**

(i) If $f : X \to Y$ is injective and $g : Y \to Z$ is injective, then $g \circ f : X \to Z$ is injective.

(ii) If $f : X \to Y$ is surjective and $g : Y \to Z$ is surjective, then $g \circ f : X \to Z$ is surjective.

Proof.

(i) Let $f : X \to Y$ and $g : Y \to Z$ be injective. To prove that $g \circ f : X \to Z$ is injective, we need to prove: for all $x, x' \in X$,

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

Suppose that $(g \circ f)(x) = (g \circ f)(x')$. Then by definition

$$g\left(f(x)\right) = g\left(f(x')\right).$$

Injectivity of g implies

$$f(x) = f(x'),$$

and injectivity of f implies

$$x = x'.$$

(ii) Let $f : X \to Y$ and $g : Y \to Z$ be surjective. To prove that $g \circ f : X \to Z$ is surjective, we need to prove that for any $z \in Z$, there exists $x \in X$ such that $(g \circ f)(x) = z$.

Let $z \in Z$. By surjectivity of $g : Y \to Z$, there exists $y \in Y$ such that $g(y) = z$. By surjectivity of $f : X \to Y$, there exists $x \in X$ such that $f(x) = y$. This means that there exists $x \in X$ such that $g\left(f(x)\right) = g(y) = z$, as desired.

□

Proposition 2.28. $f : X \to Y$ is injective if and only if for any set Z and any functions $g_1, g_2 : Z \to X$,

$$f \circ g_1 = f \circ g_2 \implies g_1 = g_2.$$

Proof.

$\boxed{\implies}$ Suppose f is injective, and suppose $f \circ g_1 = f \circ g_2$. Let $z \in Z$. Then we have

$$f\left(g_1(z)\right) = f\left(g_2(z)\right).$$

Injectivity of f implies

$$g_1(z) = g_2(z),$$

so $g_1 = g_2$ (since the choice of $z \in Z$ is arbitrary).

$\boxed{\impliedby}$ Pick $Z = \{1\}$, basically some random one-element set. Then for $x, y \in X$, define

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

Then for $x, y \in X$,

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

which shows that f is injective. □

Proposition 2.29. $f : X \to Y$ is surjective if and only if for any set Z and any functions $g_1, g_2 : Y \to Z$,

$$g_1 \circ f = g_2 \circ f \implies g_1 = g_2.$$

Proof.

$\boxed{\implies}$ Suppose that f is surjective. Let $y \in Y$. Surjectivity of f means there exists $x \in X$ such that $f(x) = y$. Then

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

so $g_1 = g_2$.

$\boxed{\impliedby}$ We prove the contrapositive. Suppose 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 $g_1 \neq g_2$.

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

Then we define

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

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. □

Invertibility

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

Definition 2.30 (Invertibility)**.** Let $f : X \to Y$. We say that

(i) f is ***left-invertible*** if there exists $g : Y \to X$ such that $g \circ f = \mathrm{id}_X$; g is a *left-inverse* of f;

(ii) f is ***right-invertible*** if there exists $h : Y \to X$ such that $f \circ h = \mathrm{id}_Y$; h is a *right-inverse* of h;

(iii) f is ***invertible*** if there exists $k : Y \to X$ which is a left and right inverse of f; k is an *inverse* of f.

Remark. Notice that if g is left-inverse to f then f is right-inverse to g. A function can have more than one left-inverse, or more than one right-inverse.

Example

Let

$$f : \mathbf{R} \to [0, \infty), \quad f(x) = x^2$$
$$g : [0, \infty) \to \mathbf{R}, \quad g(x) = \sqrt{x}$$

- f is not left-invertible. Suppose otherwise, for a contradiction, that h is a left inverse of f, so that $hf = \mathrm{id}_{\mathbf{R}}$. Then

Proposition 2.31 (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

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

□

Since the inverse is unique, we can give it a notation.

Notation. The inverse of f is denoted by f^{-1}

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

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

Lemma 2.32 (Invertibility criterion)**.** Let $f : X \to Y$. Then

(i) f is left-invertible if and only if f is injective;

(ii) f is right-invertible if and only if f is surjective;

(iii) f is invertible if and only if f is bijective.

Proof.

(i) $\boxed{\Longrightarrow}$ Suppose f is left-invertible; let g be a left-inverse of f, so $g \circ f = \mathrm{id}_X$.

Now suppose $f(a) = f(b)$. Then applying g to both sides gives $g\left(f(a)\right) = g\left(f(b)\right)$, so $a = b$.

$\boxed{\Longleftarrow}$ Let f be injective. Choose any x_0 in the domain of f. Define $g : Y \to X$ as follows; note that each $y \in Y$ is either in the image of f or not.

- If y is in the image of f, it equals $f(x)$ for a *unique* $x \in X$ (uniqueness is because of the injectivity of f), so define $g(y) = x$.
- If y is not in the image of f, define $g(y) = x_0$.

Clearly $g \circ f = \mathrm{id}_X$.

(ii) $\boxed{\Longrightarrow}$ Suppose f is right-invertible; let g be a right-inverse of f, so $f \circ g = \mathrm{id}_Y$.

Let $y \in Y$. Then $f\left(g(y)\right) = \mathrm{id}_Y(y) = y$ so $y \in f(X)$. Thus $f(X) = Y$ so f is surjective.

$\boxed{\Longleftarrow}$ Suppose f is surjective. Let $y \in Y$, then y is in the image of f, so we can choose an element $g(y) \in X$ such that $f\left(g(y)\right) = y$. This defines a function $g : Y \to X$ which is evidently a right-inverse of f.

(iii) $\boxed{\Longrightarrow}$ Suppose f is invertible. Then f is left-invertible and right-invertible. By (i) and (ii), f is injective and surjective, so f is bijective.

$\boxed{\Longleftarrow}$ Suppose f is bijective. Then by (i) and (ii), f has a left-inverse $g : Y \to X$ and a right-inverse $h : Y \to X$. But "invertible" requires a single function to be *both* a left and right inverse, so we need to show that $g = h$:

$$g = g \circ \mathrm{id}_Y = g \circ (f \circ h) = (g \circ f) \circ h = \mathrm{id}_X \circ h = h$$

so $g = h$ is an inverse of f.

□

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

Proposition 2.33 (Inverse of composition)**.** Let $f : X \to Y$, $g : Y \to Z$. 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 $g \circ f$. □

Corollary 2.34. If $f_1, \ldots, f_n$ are invertible and the composition $f_1 \circ \cdots \circ f_n$ makes sense, then it is also invertible and its inverse is

$$f_n^{-1} \circ \cdots \circ f_1^{-1}.$$

Proposition 2.35. $\sim$ is an equivalence relation between sets.

Proof. We need to prove (i) reflexivity, (ii) symmetry, and (iii) transitivity.

(i) The identity map gives a bijection from a set to itself.

(ii) Suppose $f : X \to Y$ is a bijection. Then f is invertible, with inverse $f^{-1} : Y \to X$. Since f^{-1} is invertible (with inverse f), it is bijective.

(iii) Suppose $f : X \to Y$ and $g : Y \to Z$ are bijections, and thus they are invertible. Then by the previous result, $g \circ f$ is invertible and thus bijective.

□

Theorem 2.36 (Cantor–Schröder–Bernstein)**.** If $f : X \to Y$ and $g : Y \to X$ are injective, then $A \sim B$.

Monotonicity

Definition 2.37 (Monotonicity)**.** $f : [a, b] \to \mathbf{R}$ is called

(i) ***increasing***, if any $a < x_1 \leqslant x_2 < b$, there is $f(x_1) \leqslant f(x_2)$;

(ii) ***decreasing***, if any $a < x_1 \leqslant x_2 < b$, there is $f(x_1) \geqslant f(x_2)$;

f is ***monotonic*** if it is increasing or decreasing.

Suppose $f(x)$ is continuous in $[a, b]$. To locate the roots of $f(x) = 0$:

- 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 $[a, b]$.

Definition 2.38 (Convexity)**.** A function f is ***convex*** if for all $x_1, x_2 \in D_f$ and $0 \leqslant t \leqslant 1$, we have

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

f is *strictly convex* if the $\leqslant$ sign above is replaced with a strict inequality $<$.

Similarly, f is ***concave*** if for all $x_1, x_2 \in D_f$ and $0 \leqslant t \leqslant 1$, we have

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

f is *strictly concave* if the $\geqslant$ sign above is replaced with a strict inequality $>$.

§2.4 Cardinality

This section is about formalising the notion of the "size" of a set.

Definition 2.39. A and B said to be ***equivalent*** (or have the same *cardinality*), denoted by $A \sim B$, if there exists a bijection $f : A \to B$.

Notation. For $n \in \mathbf{N}$, denote

$$\mathbf{N}_n = \{k \in \mathbf{N} \mid 1 \leqslant k \leqslant n\},$$
$$n\mathbf{N} = \{nk \mid k \in \mathbf{N}\}.$$

Definition 2.40. For any set A, we say

(i) A is ***finite*** if $A \sim \mathbf{N}_n$ for some integer $n \in \mathbf{N}$, then the *cardinality* of A is $|A| = n$; A is *infinite* if A is not finite;

(ii) A is ***countable*** if $A \sim \mathbf{N}$; A is *uncountable* if A is neither finite nor countable; A is *at most countable* if A is finite or countable.

Remark. Any countable set can be "listed" in a sequence $a_1, a_2, \ldots$ of distinct terms. This technique is particularly useful when there is not possible to deduce an explicit formula for a bijection.

Proposition 2.41. $\mathbf{N}$ is infinite.

Proof. We want to show that there does not exist a bijection from $\mathbf{N}_n$ to $\mathbf{N}$, for all $n \in \mathbf{N}$. We prove by induction on n.

For the base case $n = 1$, if there exists a function $f_1 : \{1\} \to \mathbf{N}$, consider the set $\mathbf{N} \setminus f_1(\{1\})$. It is not empty, so f_1 is not surjective, thus it is not bijective.

For the inductive step, we want to show if there does not exist a bijection from $\mathbf{N}_k$ to $\mathbf{N}$, then there does not exist a bijection from $\mathbf{N}_{k+1}$ to $\mathbf{N}$. We prove the contrapositive: if there exists a bijection from $\mathbf{N}_{k+1} \to \mathbf{N}$, then there exists a bijection from $\mathbf{N}_k$ to $\mathbf{N}$.

Suppose $h : \mathbf{N}_{k+1} \to \mathbf{N}$ is a bijection. If remove the element $k+1$, then there exists a bijection from $\mathbf{N}_k$ to $\mathbf{N} \setminus \{h(k+1)\}$. But $\mathbf{N} \setminus \{h(k+1)\} \sim \mathbf{N}$ so $\mathbf{N}_k \sim \mathbf{N}$. □

Corollary 2.42. Any countable set is infinite.

> **Example**
>
> $\mathbf{N}$ is countable since the identity map from $\mathbf{N}$ to $\mathbf{N}$ is a bijection.

> **Example**
>
> $n\mathbf{N}$ is countable.
>
> *Proof.* Let $f : \mathbf{N} \to n\mathbf{N}$ which sends $k \mapsto nk$. We now need to show that f is (i) injective, and (ii) surjective.

(i) For any $k_1, k_2 \in \mathbf{N}$, $nk_1 = nk_2$ implies $k_1 = k_2$ so f is injective.

(ii) For any $x \in n\mathbf{N}$, $x = nk$ for some $k \in \mathbf{N}$, thus $\frac{x}{n} = k \in \mathbf{N}$ so f is surjective.

Hence f is bijective, so $n\mathbf{N} \sim \mathbf{N}$ and we are done. □

Example

Z is countable.

Proof. Consider the following arrangement of the elements of **Z** and **N**:

$$\begin{aligned} \mathbf{Z}: &\quad 0, 1, -1, 2, -2, 3, -3, \dots \\ \mathbf{N}: &\quad 1, 2, 3, 4, 5, 6, 7, \dots \end{aligned}$$

In fact we can write an explicit formula for a bijection $f : \mathbf{N} \to \mathbf{Z}$ where

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

□

Proposition 2.43. Every infinite subset of a countable set is countable.

Proof. Let S be the countable set. Then we can arrange the elements of S in a sequence (s_n) of distinct elements:

$$s_1, s_2, \dots$$

Suppose $E \subset S$ is infinite. The main idea is to show that we can list out the elements of E in a sequence. We now construct a sequence (n_k) as follows: Let

$$\begin{aligned} n_1 &= \min\{i \mid s_i \in E\} \\ n_2 &= \min\{i \mid s_i \in E, i > n_1\} \\ &\vdots \\ n_k &= \min\{i \mid s_i \in E, i > n_{k-1}\}. \end{aligned}$$

Then

$$E = \{s_{n_1}, s_{n_2}, \dots\},$$

where we note that the function $f(k) = s_{n_k}$ $(k = 1, 2, \dots)$ is bijective. Hence $E \sim \mathbf{N}$, as desired. □

Remark. This shows that countable sets represent the "smallest" infinity: No uncountable set can be a subset of a countable set.

Proposition 2.44. The countable union of countable sets is countable.

Proof. Let $\{A_n \mid n \in \mathbf{N}\}$ be a fanily of countable sets; clearly this is a countable collection of sets (indexed by $\mathbf{N}$). Then we want to show that the union

$$S = \bigcup_{n=1}^{\infty} A_n$$

is countable.

Since every set A_n is countable, we can list its elements in a sequence (a_{nk}) $(k = 1, 2, 3, \dots)$. Arrange the elements of all the sets in $\{A_n\}$ in the form of an infinite array, containing all elements of S, where the elements of A_n form the n-th row.

$$\begin{array}{ll} A_1\colon & \cancel{a_{11}} \quad \cancel{a_{12}} \quad \cancel{a_{13}} \quad \cancel{a_{14}} \quad \cdots \\ A_2\colon & \cancel{a_{21}} \quad \cancel{a_{22}} \quad \cancel{a_{23}} \quad \cancel{a_{24}} \quad \cdots \\ A_3\colon & \cancel{a_{31}} \quad \cancel{a_{32}} \quad \cancel{a_{33}} \quad \cancel{a_{34}} \quad \cdots \\ A_4\colon & \cancel{a_{41}} \quad \cancel{a_{42}} \quad \cancel{a_{43}} \quad \cancel{a_{44}} \quad \cdots \\ \vdots & \end{array}$$

We then zigzag our way through the array, and arrange these elements in a sequence

$$a_{11}, a_{21}, a_{12}, a_{31}, a_{22}, a_{13}, a_{41}, a_{32}, a_{23}, a_{14}, \dots$$

thus S is countable, and we are almost done!

A small problem is that if any two of the sets A_n have elements in common, these will appear more than once in the above sequence. Then we take a subset $T \subset S$, where every element only appears once. Note that T is an infinite subset, since $A_1 \subset T$ is infinite. Then since T is an infinite subset of a countable set S, by Proposition 2.43, T is countable. □

Remark. If we were to instead start by going down by the first row of the above array, then we would not get to the second row (and beyond); all that would show is the first row is countable. Instead, we wind our way through diagonally, ensuring that we hit every number of the array.

Corollary 2.45. Suppose A is an indexing set that is at most countable. Let $\{B_\alpha \mid \alpha \in A\}$ be a family of sets that are at most countable. Then the union

$$\bigcup_{\alpha \in A} B_\alpha$$

is at most coutable.

Proposition 2.46. Let A be a countable set. For $n \in \mathbf{N}$, let

$$B_n = \{(a_1, \dots, a_n) \mid a_i \in A\}.$$

Then B_n is countable.

Proof. We prove by induction on n. That B_1 is countable is evident, since $B_1 = A$.

Now suppose B_{n-1} is countable. The elements of B_n are of the form

$$(b, a) \quad (b \in B_{n-1}, a \in A)$$

For every fixed b, the set of ordered pairs (b, a) is equivalent to A, and hence countable. Thus B_n is a union of countable sets. By Proposition 2.44, B_n is countable. □

Corollary 2.47. $\mathbf{Q}$ is countable.

Proof. Note that every $x \in \mathbf{Q}$ is of the form $\frac{b}{a}$, where $a, b \in \mathbf{Z}$. By the previous result, taking $n = 2$, the set of pairs (a, b) and therefore the set of fractions $\frac{b}{a}$ is countable. □

That not all infinite sets are, however, countable, is shown by the next result.

Proposition 2.48. Let A be the set of all sequences whose elements are the digits 0 and 1. Then A is uncountable.

Proof. Let $E \subset A$ be countable, consisting of the sequences $s_1, s_2, s_3, \ldots$.

We construct a new sequence s as follows:

$$n\text{-th digit of } s = \begin{cases} 0 & \text{if } n\text{-th digit in } s_n \text{ is } 1, \\ 1 & \text{if } n\text{-th digit in } s_n \text{ is } 0. \end{cases}$$

Then the sequence s differs from every member of E in at least one place, so $s \notin E$. But clearly $s \in A$; hence $E \subsetneq A$.

We have shown that every countable subset of A is a proper subset of A. It follows that A is uncountable (for otherwise A would be a proper subset of A, which is absurd). □

Remark. The idea of the above proof is called *Cantor's diagonal process*, first used by Cantor. This is because if elements of the sequences $s_1, s_2, s_3, \ldots$ are listed out in an array, it is the elements on the diagonal which are involved in the construction of the new sequence.

Corollary 2.49. $\mathbf{R}$ is uncountable.

Proof. This follows from the binary representation of the real numbers. □

Theorem 2.50 (Cantor's theorem)**.** For any set A, we have $A \not\sim \mathcal{P}(A)$.

Proof. Suppose otherwise, for a contradiction, that $A \sim \mathcal{P}(A)$. Then there exists a bijection $f : A \to \mathcal{P}(A)$. Then for each $x \in A$, $f(x)$ is a subset of A. Now consider the "anti-diagonal" set

$$B = \{x \in A \mid x \notin f(x)\}.$$

That is, B is the subset of A containing all $x \in A$ such that x is not in the set $f(x)$. Since $B \subset A$, we have $B \in \mathcal{P}(A)$. Since f is bijective (in particular surjective), there exists $x \in A$ such that $f(x) = B$. Now there are two cases: (i) $x \in B$, or (ii) $x \notin B$.

(i) If $x \in B$, then by definition of the set B it must be the case that $x \notin f(x)$. But since $f(x) = B$, we then have $x \notin D$. This is absurd since we cannot have $x \in B$ and $x \notin B$ simultaneously.

(ii) If $x \notin B$, by definition of the set B, this implies that $x \in f(x)$. But $f(x) = B$. So we have $x \in B$ and $x \notin B$, which is again absurd.

In either case, we have reached a contradiction. Hence there cannot exist a surjective (and thus bijective) function $A \to \mathcal{P}(A)$. □

Exercises

Problem 2.1. Prove the following statements:

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

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

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

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

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

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

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

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

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

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

Problem 2.2. 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

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

A subset $Y \in \mathbf{C}^n$ 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 $T = \{f_1, f_2, \ldots\}$ 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 2.3. Let $A = \mathbf{R}$ and for any $x, y \in A$, $x \sim y$ if and only if $x - y \in \mathbf{Z}$. 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 $\mathbf{C}$, 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)

$$(x' + y') - (x + y) = (x' - x) + (y' - y) \in \mathbf{Z}$$

Thus $[x' + y'] = [x + y]$

$$(-x') - (-x) = -(x' - x) \in \mathbf{Z}$$

Thus $[-x'] = [-x]$.

(b) Complex numbers in the polar form: $z = re^{i\theta}$

Then the correspondence is given by $\phi([x]) = e^{2\pi ix}$

$$[x] = [y] \iff x - y \in \mathbf{Z} \iff e^{2\pi i(x-y)} = 1 \iff e^{2\pi ix} = e^{2\pi iy}$$

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

$$\phi([x]) = e^{2\pi ix'} = e^{2\pi ix} \cdot e^{2\pi i(x'-x)} = e^{2\pi ix}$$

(c) You can either refer to the specific correspondence $\phi([x]) = e^{2\pi ix}$ or use its properties.

$$\phi(-[x])\phi([x]) = \phi([-x])\phi([x]) = \phi([-x + x]) = \phi([0]) = 1$$

□

Problem 2.4 (Complex Numbers)**.** Let $\mathbf{R}[x]$ denote the set of real polynomials. Define

$$\mathbf{C} = \mathbf{R}[x]/(x^2+1)\mathbf{R}[x]$$

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.

Problem 2.5. The set of all algebraic numbers is countable. (Exercise 2)

II

Abstract Algebra

3 Groups

§3.1 Introduction to Groups

Definitions and Properties

Definition 3.1 (Binary operation)**.** A ***binary operation*** $*$ on a set G is a function $* : G \times G \to G$. For any $a, b \in G$, we write $a * b$ for the image of (a, b) under $*$.

$*$ is *associative* on G if $(a * b) * c = a * (b * c)$ for all $a, b, c \in G$.

$*$ is *commutative* on G if $a * b = b * a$ for all $a, b \in G$.

Definition 3.2 (Group)**.** A ***group*** $(G, *)$ consists of a set G and a binary operation $*$ on G satisfying the following group axioms:

(i) Associativity: $a * (b * c) = (a * b) * c$ for all $a, b, c \in G$.

(ii) Identity: there exists identity element $e \in G$ such that $a * e = e * a = a$ for all $a \in G$.

(iii) Invertibility: for all $a \in G$, there exists inverse $c \in G$ such that $a * c = c * a = e$.

G is ***abelian***[1] if the operation is commutative; it is *non-abelian* if otherwise.

Remark. When verifying that $(G, *)$ is a group we have to check (i), (ii), (iii) above and also that $*$ is a binary operation closed in G—that is, $a * b \in G$ for all $a, b \in G$.

Notation. We simply denote a group $(G, *)$ by G if the operation is clear.

Notation. We abbreviate $a * b$ to just ab if the operation is clear.

Notation. Since $*$ is associative, we omit unnecessary parentheses and write $(ab)c = a(bc) = abc$.

Notation. For any $a \in G$, $n \in \mathbf{Z}^+$ we denote $a^n = \underbrace{a \cdot a \cdots a}_{n \text{ times}}$.

Notation. We write $(\mathbf{Z}, +)$, $(\mathbf{Q}, +)$, $(\mathbf{R}, +)$, $(\mathbf{C}, +)$ as simply $\mathbf{Z}$, $\mathbf{Q}$, $\mathbf{R}$, $\mathbf{C}$.

Example

- $\mathbf{Z}$, $\mathbf{Q}$, $\mathbf{R}$, $\mathbf{C}$ are groups, with identity 0 and (additive) inverse $-a$ for all a.
- $\mathbf{Q} \setminus \{0\}$, $\mathbf{R} \setminus \{0\}$, $\mathbf{C} \setminus \{0\}$, $\mathbf{Q}^+$, $\mathbf{R}^+$ are groups under $\times$, with identity 1 and (multiplicative) inverse $\frac{1}{a}$ for all a; $\mathbf{Z} \setminus \{0\}$ is not a group under $\times$, because all elements except for ± 1 do not have an inverse in $\mathbf{Z} \setminus \{0\}$.
- For $n \in \mathbf{Z}^+$, $\mathbf{Z}_n$ is an abelian group under $+$.

[1] after the Norwegian mathematician Niels Abel (1802–1829)

- For $n \in \mathbf{Z}^+$, $(\mathbf{Z}_n)^\times$ is an abelian group under multiplication.

Proposition 3.3. Let G be a group. Then

(i) the identity of G is unique,

(ii) for each $a \in G$, a^{-1} is unique,

(iii) $(a^{-1})^{-1} = a$ for all $a \in G$,

(iv) $(ab)^{-1} = b^{-1}a^{-1}$,

(v) for any $a_1, \ldots, a_n \in G$, $a_1 \cdots a_n$ is independent of how we arrange the parantheses (generalised associative law).

Proof.

(i) Suppose otherwise, then e and e' are identites of G. We have

$$e = ee' = e'$$

where the first equality holds as e' is an identity, and the second equality holds as e is an identity. Since $e = e'$, the identity is unique.

(ii) Suppose otherwise, then b and c are both inverses of a. Let e be the identity of G. Then $ab = e$, $ca = e$. Thus

$$c = ce = c(ab) = (ca)b = eb = b.$$

Hence the inverse is unique.

(iii) To show $(a^{-1})^{-1} = a$ is exactly the problem of showing that a is the inverse of a^{-1}, which is by definition of the inverse (with the roles of a and a^{-1} interchanged).

(iv) Let $c = (ab)^{-1}$. Then $(ab)c = e$, or $a(bc) = e$ by associativity, which gives $bc = a^{-1}$ and thus $c = b^{-1}a^{-1}$ by multiplying b^{-1} on both sides.

(v) The result is trivial for $n = 1, 2, 3$. For all $k < n$ assume that any $a_1 \cdots a_k$ is independent of parantheses. Then

$$(a_1 \cdots a_n) = (a_1 \cdots a_k)(a_{k+1} \cdots a_n).$$

Then by assumption both are independent of parentheses since $k, n - k < n$ so by induction we are done.

□

Notation. Since the inverse is unique, we denote the inverse of $a \in G$ by a^{-1}.

Proposition 3.4 (Cancellation law)**.** Let $a, b \in G$. Then the equations $ax = b$ and $ya = b$ have unique solutions for $x, y \in G$. In particular, we can cancel on the left and right.

Proof. We can solve $ax = b$ by applying a^{-1} to both sides of the equation to get $x = a^{-1}b$. The uniqueness of x follows because a^{-1} is unique. A similar case holds for $ya = b$. □

Definition 3.5 (Order of a group)**.** Let G be a group. Its cardinality $|G|$ is called the ***order*** of G. We say that a group G is a *finite group* if $|G| < \infty$.

One way to represent a finite group is by means of the group table or Cayley table[2]. Let $G = \{e, g_2, g_3, \ldots, g_n\}$ 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.

Examples

Example (Product group)

Let $(G, *_G)$ and $(H, *_H)$ be groups. Then the operation $*$ is defined on $G \times H$ by

$$(g_1, h_1) * (g_2, h_2) = (g_1 *_G g_2, h_1 *_H h_2)$$

for all $g_1, g_2 \in G$, $h_1, h_2 \in H$. $(G \times H, *)$ is called the *product group* of G and H.

We check that the product group is a group:

(i) Since $*_G$ and $*_H$ are both associative binary operations, it follows that $*$ is also an associative binary operation on $G \times H$.

(ii) We also note

$$e_{G\times H} = (e_G, e_H), \quad (g, h)^{-1} = (g^{-1}, h^{-1})$$

as for any $g \in G$, $h \in H$,

$$(e_G, e_H) * (g, h) = (g, h) = (g, h) * (e_G, e_H).$$

(iii) As for identity,

$$(g^{-1}, h^{-1}) * (g, h) = (e_G, e_H) = (g, h) * (g^{-1}, h^{-1}).$$

Example (Dihedral groups)

An important family of groups is the ***dihedral groups***. For $n \in \mathbf{Z}^+$, $n \geqslant 3$, let D_{2n} be the set of symmetries[a] of a regular n-gon.

Remark. Here "D" stands for "dihedral", meaning two-sided.

To visualise this, we first choose a labelling of the n vertices. Then each symmetry S can be

[2] after the English mathematician Arthur Cayley (1821 – 1895)

described uniquely by the corresponding permutation σ of $\{1, 2, \ldots, n\}$ where if the symmetry s puts vertex i in the place where vertex j was originally, then σ is the permutation sending i to j.

We now make D_{2n} into a group. For $S, T \in D_{2n}$, define the binary operation ST to be the symmetry obtained by first applying T then S to the n-gon (this is analagous to function composition). If S and T effect the permutations σ and τ respectively on the vertices, then ST effects $\sigma \circ \tau$.

(i) The binary operation on D_{2n} is associative since the composition of functions is associative.

(ii) The identity of D_{2n} is the identity symmetry, which leaves all vertices fixed, denoted by 1.

(iii) The inverse of $S \in D_{2n}$ is the symmetry which reverses all rigid motions of S (so if S effects permutation σ on the vertices, S^{-1} effects σ^{-1}).

Let r be the rotation clockwise about the origin by $\frac{2\pi}{n}$ radians, let s be the reflection about the line of symmetry through the first labelled vertex and the origin.

Proposition.

(i) $|r| = n$

(ii) $|s| = 2$

(iii) $s \neq r^i$ for all i

(iv) $sr^i \neq sr^j$ for all $i \neq j$ $(0 \leqslant i, j \leqslant n-1)$, so

$$D_{2n} = \{1, r, \ldots, r^{n-1}, s, sr, \ldots, sr^{n-1}\}$$

and thus $|D_{2n}| = 2n$.

(v) $rs = sr^{-1}$

(vi) $r^i s = sr^{-i}$

Proof.

(i) It is obvious that $1, r, r^2, \ldots, r^{n-1}$ are all distinct and $r^n = 1$, so $|r| = n$.

(ii) This is fairly obvious: either reflect or do not reflect.

(iii) This is also obvious: the effect of any reflection cannot be obtained from any form of rotation.

(iv) Just cancel on the left and use the fact that $|r| = n$. We assume that $i \not\equiv j \pmod{n}$.

(v) Omitted.

(vi) By (5), this is true for $i = 1$. Assume it holds for $k < n$. Then $r^{k+1}s = r(r^k s) = rsr^{-k}$. Then $rs = sr^{-1}$ so $rsr^{-k} = sr^{-1}r^{-k} = sr^{-k-1}$ so we are done.

□

A presentation for the dihedral group D_{2n} using generators and relations is

$$D_{2n} = \langle r, s \mid r^n = s^2 = 1, rs = sr^{-1} \rangle.$$

[a] a symmetry is any rigid motion of the n-gon which can be effected by taking a copy of the n-gon, moving this copy in any fashion in 3-space and then placing the copy back on the original n-gon so it exactly covers it. A symmetry can be a reflection or a rotation.

Example (Permutation groups)

Let S be a non-empty set. A bijection $S \to S$ is called a *permutation* of S; the set of permutations of S is denoted by $\mathrm{Sym}(S)$.

We now show that $\mathrm{Sym}(S)$ is a group under function composition $\circ$; $(\mathrm{Sym}(S), \circ)$ is the ***symmetric group*** on S. Note that $\circ$ is a binary operation on $\mathrm{Sym}(S)$ since if $\sigma : S \to S$ and $\tau : S \to S$ are both bijections, then $\sigma \circ \tau$ is also a bijection from S to S.

(i) Function composition is associative so $\circ$ is associative.

(ii) The identity of $\mathrm{Sym}(S)$ is the identity map 1, defined by $1(a) = a$ for all $a \in S$.

(iii) For every permutation σ, σ is bijective and thus invertible, so there exists a (2-sided) inverse $\sigma^{-1} : S \to S$ satisfying $\sigma \circ \sigma^{-1} = \sigma^{-1} \circ \sigma = 1$.

In the special case where $S = \{1, 2, \ldots, n\}$, the symmetric group on S is called the *symmetric group of degree* n, denoted by S_n.

Proposition. If $|S| \geqslant 3$ then $\mathrm{Sym}(S)$ is non-abelian.

Proof. Let $S = \{x_1, x_2, x_3\}$ where three elements are distinct. □

Proposition. $|S_n| = n!$

Proof. Obvious, since there are $n!$ permutations of $\{1, 2, \ldots, n\}$. □

Example (Matrix groups)

For $n \in \mathbf{Z}^+$, let $GL_n(\mathbf{F})$ be the set of all $n \times n$ invertible matrices whose entries are in $\mathbf{F}$:

$$GL_n(\mathbf{F}) = \{A \in M_{n\times n}(\mathbf{F}) \mid \det(A) \neq 0\}.$$

We show that $GL_n(\mathbf{F})$ is a group under matrix multiplication; $GL_n(\mathbf{F})$ is the ***general linear group*** of degree n.

(i) Since $\det(AB) = \det(A) \cdot \det(B)$, it follows that if $\det(A) \neq 0$ and $\det(B) \neq 0$, then $\det(AB) \neq 0$, so $GL_n(\mathbf{F})$ is closed under matrix multiplication.

(ii) Matrix multiplication is associative.

(iii) $\det(A) \neq 0$ if and only if A has an inverse matrix, so each $A \in GL_n(\mathbf{F})$ has an inverse $A^{-1} \in GL_n(\mathbf{F})$ such that

$$AA^{-1} = A^{-1}A = I$$

where I is the $n \times n$ identity matrix.

Example (Quaternion group)

The ***Quaternion group*** Q_8 is defined by

$$Q_8 = \{1, -1, i, -i, j, -j, k, -k\}$$

with product $\cdot$ computed as follows:

- $1 \cdot a = a \cdot 1 = a$ for all $a \in Q_8$
- $(-1) \cdot (-1) = 1$
- $(-1) \cdot a = a \cdot (-1) = -a$ for all $a \in Q_8$
- $i \cdot i = j \cdot j = k \cdot k = -1$
- $i \cdot j = k$, $j \cdot i = -k$, $j \cdot k = i$, $k \cdot j = -i$, $k \cdot i = j$, $i \cdot k = -j$

Note that Q_8 is a non-abelian group of order 8.

Cyclic Groups and Order

Definition 3.6 (Cyclic group)**.** Let G be a group, $g \in G$. If for every $x \in G$, there exists $n \in \mathbf{Z}$ such that $g^n = x$ then g is the ***generator*** of G; denote $G = \langle g \rangle$.

G is ***cyclic*** if there is a generator for G in G.

Remark. A cyclic group may have more than one generator. For example, if $G = \langle g \rangle$, then also $G = \langle g^{-1} \rangle$ because $(g^{-1})^n = g^{-n} \in G$ for $n \in \mathbf{Z}$ so does $-n$, thus

$$\{g^n \mid n \in \mathbf{Z}\} = \{(g^{-1})^n \mid n \in \mathbf{Z}\}.$$

Example

$\mathbf{Z}$ is a cyclic group with generators 1 and -1.

Notation. For each $n \in \mathbf{Z}^+$, C_n denotes the cyclic group of order n:

$$C_n = \{e, g, g^2, \ldots, g^{n-1}\}$$

which satisfy $g^n = e$. Thus given two elements in C_n, we define

$$g^i * g^j = \begin{cases} g^{i+j} & (0 \leqslant i+j < n) \\ g^{i+j-n} & (n \leqslant i+j \leqslant 2n-2) \end{cases}$$

Proposition 3.7. Cyclic groups are abelian.

Proof. Let G be a cyclic group. For $g^i, g^j \in G$, by the laws of exponents,

$$g^i g^j = g^{i+j} = g^j g^i.$$

□

Definition 3.8 (Order)**.** Let G be a group, $g \in G$. If there is a positive integer k such that $g^k = e$, then the ***order*** of g is defined as

$$o(g) \coloneqq \min\{m > 0 \mid g^m = e\}.$$

Otherwise we say that the order of g is infinite.

Proposition 3.9. If G is finite, then $o(g)$ is finite for each $g \in G$.

Proof. Consider the list

$$g, g^2, g^3, g^4, \cdots \in G.$$

As G is finite, then this list must have repeats. Hence there are integers $i > j$ such that $g^i = g^j$. So $g^{i-j} = e$ showing that $\{m > 0 \mid g^m = e\}$ is non-empty and so has a minimal element. □

Proposition 3.10. If $g \in G$ and $o(g)$ is finite, then $g^n = e$ if and only if $o(g) \mid n$.

Proof.

$\boxed{\Longleftarrow}$ Suppose $o(g) \mid n$. Then $n = ko(g)$ for some $k \in \mathbf{Z}$, so

$$g^n = \left(g^{o(g)}\right)^k = e^k = e.$$

$\boxed{\Longrightarrow}$ Suppose $g^n = e$. By the division algorithm, there exists integers q, r such that $n = qo(g) + r$, where $0 \leqslant r < o(g)$. Then

$$g^r = g^{n-qo(g)} = g^n \left(g^{o(g)}\right)^{-q} = e.$$

By the minimality of $o(g)$, we must have $r = 0$, and so $n = qo(g)$ implies $o(g) \mid n$. □

Corollary 3.11. Let G be a cyclic group, $g \in G$. Then $g^k = g^m$ if and only if $m \equiv k \pmod{o(g)}$.

Proposition 3.12. If $G = \langle g \rangle$, then $|G| = o(g)$ (where if one side of this equality is infinite, so is the other). More specifically,

(i) if $|G| = n < \infty$, then $g^n = e$ and $e, g, g^2, \ldots, g^{n-1}$ are all the distinct elements of G;

(ii) if $|G| = \infty$, then $g^n \neq e$ for all $n \neq 0$, and $g^a \neq g^b$ for all $a, b \in \mathbf{Z}$, $a \neq b$.

Proof.

(i) We first show that all the elements are distinct. If $g^a = g^b$ for $0 \leqslant a < b < n$, then $g^{b-a} = e$, which contradicts the minimality of $o(g)$. Thus G has at least n elements and it remains to show that these are all of them.

For the element g^t, by the division algorithm, we can write $t = qn + r$ where $0 \leqslant r < n$. Then

$$g^t = g^{qn+r} = (g^n)^q \, g^r = g^r \in \{e, g, g^2, \ldots, g^{n-1}\}$$

since $0 \leqslant r < n$.

(ii) Suppose $o(g) = \infty$, so no positive power of g is the identity. If $g^a = g^b$ for some $a, b \in \mathbf{Z}$, $a < b$, then $g^{b-a} = e$, contradicting the previous statement. Thus distinct powers of g are distinct elements of G, so $|G| = \infty$.

□

Note that a given cyclic group may have more than one generator. The next results determine precisely which powers of g generate the group $\langle g \rangle$.

Proposition 3.13. Let G be a group, $g \in G$. Let $a \in \mathbf{Z} \setminus \{0\}$.

(i) If $o(g) = \infty$, then $o(g^a) = \infty$.

(ii) If $o(g) = n < \infty$, then $o(g) = \dfrac{n}{\gcd(n, a)}$. In particular, if $a \mid n$, then $o(g^a) = \dfrac{n}{a}$.

Proof.

(i) Suppose, for a contradiction, that $o(g) = \infty$ but $o(g^a) = m < \infty$. Then by definition of order,

$$e = (g^a)^m = g^{am}.$$

Also,

$$g^{-am} = (g^{am})^{-1} = e^{-1} = e.$$

Now one of am or $-am$ is positive (since $a \neq 0$ and $m \neq 0$), so some positive power of g is the identity. This contradicts the hypothesis $o(g) = \infty$.

(ii)

□

Proposition 3.14. Let $G = \langle g \rangle$.

(i) If $o(g) = \infty$, then $G = \langle g^a \rangle$ if and only if $a = \pm 1$.

(ii) If $o(g) = n < \infty$, then $G = \langle x^a \rangle$ if and only if $\gcd(a, n) = 1$. In particular, the number of generators of G is $\phi(n)$ (where ϕ is Euler's totient function).

Subgroups

Definition 3.15 (Subgroup)**.** Let G be a group. A non-empty $H \subset G$ is a ***subgroup*** of G, denoted $H \leqslant G$, if H is closed under products and inverses; that is,

(i) $e \in H$;

(ii) $xy \in H$ for all $x, y \in H$;

(iii) $x^{-1} \in H$ for all $x \in H$.

Remark. If $*$ is an associative (respectively, commutative) binary operation on G and $*$ is restricted to some $H \subset G$ is a binary operation on H, then $*$ is automatically associative (respectively, commutative) on H as well.

The following result provides a convenient method to determine if a given subset of a group is a subgroup.

Lemma 3.16 (Subgroup criterion)**.** Let G be a group, $H \subset G$ is non-empty. Then $H \leqslant G$ if and only if $xy^{-1} \in H$ for all $x, y \in H$.

Proof.

$\boxed{\implies}$ If $H \leqslant G$, then we are done, by definition of subgroup.

$\boxed{\impliedby}$ We want to prove that for non-empty $H \subset G$, if $xy^{-1} \in H$ for all $x, y \in H$, then $H \leqslant G$, by checking the group axioms:

(i) Since $H \neq \emptyset$, take $x \in H$, let $y = x$, then $e = xx^{-1} \in H$, so H contains the identity of G.

(ii) Since $e \in H$, $x \in H$, then $x^{-1} \in H$ so H is closed under taking inverses.

(iii) For any $x, y \in H$, $x, y^{-1} \in H$, so by (ii), $x(y^{-1})^{-1} = xy \in H$, so H is closed under multiplication.

□

Proposition 3.17. Let $G = \langle g \rangle$ be a cyclic group. Then every subgroup of G is cyclic.

Proposition 3.18. Let G be a group, $H, K \leqslant G$. Then $H \cap K \leqslant G$.

Proof. Apply the subgroup criterion:

(i) Since $e_G \in H$ and $e_G \in K$, we have $e_G \in H \cap K$, so $H \cap K \neq \emptyset$.

(ii) Let $a, b \in H \cap K$. Then $a, b \in H$ and $a, b \in K$. Since $H, K \leqslant G$, by the subgroup criterion, $ab^{-1} \in H$ and $ab^{-1} \in K$, so $ab^{-1} \in H \cap K$.

□

Corollary 3.19. Let G be a group, $\{H_i \mid i \in I\}$ is a collection of subgroups of G. Then

$$\bigcap_{i \in I} H_i \leqslant G.$$

Thus we may make the following definition.

Definition 3.20 (Subgroup generated by subset of group)**.** Let G be a group, $S \subset G$. The ***subgroup generated by*** S, denoted by $\langle S \rangle$, is the smallest subgroup of G which contains S.

If $g \in G$, then we write $\langle g \rangle$ (rather than the more accurate but cumbersome $\langle \{g\} \rangle$).

If $\langle S \rangle = G$, then the elements of S are said to be *generators* of G.

Example

If G is abelian and $g, h \in G$ then

$$\langle g, h \rangle = \{g^r h^s \mid r, s \in \mathbf{Z}\}.$$

Proof. Certainly $\{g^r h^s \mid r, s \in \mathbf{Z}\} \subset \langle g, h \rangle$. However, when G is abelian (or indeed if just $gh = hg$), then $\{g^r h^s \mid r, s \in \mathbf{Z}\}$ is a subgroup as follows:

(i) $e = g^0 h^0 \in \{g^r h^s \mid r, s \in \mathbf{Z}\}$

(ii) $(g^k h^l)(g^K h^L) = g^{k+K} h^{l+L} \in \{g^r h^s \mid r, s \in \mathbf{Z}\}$

(iii) $(g^k h^l)^{-1} = h^{-l} g^{-k} = g^{-k} h^{-l} \in \{g^r h^s \mid r, s \in \mathbf{Z}\}$

□

§3.2 Cosets and Lagrange's Theorem

Definition 3.21 (Coset)**.** Let $H \leqslant G$. For $a \in G$, a ***left coset*** and ***right coset*** of H in G are

$$aH \coloneqq \{ah \mid h \in H\}$$
$$Ha \coloneqq \{ha \mid h \in H\}$$

Any element of a coset is called a *representative* for the coset.

The set of left cosets is given by

$$(G/H)_l \coloneqq \{aH \mid a \in G\}.$$

Similarly, the set of right cosets is given by

$$(G/H)_r \coloneqq \{Ha \mid a \in G\}.$$

Lemma 3.22. Let $H \leqslant G$. Then $aH = H$ if and only if $a \in H$. (Similarly, $Ha = H$ if and only if $a \in H$.)

Proof.

$\boxed{\Longrightarrow}$ Suppose $aH = H$. Then $ah \in H$ for some $h \in H$. Let $k = ah$, then $a = kh^{-1} \in H$.

$\boxed{\Longleftarrow}$ Let $a \in H$. Then $aH \subset H$.

Since $a^{-1} \in H$, $a^{-1}H \subset H$. Then $H = eH = (aa^{-1})H = a(a^{-1})H \subset aH$. Hence $aH = H$. □

The next result shows when two cosets are equal.

Lemma 3.23. Let $H \leqslant G$, $a, b \in G$. Then $aH = bH$ if and only if $a^{-1}b \in H$.

Proof.

$$\begin{aligned} aH = bH &\iff a^{-1}(aH) = a^{-1}bH \\ &\iff (a^{-1}a)H = (a^{-1}b)H \\ &\iff H = (a^{-1}b)H \end{aligned}$$

Note that from the previous result, $H = (a^{-1}b)H$ if and only if $a^{-1}b \in H$. □

Proposition 3.24. Let $H \leqslant G$. Then $(G/H)_l$ forms a partition of G. (Similar remarks hold for right cosets.)

We need to prove the following.

(i) For all $a \in G$, $aH \neq \emptyset$.

(ii) $\bigcup_{a \in G} aH = G$.

(iii) For every $a, b \in G$, $aH \cap bH = \emptyset$ or $aH = bH$.

Proof.

(i) Since $H \leqslant G$, $e \in H$. Thus for all $a \in G$, $a = ae \in aH$ so $aH \neq \emptyset$.

(ii) For all $a \in G$, $aH \subset G$, then $\bigcup_{a \in G} aH \subset G$. Note that $a \in G$ implies $a = ae \in aH$, and so $G = \bigcup_{a \in G} g \subset \bigcup_{a \in G} aH$. By double inclusion we are done.

(iii) If $aH \cap bH = \emptyset$, then we are done. If $aH \cap bH \neq \emptyset$ we need to show $aH = bH$. Let $x \in G$ such that $x \in aH \cap bH$. Then $x = ah_1 = bh_2$ for $h_1, h_2 \in H$ so $h_1 = a^{-1}bh_2$. Notice that $a^{-1}b = h_1 h_2^{-1} \in H$ and thus $aH = bH$.

□

Definition 3.25 (Index)**.** The number of left cosets of H in G is called the ***index*** of H in G, denoted by $|G : H|$.

The following result shows that H partitions G into equal-sized parts.

Lemma 3.26. The cosets of H in G are the same size as H; that is, for all $a \in G$, $|aH| = |H|$.

Proof. Let $f : H \to aH$ which sends $h \mapsto ah$. For $h_1, h_2 \in H$,

$$\begin{aligned} f(h_1) = f(h_2) &\implies ah_1 = ah_2 \\ &\implies a^{-1}ah_1 = a^{-1}ah_2 \\ &\implies h_1 = h_2 \end{aligned}$$

thus f is an injective mapping. Note that f is surjective by the definition of aH. Since f is bijective, $|H| = |aH|$. □

Theorem 3.27 (Lagrange's theorem)**.** Let G be a finite group, $H \leqslant G$. Then $|G| = |H|\,|G : H|$.

Proof. Let $|H| = n$, and let $|G : H| = k$. Since G is partitioned into k disjoint subsets, each of which has cardinality n, we have $|G| = kn$, or

$$|G| = |H|\,|G : H|$$

as desired. □

Theorem 3.28 (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; that is,

$$H = \{a^i \mid i \in \mathbf{Z}\}.$$

Since G is finite and $|H| < |G|$, H must be finite, so 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 $a^{-i} = \left(a^{-1}\right)^i$, 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

$$a^{|G|} = (a^k)^{\frac{|G|}{k}} = e.$$

□

Theorem 3.29 (Fermat–Euler Theorem (or Euler's totient theorem))**.** If a and N are coprime, then $a^{\phi(N)} \equiv 1 \pmod{N}$, where ϕ is Euler's totient function.

Proposition 3.30. A group of prime order is cyclic.

Definition 3.31. Let $H, K \leqslant G$, define

$$HK = \{hk \mid h \in H, k \in K\}.$$

Proposition 3.32. If $H, K \leqslant G$ are finite groups, then

$$|HK| = \frac{|H||K|}{|H \cap K|}.$$

Proof. Notice that HK is a union of left cosets of K, namely

$$HK = \bigcup_{h \in H} hK.$$

□

§3.3 Normal Subgroups, Quotient Groups

Definition 3.33 (Normal subgroup)**.** Let G be a group. $H \leqslant G$ is a ***normal subgroup*** of G, denoted by $H \triangleleft G$, if

$$aH = Ha \quad (\forall a \in G)$$

If G has no non-trivial normal subgroup, then G is a *simple group*.

Remark. This does *not* mean that $ah = ha$ for all $a \in G$, $h \in H$ or that G is abelian. Although we can easily see that all subgroups of abelian groups are normal. In general, a left coset does not equal the right coset.

Lemma 3.34. The following are equivalent.

(i) $H \triangleleft G$.

(ii) $ghg^{-1} \in H$ for all $g \in G$, $h \in H$.

(iii) $gHg^{-1} = H$ for all $g \in G$.

Proof.

$\boxed{\text{(i)} \iff \text{(ii)}}$ In the forward direction, $aH = Ha$ for all $a \in G$. Let $g \in G$, $x \in H$. Then $gH = Hg$ so $gx = h'g$ for some $h' \in H$. Then $gxg^{-1} = h'gg^{-1} = h' \in H$.

In the reverse direction, $ghg^{-1} \in H$ for all $g \in G$, $h \in H$. Fix g. Then $ghg^{-1} \in H$ implies $gh \in Hg$ for all $h \in H$. So $gH \subset Hg$. Similarly $gH \supset Hg$, so $gH = Hg$.

$\boxed{\text{(i)} \iff \text{(iii)}}$ $H \triangleleft G$ if and only if for all $g \in G$,

$$\begin{aligned} gH = Hg &\iff (gH)g^{-1} = (Hg)g^{-1} \\ &\iff gHg^{-1} = H \end{aligned}$$

□

Definition 3.35 (Quotient group)**.** Let G be a group, $H \triangleleft G$. Then the ***quotient group*** of G by H is

$$G/H \coloneqq \{aH \mid a \in G\}.$$

Proposition 3.36. G/H is a group under the following operation. Let $aH, bH \in G/H$. Then the product of aH and bH is $(aH)(bH)$.

$$(aH)(bH) = a(Hb)H = a(bH)H = abH$$

Proof. Check group axioms.

(i) For $a, b, c \in G$,

$$\begin{aligned}(aH)(bHcH) &= (aH)(bcH)\\ &= a(bc)H\\ &= (ab)cH\\ &= (aHbH)cH\end{aligned}$$

so the operation is associative.

(ii) The identity of G/H is the coset eH.

(iii) For $aH \in G/H$, the inverse of aH is $a^{-1}H$ as is immediate from the definition of the product.

□

Lemma 3.37. Let G be a finite group, $H \triangleleft G$. Then

$$|G/H| = |G : H| = \frac{|G|}{|H|}.$$

Definition 3.38 (Quotient map)**.** Let $H \triangleleft G$. The map $\pi : G \to G/H$ which sends $g \mapsto gH$ is called the ***quotient map***.

§3.4 Homomorphisms and Isomorphisms

In this section, we make precise the notion of when two groups "look the same"; that is, they have the same group-theoretic structure. This is the notion of an *isomorpism* between two groups.

Definitions and Examples

Definition 3.39 (Homomorphism)**.** Let $(G, *)$ and $(H, \diamond)$ be groups. A map $\phi : G \to H$ is called a ***homomorphism*** if, for all $x, y \in G$,

$$\phi(x * y) = \phi(x) \diamond \phi(y).$$

When the group operations for G and H are not explicitly written, we have

$$\phi(xy) = \phi(x)\phi(y).$$

Definition 3.40 (Isomorphism)**.** $\phi : G \to H$ is called an ***isomorphism*** if

(i) ϕ is a homomorphism;

(ii) ϕ is a bijection.

Then G and H are said to be ***isomorphic***, denoted by $G \cong H$.

Intuitively, G and H are the same group except that the elements and the operations may be written differently in G and H.

We also have the following terminology: An *automorphism* of a group G is an isomorphism from G to G. The automorphisms of G form a group $\mathrm{Aut}(G)$ under composition. An endomorphism of G is a homomorphism from G to G. (Rarely used) A *monomorphism* is an injective homomorphism and an *epimorphism* is a surjective homomorphism.

Example

For any group G, $G \cong G$ as the identity map provides an isomorphism from G to itself. (Exercise: prove that the identity map is the *only* isomorphism from G to itself.)

$\mathbf{Z} \cong 10\mathbf{Z}$ as the map $\phi : \mathbf{Z} \to 10\mathbf{Z}$ by $x \mapsto 10x$ is a homomorphism and a bijection.

Example

$(\mathbf{R}, +) \cong (\mathbf{R}^+, \times)$.

Proof. The exponential map $\exp : \mathbf{R} \to \mathbf{R}^+$ defined by $\exp(x) = e^x$ is an isomorphism from $(\mathbf{R}, +)$ to $(\mathbf{R}^+, \times)$.

(i) $\exp$ is a bijection since it has an inverse function (namely $\ln$).

(ii) exp preserves the group operations since $e^{x+y} = e^x e^y$.

We see that both the elements and the operations are different yet the two groups are isomorphic, that is, as groups they have identical structures. □

Proposition 3.41. Let $\phi : G \to H$ be a homomorphism. Let $g \in G$, $n \in \mathbf{Z}$. Then

(i) $\phi(e_G) = e_H$;

(ii) $\phi(g^{-1}) = (\phi(g))^{-1}$;

(iii) $\phi(g^n) = (\phi(g))^n$.

Proof.

(i) $\phi(e_G) = \phi(e_G e_G) = \phi(e_G)\phi(e_G)$, then apply $\phi(e_G)^{-1}$ to both sides to get $\phi(e_G) = e_H$.

(ii) $\phi(g)\phi(g^{-1}) = \phi(gg^{-1}) = \phi(e_G) = e_H$.

(iii) Note more generally that we can show $\phi(g^n) = (\phi(g))^n$ for $n > 0$ by induction. For $n = -k < 0$ we have

$$\phi(g^n) = \phi((g^{-1})^k) = (\phi(g^{-1}))^k = (\phi(g)^{-1})^k = \phi(g)^n.$$

□

Theorem 3.42. Any two cyclic groups of the same order are isomorphic.

Proof. Suppose $\langle x \rangle$ and $\langle y \rangle$ are both cyclic groups of order n. We first prove the case where $n < \infty$. We claim that the map $\phi : \langle x \rangle \to \langle y \rangle$ which sends $x^k \mapsto y^k$ is an isomorphism.

Lemma. Let G be a group, $g \in G$, let $m, n \in \mathbf{Z}$. Denote $d = \gcd(m, n)$. If $g^n = e$ and $g^m = e$, then $g^d = e$.

Proof. By Bezout's lemma, since $d = \gcd(m, n)$, then there exists $q, r \in \mathbf{Z}$ such that $qm + rn = d$. Thus

$$g^d = g^{qm+rn} = (g^m)^q (g^n)^r = e.$$

□

We first show that ϕ is well-defined; that is, $x^r = x^s \implies \phi(x^r) = \phi(x^s)$. Note that $x^{r-s} = e$, so by the above lemma, $n \mid r - s$. Write $r = tn + s$ for some $t \in \mathbf{Z}$, so

$$\phi(x^r) = \phi(x^{tn+s}) = y^{tn+s} = (y^n)^t y^s = y^s = \phi(x^s).$$

We then show that ϕ is a homomorphism:

$$\phi(x^a x^b) = \phi(x^{a+b}) = y^{a+b} = y^a y^b = \phi(x^a)\phi(x^b).$$

Finally we show that ϕ is bijective. Since the element y^k of $\langle y \rangle$ is in the image of x^k under ϕ, ϕ is surjective. Since both groups have the same finite order, any surjection from one to the other is a bijection. Therefore ϕ is an isomorphism.

We now prove the case where $n = \infty$. If $\langle x \rangle$ is an infinite cyclic group, let $\phi : \mathbf{Z} \to \langle x \rangle$ be defined by $\phi(k) = x^k$. (This map is well-defined since there is no ambiguity in the representation of elements in the domain.)

Since $x^a \neq x^b$ for all distinct $a, b \in \mathbf{Z}$, ϕ is injective. By definition of a cyclic group, ϕ is surjective. As above, the laws of exponents ensure ϕ is a homomorphism. Hence ϕ is an isomorphism. □

Kernel and Image

Definition 3.43 (Kernel and image)**.** Let $\phi : G \to H$ be a homomorphism. Then the ***kernel*** of ϕ is

$$\ker \phi \coloneqq \{g \in G \mid \phi(g) = e_H\} \subset G.$$

The ***image*** of G under ϕ is

$$\operatorname{im} \phi \coloneqq \phi(G) = \{\phi(g) \mid g \in G\} \subset H.$$

Remark. $\operatorname{im} \phi$ is the usual set theoretic image of ϕ.

Proposition 3.44. Let $\phi : G \to H$ be a homomorphism. Then

(i) $\ker \phi \triangleleft G$;

(ii) $\operatorname{im} \phi \leqslant H$.

Proof.

(i) Apply the subgroup criterion. Since $e_G \in \ker \phi$, $\ker \phi \neq \emptyset$. Let $x, y \in \ker \phi$; that is, $\phi(x) = \phi(y) = e_H$. Then

$$\phi(xy^{-1}) = \phi(x)\phi(y)^{-1} = e_H$$

so $xy^{-1} \in \ker \phi$. By the subgroup criterion, $\ker \phi \leqslant G$.

(ii) Since $\phi(e_G) = e_H$, $e_H \in \operatorname{im} \phi$ so $\operatorname{im} \phi \neq \emptyset$. Let $x, y \in \operatorname{im} \phi$. Then there exists $a, b \in G$ such that $\phi(a) = x$, $\phi(b) = y$. Then

$$xy^{-1} = \phi(a)\phi(b)^{-1} = \phi(ab^{-1})$$

so $xy^{-1} \in \operatorname{im} \phi$. By the subgroup criterion, $\operatorname{im} \phi \leqslant G$.

□

Proposition 3.45. Let $\phi : G \to H$ be a homomorphism. Then ϕ is injective if and only if $\ker \phi = \{e_G\}$.

Proof.

$\boxed{\Longrightarrow}$ Suppose ϕ is injective. Since $\phi(e_G) = e_H$, $e_G \in \ker\phi$ so $\{e_G\} \subset \ker\phi$.

Conversely, let $g \in \ker\phi$, so $\phi(g) = e_H$. Then $\phi(g) = e_H = \phi(e_H)$, so by injectivity, $g = e_G$ and thus $g = e_G$. Hence $\ker\phi \subset \{e_G\}$, so $\ker\phi = \{e_G\}$.

$\boxed{\Longleftarrow}$ Suppose $\ker\phi = \{e_G\}$. Suppose $\phi(a) = \phi(b)$, then

$$\begin{aligned}\phi(a) &= \phi(b)\\ \phi(a)\phi(b)^{-1} &= \phi(b)\phi(b)^{-1}\\ \phi(a)\phi(b)^{-1} &= e_H\\ \phi(ab^{-1}) &= e_H\end{aligned}$$

Hence $ab^{-1} \in \ker\phi = \{e_G\}$, so $ab^{-1} = e_G$ and thus $a = b$. Therefore ϕ is injective. □

Isomorphism Theorems

Theorem 3.46 (First isomorphism theorem)**.** Let $\phi : G \to H$ be a homomorphism. Then

$$G/\ker\phi \cong \operatorname{im}\phi(G).$$

Theorem 3.47 (Second isomorphism theorem)**.** Let $A, B \leqslant G$, $A \leqslant N_G(B)$. Then

(i) $AB \leqslant G$;

(ii) $B \triangleleft AB$;

(iii) $A \cap B \triangleleft A$;

(iv) $AB/B \cong A/A \cap B$.

Theorem 3.48 (Third isomorphism theorem)**.** Let $H, K \triangleleft G$, $H \leqslant K$. Then $K/H \triangleleft G/H$, and

$$(G/H)/(K/H) \cong G/K.$$

If we denote the quotient by H with a bar, this can be written

$$\overline{G}/\overline{K} \cong G/K.$$

Theorem 3.49 (Fourth isomorphism theorem)**.**

Theorem 3.50 (Cayley's theorem)**.**

§3.5 Group Actions

We move now, from thinking of groups in their own right, to thinking of how groups can move sets around—for example, how S_n permutes $\{1, 2, \ldots, n\}$ and matrix groups move vectors.

Definition 3.51 (Group action)**.** A ***group action*** of a group G on a set A is a map from $G \times A \to A$ (written as $g \cdot a$, for all $g \in G$, $a \in A$) satisfying the following properties:

(i) $g_1 \cdot (g_2 \cdot a) = (g_1 g_2) \cdot a$, for all $g_1, g_2 \in G$, $a \in A$;

(ii) $e_G \cdot a = a$ for all $a \in A$.

We say that G is a group acting on a set A.

Intuitively, a group action of G on a set A means that every element g in G acts as a permutation on A in a manner consistent with the group operations in G. There is also a notion of left *action* and *right action*.

For the following defintions, let G be a group, and $A \subset G$ be non-empty.

Definition 3.52 (Centraliser)**.** The ***centraliser*** of A in G is defined by

$$C_G(A) \coloneqq \{g \in G \mid \forall a \in A, gag^{-1} = a\}.$$

Since $gag^{-1} = a$ if and only if $ga = ag$, $C_G(A)$ is the set of elements of G which commute with every element of A.

We check that $C_G(A) \leqslant G$:

(i) $e \in C_G(A)$, so $C_G(A) \neq \emptyset$.

(ii) Let $x, y \in C_G(A)$; that is, for all $a \in A$, $xax^{-1} = a$ and $yay^{-1} = a$. Then

$$\begin{aligned}(xy)a(xy)^{-1} &= (xy)a(y^{-1}x^{-1}) \\ &= x(yay^{-1})x^{-1} \\ &= xax^{-1} = a\end{aligned}$$

so $xy \in C_G(A)$. Hence $C_G(A)$ is closed under products.

(iii) Let $x \in C_G(A)$; that is, for all $a \in A$, $xax^{-1} = a$. Applying x^{-1} to both sides gives $ax^{-1} = x^{-1}a$. Applying x to both sides gives $a = x^{-1}ax$, so $x^{-1} \in C_G(A)$. Hence $C_G(A)$ is closed under taking inverses.

Notation. In the special case when $A = \{a\}$ we simply write $C_G(a)$ instead of $C_G(\{a\})$. In this case $a^n \in C_G(a)$ for all $n \in \mathbf{Z}$.

Definition 3.53 (Centre)**.** The ***centre*** of G is the set of elements which commute with all the elements of G:

$$Z(G) \coloneqq \{g \in G \mid \forall x \in G, gx = xg\}.$$

Note that $Z(G) = C_G(G)$, so the argument above proves $Z(G) \leqslant G$ as a special case.

Definition 3.54 (Normaliser)**.** Define $gAg^{-1} = \{gag^{-1} \mid a \in A\}$. The ***normaliser*** of A in G is

$$N_G(A) \coloneqq \{g \in G \mid gAg^{-1} = A\}.$$

Notice that if $g \in C_G(A)$, then $gag^{-1} = a \in A$ for all $a \in A$ so $C_G(A) \leqslant N_G(A)$. The proof that $N_G(A) \leqslant G$ is similar to the one that $C_G(A) \leqslant G$.

Definition 3.55 (Stabiliser)**.** If G is a group acting on a set S, $s \in S$, then the ***stabiliser*** of s in G is

$$G_s \coloneqq \{g \in G \mid g \cdot s = s\}.$$

Notation. Denote the set of all fixed points to be $S^G = \{s \in S \mid \forall g \in G, gs = g\}$.

We check that $G_s \leqslant G$:

(i) By definition of group action, $e_G \cdot a = a$, so $e_G \in G_s$.

(ii) Let $x, y \in G_s$, then

$$\begin{aligned}(xy) \cdot s &= x \cdot (y \cdot s)\\ &= x \cdot s = s\end{aligned}$$

so $xy \in G_s$. Hence G_s is closed under products.

(iii) Let $x \in G_s$; that is, $x \cdot s = s$. Then

$$\begin{aligned}x^{-1} \cdot s &= x^{-1} \cdot (x \cdot s)\\ &= (x^{-1}x) \cdot s\\ &= e \cdot s = s\end{aligned}$$

so $x^{-1} \in G_s$. Hence G_s is closed under taking inverses.

Definition 3.56. The *kernel* of the action of G on S is

$$\{g \in G \mid \forall s \in S, g \cdot s = s\}.$$

Definition 3.57 (Orbit)**.** Let G be a group that acts on a set S. Define the ***orbit*** of a group element $s \in S$ as

$$G(s) \coloneqq \{g \cdot s \in S \mid g \in G\}.$$

Conjugation

Sylow's Theorem

Definition 3.58 (Sylow p-subgroup)**.** Let G be a group, and let p be a prime.

(i) A group of order p^α $(\alpha \geqslant 1)$ is called a *p-group*. Subgroups of G which are p-groups are called *p-subgroups*.

(ii) If $|G| = p^\alpha m$ $(p \nmid m)$, then a subgroup of order p^α is called a ***Sylow p-subgroup*** of G.

Notation. The set of Sylow p-subgroups of G is denoted by $Syl_p(G)$, and the number of Sylow p-subgroups of G is denoted by $n_p(G)$ (or just n_p when G is clear from the context).

Theorem 3.59 (Sylow's theorem)**.** Let $|G| = p^\alpha m$, where p is a prime and $p \nmid m$.

(i) Sylow p-subgroups of G exist, i.e. $Syl_p(G) \neq \emptyset$.

(ii) If P is a Sylow p-subgroup of G, and Q is any p-subgroup of G, then there exists $g \in G$ such that $Q \leqslant gPg^{-1}$, i.e. Q is contained in some conjugate of P. In particular, any two Sylow p-subgroups of G are conjugate in G.

(iii) $n_p \equiv 1 \pmod{p}$. Furthermore, n_p is the index in G of the normaliser $N_G(P)$ for any Sylow p-subgroup P, hence $n_p \mid m$.

§3.6 Group Product, Finite Abelian Groups

Definition 3.60 (Direct product)**.** The ***direct product*** $G_1 \times \cdots \times G_n$ of the groups $(G_1, *_1), \ldots, (G_n, *_n)$ is the Cartesian product

$$G_1 \times \cdots \times G_n \coloneqq \{(g_1, \ldots, g_n) \mid g_i \in G_i\}$$

with operation defined componentwise:

$$(g_1, \ldots, g_n) * (h_1, \ldots, h_n) = (g_1 *_1 h_1, \ldots, g_n *_n h_n).$$

Proposition 3.61. If $G_1, \ldots, G_n$ are groups, then

$$|G_1 \times \cdots \times G_n| = |G_1|\,|G_2| \cdots |G_n|.$$

Proof. Let $G = G_1 \times \cdots \times G_n$. The proof that the group axioms hold for G is straightforward since each axiom is a consequence of the fact that the same axiom holds for each G_i, and the operation on G defined componentwise.

The number of n-tuples in G follows from simple combinatorics. □

III

Linear Algebra

4 Vector Spaces

§4.1 Definition of Vector Space

Notation. A field is denoted by $\mathbf{F}$, which can mean either $\mathbf{R}$ or $\mathbf{C}$. $\mathbf{F}^n$ is the set of n-tuples whose elements belong to $\mathbf{F}$:

$$\mathbf{F}^n := \{(x_1, \ldots, x_n) \mid x_i \in \mathbf{F}\}$$

For $(x_1, \ldots, x_n) \in \mathbf{F}^n$ and $i = 1, \ldots, n$, we say that x_i is the i-th coordinate of $(x_1, \ldots, x_n)$.

Definition 4.1 (Vector space)**.** V is a ***vector space*** over $\mathbf{F}$ if the following properties hold:

(i) Addition is commutative: $u + v = v + u$ for all $u, v \in V$

(ii) Addition is associative: $(u + v) + w = u + (v + w)$ for all $u, v, w \in V$

Multiplication is associative: $(ab)v = a(bv)$ for all $v \in V$, $a, b \in \mathbf{F}$

(iii) Additive identity: there exists $\mathbf{0} \in V$ such that $v + \mathbf{0} = v$ for all $v \in V$

(iv) Additive inverse: for every $v \in V$, there exists $w \in V$ such that $v + w = \mathbf{0}$

(v) Multiplicative identity: $1v = v$ for all $v \in V$

(vi) Distributive properties: $a(u + v) = au + av$ and $(a + b)v = av + bv$ for all $a, b, \in \mathbf{F}$ and $u, v \in V$

Notation. For the rest of this text, V denotes a vector space over $\mathbf{F}$.

> **Example**
>
> $\mathbf{R}^n$ is a vector space over $\mathbf{R}$, $\mathbf{C}^n$ is a vector space over $\mathbf{C}$.

Elements of a vector space are called *vectors* or *points*.

The scalar multiplication in a vector space depends on $\mathbf{F}$. Thus when we need to be precise, we will say that V is a vector space over $\mathbf{F}$ instead of saying simply that V is a vector space. For example, $\mathbf{R}^n$ is a vector space over $\mathbf{R}$, and $\mathbf{C}^n$ is a vector space over $\mathbf{C}$. A vector space over $\mathbf{R}$ is called a *real vector space*; a vector space over $\mathbf{C}$ is called a *complex vector space*.

Proposition 4.2 (Uniqueness of additive identity)**.** A vector space has a unique additive identity.

Proof. Suppose otherwise, then $\mathbf{0}$ and $\mathbf{0}'$ are additive identities of V. Then

$$\mathbf{0}' = \mathbf{0}' + \mathbf{0} = \mathbf{0} + \mathbf{0}' = \mathbf{0}$$

where the first equality holds because $\mathbf{0}$ is an additive identity, the second equality comes from commutativity, and the third equality holds because $\mathbf{0}'$ is an additive identity. Thus $\mathbf{0}' = \mathbf{0}$. □

Proposition 4.3 (Uniqueness of additive inverse)**.** Every element in a vector space has a unique additive inverse.

Proof. Suppose otherwise, then for $v \in V$, w and w' are additive inverses of v. Then

$$w = w + \mathbf{0} = w + (v + w') = (w + v) + w' = \mathbf{0} + w' = w'.$$

Thus $w = w'$. □

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 $w + (-v)$.

We now prove some seemingly trivial facts.

Proposition 4.4.

(i) For every $v \in V$, $0v = \mathbf{0}$.

(ii) For every $a \in \mathbf{F}$, $a\mathbf{0} = \mathbf{0}$.

(iii) For every $v \in V$, $(-1)v = -v$.

Proof.

(i) For $v \in V$, we have

$$0v = (0 + 0)v = 0v + 0v.$$

Adding the additive inverse of $0v$ to both sides of the equation gives $\mathbf{0} = 0v$.

(ii) For $a \in \mathbf{F}$, we have

$$a\mathbf{0} = a(\mathbf{0} + \mathbf{0}) = a\mathbf{0} + a\mathbf{0}.$$

Adding the additive inverse of $a\mathbf{0}$ to both sides of the equation gives $\mathbf{0} = a\mathbf{0}$.

(iii) For $v \in V$, we have

$$v + (-1)v = 1v + (-1)v = (1 + (-1))v = 0v = \mathbf{0}.$$

Since $v + (-1)v = \mathbf{0}$, $(-1)v$ is the additive inverse of v.

□

Example

$\mathbf{F}^\infty$ is defined to be the set of all sequences of elements of $\mathbf{F}$:

$$\mathbf{F}^\infty \coloneqq \{(x_1, x_2, \dots) \mid x_i \in \mathbf{F}\}$$

- Addition on $\mathbf{F}^\infty$ is defined by

$$(x_1, x_2, \dots) + (y_1, y_2, \dots) = (x_1 + y_1, x_2 + y_2, \dots)$$

- Scalar multiplication on $\mathbf{F}^\infty$ is defined by

$$\lambda(x_1, x_2, \dots) = (\lambda x_1, \lambda x_2, \dots)$$

Verify that $\mathbf{F}^\infty$ becomes a vector space over $\mathbf{F}$. Also verify that the additive identity in $\mathbf{F}^\infty$ is $\mathbf{0} = (0, 0, \dots)$.

Our next example of a vector space involves a set of functions.

Example

If S is a set, $\mathbf{F}^S \coloneqq \{f \mid f : S \to \mathbf{F}\}$.

- Addition on $\mathbf{F}^S$ is defined by

$$(f + g)(x) = f(x) + g(x) \quad (\forall x \in S)$$

  for all $f, g \in \mathbf{F}^S$.

- Multiplication on $\mathbf{F}^S$ is defined by

$$(\lambda f)(x) = \lambda f(x) \quad (\forall x \in S)$$

  for all $\lambda \in \mathbf{F}$, $f \in \mathbf{F}^S$.

Verify that if S is a non-empty set, then $\mathbf{F}^S$ is a vector space over $\mathbf{F}$.

Also verify that the additive identity of $\mathbf{F}^S$ is the function $0 : S \to \mathbf{F}$ defined by

$$0(x) = 0 \quad (\forall x \in S)$$

and for $f \in \mathbf{F}^S$, additive inverse of f is the function $-f : S \to \mathbf{F}$ defined by

$$(-f)(x) = -f(x) \quad (\forall x \in S)$$

Remark. $\mathbf{F}^n$ and $\mathbf{F}^\infty$ are special cases of the vector space $\mathbf{F}^S$; think of $\mathbf{F}^n$ as $\mathbf{F}^{\{1,2,\dots,n\}}$, and $\mathbf{F}^\infty$ as $\mathbf{F}^{\{1,2,\dots\}}$.

Example (Complexification)

Suppose V is a real vector space. The *complexifcation* of V, denoted by $V_{\mathbf{C}}$, equals $V \times V$. An element of $V_{\mathbf{C}}$ is an ordered pair (u, v), where $u, v \in V$, which we write as $u + iv$.

- Addition on $V_{\mathbf{C}}$ is defined by

$$(u_1 + iv_1) + (u_2 + iv_2) = (u_1 + u_2) + i(v_1 + v_2)$$

  for all $u_1, v_2, u_2, v_2 \in V$.

- Complex scalar multiplication on $V_{\mathbf{C}}$ is defined by

$$(a + bi)(u + iv) = (au - bv) + i(av + bu)$$

  for all $a, b \in \mathbf{R}$ and all $u, v \in V$.

You should verify that with the defnitions of addition and scalar multiplication as above, $V_{\mathbf{C}}$ is a (complex) vector space.

§4.2 Subspaces

Whenever we have a mathematical object with some structure, we want to consider subsets that also have the same structure.

Definition 4.5 (Subspace)**.** $U \subset V$ is a ***subspace*** of V if U is also a vector space (with the same addition and scalar multiplication as on V). We denote this as $U \leqslant V$.

The sets $\{\mathbf{0}\}$ and V are always subspaces of V. The subspace $\{\mathbf{0}\}$ is called the *zero subspace* or *trivial subspace*. Subspaces other than V are called *proper subspaces*.

The following result is useful in determining whether a given subset of V is a subspace of V.

Lemma 4.6 (Subspace test)**.** Suppose $U \subset V$. Then $U \leqslant V$ if and only if U satisfies the following conditions:

(i) Additive identity: $\mathbf{0} \in U$

(ii) Closed under addition: $u + w \in U$ for all $u, w \in U$

(iii) Closed under scalar multiplication: $\lambda u \in U$ for all $\lambda \in \mathbf{F}$, $u \in U$

Proof.

$\boxed{\Longrightarrow}$ If $U \leqslant V$, then U satisfies the three conditions above by the definition of vector space.

$\boxed{\Longleftarrow}$ Suppose U satisfies the three conditions above. (i) ensures that the additive identity of V is in U. (ii) ensures that addition makes sense on U. (iii) ensures that scalar multiplication makes sense on U.

If $u \in U$, then $-u = (-1)u \in U$ by (iii). Hence every element of U has an additive inverse in U.

The other parts of the definition of a vector space, such as associativity and commutativity, are automatically satisfied for U because they hold on the larger space V. Thus U is a vector space and hence is a subspace of V. □

Proposition 4.7. Suppose $U \leqslant V$. Then

(i) U is a vector space over $\mathbf{F}$. In fact, the only subsets of V that are vector spaces over $\mathbf{F}$ are the subspaces of V;

(ii) if $W \leqslant U$, then $W \leqslant V$ ("a subspace of a subspace is a subspace").

Proof.

(i) We first check that we have legitimate operations. Since U is closed under addition, the operation $+$ restricted to U gives a map $U \times U \to U$. Likewise since U is closed under scalar multiplication, that operation restricted to U gives a map $\mathbf{F} \times U \to U$.

We now check that U satisfies the vector space axioms.

(i) Commutativity and associativity of addition are inherited from V.

(ii) There is an additive identity (by the subspace test).

(iii) There are additive inverses: if $u \in U$ then multiplying by $-1 \in \mathbf{F}$ and shows that $-u = (-1)u \in U$.

(iv) The remaining four properties are all inherited from V. That is, they apply to general vectors of V and vectors in U are vectors in V.

(ii) This is immediate from the definition of a subspace.

□

Definition 4.8 (Sum of subsets)**.** Suppose $U_1, \dots, U_n \subset V$. The sum of $U_1, \dots, U_n$ is the set of all possible sums of elements of $U_1, \dots, U_n$:

$$U_1 + \cdots + U_n \coloneqq \{u_1 + \cdots + u_n \mid u_i \in U_i\}.$$

Example

Suppose that $U = \{(x, 0, 0) \in \mathbf{F}^3 \mid x \in F\}$ and $W = \{(0, y, 0) \in \mathbf{F}^3 \mid y \in \mathbf{F}\}$. Then

$$U + W = \{(x, y, 0) \mid x, y \in \mathbf{F}\}.$$

Suppose that $U = \{(x, x, y, y) \in \mathbf{F}^4 \mid x, y \in \mathbf{F}\}$ and $W = \{(x, x, x, y) \in \mathbf{F}^4 \mid x, y \in \mathbf{F}\}$. Then

$$U + W = \{(x, x, y, z) \in \mathbf{F}^4 \mid x, y, z \in \mathbf{F}\}.$$

The next result states that the sum of subspaces is a subspace, and is in fact the smallest subspace containing all the summands.

Proposition 4.9. Suppose $U_1, \dots, U_n \leqslant V$. Then $U_1 + \cdots + U_n$ is the smallest subspace of V containing $U_1, \dots, U_n$.

Proof. It is easy to see that $\mathbf{0} \in U_1 + \cdots + U_n$ and that $U_1 + \cdots + U_n$ is closed under addition and scalar multiplication. Hence by the subspace test, $U_1 + \cdots + U_n \leqslant V$.

Let M be the smallest subspace of V containing $U_1, \dots, U_n$. We want to show that $U_1 + \cdots + U_n = M$. To do so, we show double inclusion: $U_1 + \cdots + U_n \subset M$ and $M \subset U_1 + \cdots + U_n$.

(i) For all $u_i \in U_i$ $(1 \leqslant i \leqslant n)$,

$$u_i = \mathbf{0} + \cdots + \mathbf{0} + u_i + \mathbf{0} + \cdots + \mathbf{0} \in U_1 + \cdots + U_n,$$

where all except one of the u's are $\mathbf{0}$. Thus $U_i \subset U_1 + \cdots + U_n$ for $1 \leqslant i \leqslant n$. Hence $M \subset U_1 + \cdots + U_n$.

(ii) Conversely, every subspace of V containing $U_1, \dots, U_n$ contains $U_1 + \cdots + U_n$ (because subspaces must contain all finite sums of their elements). Hence $U_1 + \cdots + U_n \subset M$.

□

Definition 4.10 (Direct sum)**.** Suppose $U_1, \ldots, U_n \leqslant V$. If each element of $U_1 + \cdots + U_n$ can be written in only one way as a sum $u_1 + \cdots + u_n$, $u_i \in U_i$, then $U_1 + \cdots + U_n$ is called a ***direct sum***. In this case, we denote the sum as

$$U_1 \oplus \cdots \oplus U_n.$$

Example

Suppose that $U = \{(x, y, 0) \in \mathbf{F}^3 \mid x, y \in \mathbf{F}\}$ and $W = \{(0, 0, z) \in \mathbf{F}^3 \mid z \in \mathbf{F}\}$. Then $\mathbf{F}^3 = U \oplus W$.

Suppose U_i is the subspace of $\mathbf{F}^n$ of those vectors whose coordinates are all 0 except for the i-th coordinate; that is, $U_i = \{(0, \ldots, 0, x, 0, \ldots, 0) \in \mathbf{F}^n \mid x \in \mathbf{F}\}$. Then $\mathbf{F}^n = U_1 \oplus \cdots \oplus U_n$.

Lemma 4.11 (Condition for direct sum)**.** Suppose $V_1, \ldots, V_n \leqslant V$, let $W = V_1 + \cdots + V_n$. Then the following are equivalent:

(i) Any element in W can be uniquely expressed as the sum of vectors in $V_1, \ldots, V_n$.

(ii) If $v_i \in V_i$ satisfies $v_1 + \cdots + v_n = \mathbf{0}$, then $v_1 = \cdots = v_n = \mathbf{0}$.

(iii) For $k = 2, \ldots, n$, $(V_1 + \cdots + V_{k-1}) \cap V_k = \{\mathbf{0}\}$.

Proof.

(i) $\iff$ (ii) First suppose W is a direct sum. Then by the definition of direct sum, the only way to write $\mathbf{0}$ as a sum $u_1 + \cdots + u_n$ is by taking $u_i = \mathbf{0}$.

Now suppose that the only way to write $\mathbf{0}$ as a sum $v_1 + \cdots + v_n$ by taking $v_1 = \cdots = v_n = \mathbf{0}$. For $v \in V_1 + \cdots + V_n$, suppose that there is more than one way to represent v:

$$\begin{aligned} v &= v_1 + \cdots + v_n \\ v &= v_1' + \cdots + v_n' \end{aligned}$$

for some $v_i, v_i' \in V_i$. Substracting the above two equations gives

$$\mathbf{0} = (v_1 - v_1') + \cdots + (v_n - v_n').$$

Since $v_i - v_i' \in V_i$, we have $v_i - v_i' = \mathbf{0}$ so $v_i = v_i'$. Hence there is only one unique way to represent $v_1 + \cdots + v_n$, thus W is a direct sum.

(ii) $\iff$ (iii) First suppose if $v_i \in V_i$ satisfies $v_1 + \cdots + v_n = \mathbf{0}$, then $v_1 = \cdots = v_n = \mathbf{0}$. Let $v_k \in (V_1 + \cdots + V_{k-1}) \cap V_k$. Then $v_k = v_1 + \cdots + v_{k-1}$ where $v_i \in V_i$ $(1 \leqslant i \leqslant k-1)$. Thus

$$\begin{aligned} v_1 + \cdots + v_{k-1} - v_k &= \mathbf{0} \\ v_1 + \cdots + v_{k-1} + (-v_k) + \mathbf{0} + \cdots + \mathbf{0} &= \mathbf{0} \end{aligned}$$

by taking $v_{k+1} = \cdots = v_n = \mathbf{0}$. Then $v_1 = \cdots = v_k = \mathbf{0}$.

Now suppose that for $k = 2, \ldots, n$, $(V_1 + \cdots + V_{k-1}) \cap V_k = \{\mathbf{0}\}$.

$$v_1 + \cdots + v_n = \mathbf{0}$$
$$v_1 + \cdots + v_{n-1} = -v_n$$

where $v_1 + \cdots + v_{n-1} \in V_1 + \cdots + V_{n-1}$, $-v_n \in V_n$. Thus

$$v_1 + \cdots + v_{n-1} = -v_n \in (V_1 + \cdots + V_{n-1}) \cap V_n = \{\mathbf{0}\}$$

so $v_1 + \cdots + v_{n-1} = \mathbf{0}$, $v_n = \mathbf{0}$. Induction on n gives $v_1 = \cdots = v_{n-1} = v_n = \mathbf{0}$. □

Proposition 4.12. Suppose $U, W \leqslant V$. Then $U + W$ is a direct sum if and only if $U \cap W = \{\mathbf{0}\}$.

Proof.

$\boxed{\Longrightarrow}$ Suppose that $U + W$ is a direct sum. If $v \in U \cap W$, then $\mathbf{0} = v + (-v)$, where $v \in U$, $-v \in W$. By the unique representation of $\mathbf{0}$ as the sum of a vector in U and a vector in W, we have $v = \mathbf{0}$. Thus $U \cap W = \{\mathbf{0}\}$.

$\boxed{\Longleftarrow}$ Suppose $U \cap W = \{\mathbf{0}\}$. Suppose $u \in U$, $w \in W$, and $0 = u + w$. $u = -w \in W$, thus $u \in U \cap W$, so $u = w = \mathbf{0}$. By Lemma 4.11, $U + W$ is a direct sum. □

§4.3 Span and Linear Independence

Definition 4.13 (Linear combination)**.** v is a ***linear combination*** of vectors $v_1, \ldots, v_n \in V$ if there exists $a_1, \ldots, a_n \in \mathbf{F}$ such that

$$v = a_1 v_1 + \cdots + a_n v_n.$$

Definition 4.14 (Span)**.** The ***span*** of $\{v_1, \ldots, v_n\}$ is the set of all linear combinations of $v_1, \ldots, v_n$:

$$\operatorname{span}(v_1, \ldots, v_n) \coloneqq \{a_1 v_1 + \cdots + a_n v_n \mid a_i \in \mathbf{F}\}.$$

The span of the empty set $\{\ \}$ is defined to be $\{\mathbf{0}\}$.

We say that $v_1, \ldots, v_n$ *spans* V if $\operatorname{span}(v_1, \ldots, v_n) = V$.

If $S \subset V$ is such that $\operatorname{span}(S) = V$, then we say that S *spans* V, and that S is a *spanning set* for V:

$$\operatorname{span}(S) \coloneqq \{a_1 v_1 + \cdots + a_n v_n \mid v_i \in S, a_i \in \mathbf{F}\}.$$

Proposition 4.15. $\operatorname{span}(v_1, \ldots, v_n)$ in V is the smallest subspace of V containing $v_1, \ldots, v_n$.

Proof. First we show that $\operatorname{span}(v_1, \ldots, v_n) \leqslant V$, using the subspace test.

(i) $\mathbf{0} = 0v_1 + \cdots + 0v_n \in \operatorname{span}(v_1, \ldots, v_n)$

(ii) $(a_1 v_1 + \cdots + a_n v_n) + (c_1 v_1 + \cdots + c_n v_n) = (a_1 + c_1)v_1 + \cdots + (a_n + c_n)v_n \in \operatorname{span}(v_1, \ldots, v_n)$, so $\operatorname{span}(v_1, \ldots, v_n)$ is closed under addition.

(iii) $\lambda(a_1 v_1 + a_n v_n) = (\lambda a_1)v_1 + \cdots + (\lambda a_n)v_n \in \operatorname{span}(v_1, \ldots, v_n)$, so $\operatorname{span}(v_1, \ldots, v_n)$ is closed under scalar multiplication.

Let M be the smallest vector subspace of V containing $v_1, \ldots, v_n$. We claim that $M = \operatorname{span}(v_1, \ldots, v_n)$. To show this, we show that (i) $M \subset \operatorname{span}(v_1, \ldots, v_n)$ and (ii) $M \supset \operatorname{span}(v_1, \ldots, v_n)$.

(i) Each v_i is a linear combination of $v_1, \ldots, v_n$, as

$$v_i = 0 \cdot v_1 + \cdots + 0 \cdot v_{i-1} + 1 \cdot v_i + 0 \cdot v_{i+1} + \cdots + 0 \cdot v_n,$$

so by the definition of the span as the collection of all linear combinations of $v_1, \ldots, v_n$, we have that $v_i \in \operatorname{span}(v_1, \ldots, v_n)$. But M is the smallest vector subspace containing $v_1, \ldots, v_n$, so

$$M \subset \operatorname{span}(v_1, \ldots, v_n).$$

(ii) Since $v_i \in M$ $(1 \leqslant i \leqslant n)$ and M is a vector subspace (closed under addition and scalar multiplication), it follows that

$$a_1 v_1 + \cdots + a_n v_n \in M$$

for all $a_i \in \mathbf{F}$ (i.e. M contains all linear combinations of $v_1, \ldots, v_n$). So

$$\operatorname{span}(v_1, \ldots, v_n) \subset M.$$

□

Definition 4.16 (Finite-dimensional vector space)**.** V is ***finite-dimensional*** if there exists some list of vector $\{v_1, \ldots, v_n\}$ that spans V; otherwise, it is *infinite-dimensional*.

Remark. Recall that by defnition every list of vectors has finite length.

Remark. From this definition, infinite-dimensionality is the negation of finite-dimensionality (i.e. *not* finite-dimensional). Hence to prove that a vector space is infinite-dimensional, we prove by contradiction; that is, first assume that the vector space is finite-dimensional, then try to come to a contradiction.

Exercise

For positive integer n, $\mathbf{F}^n$ is finite-dimensional.

Proof. Suppose $(x_1, x_2, \ldots, x_n) \in \mathbf{F}^n$, then

$$(x_1, x_2, \ldots, x_n) = x_1(1, 0, \ldots, 0) + x_2(0, 1, \ldots, 0) + \cdots + x_n(0, 0, \ldots, 1)$$

so

$$(x_1, \ldots, x_n) \in \operatorname{span}\left((1, 0, \ldots, 0), (0, 1, \ldots, 0), \ldots, (0, \ldots, 0, 1)\right).$$

The vectors $(1, 0, \ldots, 0), (0, 1, \ldots, 0), \ldots, (0, \ldots, 0, 1)$ spans $\mathbf{F}^n$, so $\mathbf{F}^n$ is finite-dimensional. □

Definition 4.17 (Linear independence)**.** A list of vectors $v_1, \ldots, v_n$ is ***linearly independent*** in V if the only choice of $a_1, \ldots, a_n \in \mathbf{F}$ that makes

$$a_1 v_1 + \cdots + a_n v_n = \mathbf{0}$$

is $a_1 = \cdots = a_n = 0$; otherwise, it is *linearly dependent*.

We say that $S \subset V$ is linearly independent if every finite subset of S is linearly independent.

Proposition 4.18 (Compare coefficients)**.** Let $v_1, \ldots, v_n$ be linearly independent in V. Then

$$a_1 v_1 + \cdots + a_n v_n = b_1 v_1 + \cdots + b_n v_n$$

if and only if $a_i = b_i$ $(1 \leqslant i \leqslant n)$.

Proof. Exercise. □

The following result will often be useful; it states that given a linearly dependent set of vectors, one of the vectors is in the span of the previous ones; furthermore we can throw out that vector without changing the span of the original set.

Lemma 4.19 (Linear dependence lemma)**.** Suppose $v_1, \ldots, v_n$ are linearly dependent in V. Then there exists v_k such that the following hold:

(i) $v_k \in \operatorname{span}(v_1, \ldots, v_{k-1})$

(ii) $\operatorname{span}(v_1, \ldots, v_{k-1}, v_{k+1}, \ldots, v_n) = \operatorname{span}(v_1, \ldots, v_n)$

Proof.

(i) Since $v_1, \ldots, v_n$ are linearly dependent, there exists $a_1, \ldots, a_n \in \mathbf{F}$, not all 0, such that

$$a_1 v_1 + \cdots + a_n v_n = 0.$$

Take $k = \max\{1, \ldots, n\}$ such that $a_k \neq 0$. Then

$$v_k = -\frac{a_1}{a_k} v_1 - \cdots - \frac{a_{k-1}}{a_k} v_{k-1},$$

which means that v_k can be written as a linear combination of $v_1, \ldots, v_{k-1}$, so $v_k \in \operatorname{span}(v_1, \ldots, v_{k-1})$ by definition of span.

(ii) Now suppose k is such that $v_k \in \operatorname{span}(v_1, \ldots, v_{k-1})$. Then there exists $b_1, \ldots, b_{k-1} \in \mathbf{F}$ be such that

$$v_k = b_1 v_1 + \cdots + b_{k-1} v_{k-1}. \tag{1}$$

Suppose $u \in \operatorname{span}(v_1, \ldots, v_n)$. Then there exists $c_1, \ldots, c_n \in \mathbf{F}$ such that

$$u = c_1 v_1 + \cdots + c_n v_n. \tag{2}$$

In (2), we can replace v_k with the RHS of (1), which gives

$$\begin{aligned} u &= c_1 v_1 + \cdots + c_{k-1} v_{k-1} + c_k v_k + c_{k+1} v_{k+1} + \cdots + c_n v_n \\ &= c_1 v_1 + \cdots + c_{k-1} v_{k-1} + c_k (b_1 v_1 + \cdots + b_{k-1} v_{k-1}) + c_{k+1} v_{k+1} + \cdots + c_n v_n \\ &= c_1 v_1 + \cdots + c_{k-1} v_{k-1} + c_k b_1 v_1 + \cdots + c_k b_{k-1} v_{k-1} + c_{k+1} v_{k+1} + \cdots + c_n v_n \\ &= (c_1 + b c_k) v_1 + \cdots + (c_{k-1} + b_{k-1} c_k) v_{k-1} + c_{k+1} v_{k+1} + \cdots + c_n v_n. \end{aligned}$$

Thus $u \in \operatorname{span}(v_1, \ldots, v_{k-1}, v_{k+1}, \ldots, v_n)$. This shows that removing v_k from $v_1, \ldots, v_n$ does not change the span of the list.

□

The following result says that no linearly independent set in V is longer than a spanning set in V.

Proposition 4.20. In a finite-dimensional vector space, the length of every linearly independent set of vectors is less than or equal to the length of every spanning set of vectors.

Proof. Suppose $A = \{u_1, \ldots, u_m\}$ is linearly independent in V, $B = \{w_1, \ldots, w_n\}$ spans V. We want to prove that $m \leqslant n$.

Since B spans V, if we add any other vector from V to the list B, we will get a linearly dependent list, since this newly added vector can, by the definition of a span, be expressed as a linear combination of the vectors in B. In particular, if we add $u_1 \in A$ to B, then the new list

$$\{u_1, w_1, \ldots, w_n\}$$

is linearly dependent. By the linear independence lemma, we can remove one of the w_i's from B, so that the remaining list of n vectors still spans V. For the sake of argument, let's say we remove w_n (we can always order the w_i's in the list so that the element we remove is at the end). Then we are left with the revised list

$$B_1 = \{u_1, w_1, \ldots, w_{n-1}\}.$$

We can repeat this process m times, each time adding the next element u_i from list A and removing the last w_i. Because of the linear dependence lemma, we know that there must always be a w_i that can be removed each time we add a u_i, so there must be at least as many w_i's as u_i's. In other words, $m \leqslant n$ which is what we wanted to prove. □

Remark. We can use this result to show, without any computations, that certain lists are not linearly independent and that certain lists do not span a given vector space.

Our intuition suggests that every subspace of a finite-dimensional vector space should also be finite-dimensional. We now prove that this intuition is correct.

Proposition 4.21. Every subspace of a finite-dimensional vector space is finite-dimensional.

Proof. Suppose V is finite-dimensional, $U \leqslant V$. To show that U is finite-dimensional, we need to find a spanning set of vectors in U. We prove by construction of this spanning set.

Step 1 If $U = \{\mathbf{0}\}$, then U is finite-dimensional and we are done. Otherwise, choose $v_1 \in U$, $v_1 \neq \mathbf{0}$ and add it to our list of vectors.

Step k Our list so far is $\{v_1, \ldots, v_{k-1}\}$. If $U = \operatorname{span}(v_1, \ldots, v_{k-1})$, then U is finite-dimensional and we are done. Otherwise, choose $v_k \in U$ such that $v_k \notin \operatorname{span}(v_1, \ldots, v_{k-1})$ and add it to our list.

After each step, we have constructed a list of vectors such that no vector in this list is in the span of the previous vectors; by the linear dependence lemma, our constructed list is a linearly independent set.

By Proposition 4.20, this linearly independent set cannot be longer than any spanning set of V. Thus the process must terminate after a finite number of steps, and we have constructed a spanning set of U. Hence U is finite-dimensional. □

§4.4 Bases

Definition 4.22 (Basis)**.** $B = \{v_1 \ldots, v_n\}$ is a ***basis*** of V if

(i) B is linearly independent in V;

(ii) B is a spanning set of V.

Example (Standard basis)

Let $\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ where the i-th coordinate is 1. $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is a basis of $\mathbf{F}^n$, known as the *standard basis* of $\mathbf{F}^n$.

Lemma 4.23 (Criterion for basis)**.** Let $B = \{v_1, \ldots, v_n\}$ be a list of vectors in V. Then B is a basis of V if and only if every $v \in V$ can uniquely expressed as a linear combination of $v_1, \ldots, v_n$.

Proof.

$\boxed{\Longrightarrow}$ Let $v \in V$. Since B is a basis of V, there exists $a_1, \ldots, a_n \in \mathbf{F}$ such that

$$v = a_1 v_1 + \cdots + a_n v_n. \tag{1}$$

To show that the representation is unique, suppose that $c_1, \ldots, c_n \in \mathbf{F}$ also satisfy

$$v = c_1 v_1 + \cdots + c_n v_n. \tag{2}$$

Subtracting (2) from (1) gives

$$\mathbf{0} = (a_1 - c_1)v_1 + \cdots + (a_n - c_n)v_n.$$

Since $v_1, \ldots, v_n$ are linearly independent, we have $a_i - c_i = 0$, or $a_i = c_i$ for all i $(1 \leqslant i \leqslant n)$. Thus the representation of v as a linear combination of $v_1, \ldots, v_n$ is unique.

$\boxed{\Longleftarrow}$ Suppose that every $v \in V$ can be uniquely expressed as a linear combination of $v_1, \ldots, v_n$. This implies that B spans V. To show that B is linearly independent, suppose that $a_1, \ldots, a_n \in \mathbf{F}$ satisfy

$$a_1 v_1 + \cdots + a_n v_n = \mathbf{0}.$$

Since $\mathbf{0}$ can be uniquely expressed as a linear combination of $v_1, \ldots, v_n$, we have $a_1 = \cdots = a_n = 0$, thus B is linearly independent. Since B is linearly independent and spans V, B is a basis of V. □

A spanning set in a vector space may not be a basis because it is not linearly independent. Our next result says that given any spanning set, some (possibly none) of the vectors in it can be discarded so that the remaining list is linearly independent and still spans the vector space.

Lemma 4.24. Every spanning set in a vector space can be reduced to a basis of the vector space.

Proof. Suppose $B = \{v_1, \ldots, v_n\}$ spans V. We want to remove some vectors from B so that the remaining vectors form a basis of V. We do this through the multistep process described below.

Step 1 If $v_1 = \mathbf{0}$, delete v_1 from B. If $v_1 \neq \mathbf{0}$, leave B unchanged.

Step k If $v_k \in \operatorname{span}(v_1, \ldots, v_{k-1})$, delete v_k from B. If $v_k \notin \operatorname{span}(v_1, \ldots, v_{k-1})$, leave B unchanged.

Stop the process after step n, getting a list B. Since we only delete vectors from B that are in the span of the previous vectors, by the linear dependence lemma, the list B still spans V.

The process ensures that no vector in B is in the span of the previous ones. By the linear dependence lemma, B is linearly independent.

Since B is linearly independent and spans V, B is a basis of V. □

Corollary 4.25. Every finite-dimensional vector space has a basis.

Proof. We prove by construction. Suppose V is finite-dimensional. By definition, there exists a spanning set of vectors in V. By Lemma 4.24, the spanning set can be reduced to a basis. □

Now we show that given any linearly independent set, we can adjoin some additional vectors so that the extended list is still linearly independent but also spans the space.

Lemma 4.26. Every linearly independent set of vectors in a finite-dimensional vector space can be extended to a basis of the vector space.

Proof. Suppose $u_1, \ldots, u_m$ are linearly independent in V, $w_1, \ldots, w_n$ span V. Then the list

$$\{u_1, \ldots, u_m, w_1, \ldots, w_n\}$$

spans V. By Lemma 4.24, we can reduce this list to a basis of V consisting $u_1, \ldots, u_m$ (since $u_1, \ldots, u_m$ are linearly independent, $u_i \notin \operatorname{span}(u_1, \ldots, u_{i-1})$ for all i, so none of the u_i's are deleted in the process), and some of the w_i's. □

We now show that every subspace of a finite-dimensional vector space can be paired with another subspace to form a direct sum of the whole space.

Corollary 4.27. Suppose V is finite-dimensional, $U \leqslant V$. Then there exists $W \leqslant V$ such that $V = U \oplus W$.

Proof. Since V is finite-dimensional and $U \leqslant V$, by Proposition 4.21, U is finite-dimensional, so U has a basis B, by Corollary 4.25; let $B = \{u_1, \ldots, u_n\}$. Since B is linearly independent, by Lemma 4.26, B can be extended to a basis of V, say

$$\{u_1, \ldots, u_n, w_1, \ldots, w_n\}.$$

Take $W = \operatorname{span}(w_1, \ldots, w_n)$. We claim that $V = U \oplus W$. To show this, by Lemma 4.11, we need to show that (i) $V = U + W$, and (ii) $U \cap W = \{\mathbf{0}\}$.

(i) Suppose $v \in V$. Since $\{u_1, \ldots, u_n, w_1, \ldots, w_n\}$ spans V, there exists $a_1, \ldots, a_n, b_1, \ldots, b_n \in \mathbf{F}$ such that

$$v = a_1u_1 + \cdots + a_nu_n + b_1w_1 + \cdots + b_nw_n.$$

Take $u = a_1u_1 + \cdots + a_nu_n \in U$, $w = b_1w_1 + \cdots + b_nw_n \in W$. Then $v = u + w \in U + W$, so $V = U + W$.

(ii) Suppose $v \in U \cap W$. Since $v \in U$, v can be written as a linear combination of $u_1, \ldots, u_n$:

$$v = a_1u_1 + \cdots + a_nu_n. \tag{1}$$

Since $v \in W$, v can be written as a linear combination of $w_1, \ldots, w_n$:

$$v = b_1w_1 + \cdots + b_nw_n. \tag{2}$$

Subtracting (2) from (1) gives

$$\mathbf{0} = a_1u_1 + \cdots + a_nu_n - b_1w_1 - \cdots - b_nw_n.$$

Since $u_1, \ldots, u_n, w_1, \ldots, w_n$ are linearly independent, we have $a_i = b_i = 0$ for all i $(1 \leqslant i \leqslant n)$. Thus $v = \mathbf{0}$, so $U \cap W = \{\mathbf{0}\}$.

□

§4.5 Dimension

Lemma 4.28. Any two bases of a finite-dimensional vector space have the same length.

Proof. Suppose V is finite-dimensional, let B_1 and B_2 be two bases of V. By definition, B_1 is linearly independent in V, and B_2 spans V, so by Proposition 4.20, $|B_1| \leqslant |B_2|$.

Similarly, by definition, B_2 is linearly independent in V and B_1 spans V, so $|B_2| \leqslant |B_1|$.

Since $|B_1| \leqslant |B_2|$ and $|B_2| \leqslant |B_1|$, we have $|B_1| = |B_2|$, as desired. □

Since any two bases of a finite-dimensional vector space have the same length, we can formally define the dimension of such spaces.

Definition 4.29 (Dimension)**.** The ***dimension*** of V is the length of any basis of V, denoted by $\dim V$.

Proposition 4.30. Suppose V is finite-dimensional, $U \leqslant V$. Then $\dim U \leqslant \dim V$.

Proof. Since V is finite-dimensional and $U \leqslant V$, U is finite-dimensional. Let B_U be a basis of U, and B_V be a basis of V.

By definition, B_U is linearly independent in V, and B_V spans V. By Proposition 4.20, $|B_U| \leqslant |B_V|$, so

$$\dim U = |B_U| \leqslant |B_V| = \dim V,$$

since $|B_U| = \dim U$ and $|B_V| = \dim V$ by definition. □

To check that a list of vectors is a basis, we must show that it is linearly independent and that it spans the vector space. The next result shows that if the list in question has the right length, then we only need to check that it satisfies one of the two required properties.

Proposition 4.31. Suppose V is finite-dimensional. Then

(i) every linearly independent set of vectors in V with length $\dim V$ is a basis of V;

(ii) every spanning set of vectors in V with length $\dim V$ is a basis of V.

Proof.

(i) Suppose $\dim V = n$, $\{v_1, \ldots, v_n\}$ is linearly independent in V. By Lemma 4.26, $\{v_1, \ldots, v_n\}$ can be extended to a basis of V. However, every basis of V has length n (by definition of dimension), which means that no elements are adjoined to $\{v_1, \ldots, v_n\}$. Hence $\{v_1, \ldots, v_n\}$ is a basis of V, as desired.

(ii) Suppose $\dim V = n$, $\{v_1, \ldots, v_n\}$ spans V. By Lemma 4.24, $\{v_1, \ldots, v_n\}$ can be reduced to a basis of V. However, every basis of V has length n, which means that no elements are deleted from $\{v_1, \ldots, v_n\}$. Hence $\{v_1, \ldots, v_n\}$ is a basis of V, as desired.

□

Corollary 4.32. Suppose V is finite-dimensional, $U \leqslant V$. If $\dim U = \dim V$, then $U = V$.

Proof. Let $\dim U = \dim V = n$, let $\{u_1, \ldots, u_n\}$ be a basis of U. Then $\{u_1, \ldots, u_n\}$ is linearly independent in V (because it is a basis of U) of length $\dim V$. From Proposition 4.31, $\{u_1, \ldots, u_n\}$ is a basis of V. In particular every vector in V is a linear combination of $u_1, \ldots, u_n$. Thus $U = V$. □

Lemma 4.33 (Dimension of sum)**.** Suppose V is finite-dimensional, $U_1, U_2 \leqslant V$. Then

$$\dim(U_1 + U_2) = \dim U_1 + \dim U_2 - \dim(U_1 \cap U_2).$$

Proof. Let $\{u_1, \ldots, u_m\}$ be a basis of $U_1 \cap U_2$; thus $\dim(U_1 \cap U_2) = m$. Since $\{u_1, \ldots, u_m\}$ is a basis of $U_1 \cap U_2$, it is linearly independent in U_1. By Lemma 4.26, $\{u_1, \ldots, u_m\}$ can be extended to a basis $\{u_1, \ldots, u_m, v_1, \ldots, v_j\}$ of U_1; thus $\dim U_1 = m + j$. Similarly, extend $\{u_1, \ldots, u_m\}$ to a basis $\{u_1, \ldots, u_m, v_1, \ldots, v_k\}$ of U_2; thus $\dim U_2 = m + k$.

We will show that

$$\{u_1, \ldots, u_m, v_1, \ldots, v_j, w_1, \ldots, w_k\}$$

is a basis of $U_1 + U_2$. This will complete the proof because then we will have

$$\begin{aligned} \dim(U_1 + U_2) &= m + j + k \\ &= (m + j) + (m + k) - m \\ &= \dim U_1 + \dim U_2 - \dim(U_1 \cap U_2). \end{aligned}$$

We just need to show that $\{u_1, \ldots, u_m, v_1, \ldots, v_j, w_1, \ldots, w_k\}$ is linearly independent. To prove this, suppose

$$a_1 u_1 + \cdots + a_m u_m + b_1 v_1 + \cdots + b_j v_j + c_1 w_1 + \cdots + c_k w_k = \mathbf{0}, \tag{1}$$

where $a_i, b_i, c_i \in \mathbf{F}$. We need to show that $a_i = b_i = c_i = 0$ for all i. (1) can be rewritten as

$$c_1 w_1 + \cdots + c_k w_k = -a_1 u_1 - \cdots - a_m u_m - b_1 v_1 - \cdots - b_j v_j,$$

which shows that $c_1 w_1 + \cdots + c_k w_k \in U_1$. But actually all the w_i's are in U_2, so $c_1 w_1 + \cdots + c_k w_k \in U_2$, thus $c_1 w_1 + \cdots + c_k w_k \in U_1 \cap U_2$. Then we can write

$$c_1 w_1 + \cdots + c_k w_k = d_1 u_1 + \cdots + d_m u_m$$

for some $d_i \in \mathbf{F}$. But $u_1, \ldots, u_m, w_1, \ldots, w_k$ are linearly independent, so $c_i = d_i = 0$ for all i. Thus our original equation (1) becomes

$$a_1 u_1 + \cdots + a_m u_m + b_1 v_1 + \cdots + b_j v_j = \mathbf{0}.$$

Since $u_1, \ldots, u_m, v_1, \ldots, v_j$ are linearly independent, we have $a_i = b_i = 0$ for all i, as desired. □

Exercises

Problem 4.1 ([Axl24] 1C Q12)**.** Suppose W is a vector space over $\mathbf{F}$, V_1 and V_2 are subspaces of W. Show that $V_1 \cup V_2$ is a vector space over $\mathbf{F}$ if and only if $V_1 \subset V_2$ or $V_2 \subset V_1$.

Solution. The backward direction is trivial. We focus on proving the forward direction.

Supppse otherwise, then $V_1 \setminus V_2 \neq \emptyset$ and $V_2 \setminus V_1 \neq \emptyset$. Pick $v_1 \in V_1 \setminus V_2$ and $v_2 \in V_2 \setminus V_1$. Then

$$\begin{aligned} v_1, v_2 \in V_1 \cup V_2 &\implies v_1 + v_2 \in V_1 \cup V_2 \\ &\implies v_2, v_1 + v_2 \in V_2 \\ &\implies v_1 = (v_1 + v_2) - v_2 \in V_2 \end{aligned}$$

which is a contradiction. □

Problem 4.2 ([Axl24] 1C Q13)**.** Suppose W is a vector space over $\mathbf{F}$, V_1, V_2, V_3 are subspaces of W. Then $V_1 \cup V_2 \cup V_3$ is a vector space over $\mathbf{F}$ if and only if one of the V_i contains the other two.

Solution. We prove the forward direction. Suppose otherwise, then $v_1 \in V_1 \setminus (V_2 + V_3)$, $v_2 \in V_2 \setminus (V_1 + V_3)$, $v_3 \in V_3 \setminus (V_1 + V_2)$. Consider

$$\{v_1 + v_2 + v_3, v_1 + v_2 + 2v_3, v_1 + 2v_2 + v_3, v_1 + 2v_2 + 2v_3\} \subset V_1 \cup V_2 \cup V_3$$

Then

$$\begin{aligned} &(v_1 + v_2 + 2v_3) - (v_1 + v_2 + v_3) = v_3 \notin V_1 + V_2 \\ &\implies v_1 + v_2 + v_3 \notin V_1 + V_2 \quad \text{or} \quad v_1 + v_2 + 2v_3 \notin V_1 + V_2 \\ &\implies v_1 + v_2 + v_3 \in V_3 \quad \text{or} \quad v_1 + v_2 + 2v_3 \in V_3 \\ &\implies v_1 + v_2 \in V_3 \end{aligned}$$

Similarly,

$$\begin{aligned} &(v_1 + 2v_2 + 2v_3) - (v_1 + 2v_2 + v_3) = v_3 \notin V_1 + V_2 \\ &\implies v_1 + 2v_2 + v_3 \notin V_1 + V_2 \quad \text{or} \quad v_1 + 2v_2 + 2v_3 \notin V_1 + V_2 \\ &\implies v_1 + 2v_2 + v_3 \in V_3 \quad \text{or} \quad v_1 + 2v_2 + 2v_3 \in V_3 \\ &\implies v_1 + 2v_2 \in V_3 \end{aligned}$$

This implies $(v_1 + 2v_2) - (v_1 + v_2) = v_2 \in V_3$, a contradiction. □

Problem 4.3 ([Axl24] 2A Q12)**.** Suppose $\{v_1, \ldots, v_n\}$ is linearly independent in V, $w \in V$. Prove that if $\{v_1 + w, \ldots, v_n + w\}$ is linearly dependent, then $w \in \operatorname{span}(v_1, \ldots, v_n)$.

Solution. If $\{v_1 + w, \ldots, v_n + w\}$ is linearly dependent, then there exists $a_1, \ldots, a_n \in \mathbf{F}$, not all zero,

such that

$$a_1(v_1 + w) + \cdots + a_n(v_n + w) = 0,$$

or

$$a_1v_1 + \cdots + a_nv_n = -(a_1 + \cdots + a_n)w.$$

Suppose otherwise, that $a_1 + \cdots + a_n = 0$. Then

$$a_1v_1 + \cdots + a_nv_n = \mathbf{0},$$

but the linear independence of $\{v_1, \ldots, v_n\}$ implies that $a_1 = \cdots = a_n = 0$, which is a contradiction. Hence we must have $a_1 + \cdots + a_n \neq 0$, so we can write

$$w = -\frac{a_1}{a_1 + \cdots + a_n}v_1 - \cdots - \frac{a_n}{a_1 + \cdots + a_n}v_n,$$

which is a linear combination of $v_1, \ldots, v_n$. Thus by definition of span, $w \in \operatorname{span}(v_1, \ldots, v_n)$. □

Problem 4.4 ([Axl24] 2A Q14)**.** Suppose $\{v_1, \ldots, v_n\} \subset V$. Let

$$w_i = v_1 + \cdots + v_i \quad (i = 1, \ldots, n)$$

Show that $\{v_1, \ldots, v_n\}$ is linearly independent if and only if $\{w_1, \ldots, w_n\}$ is linearly independent.

Solution. Write

$$\begin{aligned} v_1 &= w_1 \\ v_2 &= w_2 - w_1 \\ v_3 &= w_3 - w_2 \\ &\vdots \\ v_n &= w_n - w_{n-1}. \end{aligned}$$

$\boxed{\Longrightarrow}$

$$a_1w_1 + \cdots + a_nw_n = \mathbf{0}$$

for some $a_i \in \mathbf{F}$. Expressing w_i's as v_i's,

$$a_1v_1 + a_2(v_1 + v_2) + \cdots + a_n(v_1 + \cdots + v_n) = 0,$$

or

$$(a_1 + \cdots + a_n)v_1 + (a_2 + \cdots + a_n)v_2 + \cdots + a_nv_n = \mathbf{0}.$$

Since $v_1, \dots, v_n$ are linearly independent,

$$\begin{aligned} a_1 + a_2 + \cdots + a_n &= 0 \\ a_2 + \cdots + a_n &= 0 \\ &\vdots \\ a_n &= 0 \end{aligned}$$

on solving simultaneously gives $a_1 = \cdots = a_n = 0$.

$\boxed{\Longleftarrow}$ Similar to the above. □

Problem 4.5 ([Axl24] 2A Q18)**.** Prove that $\mathbf{F}^\infty$ is infinite-dimensional.

Solution. To prove that $\mathbf{F}^\infty$ has no finite spanning sets, we prove by contradiction. Suppose otherwise, that there exists a finite spanning set of $\mathbf{F}^\infty$, say $\{v_1, \dots, v_n\}$.

Let

$$\begin{aligned} e_1 &= (1, 0, \dots) \\ e_2 &= (0, 1, 0, \dots) \\ e_3 &= (0, 0, 1, 0, \dots) \\ &\vdots \\ e_{n+1} &= (0, \dots, 0, 1, 0, \dots) \end{aligned}$$

where e_i has a 1 at the i-th coordinate, and 0's for the remaining coordinates. Let

$$a_1 e_1 + \cdots + a_{n+1} e_{n+1} = \mathbf{0}$$

for some $a_i \in \mathbf{F}$. Then

$$(a_1, a_2, \dots, a_{n+1}, 0, 0, \dots) = \mathbf{0}$$

so $a_1 = a_2 = \cdots = a_{n+1} = 0$. Thus $\{e_1, \dots, e_{n+1}\}$ is a linearly independent set, of length $n+1$. However, $\{v_1, \dots, v_n\}$ is a spanning set of length n. By Proposition 4.20, we have reached a contradiction. □

Problem 4.6 ([Axl24] 2B Q5)**.** Suppose V is finite-dimensional, $U, W \leqslant V$ such that $V = U + W$. Prove that V has a basis in $U \cup W$.

Solution. Let $\{v_i\}_{i=1}^n$ denote the basis for V. By definition we have $v_i = u_i + w_i$ for some $u_i \in U$, $w_i \in W$. Then we have the spanning set of the vector space V $\sum_{i=1}^n a_i(u_i + w_i)$, which can be reduced to a basis by the lemma. □

Problem 4.7 ([Axl24] 2B Q7)**.** Suppose $\{v_1, v_2, v_3, v_4\}$ is a basis of V. Prove that

$$\{v_1 + v_2, v_2 + v_3, v_3 + v_4, v_4\}$$

is also a basis of V.

Solution. We know that $\{v_1, v_2, v_3, v_4\}$ is linearly independent and spans V. Then there exist $a_i \in \mathbf{F}$ such that

$$a_1(v_1 + v_2) + a_2(v_2 + v_3) + a_3(v_3 + v_4) + a_4 v_4 = 0 \implies a_1 = a_2 = a_3 = a_4 = 0.$$

Write

$$\begin{aligned} &a_1(v_1 + v_2) + a_2(v_2 + v_3) + a_3(v_3 + v_4) + a_4 v_4 \\ &= a_1 v_1 + (a_1 + a_2)v_2 + (a_2 + a_3)v_3 + (a_3 + a_4)v_4, \end{aligned}$$

this shows the linear independence. To prove spanning, let $v \in V$, then

$$\begin{aligned} v &= a_1 v_1 + a_2 v_2 + a_3 v_3 + a_4 v_4 \\ &= a_1(v_1 + v_2) + (a_2 - a_1)(v_2 + v_3) + (a_3 - a_2)(v_3 + v_4) + (a_4 - a_3)v_4, \end{aligned}$$

which is a linear combination of $v_1 + v_2, v_2 + v_3, v_3 + v_4, v_4$. □

Problem 4.8 ([Axl24] 2B Q10)**.** Suppose $U, W \leqslant V$ such that $V = U \oplus W$. Suppose also that $\{u_1, \ldots, u_m\}$ is a basis of U, $\{w_1, \ldots, w_n\}$ is a basis of W. Prove that

$$\{u_1, \ldots, u_m, w_1, \ldots, w_n\}$$

is a basis of V.

Solution. We know that this set is linearly independent (otherwise violating the direct sum assumption) so it suffices to prove the spanning. Let $v \in V$, then

$$v = u + w = \sum_{i=1}^{m} a_i u_i + \sum_{i=1}^{n} b_j w_j.$$

□

Problem 4.9 ([Axl24] 2C Q8)**.**

Problem 4.10 ([Axl24] 2C Q16)**.**

Problem 4.11 ([Axl24] 2C Q17)**.** Suppose that $V_1, \ldots, V_n \leqslant V$ are finite-dimensional. Prove that $V_1 + \cdots + V_n$ is finite-dimensional, and

$$\dim(V_1 + \cdots + V_n) \leqslant \dim V_1 + \cdots + \dim V_n.$$

Solution. We prove by induction on n. The base case is trivial. Assume the statement holds for k. Then for $k + 1$, denoting $V_1 + \cdots + V_k = M_k$, we have that

$$\dim(M_k + V_{k+1}) \leqslant \dim V_1 + \cdots + \dim V_{k+1},$$

which is finite. □

5 Linear Maps

§5.1 Vector Space of Linear Maps

Definition 5.1 (Linear map)**.** A ***linear map*** from V to W is a function $T : V \to W$ satisfying the following properties:

(i) Additivity: $T(v + w) = Tv + Tw$ for all $v, w \in V$

(ii) Homogeneity: $T(\lambda v) = \lambda T(v)$ for all $\lambda \in \mathbf{F}$, $v \in V$

Notation. The set of linear maps from V to W is denoted by $\mathcal{L}(V, W)$; the set of linear maps on V (from V to V) is denoted by $\mathcal{L}(V)$.

The existence part of the next result means that we can find a linear map that takes on whatever values we wish on the vectors in a basis. The uniqueness part of the next result means that a linear map is completely determined by its values on a basis.

Lemma 5.2 (Linear map lemma)**.** Suppose $\{v_1, \ldots, v_n\}$ is a basis of V, and $w_1, \ldots, w_n \in W$. Then there exists a unique linear map $T : V \to W$ such that

$$Tv_i = w_i \quad (i = 1, \ldots, n)$$

Proof. First we show the existence of a linear map T with the desired property. Define $T : V \to W$ by

$$T(c_1 v_1 + \cdots + c_n v_n) = c_1 w_1 + \cdots + c_n w_n,$$

for some $c_i \in \mathbf{F}$. Since $\{v_1, \ldots, v_n\}$ is a basis of V, by Lemma 4.23, each $v \in V$ can be uniquely expressed as a linear combination of $v_1, \ldots, v_n$, thus the equation above does indeed define a function $T : V \to W$. For i $(1 \leqslant i \leqslant n)$, take $c_i = 1$ and the other c's equal to 0, then

$$T(0v_1 + \cdots + 1v_i + \cdots + 0v_n) = 0w_1 + \cdots + 1w_i + \cdots + 0w_n$$

which shows that $Tv_i = w_i$.

We now show that $T : V \to W$ is a linear map:

(i) For $u, v \in V$ with $u = a_1v_1 + \cdots + a_nv_n$ and $c_1v_1 + \cdots + c_nv_n$,

$$\begin{aligned} T(u+v) &= T\left((a_1+c_1)v_1 + \cdots + (a_n+c_n)v_n\right) \\ &= (a_1+c_1)w_1 + \cdots + (a_n+c_n)w_n \\ &= (a_1w_1 + \cdots + a_nw_n) + (c_1w_1 + \cdots + c_nw_n) \\ &= Tu + Tv. \end{aligned}$$

(ii) For $\lambda \in \mathbf{F}$ and $v = c_1v_1 + \cdots + c_nv_n$,

$$\begin{aligned} T(\lambda v) &= T(\lambda c_1v_1 + \cdots + \lambda c_nv_n) \\ &= \lambda c_1w_1 + \cdots + \lambda c_nw_n \\ &= \lambda(c_1w_1 + \cdots + c_nw_n) \\ &= \lambda Tv. \end{aligned}$$

To prove uniqueness, now suppose that $T \in \mathcal{L}(V, W)$ and $Tv_i = w_i$ for $i = 1, \ldots, n$. Let $c_i \in \mathbf{F}$. The homogeneity of T implies that $T(c_iv_i) = c_iw_i$. The additivity of T now implies that

$$T(c_1v_1 + \cdots + c_nv_n) = c_1w_1 + \cdots + c_nw_n.$$

Thus T is uniquely determined on $\operatorname{span}\{v_1, \ldots, v_n\}$. Since $\{v_1, \ldots, v_n\}$ is a basis of V, this implies that T is uniquely determined on V. □

Proposition 5.3. $\mathcal{L}(V, W)$ is a vector space, with the operations addition and scalar multiplication defined as follows: suppose $S, T \in \mathcal{L}(V, W)$, $\lambda \in \mathbf{F}$,

(i) $(S+T)(v) = Sv + Tv$

(ii) $(\lambda T)(v) = \lambda(Tv)$

for all $v \in V$.

Proof. Exercise. □

Definition 5.4 (Product of linear maps)**.** $T \in \mathcal{L}(U, V)$, $S \in \mathcal{L}(V, W)$, then the ***product*** $ST \in \mathcal{L}(U, W)$ is defined by

$$(ST)(u) = S(Tu) \quad (\forall u \in U)$$

Remark. In other words, ST is just the usual composition $S \circ T$ of two functions.

Remark. ST is defined only when T maps into the domain of S.

Proposition 5.5 (Algebraic properties of products of linear maps)**.**

(i) Associativity: $(T_1T_2)T_3 = T_1(T_2T_3)$ for all linear maps T_1, T_2, T_3 such that the products make sense (meaning that T_3 maps into the domain of T_2, T_2 maps into the domain of T_1)

(ii) Iidentity: $TI = IT = T$ for all $T \in \mathcal{L}(V, W)$ (the first I is the identity map on V, and the second I is the identity map on W)

(iii) Distributive: $(S_1 + S_2)T = S_1T + S_2T$ and $S(T_1 + T_2) = ST_1 + ST_2$ for all $T, T_1, T_2 \in \mathcal{L}(U, V)$ and $S, S_1, S_2 \in \mathcal{L}(V, W)$

Proof. Exercise. □

Proposition 5.6. Suppose $T \in \mathcal{L}(V, W)$. Then $T(\mathbf{0}) = \mathbf{0}$.

Proof. By additivity, we have

$$T(\mathbf{0}) = T(\mathbf{0} + \mathbf{0}) = T(\mathbf{0}) + T(\mathbf{0}).$$

Add the additive inverse of $T(\mathbf{0})$ to each side of the equation to conclude that $T(\mathbf{0}) = \mathbf{0}$. □

§5.2 Kernel and Image

Definition 5.7 (Kernel)**.** Suppose $T \in \mathcal{L}(V, W)$. The ***kernel*** of T is the subset of V consisting of those vectors that T maps to $\mathbf{0}$:

$$\ker T \coloneqq \{v \in V \mid Tv = \mathbf{0}\} \subset V.$$

Proposition 5.8. Suppose $T \in \mathcal{L}(V, W)$. Then $\ker T \leqslant V$.

Proof. By Lemma 4.6, we check the conditions of a subspace:

(i) By Proposition 5.6, $T(\mathbf{0}) = \mathbf{0}$, so $\mathbf{0} \in \ker T$.

(ii) For all $v, w \in \ker T$,

$$T(v + w) = Tv + Tw = \mathbf{0} \implies v + w \in \ker T$$

so $\ker T$ is closed under addition.

(iii) For all $v \in \ker T$, $\lambda \in \mathbf{F}$,

$$T(\lambda v) = \lambda Tv = \mathbf{0} \implies \lambda v \in \ker T$$

so $\ker T$ is closed under scalar multiplication.

□

Definition 5.9 (Injectivity)**.** Suppose $T \in \mathcal{L}(V, W)$. T is ***injective*** if

$$Tu = Tv \implies u = v.$$

Proposition 5.10. Suppose $T \in \mathcal{L}(V, W)$. Then T is injective if and only if $\ker T = \{\mathbf{0}\}$.

Proof.

$\boxed{\implies}$ Suppose T is injective. Let $v \in \ker T$, then

$$Tv = \mathbf{0} = T(\mathbf{0}) \implies v = \mathbf{0}$$

by the injectivity of T. Hence $\ker T = \{\mathbf{0}\}$ as desired.

$\boxed{\impliedby}$ Suppose $\ker T = \{\mathbf{0}\}$. Let $u, v \in V$ such that $Tu = Tv$. Then

$$T(u - v) = Tu - Tv = \mathbf{0}.$$

By definition of kernel, $u - v \in \ker T = \{\mathbf{0}\}$, so $u - v = \mathbf{0}$, which implies that $u = v$. Hence T is injective, as desired. □

Definition 5.11 (Image)**.** Suppose $T \in \mathcal{L}(V, W)$. The ***image*** of T is the subset of W consisting of those vectors that are of the form Tv for some $v \in V$:

$$\operatorname{im} T \coloneqq \{Tv \mid v \in V\} \subset W.$$

Proposition 5.12. Suppose $T \in \mathcal{L}(V, W)$. Then $\operatorname{im} T \leqslant W$.

Proof.

(i) $T(\mathbf{0}) = \mathbf{0}$ implies that $\mathbf{0} \in \operatorname{im} T$.

(ii) For $w_1, w_2 \in \operatorname{im} T$, there exist $v_1, v_2 \in V$ such that $Tv_1 = w_1$ and $Tv_2 = w_2$. Then

$$w_1 + w_2 = Tv_1 + Tv_2 = T(v_1 + v_2) \in \operatorname{im} T \implies w_1 + w_2 \in \operatorname{im} T.$$

(iii) For $w \in \operatorname{im} T$ and $\lambda \in \mathbf{F}$, there exists $v \in V$ such that $Tv = w$. Then

$$\lambda w = \lambda T v = T(\lambda v) \in \operatorname{im} T \implies \lambda w \in \operatorname{im} T.$$

□

Definition 5.13 (Surjectivity)**.** Suppose $T \in \mathcal{L}(V, W)$. T is ***surjective*** if $\operatorname{im} T = W$.

Fundamental Theorem of Linear Maps

Theorem 5.14 (Fundamental theorem of linear maps)**.** Suppose V is finite-dimensional, $T \in \mathcal{L}(V, W)$. Then $\operatorname{im} T$ is finite-dimensional, and

$$\dim V = \dim \ker T + \dim \operatorname{im} T. \tag{5.1}$$

Proof. Let $\{u_1, \ldots, u_m\}$ be basis of $\ker T$, then $\dim \ker T = m$. The linearly independent list $u_1, \ldots, u_m$ can be extended to a basis

$$\{u_1, \ldots, u_m, v_1, \ldots, v_n\}$$

of V, thus $\dim V = m + n$. To simultaneously show that $\operatorname{im} T$ is finite-dimensional and $\dim \operatorname{im} T = n$, we prove that $\{Tv_1, \ldots, Tv_n\}$ is a basis of $\operatorname{im} T$. Thus we need to show that the set (i) spans $\operatorname{im} T$, and (ii) is linearly independent.

(i) Let $v \in V$. Since $\{u_1, \ldots, u_m, v_1, \ldots, v_n\}$ spans V, we can write

$$v = a_1 u_1 + \cdots + a_m u_m + b_1 v_1 + \cdots + b_n v_n,$$

for some $a_i, b_i \in \mathbf{F}$. Applying T to both sides of the equation, and noting that $Tu_i = \mathbf{0}$ since $u_i \in \ker T$,

$$\begin{aligned} Tv &= T\left(a_1 u_1 + \cdots + a_m u_m + b_1 v_1 + \cdots + b_n v_n\right) \\ &= a_1 \underbrace{Tu_1}_{\mathbf{0}} + \cdots + a_m \underbrace{Tu_m}_{\mathbf{0}} + b_1 Tv_1 + \cdots + b_n v_n \\ &= b_1 Tv_1 + \cdots + b_n Tv_n \in \operatorname{im} T. \end{aligned}$$

Since every element of $\operatorname{im} T$ can be expressed as a linear combination of $Tv_1, \ldots, Tv_n$, we have that $\{Tv_1, \ldots, Tv_n\}$ spans $\operatorname{im} T$.

Moreover, since there exists a set of vectors that spans $\operatorname{im} T$, $\operatorname{im} T$ is finite-dimensional.

(ii) Suppose there exist $c_1, \ldots, c_n \in \mathbf{F}$ such that

$$c_1 Tv_1 + \cdots + c_n Tv_n = \mathbf{0}.$$

Then

$$T(c_1 v_1 + \cdots + c_n v_n) = T(\mathbf{0}) = \mathbf{0},$$

which implies $c_1 v_1 + \cdots + c_n v_n \in \ker T$. Since $\{u_1, \ldots, u_m\}$ is a spanning set of $\ker T$, we can write

$$c_1 v_1 + \cdots + c_n v_n = d_1 u_1 + \cdots + d_m u_m$$

for some $d_i \in \mathbf{F}$, or

$$c_1 v_1 + \cdots + c_n v_n - d_1 u_1 - \cdots - d_m u_m = \mathbf{0}.$$

Since $u_1, \ldots, u_m, v_1, \ldots, v_n$ are linearly independent, $c_i = d_i = 0$. Since $c_i = 0$, $\{Tv_1, \ldots, Tv_n\}$ is linearly independent.

□

We now show that no linear map from a finite-dimensional vector space to a "smaller" vector space can be injective, where "smaller" is measured by dimension.

Proposition 5.15. Suppose V and W are finite-dimensional vector spaces, $\dim V > \dim W$. Then there does not exist $T \in \mathcal{L}(V, W)$ such that T is injective.

Proof. Since W is finite-dimensional and $\operatorname{im} T \leqslant W$, by Proposition 4.30, we have that $\dim \operatorname{im} T \leqslant \dim W$.

Let $T \in \mathcal{L}(V, W)$. Then

$$\begin{aligned} \dim \ker T &= \dim V - \dim \operatorname{im} T && (1) \\ &\geqslant \dim V - \dim W && (2) \\ &> 0 \end{aligned}$$

where (1) follows from the fundamental theorem of linear maps, (2) follows from the above claim.

Since $\dim \ker T > 0$. This means that $\ker T$ contains some $v \in V \setminus \{\mathbf{0}\}$. Since $\ker T \neq \{\mathbf{0}\}$, T is not injective. □

The next result shows that no linear map from a finite-dimensional vector space to a "bigger" vector space can be surjective, where "bigger" is also measured by dimension.

Proposition 5.16. Suppose V and W are finite-dimensional vector spaces, $\dim V < \dim W$. Then there does not exist $T \in \mathcal{L}(V, W)$ such that T is surjective.

Proof. Let $T \in \mathcal{L}(V, W)$. Then

$$\begin{aligned} \dim \operatorname{im} T &= \dim V - \dim \ker T && (1) \\ &\leqslant \dim V && (2) \\ &< \dim W, \end{aligned}$$

where (1) follows from the fundamental theorem of linear maps, (2) follows since the dimension of a vector space is non-negative so $\dim \ker T \geqslant 0$.

Since $\dim \operatorname{im} T < \dim W$, $\operatorname{im} T \neq W$ so T is not surjective. □

Example (Homogeneous system of linear equations)
Consider the homogeneous system of linear equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= 0 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= 0 \end{aligned} \qquad (*)$$

where $a_{ij} \in \mathbf{F}$.

Define $T : \mathbf{F}^n \to \mathbf{F}^m$ by

$$T(x_1, \ldots, x_n) = \left(\sum_{i=1}^{n} a_{1i}x_i, \ldots, \sum_{i=1}^{m} a_{mi}x_i \right).$$

The solution set of $(*)$ is given by

$$\ker T = \left\{ (x_1, \ldots, x_n) \in \mathbf{F}^n \;\middle|\; \sum_{i=1}^{n} a_{1i}x_i = 0, \ldots, \sum_{i=1}^{n} a_{mi}x_i = 0 \right\}.$$

Proposition. A homogeneous system of linear equations with more variables than equations has non-zero solutions.

Proof. If $n > m$, then

$$\begin{aligned} \dim \mathbf{F}^n > \dim \mathbf{F}^m &\implies T \text{ is not injective} \\ &\implies \ker T \neq \{\mathbf{0}\} \\ &\implies (*) \text{ has non-zero solutions} \end{aligned}$$

□

Proposition. A system of linear equations with more equations than variables has no solution for some choice of the constant terms.

Proof. If $n < m$, then

$$\begin{aligned}\dim \mathbf{F}^n < \dim \mathbf{F}^m &\implies T \text{ is not surjective} \\ &\implies \exists (c_1, \ldots, c_m) \in \mathbf{F}^m, \forall (x_1, \ldots, x_n) \in \mathbf{F}^n, T(x_1, \ldots, x_n) \neq (c_1, \ldots, c_m)\end{aligned}$$

Thus the choice of constant terms $(c_1, \ldots, c_m)$ is such that the system of linear equations

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= c_1 \\ &\vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= c_m \end{aligned}$$

has no solutions $(x_1, \ldots, x_n)$. □

§5.3 Matrices

Representing a Linear Map by a Matrix

Definition 5.17 (Matrix)**.** Suppose $m, n \in \mathbf{N}$. An $m \times n$ ***matrix*** A is a rectangular array with m rows and n columns:

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

where $a_{ij} \in \mathbf{F}$ denotes the entry in row i, column j. We also denote $A = (a_{ij})_{m\times n}$, and drop the subscript if there is no ambiguity.

Notation. i is used for indexing across the m rows, j is used for indexing across the n columns.

Notation. $\mathcal{M}_{m\times n}(\mathbf{F})$ denotes the set of $m \times n$ matrices with entries in $\mathbf{F}$.

As we will soon see, matrices provide an efficient method of recording the values of Tv_j's in terms of a basis of W.

Definition 5.18 (Matrix of linear map)**.** Suppose $T \in \mathcal{L}(V, W)$, $\mathcal{V} = \{v_1, \ldots, v_n\}$ is a basis of V, $\mathcal{W} = \{w_1, \ldots, w_m\}$ is a basis of W. The matrix of T with respect to these bases is the $m \times n$ matrix $\mathcal{M}(T)$, whose entries a_{ij} are defined by

$$Tv_j = \sum_{i=1}^{m} a_{ij} w_i.$$

That is, the j-th column of $\mathcal{M}(T)$ consists of the scalars $a_{1j}, \ldots, a_{mj}$ needed to write Tv_j as a linear combination of the bases of W.

Notation. If the bases of V and W are not clear from the context, we adopt the notation $\mathcal{M}(T; \mathcal{V}, \mathcal{W})$.

Addition and Scalar Multiplication of Matrices

Definition 5.19 (Matrix operations)**.**

(i) Addition: the sum of two matrices of the same size is the matrix obtained by adding corresponding entries in the matrices:

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} + \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & & \vdots \\ c_{m1} & \cdots & c_{mn} \end{pmatrix} = \begin{pmatrix} a_{11}+c_{11} & \cdots & a_{1n}+c_{1n} \\ \vdots & & \vdots \\ a_{m1}+c_{m1} & \cdots & a_{mn}+c_{mn} \end{pmatrix}.$$

(ii) Scalar multiplication: the product of a scalar and a matrix is the matrix obtained by multiplying each entry in the matrix by the scalar:

$$\lambda \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} = \begin{pmatrix} \lambda a_{11} & \cdots & \lambda a_{1n} \\ \vdots & & \vdots \\ \lambda a_{m1} & \cdots & \lambda a_{mn} \end{pmatrix}.$$

Proposition 5.20. Suppose $S, T \in \mathcal{L}(V, W)$. Then

(i) $\mathcal{M}(S+T) = \mathcal{M}(S) + \mathcal{M}(T)$;

(ii) $\mathcal{M}(\lambda T) = \lambda \mathcal{M}(T)$ for $\lambda \in \mathbf{F}$.

Proof. Suppose $S, T \in \mathcal{L}(V, W)$, $\{v_1, \ldots, v_n\}$ is a basis of V, $\{w_1, \ldots, w_m\}$ is a basis of W.

(i) By definition, $\mathcal{M}(S)$ is the matrix whose entries a_{ij} are defined by

$$Sv_j = \sum_{i=1}^{m} a_{ij} w_i.$$

Similarly, $\mathcal{M}(T)$ is the matrix whose entries b_{ij} are defined by

$$Tv_j = \sum_{i=1}^{m} b_{ij} w_i.$$

$\mathcal{M}(S+T)$ is the matrix whose entries c_{ij} are defined by

$$\begin{aligned} (S+T)v_j &= \sum_{i=1}^{m} c_{ij} w_i \\ Sv_j + Tv_j &= \sum_{i=1}^{m} c_{ij} w_i \\ \sum_{i=1}^{m} a_{ij} w_i + \sum_{i=1}^{m} b_{ij} w_i &= \sum_{i=1}^{m} c_{ij} w_i \\ \sum_{i=1}^{m} (a_{ij} + b_{ij}) w_i &= \sum_{i=1}^{m} c_{ij} w_i \\ a_{ij} + b_{ij} &= c_{ij}. \end{aligned}$$

(ii) By definition, $\mathcal{M}(T)$ is the matrix whose entries a_{ij} are defined by

$$Tv_j = \sum_{i=1}^{m} a_{ij} w_i.$$

Then for $\lambda \in \mathbf{F}$, $\mathcal{M}(\lambda T)$ is the matrix whose entries b_{ij} are defined by

$$\begin{aligned} \lambda T v_j &= \sum_{i=1}^{m} b_{ij} w_i \\ \lambda \sum_{i=1}^{m} a_{ij} w_i &= \sum_{i=1}^{m} b_{ij} w_i \\ \lambda a_{ij} &= b_{ij}. \end{aligned}$$

□

Proposition 5.21. With addition and scalar multiplication defined as above, $\mathcal{M}_{m\times n}(\mathbf{F})$ is a vector space of dimension mn.

Proof. The verification that $\mathcal{M}_{m\times n}(\mathbf{F})$ is a vector space is left to the reader. Note that the additive identity of $\mathcal{M}_{m\times n}(\mathbf{F})$ is the $m \times n$ matrix all of whose entries equal 0.

The reader should also verify that the list of distinct $m \times n$ matrices that have 0 in all entries except for a 1 in one entry is a basis of $\mathcal{M}_{m\times n}(\mathbf{F})$. There are mn such matrices, so the dimension of $\mathcal{M}_{m\times n}(\mathbf{F})$ equals mn. □

Matrix Multiplication

Definition 5.22 (Matrix multiplication)**.** Suppose $A = (a_{ij})_{m\times n}$, $B = (b_{jk})_{n\times p}$. Then $AB = (c_{ik})_{m\times p}$ has entries defined by

$$c_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk}.$$

Remark. Thus the entry in row j, column k of AB is computed by taking row j of A and column k of B, multiplying together corresponding entries, and then summing.

Remark. Note that we define the product of two matrices only when the number of columns of the first matrix equals the number of rows of the second matrix.

In the next result, we assume that the same basis of V is used in considering $T \in \mathcal{L}(U, V)$ and $S \in \mathcal{L}(V, W)$, the same basis of W is used in considering $S \in \mathcal{L}(V, W)$ and $ST \in \mathcal{L}(U, W)$, and the same basis of U is used in considering $T \in \mathcal{L}(U, V)$ and $ST \in \mathcal{L}(U, W)$.

Proposition 5.23 (Matrix of product of linear maps)**.** If $T \in \mathcal{L}(U, V)$ and $S \in \mathcal{L}(V, W)$, then $\mathcal{M}(ST) = \mathcal{M}(S)\mathcal{M}(T)$.

Proof. Suppose $\{v_1, \ldots, v_n\}$ is a basis of V, $\{w_1, \ldots, w_m\}$ is a basis of W, $\{u_1, \ldots, u_p\}$ is a basis of U.

Let $\mathcal{M}(S) = (a_{ij})_{m\times n}$, $\mathcal{M}(T) = (b_{jk})_{n\times p}$, where

$$\begin{aligned} Sv_j &= \sum_{i=1}^{m} a_{ij} w_i \\ Tu_k &= \sum_{j=1}^{n} b_{jk} v_j. \end{aligned}$$

For $k = 1, \ldots, p$, we have

$$\begin{aligned} (ST)u_k &= S(Tu_k) \\ &= S\left(\sum_{j=1}^{n} b_{jk} v_j\right) \\ &= \sum_{j=1}^{n} b_{jk} S v_j \\ &= \sum_{j=1}^{n} b_{jk} \left(\sum_{i=1}^{m} a_{ij} w_i\right) \\ &= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij} b_{jk}\right) w_i. \end{aligned}$$

□

Notation. $A_{i,\cdot}$ denotes the row vector corresponding to the i-th row of A; $A_{\cdot,j}$ denotes the column vector corresponding to the j-th column of A.

Proposition 5.24. Suppose $A = (a_{ij})_{m\times n}$, $B = (b_{jk})_{n\times p}$. Let $AB = (c_{ik})_{m\times p}$. Then

$$c_{ik} = A_{i,\cdot} B_{\cdot,k}$$

That is, the entry in row i, column k of AB equals (row i of A) times (column k of B).

Proof. By definition,

$$\begin{aligned} A_{i,\cdot} B_{\cdot,k} &= \begin{pmatrix} a_{i1} & \cdots & a_{in} \end{pmatrix} \begin{pmatrix} b_{1k} \\ \vdots \\ b_{nk} \end{pmatrix} \\ &= a_{i1} b_{1k} + \cdots + a_{in} b_{nk} \\ &= \sum_{j=1}^{n} a_{ij} b_{jk} \\ &= c_{ik}. \end{aligned}$$

□

Proposition 5.25. Suppose $A = (a_{ij})_{m\times n}$, $B = (b_{jk})_{n\times p}$. Then

$$(AB)_{\cdot,k} = AB_{\cdot,k}$$

That is, column k of AB equals A times column k of B.

Proof. Using the previous result,

$$AB_{\cdot,k} = \begin{pmatrix} A_{1,\cdot}B_{\cdot,k} \\ \vdots \\ A_{n,\cdot}B_{\cdot,k} \end{pmatrix} = \begin{pmatrix} c_{1k} \\ \vdots \\ c_{nk} \end{pmatrix} = (AB)_{\cdot,k}$$

□

Proposition 5.26 (Linear combination of columns)**.** Suppose $A = (a_{ij})_{m\times n}$, $b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$. Then

$$Ab = b_1 A_{\cdot,1} + \cdots + b_n A_{\cdot,n}.$$

That is, Ab is a linear combination of the columns of A, with the scalars that multiply the columns coming from b.

Proof.

$$\begin{aligned} Ab &= \begin{pmatrix} a_{11}b_1 + \cdots + a_{1n}b_n \\ \vdots \\ a_{m1}b_1 + \cdots + a_{mn}b_n \end{pmatrix} \\ &= \begin{pmatrix} a_{11}b_1 \\ \vdots \\ a_{m1}b_1 \end{pmatrix} + \cdots + \begin{pmatrix} a_{1n}b_n \\ \vdots \\ a_{mn}b_n \end{pmatrix} \\ &= b_1 \begin{pmatrix} a_{11} \\ \vdots \\ a_{m1} \end{pmatrix} + \cdots + b_n \begin{pmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{pmatrix} \\ &= b_1 A_{\cdot,1} + \cdots + b_n A_{\cdot,n}. \end{aligned}$$

□

The following result states that matrix multiplication can be expressed as linear combinations of columns or rows.

Proposition 5.27. Suppose $C = (c_{ij})_{m\times c}$, $R = (r_{jk})_{c\times n}$. Then

(i) Columns: for $k = 1, \ldots, n$, $(CR)_{\cdot,k}$ is a linear combination of $C_{\cdot,1}, \ldots, C_{\cdot,c}$, with coefficients coming from $R_{\cdot,k}$.

(ii) Rows: for $i = 1, \ldots, m$, $(CR)_{i,\cdot}$ is a linear combination of $R_{1,\cdot}, \ldots, R_{c,\cdot}$, with coefficients coming from $C_{j,\cdot}$.

Proof.

(i)

(ii)

□

Rank of a Matrix

Definition 5.28. Suppose $A \in \mathcal{M}_{m\times n}(\mathbf{F})$. Then the *row space* of A is the span of its rows, and the *column space* of A is the span of its columns:

$$\begin{aligned}\mathrm{Row}(A) &\coloneqq \mathrm{span}\left(A_{i,\cdot} \mid 1 \leqslant i \leqslant m\right),\\ \mathrm{Col}(A) &\coloneqq \mathrm{span}\left(A_{\cdot,j} \mid 1 \leqslant j \leqslant n\right).\end{aligned}$$

The ***row rank*** and ***column rank*** of A are defined as

$$\begin{aligned}r(A) &\coloneqq \dim \mathrm{Row}(A),\\ c(A) &\coloneqq \dim \mathrm{Col}(A).\end{aligned}$$

Definition 5.29 (Transpose)**.** Suppose $A = (a_{ij})_{m\times n}$. Then the ***transpose*** of A is the matrix $A^T = (b_{ij})_{n\times m}$, whose entries are defined by

$$b_{ij} = a_{ji}.$$

Proposition 5.30 (Properties of transpose)**.** Suppose $A, B \in \mathcal{M}_{m\times n}(\mathbf{F})$, $C \in \mathcal{M}_{n\times p}(\mathbf{F})$. Then

(i) $(A+B)^T = A^T + B^T$;

(ii) $(\lambda A)^T = \lambda A^T$ for $\lambda \in \mathbf{F}$;

(iii) $(AC)^T = C^T A^T$.

Lemma 5.31 (Column-row factorisation)**.** Suppose $A \in \mathcal{M}_{m\times n}(\mathbf{F})$, $c(A) \geqslant 1$. Then there exist $C \in M_{m\times c(A)}(\mathbf{F})$, $R \in M_{c(A)\times n}(\mathbf{F})$ such that $A = CR$.

Proof. We prove by construction, i.e. construct the required matrices C and R.

Each column of A is a $m \times 1$ matrix. The set of columns of A

$$\{A_{\cdot,1}, \ldots, A_{\cdot,n}\}$$

can be reduced to a basis of $\mathrm{Col}(A)$, which has length $c(A)$, by the definition of column rank. The $c(A)$ columns in this basis can be put together to form a $m \times c(A)$ matrix, which we call C.

For $k = 1, \ldots, n$, the k-th column of A is a linear combination of the columns of C. Make the coefficients of this linear combination into column k of a $c \times n$ matrix, which we call R. By , it follows that $A = CR$. □

Lemma 5.32 (Column rank equals row rank)**.** The column rank of a matrix equals to its row rank.

Proof. Suppose $A \in \mathcal{M}_{m\times n}(\mathbf{F})$. Let $A = CR$ be the column-row factorisation of A, where $C \in \mathcal{M}_{m\times c(A)}(\mathbf{F})$, $R \in \mathcal{M}_{c(A)\times n}(\mathbf{F})$.

□

Since column rank equals row rank, we can dispense with the terms “column rank” and “row rank”, and just use the simpler term “rank”.

Definition 5.33 (Rank)**.** The ***rank*** of a matrix A is defined as

$$\operatorname{rank} A \coloneqq r(A) = c(A).$$

§5.4 Invertibility and Isomorphism

Invertibility

Notation. $I_V \in \mathcal{L}(V)$ denotes the identity map on V:

$$Iv = v \quad (\forall v \in V)$$

The subscript is omitted if there is no ambiguity.

Definition 5.34 (Invertibility)**.** $T \in \mathcal{L}(V, W)$ is ***invertible*** if there exists $S \in \mathcal{L}(W, V)$ such that $ST = I_V$, $TS = I_W$; S is known as an *inverse* of T.

Proposition 5.35 (Uniqueness of inverse)**.** The inverse of an invertible linear map is unique.

Proof. Suppose $T \in \mathcal{L}(V, W)$ is invertible, $S_1, S_2 \in \mathcal{L}(W, V)$ are inverses of T. Then

$$S_1 = S_1 I_W = S_1(TS_2) = (S_1 T)S_2 = I_V S_2 = S_2.$$

Thus $S_1 = S_2$. □

Now that we know that the inverse is unique, we can give it a notation.

Notation. If T is invertible, then its inverse is denoted by T^{-1}.

The following result is useful in determing if a linear map is invertible.

Lemma 5.36 (Invertibility criterion)**.** Suppose $T \in \mathcal{L}(V, W)$.

(i) T is invertible $\iff$ T is injective and surjective.

(ii) If $\dim V = \dim W$, T is invertible $\iff$ T is injective $\iff$ T is surjective.

Proof.

(i) $\boxed{\implies}$ Suppose $T \in \mathcal{L}(V, W)$ is invertible, which has inverse T^{-1}. Suppose $Tu = Tv$. Applying T^{-1} to both sides of the equation gives

$$u = T^{-1}Tu = T^{-1}Tv = v$$

so T is injective.

We now show T is surjective. Let $w \in W$. Then $w = T\left(T^{-1}w\right)$, which shows that $w \in \operatorname{im} T$, so $\operatorname{im} T = W$. Hence T is surjective.

$\boxed{\impliedby}$ Suppose T is injective and surjective.

Define $S \in \mathcal{L}(W, V)$ such that for each $w \in W$, $S(w)$ is the unique element of V such that $T(S(w)) = w$ (we can do this due to injectivity and surjectivity). Then we have that $T(ST)v = (TS)Tv = Tv$ and thus $STv = v$ so $ST = I$. It is easy to show that S is a linear map.

(ii) It suffices to only prove T is injective $\iff$ T is surjective. Then apply the previous result.

$\boxed{\implies}$ Suppose T is injective. Then $\dim \ker T = 0$. By the fundamental theorem of linear maps,

$$\begin{aligned}\dim \operatorname{im} T &= \dim V - \dim \ker T \\ &= \dim V \\ &= \dim W\end{aligned}$$

which implies that T is surjective.

$\boxed{\impliedby}$ Suppose T is surjective, then $\dim \operatorname{im} T = \dim W$. By the fundamental theorem of linear maps,

$$\begin{aligned}\dim \ker T &= \dim V - \dim \operatorname{im} T \\ &= \dim V - \dim W \\ &= 0\end{aligned}$$

which implies that T is injective.

□

Corollary 5.37. Suppose V and W are finite-dimensional, $\dim V = \dim W$, $S \in \mathcal{L}(W, V)$, $T = \mathcal{L}(V, W)$. Then $ST = I$ if and only if $TS = I$.

Proof.

$\boxed{\implies}$ Suppose $ST = I$. Let $v \in \ker T$. Then

$$v = Iv = (ST)v = S(Tv) = S(\mathbf{0}) = \mathbf{0} \implies \ker T = \{\mathbf{0}\}$$

so T is injective. Since $\dim V = \dim W$, by the previous result, T is invertible.

Since $ST = I$, then

$$S = STT^{-1} = IT^{-1} = T^{-1}$$

so $TS = TT^{-1} = I$, as desired.

$\boxed{\impliedby}$ Similar to above; reverse the roles of S and T (and V and W) to show that if $TS = I$ then $ST = I$. □

Isomorphism

Definition 5.38 (Isomorphism)**.** An ***isomorphism*** is an invertible linear map. V and W are ***isomorphic***, denoted by $V \cong W$, if there exists an isomorphism $T \in \mathcal{L}(V, W)$.

The following result shows that we need to look at only at the dimension to determine whether two vector spaces are isomorphic.

Lemma 5.39. Suppose V and W are finite-dimensional. Then

$$V \cong W \iff \dim V = \dim W.$$

Proof.

$\boxed{\implies}$ Suppose $V \cong W$, then there exists an isomorphism $T \in \mathcal{L}(V, W)$, which is invertible, so T is both injective and surjective, thus $\ker T = \{\mathbf{0}\}$ and $\operatorname{im} T = W$, implying $\dim \ker T = 0$ and $\dim \operatorname{im} T = \dim W$.

By the fundamental theorem of linear maps,

$$\begin{aligned}\dim V &= \dim \ker T + \dim \operatorname{im} T \\ &= 0 + \dim W = \dim W.\end{aligned}$$

$\boxed{\impliedby}$ Suppose V and W are finite-dimensional, $\dim V = \dim W = n$. Let $\{v_1, \ldots, v_n\}$ be a basis of V, $\{w_1, \ldots, w_n\}$ be a basis of W.

It suffices to construct an surjective $T \in \mathcal{L}(V, W)$. By the linear map lemma, there exists a linear map $T \in \mathcal{L}(V, W)$ such that

$$Tv_i = w_i \quad (i = 1, \ldots, n)$$

Let $w \in W$. Then there exist $a_i \in \mathbf{F}$ such that $w = a_1 w_1 + \cdots + a_n w_n$. Then

$$\begin{aligned}T(a_1 v_1 + \cdots + a_n v_n) = w &\implies w \in \operatorname{im} T \\ &\implies W = \operatorname{im} T \\ &\implies T \text{ is surjective} \\ &\implies T \text{ is invertible.}\end{aligned}$$

□

Proposition 5.40. Suppose $\{v_1, \ldots, v_n\}$ is a basis of V, $\{w_1, \ldots, w_m\}$ is a basis of W. Then

$$\mathcal{L}(V, W) \cong \mathcal{M}_{m \times n}(\mathbf{F}).$$

Proof. We claim that $\mathcal{M}$ is an isomorphism between $\mathcal{L}(V, W)$ and $\mathcal{M}_{m \times n}(\mathbf{F})$.

We already noted that $\mathcal{M}$ is linear. We need to prove that $\mathcal{M}$ is (i) injective and (ii) surjective.

(i) Given $T \in \mathcal{L}(V, W)$, if $\mathcal{M}(T) = 0$, then

$$Tv_j = 0 \quad (j = 1, \ldots, n)$$

Since $v_1, \ldots, v_n$ is a basis of V, this implies $T = \mathbf{0}$, so $\ker \mathcal{M} = \{\mathbf{0}\}$. Thus $\mathcal{M}$ is injective.

(ii) Suppose $A \in \mathcal{M}_{m\times n}(\mathbf{F})$. By the linear map lemma, there exists $T \in \mathcal{L}(V,W)$ such that

$$Tv_j = \sum_{i=1}^{m} a_{ij} w_i \quad (j = 1, \ldots, n)$$

Since $\mathcal{M}(T) = A$, $\operatorname{im} \mathcal{M} = \mathcal{M}_{m\times n}(\mathbf{F})$ so $\mathcal{M}$ is surjective.

□

Corollary 5.41. Suppose V and W are finite-dimensional. Then $\mathcal{L}(V,W)$ is finite-dimensional and

$$\dim \mathcal{L}(V,W) = (\dim V)(\dim W).$$

Proof. Since $\mathcal{L}(V,W) \cong \mathcal{M}_{m\times n}(\mathbf{F})$,

$$\dim \mathcal{L}(V,W) = \dim \mathcal{M}_{m\times n}(\mathbf{F}) = mn = (\dim V)(\dim W).$$

□

Linear Maps Thought of as Matrix Multiplication

Previously we defned the matrix of a linear map. Now we defne the matrix of a vector.

Definition 5.42 (Matrix of a vector)**.** Suppose $v \in V$, $\{v_1, \ldots, v_n\}$ is a basis of V. The matrix of v with respect to this basis is

$$\mathcal{M}(v) = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

where $b_1, \ldots, b_n \in \mathbf{F}$ are such that

$$v = b_1 v_1 + \cdots + b_n v_n.$$

Example

If $x = (x_1, \ldots, x_n) \in \mathbf{F}^n$, then the matrix of the vector x with respect to the standard basis is

$$\mathcal{M}(x) = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

Proposition 5.43. Suppose $T \in \mathcal{L}(V,W)$. Let $\{v_1, \ldots, v_n\}$ be a basis of V, $\{w_1, \ldots, w_m\}$ be a basis of W. Then

$$\mathcal{M}(T)_{\cdot,j} = \mathcal{M}(Tv_j) \quad (j = 1, \ldots, n)$$

Proof. By definition, the entries of $\mathcal{M}(T)$ are defined such that

$$Tv_j = \sum_{i=1}^{m} a_{ij} w_i \quad (j = 1, \ldots, n)$$

Then since $Tv_j \in W$, by definition, the matrix of Tv_j with respect to the basis $\{w_1, \ldots, w_m\}$ is

$$\mathcal{M}(Tv_j) = \begin{pmatrix} a_1 j \\ \vdots \\ a_{mj} \end{pmatrix}$$

which is precisely the j-th column of $\mathcal{M}(T)_{\cdot,j}$, for $j = 1, \ldots, n$. □

The following result shows that linear maps act like matrix multiplication.

Proposition 5.44. Suppose $T \in \mathcal{L}(V, W)$. Let $\{v_1, \ldots, v_n\}$ be a basis of V, $\{w_1, \ldots, w_m\}$ be a basis of W. Let $v \in V$, then

$$\mathcal{M}(Tv) = \mathcal{M}(T)\mathcal{M}(v).$$

Proof. Suppose $v = b_1 v_1 + \cdots + b_n v_n$ for some $b_1, \ldots, b_n \in \mathbf{F}$. Then

$$\begin{aligned} \mathcal{M}(Tv) &= \mathcal{M}\left(T(b_1 v_1 + \cdots + b_n v_n)\right) \\ &= b_1 \mathcal{M}(Tv_1) + \cdots + b_n \mathcal{M}(Tv_n) \\ &= b_1 \mathcal{M}(T)_{\cdot,1} + \cdots + b_n \mathcal{M}(T)_{\cdot,n} \\ &= \begin{pmatrix} \mathcal{M}(T)_{\cdot,1} & \cdots & \mathcal{M}(T)_{\cdot,n} \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \\ &= \mathcal{M}(T)\mathcal{M}(v). \end{aligned}$$

□

Notice that no bases are in sight in the statement of the next result. Although $\mathcal{M}(T)$ in the next result depends on a choice of bases of V and W, the next result shows that the column rank of $\mathcal{M}(T)$ is the same for all such choices (because $\operatorname{im} T$ does not depend on a choice of basis).

Proposition 5.45. Suppose V and W are finite-dimensional, $T \in \mathcal{L}(V, W)$. Then

$$\dim \ker T = \operatorname{rank} \mathcal{M}(T).$$

Proof. Suppose $\{v_1, \ldots, v_n\}$ is a basis of V, $\{w_1, \ldots, w_m\}$ is a basis of W.

The linear map that takes $w \in W$ to $\mathcal{M}(w)$ is an isomorphism from W to $\mathcal{M}_{m\times 1}(\mathbf{F})$ (consisting of $m \times 1$ column vectors).

The restriction of this isomorphism to $\operatorname{im} T$ [which equals $\operatorname{span}(Tv_1, \ldots, Tv_n)$] is an isomorphism from $\operatorname{im} T$ to $\operatorname{span}(\mathcal{M}(Tv_1), \ldots, \mathcal{M}(Tv_n))$. For $j = 1, \ldots, n$, the $m \times 1$ matrix $\mathcal{M}(Tv_j)$ equals column k of $\mathcal{M}(T)$. Thus

$$\dim \ker T = \operatorname{rank} \mathcal{M}(T),$$

as desired. □

Change of Basis

Definition 5.46 (Identity matrix)**.** For $n \in \mathbf{N}$, the $n \times n$ ***identity matrix*** is

$$I_n = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}.$$

Remark. Note that the symbol I is used to denote both the identity operator and the identity matrix. The context indicates which meaning of I is intended. For example, consider the equation $\mathcal{M}(I) = I$; on LHS I denotes the identity operator, and on RHS I denotes the identity matrix.

Proposition 5.47. Suppose $A \in \mathcal{M}_{n\times n}(\mathbf{F})$. Then $AI_n = I_n A = A$.

Proof. Exercise. □

Definition 5.48 (Invertible matrix)**.** $A \in \mathcal{M}_{n\times n}(\mathbf{F})$ is called *invertible* if there exists $B \in \mathcal{M}_{n\times n}(\mathbf{F})$ such that $AB = BA = I$; we call B an *inverse* of A

Proposition 5.49 (Uniqueness of inverse)**.** Suppose A is an invertible square matrix. Then there exists a unique matrix B such that $AB = BA = I$.

Proof. Suppose otherwise, for a contradiction, that A does not have a unique inverse. Let B and C be inverses of A; that is,

$$AB = BA = I,$$
$$AC = CA = I.$$

Then

$$B = BI = BAC = IC = C.$$

□

Since the inverse of a matrix is unique, we can give it a notation.

Notation. The inverse of a matrix A is denoted by A^{-1}.

Proposition 5.50.

(i) Suppose A is an invertible square matrix. Then $(A^{-1})^{-1} = A$.

(ii) Suppose A and C are invertible square matrices of the same size. Then AC is invertible, and $(AC)^{-1} = C^{-1}A^{-1}$.

Proof.

(i)
$$A^{-1}A = AA^{-1} = I,$$
so the inverse of A^{-1} is A.

(ii)
$$\begin{aligned}(AC)(C^{-1}A^{-1}) &= A(CC^{-1})A^{-1}\\ &= AIA^{-1}\\ &= AA^{-1}\\ &= I,\end{aligned}$$
and similarly $(C^{-1}A^{-1})(AC) = I$.

□

Proposition 5.51 (Matrix of product of linear maps)**.** Suppose $T \in \mathcal{L}(U,V)$, $S \in \mathcal{L}(V,W)$. Let $\mathcal{U} = \{u_1,\dots,u_m\}$ be a basis of U, $\mathcal{V} = \{v_1,\dots,v_n\}$ be a basis of V, $\mathcal{W} = \{w_1,\dots,w_p\}$ be a basis of W. Then
$$\mathcal{M}(ST;\mathcal{U},\mathcal{W}) = \mathcal{M}(S;\mathcal{V},\mathcal{W})\,\mathcal{M}(T;\mathcal{U},\mathcal{V}).$$

Proof. Refer to previous section. Now we are just being more explicit about the bases involved. □

Corollary 5.52. Suppose that $\mathcal{U} = \{u_1,\dots,u_n\}$ and $\mathcal{V} = \{v_1,\dots,v_n\}$ are bases of V. Then the matrices
$$\mathcal{M}(I;\mathcal{U},\mathcal{V}) \quad \text{and} \quad \mathcal{M}(I;\mathcal{V},\mathcal{U})$$
are invertible, and each is the inverse of the other.

Proof. □

Theorem 5.53 (Change-of-basis formula)**.** Suppose $T \in \mathcal{L}(V)$. Let $\mathcal{U} = \{u_1,\dots,u_n\}$ and $\mathcal{V} = \{v_1,\dots,v_n\}$ be bases of V. Let
$$A = \mathcal{M}(T;\mathcal{U}), \quad B = \mathcal{M}(T;\mathcal{V}),$$

and $C = \mathcal{M}(I; \mathcal{U}, \mathcal{V})$. Then

$$A = C^{-1}BC. \tag{5.2}$$

Proof. Note that

$$\begin{aligned}\mathcal{M}(T; \mathcal{U}, \mathcal{V}) &= \underbrace{\mathcal{M}(T; \mathcal{V})}_{B} \underbrace{\mathcal{M}(I; \mathcal{U}, \mathcal{V})}_{C} \\ &= \underbrace{\mathcal{M}(I; \mathcal{U}, \mathcal{V})}_{C} \underbrace{\mathcal{M}(T; \mathcal{U})}_{A}\end{aligned}$$

Hence $BC = CA$, and the desired result follows. □

Proposition 5.54. Suppose $\{v_1, \ldots, v_n\}$ is a basis of V, $T \in \mathcal{L}(V)$ is invertible. Then

$$\mathcal{M}\left(T^{-1}\right) = (\mathcal{M}(T))^{-1},$$

where both matrices are with respect to the basis $\{v_1, \ldots, v_n\}$.

Proof. We have that

$$\mathcal{M}\left(T^{-1}\right) \mathcal{M}(T) = \mathcal{M}\left(T^{-1}T\right) = \mathcal{M}(I) = I.$$

□

§5.5 Products and Quotients of Vector Spaces

Products of Vector Spaces

Definition 5.55 (Product)**.** Suppose $V_1, \ldots, V_n$ are vector spaces over $\mathbf{F}$. The ***product*** $V_1 \times \cdots \times V_n$ is defined by

$$V_1 \times \cdots \times V_n \coloneqq \{(v_1, \ldots, v_n) \mid v_i \in V_i\}.$$

Remark. This is analagous to the Cartesian product of sets.

Proposition 5.56. $V_1 \times \cdots \times V_n$ is a vector space over $\mathbf{F}$, with addition and scalar multiplication defined by

$$(u_1, \ldots, u_n) + (v_1, \ldots, v_n) = (u_1 + v_1, \ldots, u_n + v_n)$$
$$\lambda(v_1, \ldots, v_n) = (\lambda v_1, \ldots, \lambda v_n)$$

The following result shows that the dimension of a product is the sum of dimensions.

Proposition 5.57. Suppose $V_1, \ldots, V_n$ are finite-dimensional. Then $V_1 \times \cdots \times V_n$ is finite-dimensional, and

$$\dim(V_1 \times \cdots \times V_n) = \dim V_1 + \cdots + \dim V_n.$$

Proof. For each V_k $(k = 1, \ldots, n)$, choose a basis:

$$\mathcal{B}_k = \{e_{k1}, \ldots, e_{k \dim V_k}\}.$$

For each basis vector of each V_k, consider the set consisting of elements of $V_1 \times \cdots \times V_n$ that equal the basis vector in the k-th slot and 0 in the other slots:

$$\mathcal{B} = \{(0, \ldots, \underbrace{e_{ki}}_{k\text{-th slot}}, \ldots, 0) \mid 1 \leqslant i \leqslant \dim V_k, 1 \leqslant k \leqslant n\}.$$

We want to show that $\mathcal{B}$ is a basis of $V_1 \times \cdots \times V_n$. Thus we need to show that it is (i) a spanning set, and (ii) linearly independent.

(i) Let $(v_1, \ldots, v_n) \in V_1 \times \cdots \times V_n$. For $k = 1, \ldots, n$, since $\mathcal{B}_k$ is a basis for V_k, we can write

$$v_k = \sum_{i=1}^{\dim V_k} a_{ki} e_{ki}.$$

for some $a_{k1}, \ldots, a_{k \dim V_k} \in \mathbf{F}$. Then

$$\begin{aligned}(v_1, \ldots, v_n) &= \sum_{k=1}^{n} (0, \ldots, v_k, \ldots, 0) \\ &= \sum_{k=1}^{n} \left(0, \ldots, \sum_{i=1}^{\dim V_k} a_{ki} e_{ki}, \ldots, 0\right) \\ &= \sum_{k=1}^{n} \sum_{i=1}^{\dim V_k} a_{ki} \left(0, \ldots, e_{ki}, \ldots, 0\right)\end{aligned}$$

which is a linear combination of vectors in $\mathcal{B}$. Hence $\mathcal{B}$ spans $V_1 \times \cdots \times V_n$.

(ii) Suppose there exist $a_{ki} \in \mathbf{F}$ such that

$$\begin{aligned}\sum_{k=1}^{n} \sum_{i=1}^{\dim V_k} a_{ki} \left(0, \ldots, e_{ki}, \ldots, 0\right) &= \mathbf{0} \\ \sum_{k=1}^{n} \left(0, \ldots, \sum_{i=1}^{\dim V_k} a_{ki} e_{ki}, \ldots, 0\right) &= \mathbf{0} \\ \left(\sum_{i=1}^{\dim V_1} a_{1i} e_{1i}, \sum_{i=1}^{\dim V_2} a_{2i} e_{2i}, \ldots, \sum_{i=1}^{\dim V_n} a_{ni} e_{ni}\right) &= \mathbf{0}\end{aligned}$$

so for $k = 1, \ldots, n$,

$$\sum_{i=1}^{\dim V_k} a_{ki} e_{ki} = \mathbf{0}.$$

By the linear independence of vectors in $\mathcal{B}_k$, we have that

$$a_{k1} = \cdots = a_{k \dim V_k} = 0$$

for $k = 1, \ldots, n$.

Hence

$$\begin{aligned}\dim(V_1 \times \cdots \times V_n) &= |\mathcal{B}| \\ &= |\mathcal{B}_1| + \cdots + |\mathcal{B}_n| \\ &= \dim V_1 + \cdots + \dim V_n.\end{aligned}$$

□

Products are also related to direct sums, by the following result.

Proposition 5.58. Suppose that $V_1, \ldots, V_n \leqslant V$. Define a linear map

$$\begin{aligned}\Gamma : V_1 \times \cdots \times V_n &\to V_1 + \cdots + V_n \\ (v_1, \ldots, v_n) &\mapsto v_1 + \cdots + v_n\end{aligned}$$

Then $V_1 + \cdots + V_n$ is a direct sum if and only if Γ is injective.

Proof.

(i) $\iff$ (ii) Suppose $V_1 + \cdots + V_n$ is a direct sum. Let $(v_1, \ldots, v_n) \in \ker \Gamma$. Then

$$\begin{aligned} \Gamma(v_1, \ldots, v_n) &= \mathbf{0} \\ v_1 + \cdots + v_n &= \mathbf{0} \\ v_1 = \cdots = v_n &= \mathbf{0} \end{aligned}$$

so $(v_1, \ldots, v_n) = 0$. Hence $\ker \Gamma = 0$, thus Γ is injective.

(ii) $\iff$ (i) Similar to the above proof. □

The next result says that a sum is a direct sum if and only if dimensions add up.

Proposition 5.59. Suppose V is finite-dimensional, $V_1, \ldots, V_n \leqslant V$. Then $V_1 + \cdots + V_n$ is a direct sum if and only if

$$\dim(V_1 + \cdots + V_n) = \dim V_1 + \cdots + \dim V_n.$$

Proof. The map Γ defined in the previous result is surjective. Thus by the fundamental theorem of linear maps, Γ is injective if and only if

$$\dim(V_1 + \cdots + V_n) = \dim(V_1 \times V_n).$$

Then use the previous two results above. □

Quotient Spaces

Definition 5.60 (Coset)**.** Suppose $v \in V$, $U \subset V$. Then $v + U$ is called a ***coset*** of U, defined by

$$v + U \coloneqq \{v + u \mid u \in U\}.$$

Definition 5.61 (Quotient space)**.** Suppose $U \leqslant V$. Then the ***quotient space*** V/U is the set of cosets of U:

$$V/U \coloneqq \{v + U \mid v \in V\}.$$

> **Example**
> If $U = \{(x, 2x) \in \mathbf{R}^2 \mid x \in \mathbf{R}\}$, then $\mathbf{R}^2/U$ is the set of lines in $\mathbf{R}^2$ that have gradient of 2.

It is obvious that two cosets of a subspace are equal or disjoint. We shall now prove this.

Proposition 5.62. Suppose $U \leqslant V$, and $v, w \in V$. Then

$$v - w \in U \iff v + U = w + U \iff (v + U) \cap (w + U) = \emptyset.$$

Proof. First suppose $v - w \in U$. If $u \in U$, then

$$v + u = w + \big((v - w) + u\big) \in w + U.$$

Thus $v + U \subset w + U$. Similarly, $w + U \subset v + U$. Thus $v + U = w + U$, completing the proof that $v - w \in U$ implies $v + U = w + U$.

The equation $v + U = w + U$ implies that $(v + U) \cap (w + U) \neq \emptyset$.

Now suppose $(v + U) \cap (w + U) \neq \emptyset$. Thus there exist $u_1, u_2 \in U$ such that

$$v + u_1 = w + u_2.$$

Thus $v - w = u_2 - u_1$. Hence $v - w \in U$, showing that $(v + U) \cap (w + U) \neq \emptyset$ implies $v - w \in U$, which completes the proof. □

Proposition 5.63. Suppose $U \leqslant V$. Then V/U is a vector space, with addition and scalar multiplication defined by

$$(v + U) + (w + U) = (v + w) + U$$
$$\lambda(v + U) = (\lambda v) + U$$

for all $v, w \in V$, $\lambda \in \mathbf{F}$.

Proof. □

Definition 5.64 (Quotient map)**.** Suppose $U \leqslant V$. The ***quotient map*** $\pi : V \to V/U$ is the linear map defined by

$$\pi(v) = v + U$$

for all $v \in V$.

Proposition 5.65 (Dimension of quotient space)**.** Suppose V is finite-dimensional, $U \leqslant V$. Then

$$\dim V/U = \dim V - \dim U.$$

Definition 5.66. Suppose $T \in \mathcal{L}(V, W)$. Define $\tilde{T} : V/\ker T \to W$ by

$$\tilde{T}(v + \ker T) = Tv.$$

Proposition 5.67. Suppose $T \in \mathcal{L}(V, W)$. Then

(i) $\tilde{T} \circ \pi = T$, where π is the quotient map of V onto $V/\ker T$;

(ii) $\tilde{T}$ is injective;

(iii) $\operatorname{im} \tilde{T} = \operatorname{im} T$.

Theorem 5.68 (First isomorphism theorem)**.** Suppose $T \in \mathcal{L}(V, W)$ is an isomorphism. Then

$$V/\ker T \cong \operatorname{im} T. \tag{5.3}$$

§5.6 Duality

Dual Space and Dual Map

Linear maps into the scalar field $\mathbf{F}$ play a special role in linear algebra, and thus they get a special name.

Definition 5.69 (Linear funtional)**.** A ***linear functional*** on V is a linear map from V to $\mathbf{F}$; that is, a linear functional is an element of $\mathcal{L}(V, \mathbf{F})$.

The vector space $\mathcal{L}(V, \mathbf{F})$ also gets a special name and special notation.

Definition 5.70 (Dual space)**.** The ***dual space*** of V is the vector space of linear functionals on V; that is, $V^* := \mathcal{L}(V, \mathbf{F})$.

Lemma 5.71. Suppose V is finite-dimensional. Then V^* is also finite-dimensional, and

$$\dim V^* = \dim V.$$

Proof. By , we have

$$\dim V^* := \dim \mathcal{L}(V, \mathbf{F}) = (\dim V)(\dim \mathbf{F}) = \dim V$$

as desired. □

Definition 5.72 (Dual basis)**.** If $\{v_1, \dots, v_n\}$ is a basis of V, then the ***dual basis*** of $\{v_1, \dots, v_n\}$ is

$$\{\phi_1, \dots, \phi_n\} \subset V^*,$$

where each ϕ_i is the linear functional on V such that

$$\phi_i(v_j) = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j) \end{cases}$$

The following result states that dual basis gives coefficients for linear combination.

Proposition 5.73. Suppose $\{v_1, \dots, v_n\}$ is a basis of V, and $\{\phi_1, \dots, \phi_n\}$ is the dual basis. Then for each $v \in V$,

$$v = \phi_1(v)v_1 + \cdots + \phi_n(v)v_n.$$

The following result states that the dual basis is a basis of the dual space.

Proposition 5.74. Suppose V is finite-dimensional. Then the dual basis of a basis of V is a basis of V^*.

Definition 5.75 (Dual map)**.** Suppose $T \in \mathcal{L}(V, W)$. The ***dual map*** of T is the linear map $T^* \in \mathcal{L}(V, W)$ defined for each $\phi \in W^*$ by

$$T^*(\phi) = \phi \circ T.$$

Proposition 5.76 (Algebraic properties of dual map)**.** Suppose $T \in \mathcal{L}(V, W)$. Then

(1) $(S + T)^* = S^* + T^*$ for all $S \in \mathcal{L}(V, W)$

(2) $(\lambda T)^* = \lambda T^*$ for all $\lambda \in \mathbf{F}$

(3) $(ST)^* = T^* S^*$ for all $S \in \mathcal{L}(V, W)$

Kernel and Image of Dual of Linear Map

The goal of this section is to describe $\ker T^*$ and $\operatorname{im} T^*$ in terms of $\operatorname{im} T$ and $\ker T$.

Definition 5.77 (Annihilator)**.** For $U \subset V$, the ***annihilator*** of U is defined by

$$U^\circ \coloneqq \{\phi \in V^* \mid \phi(u) = 0, \forall u \in U\}.$$

Proposition 5.78. $U^\circ \leqslant V$.

Proposition 5.79 (Dimension of annihilator)**.** Suppose V is finite-dimensional, $U \leqslant V$. Then

$$\dim U^\circ = \dim V - \dim U.$$

The following are conditions for the annihilator to equal $\{\mathbf{0}\}$ or the whole space.

Proposition 5.80. Suppose V is finite-dimensional, $U \leqslant V$. Then

(i) $U^\circ = \{\mathbf{0}\} \iff U = V$;

(ii) $U^\circ = V^* \iff U = \{\mathbf{0}\}$.

The following result concerns $\ker T^*$.

Proposition 5.81. Suppose V and W are finite-dimensional, $T \in \mathcal{L}(V, W)$. Then

(i) $\ker T^* = (\operatorname{im} T)^\circ$;

(ii) $\dim \ker T^* = \dim \ker T + \dim W - \dim V$.

The next result can be useful because sometimes it is easier to verify that T^* is injective than to show directly that T is surjective.

Proposition 5.82. Suppose V and W are finite-dimensional, $T \in \mathcal{L}(V, W)$. Then

$$T \text{ is surjective} \iff T^* \text{ is injective.}$$

The following result concerns $\operatorname{im} T^*$.

Proposition 5.83. Suppose V and W finite-dimensional, $T \in \mathcal{L}(V, W)$. Then

(i) $\dim \operatorname{im} T^* = \dim \operatorname{im} T$;

(ii) $\dim T^* = (\ker T)^\circ$.

Proposition 5.84. Suppose V and W are finite-dimensional, $T \in \mathcal{L}(V, W)$. Then

$$T \text{ is injective} \iff T^* \text{ is surjective.}$$

Matrix of Dual of Linear Map

Proposition 5.85. Suppose V and W are finite-dimensional, $T \in \mathcal{L}(V, W)$. Then

$$\mathcal{M}(T^*) = (\mathcal{M}(T))^t.$$

Exercises

Problem 5.1 ([Axl24] 3A)**.** Suppose $b, c \in \mathbf{R}$. Define $T : \mathbf{R}^3 \to \mathbf{R}^2$ by

$$T(x, y, z) = (2x - 4y + 3z + b, 6x + cxyz).$$

Show that T is linear if and only if $b = c = 0$.

Problem 5.2 ([Axl24] 3A Q11)**.** Suppose V is finite-dimensional, $T \in \mathcal{L}(V)$. Prove that T is a scalar multiple of the identity if and only if $ST = TS$ for all $S \in \mathcal{L}(V)$.

Problem 5.3 ([Axl24] 3B Q9)**.** Suppose $T \in \mathcal{L}(V, W)$ is injective, $\{v_1, \dots, v_n\}$ is linearly independent in V. Prove that $\{Tv_1, \dots, Tv_n\}$ is linearly independent in W.

Solution. Suppose there exist $a_i \in \mathbf{F}$ such that

$$\begin{aligned} &a_1 Tv_1 + \cdots + a_n Tv_n = \mathbf{0} \\ \implies &T(a_1 v_1 + \cdots + a_n v_n) = 0 \\ \implies &a_1 v_1 + \cdots + a_n v_n \in \ker T \end{aligned}$$

Since T is injective,

$$\ker T = \{\mathbf{0}\} \implies a_1 v_1 + \cdots + a_n v_n = \mathbf{0} \implies a_1 = \cdots = a_n = 0$$

since $\{v_1, \dots, v_n\}$ is linearly independent. □

Problem 5.4 ([Axl24] 3B Q11)**.** Suppose that V is finite-dimensional, $T \in \mathcal{L}(V, W)$. Prove that there exists $U \leqslant V$ such that

$$U \cap \ker T = \{\mathbf{0}\} \quad \text{and} \quad \operatorname{im} T = T(U).$$

Solution. □

Problem 5.5 ([Axl24] 3B Q19)**.** Suppose W is finite-dimensional, $T \in \mathcal{L}(V, W)$. Prove that T is injective if and only if there exists $S \in \mathcal{L}(W, V)$ such that ST is the identity operator on V.

Solution. □

Problem 5.6 ([Axl24] 3B Q20)**.** Suppose W is finite-dimensional, $T \in \mathcal{L}(V, W)$. Prove that T is surjective if and only if there exists $S \in \mathcal{L}(W, V)$ such that TS is the identity operator on W.

Problem 5.7 ([Axl24] 3B 22)**.** Suppose U, V are finite-dimensional, $S \in \mathcal{L}(V, W), T \in \mathcal{L}(U, V)$. Prove that

$$\dim \ker ST \leqslant \dim \ker S + \dim \ker T.$$

Solution. □

Problem 5.8 ([Axl24] 3D)**.** Suppose $T \in \mathcal{L}(V, W)$ is invertible. Show that T^{-1} is invertible and

$$\left(T^{-1}\right)^{-1} = T.$$

Solution. T^{-1} is invertible because there exists T such that $TT^{-1} = T^{-1}T = I$. So

$$T^{-1}T = TT^{-1} = I$$

thus $\left(T^{-1}\right)^{-1} = T$. □

3D Q11,12,17,22,23,24

Problem 5.9 ([Axl24] 3D)**.** Suppose $T \in \mathcal{L}(U, V)$ and $S \in \mathcal{L}(V, W)$ are both invertible linear maps. Prove that $ST \in \mathcal{L}(U, W)$ is invertible and that $(ST)^{-1} = T^{-1}S^{-1}$.

Solution.

$$(ST)(T^{-1}S^{-1}) = S(TT^{-1})S^{-1} = I = T^{-1}S^{-1}ST.$$

□

Problem 5.10 ([Axl24] 3D)**.** Suppose V is finite-dimensional and $T \in \mathcal{L}(V, W)$. Prove that the following are equivalent:

(i) T is invertible;

(ii) $\{Tv_1, \ldots, Tv_n\}$ is a basis of V for every basis $\{v_1, \ldots, v_n\}$ of V;

(iii) $\{Tv_1, \ldots, Tv_n\}$ is a basis of V for some basis $\{v_1, \ldots, v_n\}$ of V.

Solution.

(i) $\implies$ (ii) It only suffices to prove linear independence. We can show this

$$a_1Tv_1 + \cdots + a_nTv_n = 0 \iff a_1v_1 + \cdots + a_nv_n = 0$$

since T is injective and thus the only solution is all a_i are identically zero.

(ii) $\implies$ (iii) Trivial.

(iii) $\implies$ (i) By the linear map lemma, there exists $S \in \mathcal{L}(V)$ such that $S(Tv_i) = v_i$ for all i. Such S is the inverse of T (one can verify) and thus T is invertible. □

Problem 5.11 ([Axl24] 3E)**.** Suppose $U \leqslant V$, V/U is finite-dimensional. Prove that $V \cong U \times (V/U)$.

Solution.

$$\dim V = \dim U + (\dim V - \dim U) = \dim U + \dim(V/U).$$

□

6 Polynomials

§6.1 Definitions

Definition 6.1 (Polynomial)**.** $p : \mathbf{F} \to \mathbf{F}$ is a ***polynomial*** with coefficients in $\mathbf{F}$ if there exist $a_i \in \mathbf{F}$ such that

$$p(z) = a_0 + a_1 z + \cdots + a_n z^n \quad (z \in \mathbf{F})$$

Notation. The set of polynomials with coefficients in $\mathbf{F}$ is denoted by $\mathbf{F}[z]$.

Proposition 6.2. With the usual operations of addition and scalar multiplication, $\mathbf{F}[z]$ is a vector space over $\mathbf{F}$, as you should verify. Hence $\mathbf{F}[z]$ is a subspace of $\mathbf{F}^{\mathbf{F}}$ (vector space of functions from $\mathbf{F}$ to $\mathbf{F}$).

Definition 6.3 (Degree)**.** A polynomial $p \in \mathbf{F}[z]$ is has ***degree*** n, denoted by $\deg p = n$, if there exist scalars $a_0, a_1, \ldots, a_n \in \mathbf{F}$ with $a_n \neq 0$ such that $p(z) = a_0 + a_1 z + \cdots + a_n z^n$ for all $z \in \mathbf{F}$.

Notation. For non-negative integer n, $\mathbf{F}_n[z]$ denotes the set of polynomials with coefficients in $\mathbf{F}$ and degree at most n.

Proposition 6.4. For non-negative integer n, $\mathbf{F}_n[z]$ is finite-dimensional.

Proof. $\mathbf{F}_n[z] = \operatorname{span}(1, z, z^2, \ldots, z^n)$ [here we slightly abuse notation by letting z^k denote a function]. □

Proposition 6.5. $\mathbf{F}[z]$ is infinite-dimensional.

Proof. Consider any list of elements of $\mathbf{F}[z]$. Let n denote the highest degree of the polynomials in this list. Then every polynomial in the span of this list has degree at most n. Thus z^{n+1} is not in the span of our list. Hence no list spans $\mathbf{F}[z]$. Thus $\mathbf{F}[z]$ is infnite-dimensional. □

§6.2 Zeros of Polynomials

Definition 6.6 (Zero of polynomial)**.** $\lambda \in \mathbf{F}$ is called a ***zero*** (or *root*) of a polynomial $p \in \mathbf{F}[z]$ if

$$p(\lambda) = 0.$$

The next result is the key tool that we will use to show that the degree of a polynomial is unique.

Lemma 6.7 (Factor theorem)**.** Suppose $n \in \mathbf{Z}^+$, $p \in \mathbf{F}_n[z]$. Suppose $\lambda \in \mathbf{F}$, then $p(\lambda) = 0$ if and only if there exists $q \in \mathbf{F}_{n-1}[z]$ such that

$$p(z) = (z - \lambda)q(z) \quad (\forall z \in \mathbf{F})$$

Proof.

$\boxed{\Longrightarrow}$ Suppose $p(\lambda) = 0$. Let $a_0, a_1, \ldots, a_n \in \mathbf{F}$ be such that

$$p(z) = a_n z^n + \cdots + a_1 z + a_0 \quad (\forall z \in \mathbf{F})$$

Then for all $z \in \mathbf{F}$,

$$\begin{aligned} p(z) &= p(z) - p(\lambda) \\ &= (a_n z^n + \cdots + a_1 z + a_0) - (a_n \lambda^n + \cdots + a_1 \lambda + a_0) \\ &= a_n (z^n - \lambda^n) + \cdots + a_1 (z - \lambda). \end{aligned}$$

Note that for each $k = 1, \ldots, n$, we can factorise

$$z^k - \lambda^k = (z - \lambda)\left(z^{k-1} + z^{k-2}\lambda + \cdots + \lambda^{k-1}\right).$$

Thus p equals $z - \lambda$ times some polynomial of degree $n - 1$, as desired.

$\boxed{\Longleftarrow}$ Now suppose that there exists a polynomial $q \in \mathbf{F}[z]$ such that

$$p(z) = (z - \lambda)q(z) \quad (\forall z \in \mathbf{F})$$

Then

$$p(\lambda) = (\lambda - \lambda)q(\lambda) = 0,$$

as desired. □

Now we can prove that the degree of a polynomials determines how many zeros it has.

Proposition 6.8. Suppose $n \in \mathbf{Z}^+$, $p \in \mathbf{F}_n[z]$. Then p has at most n zeros in $\mathbf{F}$.

Proof. Prove by induction on n.

The desired result holds for $n = 1$ because if $a_1 \neq 0$ then the polynomial $a_0 + a_1 z$ has only one zero (which equals $-\frac{a_0}{a_1}$).

Now assume the desired result holds for $n - 1$. If p has no zeros in $\mathbf{F}$, then the desired result holds and we are done. Thus suppose p has a zero $\lambda \in \mathbf{F}$. By Lemma 6.7, there exists $q \in \mathbf{F}[z]$ of degree $n - 1$ such that

$$p(z) = (z - \lambda)q(z) \quad (\forall z \in \mathbf{F})$$

By the induction hypothesis, q has at most $n - 1$ zeros in $\mathbf{F}$. The equation above shows that the zeros of p in $\mathbf{F}$ are exactly the zeros of q in $\mathbf{F}$ along with λ. Thus p has at most n zeros in $\mathbf{F}$. □

The result above implies that the coefficients of a polynomial are uniquely determined (because if a polynomial had two different sets of coefficients, then subtracting the two representations of the polynomial would give a polynomial with some nonzero coefficients but infinitely many zeros). In particular, the degree of a polynomial is uniquely defined.

§6.3 Division Algorithm for Polynomials

Proposition 6.9 (Division algorithm)**.** Suppose $p, s \in \mathbf{F}[z]$, $s \neq 0$. Then there exists unique polynomials $q, r \in \mathbf{F}[z]$, where $\deg r < \deg s$, such that

$$p = sq + r.$$

Proof. Let $n = \deg p$, $m = \deg s$. If $n < m$, take $q = 0$ and $r = p$ to get the desired equation.

Now assume that $n \geqslant m$. □

§6.4 Factorisation of Polynomials over C

Theorem 6.10 (Fundamental theorem of algebra, first version)**.** Every non-constant polynomial with complex coefficients has a zero in $\mathbf{C}$.

Theorem 6.11 (Fundamental theorem of algebra)**.** If $p \in \mathbf{C}[z]$ is a non-constant polynomial, then p has a unique factorisation (except for the order of the factors) of the form

$$p(z) = c(z - \lambda_1) \cdots (z - \lambda_n),$$

where $c, \lambda_1, \ldots, \lambda_n \in \mathbf{C}$.

§6.5 Factorisation of Polynomials over R

A polynomial with real coefficients may have no real zeros. For example, the polynomial $x^2 + 1$ has no real zeros.

To obtain a factorisation theorem over $\mathbf{R}$, we will use our factorisation theorem over $\mathbf{C}$. We begin with the next result.

Proposition 6.12. Suppose $p \in \mathbf{C}[z]$ is a polynomial with real coefficients. If $\lambda \in \mathbf{C}$ is a zero of p, then so is the conjugate $\overline{\lambda}$.

We want a factorisation theorem for polynomials with real coefficients. We begin with the following result.

Lemma 6.13 (Factorisation of quadratic polynomial)**.** Suppose $b, c \in \mathbf{R}$. Then there is a polynomial factorisation of the form

$$x^2 + bx + c = (x - \lambda_1)(x - \lambda_2)$$

with $\lambda_1, \lambda_2 \in \mathbf{R}$ if and only if $b^2 \geqslant 4c$.

Theorem 6.14 (Factorisation of polynomial over $\mathbf{R}$)**.** Suppose $p \in \mathbf{R}[x]$ is a non-constant polynomial. Then p has a unique factorisation (except for the order of the factors) of the form

$$p(x) = c(x - \lambda_1) \cdots (x - \lambda_n)(x^2 + b_1 x + c_1) \cdots (x^2 + b_N x + c_N),$$

where $c, \lambda_1, \ldots, \lambda_n, b_1, \ldots, b_N, c_1, \ldots, c_N \in \mathbf{R}$, with ${b_k}^2 < 4c_k$ for each k.

7 Eigenvalues and Eigenvectors

§7.1 Invariant Subspaces

Eigenvalues

Definition 7.1 (Operator)**.** A linear map from a vector space to itself is called an ***operator***.

Definition 7.2 (Invariant subspace)**.** Suppose $T \in \mathcal{L}(V)$. $U \leqslant V$ is called ***invariant*** under T if $Tu \in U$ for all $u \in U$.

> **Example**
>
> Suppose $T \in \mathcal{L}(V)$. Then the following subspaces of V are all invariant under T.
>
> (i) The subspace $\{\mathbf{0}\}$ is invariant under T because if $u \in \{\mathbf{0}\}$, then $u = \mathbf{0}$ and hence $Tu = \mathbf{0} \in \{\mathbf{0}\}$.
>
> (ii) The subspace V is invariant under T because if $u \in V$, then $Tu \in V$.
>
> (iii) The subspace $\ker T$ is invariant under T because if $u \in \ker T$, then $Tu = \mathbf{0}$, and hence $Tu \in \ker T$, since a subspace must contain $\mathbf{0}$.
>
> (iv) The subspace $\operatorname{im} T$ is invariant under T because if $u \in \operatorname{im} T$, then $Tu \in \operatorname{im} T$ by definition.

Definition 7.3 (Eigenvalue and eigenvector)**.** Suppose $T \in \mathcal{L}(V)$. $\lambda \in \mathbf{F}$ is called an ***eigenvalue*** of T if there exists $v \in V \setminus \{\mathbf{0}\}$ such that $Tv = \lambda v$; v is called an ***eigenvector*** of T corresponding to λ.

Lemma 7.4 (Equivalent conditions to be an eigenvalue)**.** Suppose V is finite-dimensional, $T \in \mathcal{L}(V)$, $\lambda \in \mathbf{F}$. Then the following are equivalent:

(i) λ is an eigenvalue of T.

(ii) $T - \lambda I$ is not injective.

(iii) $T - \lambda I$ is not surjective.

(iv) $T - \lambda I$ is not invertible.

Proof.

$\boxed{\text{(i)} \iff \text{(ii)}}$ $Tv = \lambda v$ is equivalent to the equation $(T - \lambda I)v = \mathbf{0}$, so $T - \lambda I$ is not injective.

$\boxed{\text{(ii)} \iff \text{(iii)} \iff \text{(iv)}}$ This directly follows from Lemma 5.36. □

Proposition 7.5 (Linearly independent eigenvectors)**.** Suppose $T \in \mathcal{L}(V)$. Then every list of eigenvectors of T corresponding to distinct eigenvalues of T is linearly independent.

Proof. We prove by contradiction. Suppose, for a contradiction, that the desired result is false. Then there exists a smallest positive integer m such that $v_1, \ldots, v_m$ are linearly dependent eigenvectors of T corresponding to distinct eigenvalues $\lambda_1, \ldots, \lambda_m$ of T. The linear dependence implies there exists $a_1, \ldots, a_m \in \mathbf{F}$, none of which are 0 (because of the minimality of m) such that

$$a_1 v_1 + \cdots + a_m v_m = \mathbf{0}.$$

Applying $T - \lambda_m I$ to both sides of the equation,

$$\begin{aligned} a_1(T - \lambda_m I)v_1 + \cdots + a_{m-1}(T - \lambda_m I)v_{m-1} + a_m(T - \lambda_m I)v_m &= \mathbf{0} \\ a_1(Tv_1 - \lambda_m v_1) + \cdots + a_{m-1}(Tv_{m-1} - \lambda_m v_{m-1}) + a_m(Tv_m - \lambda_m v_m) &= \mathbf{0} \\ a_1(\lambda_1 - \lambda_m)v_1 + \cdots + a_{m-1}(\lambda_{m-1} - \lambda_m)v_{m-1} &= \mathbf{0} \end{aligned}$$

Since the eigenvalues $\lambda_1, \ldots, \lambda_m$ are distinct, none of the coefficients $a_i(\lambda_i - \lambda_m)$ equal 0. Thus $v_1, \ldots, v_{m-1}$ are $m - 1$ linearly dependent eigenvectors of T corresponding to distinct eigenvalues, contradicting the minimality of m. □

Corollary 7.6. Suppose V is finite-dimensional. Then each operator on V has at most $\dim V$ distinct eigenvalues.

Proof. Let $T \in \mathcal{L}(V)$. Suppose $\lambda_1, \ldots, \lambda_m$ are distinct eigenvalues of T with corresponding eigenvectors $v_1, \ldots, v_m$.

By Proposition 7.5, the eigenvectors $v_1, \ldots, v_m$ are linearly independent. Since the length of a linearly independent set is less than or equal to the length of a spanning set, we have that $m \leqslant \dim V$, as desired. □

Polynomials Applied to Operators

Notation. Suppose $T \in \mathcal{L}(V)$, $n \in \mathbf{Z}^+$. $T^n \in \mathcal{L}(V)$ is defined by $T^n = \underbrace{T \cdots T}_{m \text{ times}}$. T^0 is defined to be the identity operator I on V. If T is invertible with inverse T^{-1}, then $T^{-n} \in \mathcal{L}(V)$ is defined by $T^{-n} = \left(T^{-1}\right)^n$.

Having defined powers of an operator, we can now define what it means to apply a polynomial to an operator.

Definition 7.7. Suppose $T \in \mathcal{L}(V)$, $p \in \mathbf{F}[z]$ is a polynomial given by

$$p(z) = a_n z^n + \cdots + a_1 z + a_0 \quad (z \in \mathbf{F})$$

Then $p(T)$ is the operator on V defined by

$$p(T) \coloneqq a_n T^n + \cdots + a_1 T + a_0.$$

Remark. If we fix an operator $T \in \mathcal{L}(V)$, then the function $\mathbf{F}[z] \to \mathcal{L}(V)$ given by $p \mapsto p(T)$ is linear.

Definition 7.8 (Product of polynomials)**.** Suppose $p, q \in \mathbf{F}[z]$. Then $pq \in \mathbf{F}[z]$ is the polynomial defined by

$$(pq)(z) = p(z)q(z) \quad (z \in \mathbf{F})$$

Proposition 7.9. Suppose $p, q \in \mathbf{F}[z]$, $T \in \mathcal{L}(V)$. Then

(i) $(pq)(T) = p(T)q(T)$;

(ii) $p(T)q(T) = q(T)p(T)$.

Remark. This means when a product of polynomials is expanded using the distributive property, it does not matter whether the symbol is z or T.

Proof.

(i) Suppose

$$p(z) = \sum_{i=0}^{m} a_i z^i, \quad q(z) = \sum_{j=0}^{n} b_j z^j \quad (z \in \mathbf{F})$$

Then

$$\begin{aligned}(pq)(z) &= p(z)q(z) \\ &= \left(\sum_{i=0}^{m} a_i z^i\right)\left(\sum_{j=0}^{n} b_j z^j\right) \\ &= \sum_{i=0}^{m}\sum_{j=0}^{n} a_i b_j z^{i+j}.\end{aligned}$$

Thus

$$\begin{aligned}(pq)(T) &= \sum_{i=0}^{m}\sum_{j=0}^{n} a_i b_j T^{i+j} \\ &= \left(\sum_{i=0}^{m} a_i T^i\right)\left(\sum_{j=0}^{n} b_j T^j\right) \\ &= p(T)q(T).\end{aligned}$$

(ii) Using (i) twice, we have

$$p(T)q(T) = (pq)(T) = (qp)(T) = q(T)p(T).$$

□

Proposition 7.10. Suppose $T \in \mathcal{L}(V)$, $p \in \mathbf{F}[z]$. Then

(i) $\ker p(T)$ is invariant under T;

(ii) $\operatorname{im} p(T)$ is invariant under T.

Proof.

(i) Suppose $u \in \ker p(T)$. Then $p(T)u = \mathbf{0}$. Thus

$$\big(p(T)\big)\,(Tu) = \big(p(T)T\big)\,(u) = \big(Tp(T)\big)\,(u) = T\,\big(p(T)u\big) = T(\mathbf{0}) = \mathbf{0}.$$

Hence $Tu \in \ker p(T)$, so $\ker p(T)$ is invariant under T.

(ii) Suppose $u \in \operatorname{im} p(T)$. Then there exists $v \in V$ such that $u = p(T)v$. Thus

$$Tu = T\,\big(p(T)v\big) = p(T)(Tv).$$

Hence $Tu \in \operatorname{im} p(T)$, so $\operatorname{im} p(T)$ is invariant under T.

□

§7.2 The Minimal Polynomial

Existence of Eigenvalues on Complex Vector Spaces

Theorem 7.11 (Existence of eigenvalues)**.** Every operator on a finite-dimensional non-zero complex vector space has an eigenvalue.

Proof. Suppose V is a finite-dimensional complex vector space, $\dim V = n > 0$, $T \in \mathcal{L}(V)$. Let $v \in V \setminus \{\mathbf{0}\}$. Consider the set

$$\{v, Tv, T^2v, \ldots, T^nv\}.$$

Since $\dim V = n$ and this set has length $n+1$, this set is not linearly independent. Thus there exist $a_i \in \mathbf{C}$ such that

$$a_0v + \cdots + a_1Tv + a_2T^2v + \cdots + a_nT^nv = \mathbf{0},$$

which we can write as

$$p(T)v = \mathbf{0},$$

where $p(z) = a_0 + a_1z + \cdots + a_nz^n$, where we pick p such that $\deg p$ is minimal.

By the fundamental theorem of algebra, there exists a root of p in $\mathbf{C}$; let $\lambda \in \mathbf{C}$ be a root of p. Then by the factor theorem,

$$p(z) = (z - \lambda)q(z) \quad (\forall z \in \mathbf{C})$$

so

$$p(T) = (T - \lambda I)q(T).$$

Then

$$\mathbf{0} = p(T)v = (T - \lambda I)q(T)v,$$

or

$$Tq(T)v = \lambda qT(v).$$

We have that $q(T)v \neq \mathbf{0}$ since we chose $p(T)v = \mathbf{0}$, $\deg p$ is minimal. Therefore λ is an eigenvalue of T, with corresponding eigenvector $q(T)v$. □

Example

Note that the hypothesis in the result above that $\mathbf{F} = \mathbf{C}$ cannot be replaced with the hypothesis that $\mathbf{F} = \mathbf{R}$.

Consider $T \in \mathcal{L}(\mathbf{R}^2)$ defined by

$$Tv = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} v. \qquad (*)$$

Then

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

Notice that T is a rotation, so there is no vector that is fixed in its original direction. Hence T does not have an eigenvalue.

In contrast, consider $T \in \mathcal{L}(\mathbf{C}^2)$ defined by $(*)$. Then

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} i \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ i \end{pmatrix} = i \begin{pmatrix} i \\ 1 \end{pmatrix},$$

so i is an eigenvalue with corresponding eigenvector $\begin{pmatrix} i \\ 1 \end{pmatrix}$.

Eigenvalues and the Minimal Polynomial

A *monic polynomial* is a polynomial whose highest-degree coefficient equals 1.

The following result shows the existence, uniqueness and degree of the *minimal polynomial*.

Lemma 7.12. Suppose V is finite-dimensional, $T \in \mathcal{L}(V)$. Then there exists a unique monic polynomial $p \in \mathbf{F}[z]$ of smallest degree such that $p(T) = \mathbf{0}$. Furthermore, $\deg p \leqslant \dim V$.

Proof.

Existence Let $\dim V = n$. We use strong induction on n.

If $n = 0$, then I is the zero operator on V; thus take p to be the constant polynomial 1.

Now assume that $n > 0$ and that the desired result holds for all operators on all vector spaces of smaller dimension. We want to construct a monic polynomial of smallest degree such that when applied to T gives the $\mathbf{0}$ operator.

Let $u \in V \setminus \{\mathbf{0}\}$, consider the set

$$\{u, Tu, T^2u, \ldots, T^nu\}.$$

This set has length $n + 1$, so it is linearly dependent. By the linear dependence lemma, there exists a smallest positive integer $m \leqslant n$ such that T^mu is a linear combination of $u, Tu, \ldots, T^{m-1}u$; thus there exist $c_i \in \mathbf{F}$ such that

$$c_0u + c_1Tu + \cdots + c_{m-1}T^{m-1}u + T^mu = \mathbf{0}.$$

Define a monic polynomial $q \in \mathbf{F}[z]$ by $q(z) = c_0 + c_1z + \cdots + c_{m-1}z^{m-1} + z^m$. Then $q(T)u = \mathbf{0}$. Thus

for non-negative integer k,

$$q(T)\left(T^k u\right) = T^k\left(q(T)u\right) = T^k(\mathbf{0}) = \mathbf{0}.$$

By the linear dependence lemma, $\{u, Tu, \ldots, T^{m-1}u\}$ is linearly independent. Thus the above equation implies that $\dim \ker q(T) \geqslant m$. Hence by the fundamental theorem of linear maps,

$$\begin{aligned}\dim \operatorname{im} q(T) &= \dim V - \dim \ker q(T)\\ &\leqslant \dim V - m.\end{aligned}$$

Since $\operatorname{im} q(T)$ is invariant under T, we can apply the induction hypothesis to the restriction $T|_{\operatorname{im} q(T)}$. Thus there exists a monic polynomial $s \in \mathbf{F}[z]$ with $\deg s \leqslant \dim V - m$ such that

$$s\left(T|_{\operatorname{im} q(T)}\right) = \mathbf{0}.$$

Hence for all $v \in V$ we have

$$((sq)(T))v = s(T)(q(T)v) = \mathbf{0}$$

because $q(T)v \in \operatorname{im} q(T)$ and $s(T)|_{\operatorname{im} q(T)} = s(T|_{\operatorname{im} q(T)}) = \mathbf{0}$. Thus sq is a monic polynomial such that $\deg sq \leqslant \dim V$ and $(sq)(T) = \mathbf{0}$, as desired.

Uniqueness Let $p \in \mathbf{F}[z]$ be a monic polynomial of smallest degree such that $p(T) = \mathbf{0}$; let $r \in \mathbf{F}[z]$ be a monic polynomial of same degree and $r(T) = \mathbf{0}$. Then $(p-r)(T) = \mathbf{0}$ and also $\deg(p-r) < \deg p$.

We claim that $p - r = \mathbf{0}$. Suppose otherwise, for a contradiction, that $p - r \neq \mathbf{0}$. Then divide $p - r$ by the coefficient of the highest-order term in $p - r$ to get a monic polynomial $s \in \mathbf{F}[z]$, which satisfies $s(T) = \mathbf{0}$ and also $\deg s = \deg(p - r) < \deg p$, a contradiction. □

Definition 7.13 (Minimal polynomial)**.** Suppose V is finite-dimensional, $T \in \mathcal{L}(V)$. The ***minimal polynomial*** of T is the unique monic polynomial $p \in \mathbf{F}[z]$ of smallest degree such that $p(T) = \mathbf{0}$.

Theorem 7.14. Suppose V is finite-dimensional, $T \in \mathcal{L}(V)$.

(i) The zeros of the minimal polynomial of T are eigenvalues of T.

(ii) If V is a complex vector space, then the minimal polynomial of T has the form

$$(z - \lambda_1) \cdots (z - \lambda_m),$$

where λ_i are eigenvalues of T.

Proposition 7.15. Suppose V is finite-dimensional, $T \in \mathcal{L}(V)$, $q \in \mathbf{F}[z]$. Then $q(T) = 0$ if and only if q is a polynomial multiple of the minimal polynomial of T.

Proposition 7.16. Suppose V is finite-dimensional, $T \in \mathcal{L}(V)$, $U \leqslant V$ is invariant under T. Then the minimal polynomial of T is a polynomial multiple of the minimal polynomial of $T|_U$.

Corollary 7.17. Suppose V is finite-dimensional, $T \in \mathcal{L}(V)$. Then T is not invertible if and only if the constant term of the minimal polynomial of T is 0.

Eigenvalues on Odd-Dimensional Real Vector Spaces

The next result will be the key tool that we use to show that every operator on an odd-dimensional real vector space has an eigenvalue.

Proposition 7.18. Suppose V is finite-dimensional, over $\mathbf{R}$. Suppose also that $T \in \mathcal{L}(V)$ and $b, c \in \mathbf{R}$ with $b^2 < 4c$. Then $\dim \ker(T^2 + bT + cI)$ is an even number.

Proposition 7.19. Every operator on an odd-dimensional vector space has an eigenvalue.

§7.3 Upper-Triangular Matrices

Definition 7.20 (Matrix of operator)**.** Suppose $T \in \mathcal{L}(V)$. The matrix of T with respect to a basis $\mathcal{V} = \{v_1, \dots, v_n\}$ of V is the $n \times n$ matrix, whose entries a_{ij} are defined by

$$Tv_j = \sum_{i=1}^{n} a_{ij} v_i.$$

Notation. The notation $\mathcal{M}_{\mathcal{V}}(T)$ is used if the basis is not clear from the context.

Remark. Operators have square matrices.

Definition 7.21 (Diagonal of matrix)**.** The ***diagonal*** of a square matrix consists of the entries on the line from the upper left corner to the bottom right corner.

Definition 7.22 (Upper-triangular matrix)**.** A square matrix is called ***upper triangular*** if all the entries below the diagonal are 0.

Lemma 7.23 (Conditions for upper-triangular matrix)**.** Suppose $T \in \mathcal{L}(V)$, $\{v_1, \dots, v_n\}$ is a basis of V. Then the following are equivalent:

(i) The matrix with respect to $\{v_1, \dots, v_n\}$ is upper triangular.

(ii) $\operatorname{span}(v_1, \dots, v_k)$ is invariant under T for each $k = 1, \dots, n$.

(iii) $Tv_k \in \operatorname{span}(v_1, \dots, v_k)$ for each $k = 1, \dots, n$.

Lemma 7.24. Suppose $T \in \mathcal{L}(V)$, V has a basis with respect to which T has an upper-triangular matrix with diagonal matrix with diagonal entries $\lambda_1, \dots, \lambda_m$. Then

$$(T - \lambda_1 I) \cdots (T - \lambda_m I) = 0.$$

Proposition 7.25. Suppose $T \in \mathcal{L}(V)$ has an upper-triangular matrix with respect to some basis of V. Then the eigenvalues of T are precisely the entries on the diagonal of that upper-triangular matrix.

The following result gives a necessary and sufficient condition to have an upper-triangular matrix.

Lemma 7.26. Suppose V is finite-dimensional, $T \in \mathcal{L}(V)$. Then T has an upper-triangular matrix with respect to some basis of V if and only if the minimal polynomial equals $(z - \lambda_1) \cdots (z - \lambda_m)$ for some $\lambda_i \in \mathbf{F}$.

Theorem 7.27. Suppose V is finite-dimensional complex vector space, $T \in \mathcal{L}(V)$. Then T has an upper-triangular matrix with respect to some basis of V.

§7.4 Diagonalisable Operators

Definition 7.28 (Diagonal matrix)**.** A ***diagonal matrix*** is a square matrix that is 0 everywhere except possibly on the diagonal.

Remark. The entries on the diagonal are precisely the eigenvalue of the operator.

Definition 7.29 (Diagonalisable)**.** An operator on V is called ***diagonalisable*** if the operator has a diagonal matrix with respect to some basis of V.

Remark. Diagonalisation may require a different basis.

Definition 7.30 (Eigenspace)**.** Suppose $T \in \mathcal{L}(V)$, $\lambda \in \mathbf{F}$. The ***eigenspace*** of T corresponding to λ is the subspace of V defined by

$$V_\lambda \coloneqq \ker(T - \lambda I) = \{v \in V \mid Tv = \lambda v\}.$$

Remark. Hence V_λ is the set of all eigenvectors of T corresponding to λ, along with the $\mathbf{0}$ vector.

Proposition 7.31. Suppose $T \in \mathcal{L}(V)$, $\lambda_1, \ldots, \lambda_m$ are distinct eigenvalues of T. Then

$$V_{\lambda_1} + \cdots + V_{\lambda_m}$$

is a direct sum. Furthermore, if V is finite-dimensional, then

$$\dim V_{\lambda_1} + \cdots + \dim V_{\lambda_m} \leqslant \dim V.$$

Lemma 7.32 (Conditions equivalent to diagonalisability)**.** Suppose V is finite-dimensional, $T \in \mathcal{L}(V)$, $\lambda_1, \ldots, \lambda_m$ are distinct eigenvalues of T. Then the following are equivalent:

(i) T is diagonalisable.

(ii) V has a basis consisting of eigenvectors of T.

(iii) $V = V_{\lambda_1} \oplus \cdots \oplus V_{\lambda_m}$.

(iv) $\dim V = \dim V_{\lambda_1} + \cdots + \dim V_{\lambda_m}$.

Corollary 7.33. Suppose V is finite-dimensional, $T \in \mathcal{L}(V)$ has $\dim V$ distinct eigenvalues. Then T is diagonalisable.

Theorem 7.34. Suppose V is finite-dimensional, $T \in \mathcal{L}(V)$. Then T is diagonalisable if and only if the minimal polynomial of T equals $(z - \lambda_1) \cdots (z - \lambda_m)$ for distinct $\lambda_1, \ldots, \lambda_m \in \mathbf{F}$.

Corollary 7.35. Suppose $T \in \mathcal{L}(V)$ is diagonalisable, $U \leqslant V$ is invariant under T. Then $T|_U$ is a diagonalisable operator on U.

Definition 7.36 (Gershgorin disks)**.** Suppose $T \in \mathcal{L}(V)$, $\{v_1, \ldots, v_n\}$ is a basis of V. Let A denote the matrix of T with respect to this basis. A ***Gershgorin disk*** of T with respect to the basis $\{v_1, \ldots, v_n\}$ is a set of the form

$$\left\{ z \in \mathbf{F} \;\middle|\; |z - a_{ii}| \leqslant \sum_{j=1, j \neq i}^{n} |a_{ij}| \right\},$$

where $i = 1, \ldots, n$.

Theorem 7.37 (Gershgorin disk theorem)**.** Suppose $T \in \mathcal{L}(V)$, $\{v_1, \ldots, v_n\}$ is a basis of V. Then each eigenvalue of T is contained in some Gershgorin disk of T with respect to the basis $\{v_1, \ldots, v_n\}$.

§7.5 Commuting Operators

Exercises

Problem 7.1 ([Axl24] 5A Q1)**.** Suppose $T \in \mathcal{L}(V)$, $U \leqslant V$. Prove that

(i) if $U \subset \ker T$, then U is invariant under T;

(ii) if $\operatorname{im} T \subset U$, then U is invariant under T.

Solution.

(i)

(ii) Let $u \in U$. Then $Tu \in \operatorname{im} T \subset U$ so $Tu \in U$.

□

IV

Real Analysis

8 Real and Complex Number Systems

Learning Outcomes

In this chapter, we will

- define the supremum and infimum of a set;
- discuss the construction and properties of the real field $\mathbf{R}$;
- discuss the construction and properties of the complex field $\mathbf{C}$;
- discuss the construction and properties of the Euclidean space $\mathbf{R}^n$.

§8.1 Ordered Sets and Boundedness

Definitions

Let S be a set.

Definition 8.1 (Order)**.** An ***order*** on S is a binary relation, denoted by $<$, with the following properties:

(i) Trichotomy: $\forall x, y \in S$, one and only one of the following statements is true:

$$x < y, \quad x = y, \quad y < x.$$

(ii) Transitivity: $\forall x, y, z \in S$, if $x < y$ and $y < z$, then $x < z$.

We call $(S, <)$ an ***ordered set***.

Notation. $x \leqslant y$ indicates that $x < y$ or $x = y$, without specifying which of these two is to hold. In other words, $x \leqslant y$ is the negation of $x > y$.

Definition 8.2 (Boundedness)**.** Suppose S is an ordered set, and $E \subset S$.

(i) E is ***bounded above*** if there exists $\beta \in S$ such that $x \leqslant \beta$ for all $x \in E$; β is called an ***upper bound*** of E.

(ii) E is ***bounded below*** if there exists $\beta \in S$ such that $x \geqslant \beta$ for all $x \in E$; β is called a ***lower bound*** of E.

E is ***bounded*** in S if it is bounded above and below.

Definition 8.3 (Supremum)**.** Suppose S is an ordered set, $E \subset S$, and E is bounded above. We call $\alpha \in S$ the ***supremum*** of E, denoted by $\alpha = \sup E$, if it satisfies the following properties:

(i) α is an upper bound for E;

(ii) if $\beta < \alpha$ then β is not an upper bound of E, i.e. $\exists x \in S$ s.t. $x > \beta$ (least upper bound).

Definition 8.4 (Infimum)**.** We cal $\alpha \in S$ the ***infimum*** of E, denoted by $\alpha = \inf E$, if it satisfies the following properties:

(i) α is a lower bound for E;

(ii) if $\beta > \alpha$ then β is not a lower bound of E, i.e. $\exists x \in S$ s.t. $x < \beta$ (greatest lower bound).

Least-upper-bound Property

Definition 8.5. An ordered set S is said to have the ***least-upper-bound property*** (l.u.b.) if the following is true: if non-empty $E \subset S$ is bounded above, then $\sup E \in S$.

Similarly, S has the greatest-lower-bound property if the following is true: if non-empty $E \subset S$ is bounded below, then $\inf E \in S$.

Proposition 8.6. Suppose S is an ordered set. If S has the least-upper-bound property, then S has the greatest-lower-bound property.

Proof. Suppose S has the least-upper-bound property. Let non-empty $B \subset S$ be bounded below. We want to show that $\inf B \in S$.

Let $L \subset S$ be the set of all lower bounds of B; that is,

$$L = \{y \in S \mid y \leqslant x \forall x \in B\}.$$

Since B is bounded below, B has a lower bound, so $L \neq \emptyset$. Since every $x \in B$ is an upper bound of L, L is bounded above. By the least-upper-bound property of S, we have that $\sup L \in S$.

Claim. $\inf B = \sup L$.

To show that $\sup L = \inf B$ (greatest lower bound), we need to show that (i) $\sup L$ is a lower bound of B, (ii) and $\sup L$ is the greatest of the lower bounds.

(i) Suppose $\gamma < \sup L$, then γ is not an upper bound of L. Since B is the set of upper bounds of L, $\gamma \notin B$. Considering the contrapositive, if $\gamma \in B$, then $\gamma \geqslant \sup L$. Hence $\sup L$ is a lower bound of B, and thus $\sup L \in L$.

(ii) If $\sup L < \beta$ then $\beta \notin L$, since $\sup L$ is an upper bound of L. In other words, $\sup L$ is a lower bound of B, but β is not if $\beta > \sup L$. This means that $\sup L$ is the greatest of the lower bounds.

Hence $\inf B = \sup L \in S$. □

Corollary 8.7. If S has the greatest-lower-bound property, then it has the least-upper-bound property.

Hence S has the least-upper-bound property if and only if S has the greatest-lower-bound property.

Properties of Suprema and Infima

This section discusses some fundamental properties of the supremum that will be useful in this text. There is a corresponding set of properties of the infimum that the reader should formulate for himself.

Proposition 8.8 (Uniqueness of suprenum)**.** If E has a supremum, then it is unique.

Proof. We prove by contradiction. Suppose otherwise, for a contradiction, that E does not have a unique supremum; let α and β be suprema of E.

Since β is a supremum, it is an upper bound for E. Since α is a supremum, then it is the least upper bound and thus $\alpha \leqslant \beta$.

Similarly, since α is a supremum, it is an upper bound for E; since β is a supremum, it is a least upper bound and thus $\beta \leqslant \alpha$.

Since $\alpha \leqslant \beta$ and $\beta \leqslant \alpha$, we have that $\alpha = \beta$. □

The next result shows that a set with a supremum contains numbers arbitrarily close to its supremum.

Proposition 8.9 (Approximation property)**.** Let $S \subset \mathbf{R}$ be non-empty, $b = \sup S$. Then for every $a < b$ there exists $x \in S$ such that

$$a < x \leqslant b.$$

Proof. We first show $x \leqslant b$. Since $b = \sup S$ is an upper bound of S, $x \leqslant b$ for all $x \in S$.

We now show there exist $x \in S$ such that $a < x$. Suppose otherwise, for a contradiction, that $x \leqslant a$ for every $x \in S$. Then a would be an upper bound for S. But since $a < b$ and b is the supremum, this means a is smaller than the least upper bound, a contradiction. □

For the rest of this section, suppose S has the least-upper-bound property.

Proposition 8.10 (Additive property)**.** Given non-empty subsets $A, B \subset S$, let

$$C = \{x + y \mid x \in A, y \in B\}.$$

If each of A and B has a supremum, then C has a supremum, and

$$\sup C = \sup A + \sup B.$$

Proof. Let $a = \sup A$, $b = \sup B$. Let $z \in C$, then $z = x + y$ for some $x \in A$, $y \in B$. Then

$$z = x + y \leqslant a + b,$$

so $a + b$ is an upper bound for C. Since C is non-empty and bounded above, by the lub property of S, C has a supremum in S.

Let $c = \sup C$. To show that $a + b = c$, we need to show that (i) $a + b \geqslant c$, and (ii) $a + b \leqslant c$.

(i) Since c is the *least* upper bound for C, and $a + b$ is an upper bound for C, we must have that $c \leqslant a + b$.

(ii) Choose any $\varepsilon > 0$. By Proposition 8.9 there exist $x \in A$ and $y \in B$ such that

$$a - \varepsilon < x, \quad b - \varepsilon < y.$$

Adding these inequalities gives

$$a + b - 2\varepsilon < x + y \leqslant c.$$

Thus $a + b < c + 2\varepsilon$ for every $\varepsilon > 0$. Hence $a + b \leqslant c$.

□

Proposition 8.11 (Comparison property)**.** Let non-empty $A, B \subset S$ such that $a \leqslant b$ for every $a \in A$, $b \in B$. If B has a supremum, then A has a supremum, and

$$\sup A \leqslant \sup B.$$

Proof. Let $\beta = \sup B$. Since β is a supremum for B, then $b \leqslant \beta$ for all $b \in B$.

Let $a \in A$ and choose any $b \in B$. Since $a \leqslant b$ and $t \leqslant \beta$, $a \leqslant \beta$. Thus β is an upper bound for A.

Since A is non-empty and bounded above, by the lub property of S, A has a supremum in S; let $\alpha = \sup A$. Since β is an upper bound for A, and α is the *least* upper bound for A, we have that $\alpha \leqslant \beta$, as desired. □

Proposition 8.12. Let $B \subset S$ be non-empty and bounded below. Let $A = \{-b \mid b \in B\}$. Then A is non-empty and bounded above. Furthermore, $\inf B$ exists, and $\inf B = -\sup A$.

Proof. Since B is non-empty, so is A. Since B is bounded below, let β be a lower bound for B. Then $b \geqslant \beta$ for all $b \in B$, which implies $-b \leqslant -\beta$ for all $b \in B$. Hence $a \leqslant -\beta$ for all $a \in A$, so $-\beta$ is an upper bound for A.

Since A is non-empty and bounded above, by the lub property of S, A has a supremum. Then $a \leqslant \sup A$ for all $a \in A$, so $b \geqslant -\sup A$ for all $b \in B$. Thus $-\sup A$ is a lower bound for B.

Also, we saw before that if β is a lower bound for B then $-\beta$ is an upper bound for A. Then $-\beta \geqslant \sup A$ (since $\sup A$ is the least upper bound), so $\beta \leqslant -\sup A$. Therefore $-\sup A$ is the greatest lower bound of B. $\square$

§8.2 Real Numbers

Problems with Q

Q has some problems, the first of which being *algebraic incompleteness*: there exists equations with coefficients in **Q** but do not have solutions in **Q** (in fact **R** has this problem too, but **C** is algebraically complete, by the Fundamental Theorem of Algebra).

Lemma 8.13. $x^2 - 2 = 0$ has no solution in **Q**.

Proof. Suppose, for a contradiction, that $x^2 - 2 = 0$ has a solution $x = \frac{p}{q}$, $q \neq 0$. We also assume $\frac{p}{q}$ is in lowest terms; that is, p, q are coprime. Squaring both sides gives $\frac{p^2}{q^2} = 2$, or $p^2 = 2q^2$. Observe that p^2 is even, so p is even; let $p = 2m$ for some integer m. Then this implies $4m^2 = 2q^2$, or $2m^2 = q^2$. Similarly, q^2 is even so q is even.

Since p and q share a common factor of 2, we have reached a contradiction. □

The second problem is *analytic incompleteness*: there exists a sequence of rational numbers that approach a point that is not in **Q**; for example, the sequence

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

tends to the the irrational number $\sqrt{2}$.

Continuing from the above lemma,

Lemma 8.14. Let

$$A = \{p \in \mathbf{Q} \mid p > 0, p^2 < 2\},$$
$$B = \{p \in \mathbf{Q} \mid p > 0, p^2 > 2\}.$$

Then A contains no largest number, and B contains no smallest number.

Proof. Prove by construction. We associate with each rational $p > 0$ the number

$$q = p - \frac{p^2 - 2}{p + 2} = \frac{2p + 2}{p + 2}$$

and so

$$q^2 - 2 = \frac{2(p^2 - 2)}{(p + 2)^2}.$$

For any $p \in A$, $q > p$ and $q \in A$ since $q^2 < 2$, so A has no largest number.

For any $p \in B$, $q < p$ and $q \in B$ since $q^2 > 2$, so B has no smallest number. □

Proposition 8.15. **Q** does not have the least-upper-bound property.

Proof. In the previous result, note that B is the set of all upper bounds of A, and B does not have a smallest element. Hence $A \subset \mathbf{Q}$ is bounded above but A has no least upper bound in $\mathbf{Q}$. □

Real Field

The sole objective of this subsection is to prove the following result.

Theorem 8.16 (Existence of real field)**.** There exists an ordered field $\mathbf{R}$ that

(i) contains $\mathbf{Q}$ as a subfield, and

(ii) has the least-upper-bound property (also known as the completeness axiom).

Proof. We prove by contruction, as follows. □

We now want to construct $\mathbf{R}$ from $\mathbf{Q}$; one method to do so is using Dedekind cuts[1].

Definition 8.17 (Dedekind cut)**.** A ***Dedekind cut*** $\alpha \subset \mathbf{Q}$ satisfies the following properties:

(i) $\alpha \neq \emptyset$, $\alpha \neq \mathbf{Q}$;

(ii) if $p \in \alpha$, $q \in \mathbf{Q}$ and $q < p$, then $q \in \alpha$;

(iii) if $p \in \alpha$, then $p < r$ for some $r \in \alpha$.

Remark. Note that (iii) simply says that α has no largest member; (ii) implies two facts which will be used freely:

- If $p \in \alpha$ and $q \notin \alpha$, then $p < q$.
- If $r \notin \alpha$ and $r < s$, then $s \notin \alpha$.

Example

Let $r \in \mathbf{Q}$ and define

$$\alpha_r \coloneqq \{p \in \mathbf{Q} \mid p < r\}.$$

We now check that this is indeed a Dedekind cut.

(i) $p = 1 + r \notin \alpha_r$ thus $\alpha_r \neq \mathbf{Q}$. $p = r - 1 \in \alpha_r$ thus $\alpha_r \neq \emptyset$.

(ii) Suppose that $q \in \alpha_r$ and $q' < q$. Then $q' < q < r$ which implies that $q' < r$ thus $q' \in \alpha_r$.

(iii) Suppose that $q \in \alpha_r$. Consider $\dfrac{q+r}{2} \in \mathbf{Q}$ and $q < \dfrac{q+r}{2} < r$. Thus $\dfrac{q+r}{2} \in \alpha_r$.

This example shows that every rational r corresponds to a Dedekind cut α_r.

[1] proposed by German mathematician Richard Dedekind in 1872.

Example

$\sqrt[3]{2}$ is not rational, but it is real. $\sqrt[3]{2}$ corresponds to the cut

$$\alpha = \{p \in \mathbf{Q} \mid p^3 < 2\}.$$

(i) Trivial.

(ii) If $q < p$, by the monotonicity of the cubic function, this implies that $q^3 < p^3 < 2$ thus $q \in \alpha$.

(iii) If $p \in \alpha$, consider $\left(p + \frac{1}{n}\right)^3 < 2$.

Definition 8.18. The set of real numbers, denoted by $\mathbf{R}$, is the set of all Dedekind cuts:

$$\mathbf{R} \coloneqq \{\alpha \subset \mathbf{Q} \mid \alpha \text{ is a Dedekind cut}\}.$$

Proposition 8.19. $\mathbf{R}$ has an order, where $\alpha < \beta$ is defined to mean that $\alpha \subsetneq \beta$.

Proof. Simply check if this is a valid order (by checking for trichotomy and transitivity). □

Proposition 8.20. The ordered set $\mathbf{R}$ has the least-upper-bound property.

Proof. Let non-empty $A \subset \mathbf{R}$ be bounded above. Let $\beta \in \mathbf{R}$ be an upper bound of A. We want to show that A has a supremum in $\mathbf{R}$.

Let

$$\gamma = \bigcup_{\alpha \in A} \alpha.$$

Then $p \in \gamma$ if and only if $p \in \alpha$ for some $\alpha \in A$.

Claim. $\gamma \in \mathbf{R}$ and $\gamma = \sup A$.

We first prove that $\gamma \in \mathbf{R}$ by checking that it is a Dedekind cut:

(i) Since $A \neq \emptyset$, there exists $\alpha_0 \in A$. Since $\alpha_0 \in \mathbf{R}$, it is a Dedekind cut so $\alpha_0 \neq \emptyset$. Since $\alpha_0 \subset \gamma$, $\gamma \neq \emptyset$.

Since $\alpha \subset \beta$ for every $\alpha \in A$, the union of $\alpha \in A$ must be a subset of β; thus $\gamma \subset \beta$. Hence $\gamma \neq \mathbf{Q}$.

(ii) Let $p \in \gamma$. Then $p \in \alpha_1$ for some $\alpha_1 \in A$. If $q < p$, then $q \in \alpha_1$ (since α_1 is a Dedekind cut). Hence $q \in \gamma$.

(iii) If $r \in \alpha_1$ is so chosen that $r > p$, we see that $r \in \gamma$ (since $\alpha_1 \subset \gamma$).

Next we prove that $\gamma = \sup A$, by checking that (i) γ is an upper bound of A, (ii) γ is the *least* of the upper bounds.

(i) It is clear that $\alpha \leqslant \gamma$ for every $\alpha \in A$.

(ii) Suppose $\delta < \gamma$. Then there exists $s \in \gamma$ such that $s \notin \delta$. Since $s \in \gamma$, $s \in \alpha$ for some $\alpha \in A$. Hence $\delta < \alpha$, so δ is not an upper bound of A.

□

Remark. The l.u.b. property of $\mathbf{R}$ is also known as the *completeness axiom* of $\mathbf{R}$.

We now define operations on $\mathbf{R}$.

Definition 8.21 (Addition)**.** Given $\alpha, \beta \in \mathbf{R}$,

$$\alpha + \beta \coloneqq \{r \in \mathbf{Q} \mid r = a + b, a \in \alpha, b \in \beta\}.$$

We first check if the above definition makes sense. We want to show that addition on $\mathbf{R}$ is closed: for all $\alpha, \beta \in \mathbf{R}$, $\alpha + \beta \in \mathbf{R}$.

Proof. We check that $\alpha + \beta$ is a Dedekind cut:

(i) Since $\alpha \neq \emptyset$ and $\beta \neq \emptyset$, there exists $a \in \alpha$ and $b \in \beta$. Hence $r = a + b \in \alpha + \beta$ so $\alpha + \beta \neq \emptyset$.

Since $\alpha \neq \mathbf{Q}$ and $\beta \neq \mathbf{Q}$, there exist $c \neq \alpha$ and $d \neq \beta$. Thus $r' = c + d > a + b$ for any $a \in \alpha, b \in \beta$, so $r' \notin \alpha + \beta$. Hence $\alpha + \beta \neq \mathbf{Q}$.

(ii) Suppose that $r \in \alpha + \beta$ and $r' < r$. We want to show that $r' \in \alpha + \beta$.

$r = a + b$ for some $a \in \alpha, b \in \beta$. Then $r' - a < b$. Since $\beta \in \mathbf{R}$, $r' - a \in \beta$ so $r' - a = b_1$ for some $b_1 \in \beta$. Hence $r' = a + b_1 \in \alpha + \beta$.

(iii) Suppose $r \in \alpha + \beta$, so $r = a + b$ for some $a \in \alpha, b \in \beta$. Since α, β are Dedekind cuts, there exist $a' \in \alpha, b' \in \beta$ with $a < a'$ and $b < b'$. Then $r = a + b < a' + b' \in \alpha + \beta$. We define $r' = a' + b' \in \alpha + \beta$ with $r < r'$.

□

Proposition 8.22.

(i) Addition on $\mathbf{R}$ is commutative: $\alpha + \beta = \beta + \alpha$ for all $\alpha, \beta \in \mathbf{R}$.

(ii) Addition on $\mathbf{R}$ is associative: $\alpha + (\beta + \gamma) = (\alpha + \beta) + \gamma$ for all $\alpha, \beta, \gamma \in \mathbf{R}$.

(iii) Additive identity: Define $0^* \coloneqq \{p \in \mathbf{Q} \mid p < 0\}$. Then $\alpha + 0^* = \alpha$ for all $\alpha \in \mathbf{R}$.

(iv) Additive inverse: Fix $\alpha \in \mathbf{R}$, define $\beta = \{p \in \mathbf{Q} \mid \exists r > 0,\ -p - r \notin \alpha\}$. Then $\alpha + \beta = 0^*$.

Remark. Recall that to prove that two sets are equal, show double inclusion.

Proof.

(i) We need to show that $\alpha + \beta \subset \beta + \alpha$ and $\beta + \alpha \subset \alpha + \beta$.

Let $r \in \alpha + \beta$. Then $r = a + b$ for $a \in \alpha$ and $b \in \beta$. Thus $r = b + a$ since $+$ is commutative on $\mathbf{Q}$. Hence $r \in \beta + \alpha$. Therefore $\alpha + \beta \subset \beta + \alpha$.

Similarly, $\beta + \alpha \subset \alpha + \beta$.

Therefore $\alpha + \beta = \beta + \alpha$.

(ii) Let $r \in \alpha + (\beta + \gamma)$. Then $r = a + (b + c)$ where $a \in \alpha, b \in \beta, c \in \gamma$. Thus $r = (a + b) + c$ by associativity of $+$ on $\mathbf{Q}$. Therefore $r \in (\alpha + \beta) + \gamma$, hence $\alpha + (\beta + \gamma) \subset (\alpha + \beta) + \gamma$.

Similarly, $(\alpha + \beta) + \gamma \subset \alpha + (\beta + \gamma)$.

(iii) It is clear that 0^* is a Dedekind cut.

Let $r \in \alpha + 0^*$. Then $r = a + p$ for some $a \in \alpha, p \in 0^*$. Thus $r = a + p < a + 0 = a$ so $r \in \alpha$. Hence $\alpha + 0^* \subset \alpha$.

Let $r \in \alpha$. Then there exists $r' \in \alpha$ where $r' > r$. Thus $r - r' < 0$, so $r - r' \in 0^*$. We see that $r = r' + (r - r')$ where $r' \in \alpha$, $r - r' \in 0^*$. Hence $\alpha \subset \alpha + 0^*$.

(iv) Fix some $\alpha \in \mathbf{R}$. We first show that β is a Dedekind cut.

(i) Let $s \notin \alpha$, let $p = -s - 1$. Then $-p - 1 \notin \alpha$. Hence $p \in \beta$, so $\beta \neq \emptyset$.
Let $q \in \alpha$. Then $-q \notin \beta$ so $\beta \neq \mathbf{Q}$.

(ii) Let $p \in \beta$. Then there exists $r > 0$ such that $-p - r \notin \alpha$. If $q < p$, then $-q - r > -p - r$ so $-q - r \notin \alpha$. Hence $q \in \beta$.

(iii) Let $t = p + \frac{r}{2}$. Then $t > p$, and $-t - \frac{r}{2} = -p - r \notin \alpha$. Hence $t \in \beta$.

Let $r \in \alpha$, $s \in \beta$. Then $-s \notin \alpha$. This implies $r < -s$ (since α is closed downwards) so $r + s < 0$. Hence $\alpha + \beta \subset 0^*$.

To prove the opposite inclusion, let $v \in 0^*$, and let $w = -\frac{v}{2}$. Then $w > 0$. By the Archimedean property on $\mathbf{Q}$, there exists $n \in \mathbf{N}$ such that $nw \in \alpha$ but $(n+1)w \notin \alpha$. Let $p = -(n+2)w$. Then

$$-p - w = (n+2)w - w = (n+1)w \notin \alpha$$

so $p \in \beta$. Since $v = nw + p$ where $nw \in \alpha$, $p \in \beta$, $v \in \alpha + \beta$. Hence $0^* \subset \alpha + \beta$.

□

Notation. β is denoted by the more familiar notation $-\alpha$.

Proposition 8.23. If $\alpha, \beta, \gamma \in \mathbf{R}$ and $\beta < \gamma$, then $\alpha + \beta < \alpha + \gamma$.

Proof. □

We say that a Dedekind cut α is *positive* if $0 \in \alpha$, and *negative* if $0 \notin \alpha$. If α is neither positive nor negative, then $\alpha = 0^*$.

Multiplication is a little more bothersome than addition in the present context, since products of negative rationals are positive. For this reason we confine ourselves first to $\mathbf{R}^+$ (the set of all $\alpha \in \mathbf{R}$ with $\alpha > 0^*$).

Definition 8.24. For all $\alpha, \beta \in \mathbf{R}^+$, we define multiplication as

$$\alpha\beta \coloneqq \{p \in \mathbf{Q} \mid p \leqslant rs,\ r \in \alpha, s \in \beta,\ r, s > 0\}.$$

We also define $1^* \coloneqq \{q \in \mathbf{Q} \mid q < 1\}$.

As again, check if the above definition makes sense. We want to show that multiplication on $\mathbf{R}^+$ is closed: for all $\alpha, \beta \in \mathbf{R}$, $\alpha\beta \in \mathbf{R}$.

Proof. Check that $\alpha\beta$ is a Dedekind cut.

(i) $\alpha \neq \emptyset$ means there exists $r \in \alpha, r > 0$. Similarly, $\beta \neq \emptyset$ means there exists $s \in \beta, s > 0$. Then $rs \in \mathbf{Q}$ and $rs \leqslant rs$, so $rs \in \alpha\beta$. Hence $\alpha\beta \neq \emptyset$.

$\alpha \neq \mathbf{Q}$ means there exists $r' \notin \alpha$ such that $r' > r$ for all $r \in \alpha$. Similarly $\beta \neq \mathbf{Q}$ means there exists $s' \in \beta$ such that $s' > s$ for all $s \in \beta$. Then $r's' > rs$ for all $r \in \alpha, s \in \beta$, so $r's' \notin \alpha\beta$. Hence $\alpha\beta \neq \mathbf{Q}$.

(ii) Let $p \in \alpha\beta$. Then $p \leqslant ab$ for some $a \in \alpha, b \in \beta$, $a, b > 0$.

If $q < p$, then $q < p \leqslant ab$ so $q \in \alpha\beta$.

(iii) Let $p \in \alpha\beta$. Then $p \leqslant ab$ for some $a \in \alpha, b \in \beta$, $a, b > 0$. Pick $a' \in \alpha$ and $b' \in \beta$ with $a' > a$ and $b' > b$. Form $a'b' > ab \geqslant p$, $a'b' \leqslant a'b'$ means $a'b' \in \alpha \cdot \beta$.

□

We now complete the definition of multiplication by setting $\alpha 0^* = 0^* = 0^*\alpha$, and by setting

$$\alpha\beta = \begin{cases} (-\alpha)(-\beta) & a < 0^*, \beta < 0^*, \\ -[(-\alpha)\beta] & a < 0^*, \beta > 0^*, \\ -[\alpha(-\beta)] & \alpha > 0^*, \beta < 0^*. \end{cases}$$

where we make negative numbers positive, multiply, and then negate them as needed.

Proposition 8.25.

(i) Multiplication on $\mathbf{R}$ is commutative: $\alpha\beta = \beta\alpha$ for all $\alpha, \beta \in \mathbf{R}$.

(ii) Multiplication on $\mathbf{R}$ is associative: $(\alpha\beta)\gamma = \alpha(\beta\gamma)$ for all $\alpha, \beta, \gamma \in \mathbf{R}$.

(iii) Multiplicative identity: $1\alpha = \alpha$ for all $\alpha \in \mathbf{R}$.

(iv) Multiplicative inverse: If $\alpha \in \mathbf{R}$, $\alpha \neq 0^*$, then there exists $\beta \in \mathbf{R}$ such that $\alpha\beta = 1^*$.

We associate each $r \in \mathbf{Q}$ with the set

$$r^* = \{p \in \mathbf{Q} \mid p < r\}.$$

It is obvious that each r^* is a cut; that is, $r^* \in \mathbf{R}$.

Proposition 8.26. The replacement of $r \in \mathbf{Q}$ by the corresponding "rational cuts" $r^* \in \mathbf{R}$ preserves sums, products, and order. That is, for all $r^*, s^* \in \mathbf{R}$,

(i) $r^* + s^* = (r + s)^*$;

(ii) $r^* s^* = (rs)^*$;

(iii) $r^* < s^*$ if and only if $r < s$.

Proof.

(i) Let $p \in r^* + s^*$. Then $p = u + v$ for some $u \in r^*$, $v \in s^*$, where $u < r$, $v < s$. Then $p < r + s$. Hence $p \in (r + s)^*$, so $r^* + s^* \subset (r + s)^*$.

Let $p \in (r + s)^*$. Then $p < r + s$. Let $t = \frac{(r+s)-p}{2}$, and let

$$r' = r - t, \quad s' = s - t.$$

Since $t > 0$, $r' < r$ so $r' \in r^*$; $s' < s$ so $s' \in s^*$. Then $p = r' + s'$, so $p \in r^* + s^*$. Hence $(r + s)^* \subset r^* + s^*$.

(ii)

(iii) Suppose $r < s$. Then $r \in s^*$, but $r \notin r^*$. Hence $r^* < s^*$.

Conversely, suppose $r^* < s^*$. Then there exists $p \in s^*$ such that $p \in r^*$. Hence $r \leqslant p < s$, so $r < s$.

□

This shows that the ordered field $\mathbf{Q}$ is isomorphic to the ordered field $\mathbf{Q}^* = \{q^* \mid q \in \mathbf{Q}\}$ whose elements are rational cuts. It is this identification of $\mathbf{Q}$ with $\mathbf{Q}^*$ which allows us to regard $\mathbf{Q}$ as a subfield of $\mathbf{R}$.

Remark. In fact, $\mathbf{R}$ is the only ordered field with the l.u.b. property. Hence any other ordered field with the l.u.b. property is isomorphic to $\mathbf{R}$.

Properties of $\mathbf{R}$

Proposition 8.27 ($\mathbf{R}$ is archimedian)**.** For any $x \in \mathbf{R}^+$, $y \in \mathbf{R}$, there exists $n \in \mathbf{N}$ such that

$$nx > y.$$

Proof. Suppose, for a contradiction, that $nx \leqslant y$ for all $n \in \mathbf{N}$. Then y is an upper bound of the set

$$A = \{nx \mid n \in \mathbf{N}\}.$$

Since $A \subset R$ is non-empty and bounded above, by the l.u.b. property of $\mathbf{R}$, A has a supremum in $\mathbf{R}$, say $\alpha = \sup A$.

Consider $\alpha - x$. Since $\alpha - x < \alpha = \sup A$, $\alpha - x$ is not an upper bound of A. Then $\alpha - x \leqslant n_0 x$ for some $n_0 \in \mathbf{N}$; rearranging gives $\alpha \leqslant (n_0 + 1)x$. This implies that α is not an upper bound of A, which contradicts the fact that α is the supremum of A. □

Corollary 8.28. Let $\varepsilon > 0$. Then there exists $n \in \mathbf{N}$ such that $0 < \frac{1}{n} < \varepsilon$.

Proof. Take $x = \varepsilon$ and $y = 1$. □

Proposition 8.29 ($\mathbf{Q}$ is dense in $\mathbf{R}$)**.** For any $x, y \in \mathbf{R}$ with $x < y$, there exists $p \in \mathbf{Q}$ such that

$$x < p < y.$$

Proof. We prove by construction; that is, construct the required p from the given x and y.

Since $x < y$, we have $y - x > 0$. By the archimedian property, there exists $n \in \mathbf{N}$ such that

$$\frac{1}{n} < y - x.$$

Consider the set comprising multiples of $\frac{1}{n}$. Since this set is unbounded, choose the first multiple $m \in \mathbf{N}$ such that $\frac{m}{n} > x$.

We now claim that $\frac{m}{n} < y$. If not, then

$$\frac{m-1}{n} < x \quad \text{and} \quad \frac{m}{n} > y,$$

where the first inequality follows from the minimality of m. But these two statements combined imply that $\frac{1}{n} > y - x$, a contradiction. □

Proposition 8.30 ($\mathbf{R}$ is closed under taking roots)**.** For every $x \in \mathbf{R}^+$ and every $n \in \mathbf{N}$, there exists a unique $y \in \mathbf{R}^+$ so that $y^n = x$.

Proof. The uniqueness of such y is clear, since $0 < y_1 < y_2$ implies ${y_1}^n < {y_2}^n$.

Claim. $y = \sup E$, where

$$E = \{t \in \mathbf{R}^+ \mid t^n < x\}.$$

We first show that E has a supremum, by showing that it is (i) non-empty, and (ii) bounded above:

(i) Let $t = \frac{x}{1-x}$. Then $0 \leqslant t < 1$, so $t^n \leqslant t < x$. Hence $t \in E$, which implies $E \neq \emptyset$.

(ii) We claim that $1 + x$ is an upper bound for E. If $t > 1 + x$ then $t^n \geqslant t > x$, so that $t \notin E$. Hence $1 + x$ is an upper bound of E.

Hence E has a supremum; let $y = \sup E$.

To prove that $y^n = x$, we show that the inequalities (i) $y^n < x$ and (ii) $y^n > x$ lead to a contradiction. Consider the identity $b^n - a^n = (b - a)\left(b^{n-1} + b^{n-2}a + \cdots + a^{n-1}\right)$, which yields the inequality

$$b^n - a^n < (b - a)nb^{n-1}$$

when $0 < a < b$.

(i) Suppose $y^n < x$. Choose h so that $0 < h < 1$ and

$$h < \frac{x - y^n}{n(y + 1)^{n-1}}.$$

let $a = y$, $b = y + h$. Then

$$(y + h)^n - y^n < hn(y + h)^{n-1} < hn(y + 1)^{n-1} < x - y^n.$$

Thus $(y + h)^n < x$, and $y + h \in E$. Since $y + h > y$, this contradicts the fact that y is an upper bound of E.

(ii) Suppose $y^n > x$. Let

$$k = \frac{y^n - x}{ny^{n-1}}.$$

Then $0 < k < y$. If $t \geqslant y - k$, we conclude that

$$y^n - t^n \leqslant y^n - (y - k)^n < kny^{n-1} = y^n - x.$$

Thus $t^n > x$, and $t \notin E$. It follows that $y - k$ is an upper bound of E. But $y - k < y$, which contradicts the fact that y is the *least* upper bound of E.

Hence $y^n = x$, and the proof is complete. □

Notation. y is denoted by $\sqrt[n]{x}$ or $x^{\frac{1}{n}}$.

Corollary 8.31. If $a, b \in \mathbf{R}^+$ and $n \in \mathbf{N}$, then

$$(ab)^{\frac{1}{n}} = a^{\frac{1}{n}} b^{\frac{1}{n}}.$$

Proof. Let $\alpha = a^{\frac{1}{n}}$, $\beta = b^{\frac{1}{n}}$. Then

$$ab = \alpha^n \beta^n = (\alpha\beta)^n$$

since multiplication is commutative. The uniqueness assertion of the previous result shows that

$$(ab)^{\frac{1}{n}} = \alpha\beta = a^{\frac{1}{n}} b^{\frac{1}{n}}.$$

□

The next result shows that real numbers can be approximated to any desired degree of accuracy by rational numbers with finite decimal representations.

Proposition 8.32. Assume $x \geqslant 0$. Then for every integer $n \geqslant 1$ there exists a finite decimal $r_n = a_0.a_1a_2\cdots a_n$ such that

$$r_n \leqslant x < r_n + \frac{1}{10^n}.$$

Proof. We prove by construction; that is, we construct the required finite decimal from x.

Let

$$S = \{k \in \mathbf{Z} \mid k \leqslant x\}.$$

S is non-empty (since $0 \in S$), and S is bounded above by x. Hence by the lub property of $\mathbf{R}$, S has a supremum in $\mathbf{R}$, say $a_0 = \sup S$. It is easily verified that $a_0 \in S$, so a_0 is a non-negative integer. We call a_0 the *greatest integer* in x, and write $a_0 = \lfloor x \rfloor$. Clearly we have

$$a_0 \leqslant x < a_0 + 1.$$

Now let $a_1 = \lfloor 10(x - a_0) \rfloor$. Since $0 \leqslant 10(x - a_0) < 10$, we have $0 \leqslant a_1 \leqslant 9$ and

$$a_1 \leqslant 10x - 10a_0 < a_1 + 1.$$

In other words, a_1 is the largest integer satisfying the inequalities

$$a_0 + \frac{a_1}{10} \leqslant x < a_0 + \frac{a_1 + 1}{10}.$$

More generally, having chosen $a_1, \ldots, a_{n-1}$ with $0 \leqslant a_i \leqslant 9$, let a_n be the largest integer satisfying the inequalities

$$a_0 + \frac{a_1}{10} + \cdots + \frac{a_n}{10^n} \leqslant a_0 + \frac{a_1}{10} + \cdots + \frac{a_n + 1}{10^n}.$$

Then $0 \leqslant a_n \leqslant 9$ and we have

$$r_n \leqslant x < r_n + \frac{1}{10^n},$$

where $r_n = a_0.a_1a_2\cdots a_n$.

It is easy to verify that

$$x = \sup\{r_1, r_2, \ldots\}.$$

□

Extended Real Number System

Definition 8.33 (Extended real number system)**.** The ***extended real number system*** is defined to be the union

$$\overline{\mathbf{R}} := \mathbf{R} \cup \{-\infty, +\infty\},$$

where we preserve the original order in $\mathbf{R}$, and define $-\infty < x < +\infty$ for all $x \in \mathbf{R}$.

Defining $\overline{\mathbf{R}}$ is convenient since the following result holds.

Proposition 8.34. Any non-empty $E \subset \overline{\mathbf{R}}$ has a supremum and infimum in $\overline{\mathbf{R}}$.

Proof. If E is bounded above in $\mathbf{R}$, then by the l.u.b. property of $\mathbf{R}$, it has a supremum in $\mathbf{R} \subset \overline{\mathbf{R}}$. If E is not bounded above in $\mathbf{R}$, then $\sup E = +\infty \in \overline{\mathbf{R}}$.

Exactly the same remarks apply to lower bounds. □

$\overline{\mathbf{R}}$ does not form a field, but it is customary to make the following conventions for arithmetic on $\overline{\mathbf{R}}$:

(i) If $x \in \mathbf{R}$ then

$$x + \infty = +\infty, \quad x - \infty = -\infty, \quad \frac{x}{+\infty} = \frac{x}{-\infty} = 0.$$

(ii) If $x > 0$ then

$$x \cdot (+\infty) = +\infty, \quad x \cdot (-\infty) = -\infty.$$

If $x < 0$ then

$$x \cdot (+\infty) = -\infty, \quad x \cdot (-\infty) = +\infty.$$

When it is desired to make the distinction between real numbers on the one hand and the symbols $+\infty$ and $-\infty$ on the other quite explicit, the former are called *finite*.

§8.3 Complex Field

Consider the Cartesian product $\mathbf{R}^2$. A *complex number* is an ordered pair $(a, b) \in \mathbf{R}^2$.

Proposition 8.35. Let $x = (a, b)$, $y = (c, d)$ be two complex numbers. We write $x = y$ if and only if $a = c$ and $b = d$. $\mathbf{R}^2$, with addition and multiplication defined as

$$\begin{aligned} x + y &= (a + c, b + d) \\ xy &= (ac - bd, ad + bc) \end{aligned}$$

is a field. Note that the additive identity is $(0, 0)$, and multiplicative identity is $(1, 0)$. We call this structure $\mathbf{C}$, the ***complex field***.

Proof. Check the field axioms. □

The next result shows that the complex numbers of the form $(a, 0)$ have the same arithmetic properties as the corresponding real numbers a. We can therefore identify $(a, 0) \in \mathbf{C}$ with $a \in \mathbf{R}$. This identification implies that $\mathbf{R}$ is a subfield of $\mathbf{C}$.

Proposition 8.36. For any $a, b \in \mathbf{R}$, we have

$$\begin{aligned} (a, 0) + (b, 0) &= (a + b, 0), \\ (a, 0)(b, 0) &= (ab, 0). \end{aligned}$$

Proof. Exercise. □

You may have noticed that we have defined the complex numbers without referring to the mysterious square root of -1. We now show that the notation (a, b) is equivalent to the more customary $a + bi$.

Definition 8.37 (Imaginary number)**.** $i = (0, 1)$.

Proposition 8.38. $i^2 = -1$.

Proof.

$$i^2 = (0, 1)(0, 1) = (-1, 0) = -1.$$

□

Proposition 8.39. For $a, b \in \mathbf{R}$, $(a, b) = a + bi$.

Proof.

$$\begin{aligned} a + bi &= (a, 0) + (b, 0)(0, 1) \\ &= (a, 0) + (0, b) \\ &= (a, b). \end{aligned}$$

□

Definition 8.40. For $a, b \in \mathbf{R}$, $z = a+bi$, we call a and b the *real part* and *imaginary part* of z respectively, denoted by $a = \operatorname{Re}(z)$, $b = \operatorname{Im}(z)$; $\overline{z} = a - bi$ is called the *conjugate* of z.

Proposition 8.41. For $z, w \in \mathbf{C}$,

(i) $\overline{z + w} = \overline{z} + \overline{w}$

(ii) $\overline{zw} = \overline{z}\,\overline{w}$

(iii) $z + \overline{z} = 2\operatorname{Re}(z)$, $z - \overline{z} = 2i\operatorname{Im}(z)$

(iv) $z\overline{z} \in \mathbf{R}$ and $z\overline{z} \geqslant 0$

Definition 8.42. For $z \in \mathbf{C}$, tbe *absolute value* of z is defined as

$$|z| \coloneqq (z\overline{z})^{\frac{1}{2}}.$$

Proposition 8.43. For $z, w \in \mathbf{C}$,

(i) $|z| \geqslant 0$

(ii) $|\overline{z}| = |z|$

(iii) $|zw| = |z||w|$

(iv) $|\operatorname{Re}(z)| \leqslant |z|$

Proof.

(i) The square root is non-negative, by definition.

(ii) The conjugate of $\overline{z}$ is z, and the rest follows by the definition of absolute value.

(iii) Let $z = a + bi$, $w = c + di$ where $a, b, c, d \in \mathbf{R}$. Then

$$\begin{aligned} |zw|^2 &= (ac - bd)^2 + (ad - bc)^2 \\ &= (a^2 + b^2)(c^2 + d^2) \\ &= |z|^2|w|^2 = \big(|z||w|\big)^2 \end{aligned}$$

and the desired result follows by taking square roots on both sides.

(iv) Let $z = a + bi$. Note that $a^2 \leqslant a^2 + b^2$, hence

$$|\operatorname{Re}(z)| = |a| = \sqrt{a^2} \leqslant \sqrt{a^2 + b^2} = |z|.$$

□

Theorem 8.44 (Triangle inequality)**.** For $z, w \in \mathbf{C}$,

$$|z + w| \leqslant |z| + |w|. \tag{8.1}$$

Proof. Let $z, w \in \mathbf{C}$. Note that the conjugate of $z\overline{w}$ is $\overline{z}w$, so $z\overline{w} + \overline{z}w = 2\,\mathrm{Re}(z\overline{w})$. Hence

$$\begin{aligned}
|z + w|^2 &= (z + w)(\overline{z + w}) \\
&= (z + w)(\overline{z} + \overline{w}) \\
&= z\overline{z} + z\overline{w} + \overline{z}w + w\overline{w} \\
&= |z|^2 + 2\,\mathrm{Re}(z\overline{w}) + |w|^2 \\
&\leqslant |z|^2 + 2|z\overline{w}| + |w|^2 \\
&= |z|^2 + 2|z||w| + |w|^2 \\
&= \big(|z| + |w|\big)^2
\end{aligned}$$

and taking square roots yields the desired result. □

Corollary 8.45 (Generalised triangle inequality)**.** For $z_1, \ldots, z_n \in \mathbf{C}$,

$$|z_1 + \cdots + z_n| \leqslant |z_1| + \cdots + |z_n|.$$

Proof. By the triangle inequality, $|z_1 + z_2| \leqslant |z_1| + |z_2|$. Assume the statement holds for $n - 1$. Then

$$|z_1 + \cdots + z_{n-1} + z_n| \leqslant |z_1 + \cdots + z_{n-1}| + |z_n| \leqslant |z_1| + \cdots + |z_n|,$$

which establishes the claim by induction. □

Theorem 8.46 (Cauchy–Schwarz inequality)**.** If $a_1, \ldots, a_n, b_1, \ldots, b_n \in \mathbf{C}$, then

$$\left|\sum_{i=1}^{n} a_i \overline{b_i}\right|^2 \leqslant \sum_{i=1}^{n} |a_i|^2 \sum_{i=1}^{n} |b_i|^2. \tag{8.2}$$

Proof. Let

$$A = \sum_{i=1}^{n} |a_i|^2, \quad B = \sum_{i=1}^{n} |b_i|^2, \quad C = \sum_{i=1}^{n} a_i \overline{b_i}.$$

If $B = 0$, then $b_1 = \cdots = b_n = 0$, and the conclusion is trivial. Now assume that $B > 0$. Then consider

the sum

$$\begin{aligned}\sum_{i=1}^{n}|Ba_i - Cb_i|^2 &= \sum_{i=1}^{n}(Ba_i - Cb_i)(\overline{Ba_i - Cb_i}) \\ &= \sum_{i=1}^{n}(Ba_i - Cb_i)(B\overline{a_i} - \overline{Cb_i}) \\ &= B^2\sum_{i=1}^{n}|a_i|^2 - B\overline{C}\sum_{i=1}^{n}a_i\overline{b_j} - BC\sum_{i=1}^{n}\overline{a_i}b_i + |C|^2\sum_{i=1}^{n}|b_i|^2 \\ &= B^2A - B|C|^2 \\ &= B(AB - |C|^2).\end{aligned}$$

Since each term $\sum_{i=1}^{n}|Ba_i - Cb_i|^2$ is non-negative, we have that $\sum_{i=1}^{n}|Ba_i - Cb_i|^2 \geqslant 0$, and so

$$B(AB - |C|^2) \geqslant 0.$$

Since $B > 0$, it follows that $AB - |C|^2 \geqslant 0$. This is the desired inequality.

(when does equality hold?) □

Define

$$\mathbf{C}^n = \{(z_1, \ldots, z_n) \mid z_i \in \mathbf{C}\}.$$

We can define an inner product on $\mathbf{C}^n$: for $\mathbf{a}, \mathbf{b} \in \mathbf{C}^n$,

$$\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i=1}^{n} a_i\overline{b_i}.$$

We can also define the norm of $\mathbf{a} \in \mathbf{C}^n$:

$$|\mathbf{a}| = \langle \mathbf{a}, \mathbf{a} \rangle^{\frac{1}{2}}.$$

§8.4 Euclidean Space

For $n \in \mathbf{N}$, define

$$\mathbf{R}^n \coloneqq \{(x_1, \ldots, x_n) \mid x_i \in \mathbf{R}\}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$, x_i's are called the coordinates of $\mathbf{x}$. The elements of $\mathbf{R}^n$ are called *points*, or *vectors*.

Proposition 8.47. Let $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{y} = (y_1, \ldots, y_n)$. $\mathbf{R}^n$, with addition and scalar multiplication defined as

$$\begin{aligned} \mathbf{x} + \mathbf{y} &= (x_1 + y_1, \ldots, x_n + y_n), \\ \alpha\mathbf{x} &= (\alpha x_1, \ldots, \alpha x_n). \end{aligned}$$

for $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$, $\alpha \in \mathbf{R}$, is a vector space over $\mathbf{R}$. Note that the zero element of $\mathbf{R}^n$ is $\mathbf{0} = (0, \ldots, 0)$.

Proof. These two operations satisfy the commutative, associatives, and distributive laws (the proof is trivial, in view of the analagous laws for the real numbers). □

We define the *inner product* of $\mathbf{x}$ and $\mathbf{y}$ by

$$\mathbf{x} \cdot \mathbf{y} \coloneqq \sum_{i=1}^{n} x_i y_i,$$

and the *norm* of $\mathbf{x}$ by

$$\|\mathbf{x}\| \coloneqq (\mathbf{x} \cdot \mathbf{x})^{\frac{1}{2}} = \left(\sum_{i=1}^{n} {x_i}^2 \right)^{\frac{1}{2}}.$$

The structure now defined (the vector space $\mathbf{R}^n$ with the above inner product and norm) is called the ***Euclidean n-space***.

Proposition 8.48. Suppose $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{R}^n$, $\alpha \in \mathbf{R}$. Then

(i) $\|\mathbf{x}\| \geqslant 0$

(ii) $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$

(iii) $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$

(iv) $\|\mathbf{x} \cdot \mathbf{y}\| \leqslant \|\mathbf{x}\|\|\mathbf{y}\|$

(v) $\|\mathbf{x} + \mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$

(vi) $\|\mathbf{x} - \mathbf{z}\| \leqslant \|\mathbf{x} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{z}\|$

Proof.

(i) Obvious from definition.

(ii)

$$\begin{aligned}\|\mathbf{x}\| = 0 &\iff \left(\sum_{i=1}^{n} {x_i}^2\right)^{\frac{1}{2}} = 0 \\ &\iff \sum_{i=1}^{n} {x_i}^2 = 0 \\ &\iff x_1 = \cdots = x_n = 0 \\ &\iff \mathbf{x} = (0, \ldots, 0) = \mathbf{0}\end{aligned}$$

since ${x_i}^2 \geqslant 0$.

(iii) Obvious from definition.

(iv) This is an immediate consequence of the Cauchy–Schwarz inequality.

(v) By (iv) we have

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\| &= (\mathbf{x}+\mathbf{y}) \cdot (\mathbf{x}+\mathbf{y}) \\ &= \mathbf{x}\cdot\mathbf{x} + 2\mathbf{x}\cdot\mathbf{y} + \mathbf{y}\cdot\mathbf{y} \\ &\leqslant \|\mathbf{x}\|^2 + 2\|\mathbf{x}\|\|\mathbf{y}\| + \|\mathbf{y}\|^2 \\ &= \big(\|\mathbf{x}\| + \|\mathbf{y}\|\big)^2 .\end{aligned}$$

(vi) This follows directly from (v) by replacing $\mathbf{x}$ by $\mathbf{x}-\mathbf{y}$ and $\mathbf{y}$ by $\mathbf{y}-\mathbf{z}$.

□

Exercises

Problem 8.1. Consider the set $\left\{ \frac{1}{n} \,\middle|\, n \in \mathbf{N} \right\}$.

(i) Show that $\max S = 1$.

(ii) Show that if d is a lower bound for S, then $d \leqslant 0$.

(iii) Use (ii) to show that $0 = \inf S$.

Problem 8.2. Find, with proof, the supremum and/or infimum of $\{\frac{1}{n}\}$.

Solution. 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 8.3. 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

$$0 \leqslant \{(2\pi - 6) \cdot m(b - a)\} - \left(2 - \frac{\pi}{2}\right) < \delta$$

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 $\left|2k\pi + \frac{\pi}{2} - n\right| < \delta$

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.)

$$\begin{aligned} \sin n &= \sin\left(n - \left(2k\pi + \frac{\pi}{2}\right) + \left(2k\pi + \frac{\pi}{2}\right)\right) \\ &= \cos\left(n - \left(2k\pi + \frac{\pi}{2}\right)\right) \\ &= \cos\theta \end{aligned}$$

$$1 - \sin n = 2\sin^2\frac{\theta}{2} = 2\sin^2\left|\frac{\theta}{2}\right| \leqslant \frac{\theta^2}{2} < \delta$$

Hence we simply pick $\delta = \varepsilon$ to ensure that $1 - \sin n < \varepsilon$, and we're done. □

Problem 8.4 ([Rud76] Ch.1 Q1)**.** If $r \in \mathbf{Q}\backslash\{0\}$ and $x \in \mathbf{R}\backslash\mathbf{Q}$, prove that $r+x \in \mathbf{R}\backslash\mathbf{Q}$ and $rx \in \mathbf{R}\backslash\mathbf{Q}$.

Solution. We prove by contradiction. Suppose $r+x$ is rational, then $r+x = \frac{m}{n}$ for $m, n \in \mathbf{Z}$, and m, n have no common factors. Then $m = n(r+x)$. Let $r = \frac{p}{q}$ for $p, q \in \mathbf{Z}$, the former equation implies that $m = n\left(\frac{p}{q} + x\right)$, i.e., $qm = n(p+qx)$, giving

$$x = \frac{mq - np}{nq},$$

which says that x can be written as the quotient of two integers, so x is rational, a contradiction.

The proof for the case rx is similar. □

Problem 8.5 ([Rud76] Ch.1 Q4)**.** Let E be a nonempty subset of an ordered set; suppose α is a lower bound of E and β is an upper bound of E. Prove that $\alpha \leqslant \beta$.

Solution. Let $x \in E$. By definition of lower and upper bounds, $\alpha \leqslant x \leqslant \beta$. □

Problem 8.6 ([Rud76] Ch.1 Q8)**.** Prove that no order can be defined in $\mathbf{C}$ that turns it into an ordered field.

Solution. By Proposition 1.18d, an ordering $<$ that makes $\mathbf{C}$ an ordered field would have to satisfy $-1 = i^2 > 0$, contradicting $1 > 0$. □

Problem 8.7 ([Rud76] Ch.1 Q9, lexicographic order)**.** Suppose $z = a+bi$, $w = c+di$. Define an order on $\mathbf{C}$ as follows:

$$z < w \iff \begin{cases} a < c, \text{ or} \\ a = c, b < d. \end{cases}$$

Prove that this turns $\mathbf{C}$ into an ordered set. Does this ordered set have the least upper bound property?

Problem 8.8 ([Rud76] Ch.1 Q10)**.** Suppose $z = a + bi$, $w = u + iv$, and

$$a = \left(\frac{|w| + u}{2}\right)^{\frac{1}{2}}, \quad b = \left(\frac{|w| - u}{2}\right)^{\frac{1}{2}}.$$

Prove that $z^2 = w$ if $v \geqslant 0$ and that $\overline{z}^2 = w$ if $v \leqslant 0$. Conclude that every complex number (with one exception!) has two complex square roots.

Solution. We have

$$a^2 - b^2 = \frac{|w| + u}{2} - \frac{|w| - u}{2} = u,$$

and

$$2ab = (|w| + u)^{\frac{1}{2}} (|w| - u)^{\frac{1}{2}} = \left(|w|^2 - u^2\right)^{\frac{1}{2}} = \left(v^2\right)^{\frac{1}{2}} = |v|.$$

Hence

$$z^2 = \left(a^2 - b^2\right) + 2abi = u + |v|i = w$$

if $v \geqslant 0$, and

$$\overline{z}^2 = \left(a^2 - b^2\right) - 2abi = u - |v|i = w$$

if $v \leqslant 0$. Hence every non-zero w has two square roots $\pm z$ or $\pm\overline{z}$. Of course, 0 has only one square root, itself. □

Problem 8.9 ([Rud76] Ch.1 Q11)**.** If $z \in \mathbf{C}$, prove that there exists $r \geqslant 0$ and $w \in \mathbf{C}$ with $|w| = 1$ such that $z = rw$. Are w and r always uniquely determined by z?

Problem 8.10 ([Rud76] Ch.1 Q13)**.** If $x, y \in \mathbf{C}$, prove that

$$||x| - |y|| \leqslant |x - y|.$$

Solution. By the triangle inequality,

$$|x| = |(x - y) + y| \leqslant |x - y| + |y|$$

so that

$$|x| - |y| \leqslant |x - y|.$$

Interchanging the roles of x and y in the above, we also have

$$|y| - |x| \leqslant |x - y|$$

so that

$$||x| - |y|| \leqslant |x - y|.$$

□

9 Basic Topology

This chapter discusses basic notions of point set topology, which focuses on the metric space and its related structures. Then we introduce compactness and prove three major results (Theorem 9.44, Theorem 9.41, Theorem 9.42). We also briefly talk about perfect sets, and connectedness of sets.

Term	Notation
metric space	X, Y
metric	$d(p, q)$
general set	E
point in a set	p, q, r
open ball	$B_r(p)$
closed ball	$\overline{B}_r(p)$
punctured ball	$B_r(p) \setminus \{p\}$
neighbourhood	N
interior	E°
closure	$\overline{E}$
boundary	∂E
induced set	E'
compact set	K
open cover	$\mathcal{U}$
n-cell	I
Cantor set	C

Table 9.1: Notation for Chapter 9

§9.1 Metric Space

Definitions and Examples

Definition 9.1 (Metric space)**.** A ***metric space*** is a set X with an associated *metric* $d : X \times X \to \mathbf{R}$, which satisfies the following properties for all $p, q, r \in X$:

(i) $d(p, q) \geqslant 0$, where equality holds if and only if $p = q$; (positive definitiveness)

(ii) $d(p,q) = d(q,p)$; (symmetry)

(iii) $d(p,q) \leqslant d(p,r) + d(r,q)$. (triangle inequality)

For the rest of the chapter, X is taken to be a metric space, unless specified otherwise.

Example (Metrics on $\mathbf{R}^n$)

Each of the following functions define metrics on $\mathbf{R}^n$.

$$d_1(x,y) = \sum_{i=1}^{n} |x_i - y_i|;$$
$$d_2(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)}$$
$$d_\infty(x,y) = \max_{i \in \{1,2,\ldots,n\}} |x_i - y_i|.$$

These are called the ℓ^1-, ℓ^2- (or Euclidean) and ℓ^∞-distances respectively.

The proof that each of d_1, d_2, d_∞ is a metric 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 $\|x\|_2$ of a vector $x = (x_1, \ldots, x_n) \in \mathbf{R}^n$ is

$$\|x\|_2 \coloneqq \left(\sum_{i=1}^{n} {x_i}^2 \right)^{\frac{1}{2}} = \langle x, x \rangle^{\frac{1}{2}},$$

where the inner product is given by

$$\langle x, y \rangle \coloneqq \sum_{i=1}^{n} x_i y_i.$$

Then $d_2(x,y) = \|x - y\|_2$, and so the triangle inequality is the statement that

$$\|w - y\|_2 \leqslant \|w - x\|_2 + \|x - y\|_2 .$$

This follows immediately by taking $u = w - x$ and $v = x - y$ in the following lemma.

Lemma. If $u, v \in \mathbf{R}^n$ then $\|u + v\|_2 \leqslant \|u\|_2 + \|v\|_2$.

Proof. Since $\|u\|_2 \geqslant 0$ for all $u \in \mathbf{R}^n$, squaring both sides of the desired inequality gives

$${\|u + v\|_2}^2 \leqslant {\|u\|_2}^2 + 2\|u\|_2\|v\|_2 + {\|v\|_2}^2.$$

But since

$${\|u + v\|_2}^2 = \langle u + v, u + v \rangle = {\|u\|_2}^2 + 2\langle u, v \rangle + \|v\|_2^2 ,$$

this inequality is immediate from the Cauchy–Schwarz inequality, that is to say the inequality

$$|\langle u, v\rangle| \leqslant \|u\|_2 \|v\|_2 .$$

□

A metric space (X, d) naturally induces a metric on any of its subsets.

Definition 9.2 (Subspace)**.** Suppose (X, d) is a metric space, $Y \subset X$. Then the restriction of d to $Y \times Y$ gives Y a metric so that $(Y, d_{Y\times Y})$ is a metric space. We call Y equipped with this metric a ***subspace***.

Balls and Boundedness

Definition 9.3 (Balls)**.**

(i) The ***open ball*** centred at $p \in X$ with radius $r > 0$ is the set

$$B_r(p) \coloneqq \{q \in X \mid d(p, q) < r\}.$$

(ii) The ***closed ball*** centred at p with radius r is

$$\overline{B}_r(p) \coloneqq \{q \in X \mid d(p, q) \leqslant r\}.$$

(iii) The ***punctured ball*** is the open ball excluding its centre:

$$B_r(p) \setminus \{p\} = \{q \in X \mid 0 < d(p, q) < r\}.$$

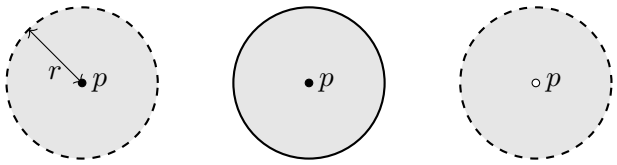


Figure 9.1: Open ball, closed ball, punctured ball

Example

Considering $\mathbf{R}^3$ with the Euclidean metric, $B_1(0)$ really is what we understand geometrically as a ball (minus its boundary, the unit sphere), whilst $\overline{B}_1(0)$ contains the unit sphere and everything inside it.

Remark. We caution that this intuitive picture of the closed ball being the open ball "together with its boundary" is totally misleading in general. For instance, in the discrete metric on a set X, the open ball $B_1(a)$ contains only the point a, whereas the closed ball $\overline{B}_1(a)$ is the whole of X.

Definition 9.4 (Bounded)**.** $E \subset X$ is said to be ***bounded*** if E is contained in some open ball; that is, there exists $M \in \mathbf{R}$ and $p \in X$ such that $E \subset B_M(p)$.

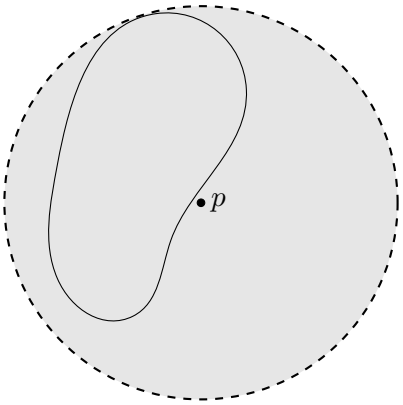


Figure 9.2: Bounded set

Proposition 9.5. Let $E \subset X$. Then the following are equivalent:

(i) E is bounded;

(ii) E is contained in some closed ball;

(iii) The set $\{d(x, y) \mid x, y \in E\}$ is a bounded subset of $\mathbf{R}$.

Proof.

$\boxed{\text{(i)} \implies \text{(ii)}}$ This is obvious.

$\boxed{\text{(ii)} \implies \text{(iii)}}$ This follows immediately from the triangle inequality.

$\boxed{\text{(iii)} \implies \text{(i)}}$ Suppose E satisfies (iii), then there exists $r \in \mathbf{R}$ such that $d(x, y) \leqslant r$ for all $x, y \in E$. If $E = \emptyset$, then E is certainly bounded. Otherwise, let $p \in E$ be an arbitrary point. Then $E \subset B_{r+1}(p)$. □

Definition 9.6 (Diameter)**.** Let non-empty $E \subset X$. Then the ***diameter*** of E is

$$\operatorname{diam} E \coloneqq \sup_{p,q \in E} d(p, q).$$

Example

Find the diameter of the open unit ball in $\mathbf{R}^n$:

$$B = \{\mathbf{x} \in \mathbf{R}^n \mid \|\mathbf{x}\| < 1\}.$$

Proof. Note that for all $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$,

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| \leqslant \|\mathbf{x}\| + \|-\mathbf{y}\| = \|\mathbf{x}\| + \|\mathbf{y}\| < 1 + 1 = 2.$$

On the other hand, for any $\varepsilon > 0$, we pick

$$\mathbf{x} = \left(1 - \frac{\varepsilon}{4}, 0, \ldots, 0\right), \quad \mathbf{y} = \left(-\left(1 - \frac{\varepsilon}{4}\right), 0, \ldots, 0\right).$$

Then $d(\mathbf{x}, \mathbf{y}) = 2 - \dfrac{\varepsilon}{2} > 2 - \varepsilon$. Since ε is arbitrary, we have that $\operatorname{diam} B = 2$. □

Proposition 9.7. $E \subset \mathbf{R}^n$ is bounded if and only if $\operatorname{diam} E < +\infty$.

Proof.

$\boxed{\Longrightarrow}$ If E is bounded, then there exists $M > 0$ such that $\|x\| \leqslant M$ for all $x \in E$.

Thus for any $x, y \in E$,

$$d(x, y) = \|x - y\| \leqslant \|x\| + \|y\| \leqslant 2M.$$

Thus $\operatorname{diam} E = \sup d(x, y) \leqslant 2M < +\infty$.

$\boxed{\Longleftarrow}$ Suppose that $\operatorname{diam} E = r$. Pick a random point $x \in E$, suppose that $\|x\| = R$.

Then for any other $y \in E$,

$$\|y\| = \|x + (y - x)\| \leqslant \|x\| + \|y - x\| \leqslant R + r.$$

Thus, by picking $M = R + r$, we obtain $\|y\| \leqslant M$ for all $y \in E$, and we are done. □

Remark. Basically we used x to confine E within a ball, which is then confined within an even bigger ball centered at the origin.

Open and Closed Sets

Definition 9.8 (Neighbourhood)**.** $N \subset X$ is a ***neighbourhood*** of $p \in X$ if there exists $\varepsilon > 0$ such that $B_\varepsilon(p) \subset N$.

Definition 9.9 (Open set)**.** $E \subset X$ is ***open*** (in X) if it is a neighbourhood of all its elements; that is, for all $p \in E$, there exists $\varepsilon > 0$ such that $B_\varepsilon(p) \subset E$.

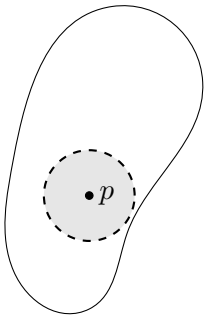


Figure 9.3: Open set

Proposition 9.10. Any open ball is open.

Proof. Let $B_r(p)$ be an open ball. Let $q \in B_r(p)$, then $d(p,q) < r$. Let $\varepsilon = r - d(p,q)$, and note that $\varepsilon > 0$.

Consider the ball $B_\varepsilon(q)$, and let $s \in B_\varepsilon(q)$. By the triangle inequality,

$$\begin{aligned} d(p,s) &\leqslant d(q,s) + d(p,q) \\ &< \varepsilon + d(p,q) \\ &= r \end{aligned}$$

and thus $s \in B_r(p)$. Since for all $q \in B_r(p)$ there exists $\varepsilon > 0$ such that $B_\varepsilon(q) \subset B_r(p)$, we have that $B_r(p)$ is open. □

Proposition 9.11. (i) Both $\emptyset$ and X are open.

(ii) For any indexing set I and collection of open sets $\{E_i \mid i \in I\}$, $\bigcup_{i\in I} E_i$ is open.

(iii) For any *finite* indexing set I and collection of open sets $\{E_i \mid i \in I\}$, $\bigcap_{i\in I} E_i$ is open.

Proof.

(i) Obvious by definition.

(ii) If $p \in \bigcup_{i\in I} E_i$, then $x \in E_i$ for some $i \in I$. Since E_i is open, there exists $\varepsilon > 0$ such that $B_\varepsilon(p) \subset E_i$ and hence $B_\varepsilon(p) \in \bigcup_{i\in I} E_i$.

(iii) Suppose that I is finite and that $p \in \bigcap_{i\in I} E_i$. For each $i \in I$, we have $p \in E_i$ and so there exists δ_i such that $B_{\delta_i}(p) \subset E_i$. Set $\delta = \min_{i\in I} \delta_i$, then $\delta > 0$ (here it is, of course, crucial that I be finite), and $B_\delta(p) \subset B_{\delta_i}(p) \subset E_i$ for all i. Therefore $B_\delta(p) \subset \bigcap_{i\in I} E_i$.

□

Remark. (i) is in fact a special case of (ii) and (iii), taking I to be the empty set.

Remark. While the indexing set I in (ii) can be arbitrary, the indexing set in (iii) must be finite. This is because in general, an arbitrary intersection of open sets is not open; for instance, the intervals $E_i = \left(-\frac{1}{i}, \frac{1}{i}\right)$ are all open in $\mathbf{R}$, but their intersection $\bigcap_{i=1}^{\infty} E_i = \{0\}$, which is not an open set.

Suppose Y is a subspace of X. We say that E is *open relative* to Y if for all $p \in E$, there exists $\varepsilon > 0$ such that $B_\varepsilon(p) \cap Y \subset E$. (Note that $B_\varepsilon(p) \cap Y$ is in the open ball in Y[1], because the metric $d' : Y \times Y \to R$ is the restriction to $Y \times Y$ of the metric $d : X \times X \to \mathbf{R}$ on X.)

Proposition 9.12. Suppose Y is a subspace of X, $E \subset Y$. Then E is open relative to Y if and only if there exists an open subset G of X such that $E = Y \cap G$.

[1] notice that the definition of an open ball depends on the metric space!

Proof.

$\boxed{\Longrightarrow}$ We prove by construction; that is, construct the required set G.

Suppose E is open relative to Y. For each $p \in E$, by openness of E, there exists $r_p > 0$ such that $B_{r_p}(p) \cap Y \subset E$. Consider the union

$$\bigcup_{p\in E} \Big(B_{r_p}(p) \cap Y\Big) \subset E.$$

Note that we can write

$$\bigcup_{p\in E} \Big(B_{r_p}(p) \cap Y\Big) = \left(\bigcup_{p\in E} B_{r_p}(p)\right) \cap Y \subset E.$$

Let

$$G = \bigcup_{p\in E} B_{r_p}(p),$$

then we have $G \cap Y \subset E$.

Since G is an intersection of open balls (which are open sets), by Proposition 9.11, G is an open subset of X.

Note for each $p \in E \subset Y$, we have $p \in Y$, and $p \in B_{r_p}(p)$ for some $r_p > 0$, so $p \in \bigcup_{p\in E} B_{r_p}(p) = G$. Hence $p \in G \cap Y$. This shows $E \subset G \cap Y$.

Hence $E = G \cap Y$.

$\boxed{\Longleftarrow}$ Suppose $E = G \cap Y$ for some open subset G of X.

Let $p \in E$. Since $p \in G$, by the openness of G, there exists $r_p > 0$ such that $B_{r_p}(p) \subset G$. Then $B_{r_p}(p) \cap Y \subset G \cap Y = E$. Thus by definition E is open relative to Y. □

The complement of an open set is a closed set.

Definition 9.13 (Closed set)**.** $E \subset X$ is ***closed*** if its complement $E^c = X \setminus E$ is open.

Proposition 9.14. Any closed ball is closed.

Proof. To prove that $\overline{B}_r(p)$ is closed, we need to show that its complement

$$\overline{B}_r(p)^c = \{q \in X \mid d(p, q) > r\}$$

is open.

Let $s \in \overline{B}_r(p)^c$. Take $\varepsilon > 0$ such that $r + \varepsilon < d(p, s)$; that is, $\varepsilon < d(p, s) - r$.

Let $q \in B_\varepsilon(s)$, then $d(q, s) < \varepsilon$. Thus $d(q, s) < d(p, s) - r$, or $r < d(p, s) - d(q, s)$. Then by the triangle inequality,

$$\begin{aligned} d(p, q) &\geqslant d(p, s) - d(q, s) \\ &> r \end{aligned}$$

Hence $q \in \overline{B}_r(p)^c$, and so $B_\varepsilon(s) \subset \overline{B}_r(p)^c$. Therefore $\overline{B}_r(p)^c$ is open, so $\overline{B}_r(p)$ is closed. □

Proposition 9.15. (i) Both $\emptyset$ and X are closed.

(ii) For any indexing set I and collection of closed sets $\{F_i \mid i \in I\}$, $\bigcap_{i\in I} F_i$ is closed.

(iii) For any *finite* indexing set I and collection of closed sets $\{F_i \mid i \in I\}$, $\bigcup_{i\in I} F_i$ is closed.

Proof. From Proposition 9.11, simply take complements and apply de Morgan's laws. □

Remark. As above, the indexing set in (iii) must be finite; for instane, the closed intervals $F_i = \left[-1 + \frac{1}{i}, 1 - \frac{1}{i}\right]$ are all closed in $\mathbf{R}$, but their union $\bigcup_{i=1}^{\infty} F_i = (-1, 1)$ is open.

Interior, Closure, Boundary

Definition 9.16. Suppose $E \subset X$.

(i) The ***interior*** of E, denoted by E°, is the union of all open subsets of X contained in E; $p \in E^\circ$ is an ***interior point*** of E. (Equivalently, E° is the set of all points in E for which E is a neighbourhood; p is an interior point if there exists $\varepsilon > 0$ such that $B_\varepsilon(p) \subset E$.)

(ii) The ***closure*** of E, denoted by $\overline{E}$, is the intersection of all closed subsets of X containing E. E is said to be ***dense*** if $\overline{E} = X$. (Equivalently, every point of X is either a limit point of E, or in E.)

(iii) The ***boundary*** of E is $\partial E = \overline{E} \setminus E^\circ$; $p \in \partial E$ is a ***boundary point*** of E.

In the figure below, the black outline represents the boundary; the grey area within represents the interior; the union represents the closure.

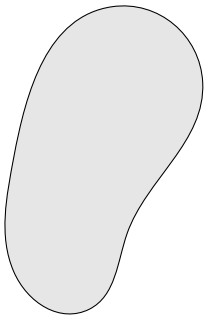


Figure 9.4: Interior, closure, boundary

Example

The interior of the closed interval $[a, b] \subset \mathbf{R}$ is the open interval (a, b).

$\mathbf{Q}$ is dense in $\mathbf{R}$.

Remark. It is obvious that the interior E° is open since the union of open sets is open; similarly, the closure $\overline{E}$ is closed since the intersection of closed sets is closed.

Proposition 9.17. Suppose $E \subset X$.

(i) E is open if and only if $E = E^\circ$.

(ii) E is closed if and only if $E = \overline{E}$.

Proof.

(i) $\boxed{\Longrightarrow}$ Suppose E is open. Then E is an open subset of X contained in E (since $E \subset E$), so $E \subset E^\circ$. Conversely, suppose $A \subset E^\circ$. Then A is an open subset of X contained in E, so $A \subset E$, and hence $E^\circ \subset E$. Therefore $E = E^\circ$.

$\boxed{\Longleftarrow}$ Since an arbitrary union of open sets is open, E° is itself an open set, and it is clearly the unique largest open subset of X contained in E.

(ii) $\boxed{\Longrightarrow}$ If E is itself closed then evidently $E = \overline{E}$.

$\boxed{\Longleftarrow}$ Since an arbitrary intersection of closed sets is closed, $\overline{E}$ is the unique smallest closed subset of X containing E.

□

Proposition 9.18. Suppose $E \subset X$. Then $p \in \overline{E}$ if and only if every open ball centred at p contains a point of E.

Proof.

$\boxed{\Longrightarrow}$ Suppose that $p \in \overline{E}$. Suppose, for a contradiction, that there exists some open ball $B_\varepsilon(p)$ that does not meet E, then $B_\varepsilon(p)^c$ is a closed set containing E. Therefore $B_\varepsilon(p)^c$ contains $\overline{E}$, and hence it contains p, which is obviously nonsense.

$\boxed{\Longleftarrow}$ Suppose that every ball $B_\varepsilon(p)$ meets E. Suppose, for a contradiction, that $p \notin \overline{E}$. Then since $\overline{E}^c$ is open, there is a ball $B_\varepsilon(p)$ contained in $\overline{E}^c$, and hence in E^c, contrary to assumption. □

Remark. A particular consequence of this is that $E \subset X$ is dense if and only if it meets every open set in X.

Proposition 9.19. Suppose $E \subset X$. Let $F \supset E$ be some closed set. Then $\overline{E} \subset F$.

Proof. Let p be a limit point of E. Then p is a limit point of F. But since F is closed, F contains all its limit points, so all the limit points of E are in F. Hence $\overline{E} \subset F$. □

Remark. This means that $\overline{E}$ is the "smallest" closed set containing E.

Limit Points

Definition 9.20.

(i) $p \in X$ (not necessarily in E) is an ***adherent point*** of E (or is *adherent* to E) if $B_\varepsilon(p) \cap E \neq \emptyset$ for all $\varepsilon > 0$.

(ii) $p \in X$ is a ***limit point*** (or *accumulation point*) of E if for all $\varepsilon > 0$, there exists $q \in E \setminus \{p\}$ such that $q \in B_\varepsilon(p)$. (In other words, p is a limit point of E if and only if p adheres to $E \setminus \{p\}$.)

The ***induced set*** of E, denoted by E', is the set of all limit points of E in X.

(iii) $p \in E$ is an ***isolated point*** of E if p is not an limit point of E (that is, there exists $\varepsilon > 0$ such that $B_\varepsilon(p) \cap E = \{p\}$).

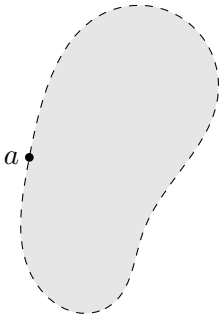


Figure 9.5: Adherent point, limit point, isolated point

Example (Adherent point)
- If $p \in E$, then p adheres to E because every ball contains p.
- If $E \subset \mathbf{R}$ is bounded above, then $\sup E$ is adherent to E.

Example (Limit point)
- The set $\left\{ \frac{1}{n} \,\middle|\, n \in \mathbf{N} \right\}$ has 0 as a limit point.
- The set of rational numbers has every real number as a limit point.
- Every point of the closed interval $[a, b]$ is a limit point of the set of numbers in the open interval (a, b).
- Consider the metric space $\mathbf{R}^2$. The limit point set of any open ball $B_r(p)$ is the closed ball $\overline{B}_r(p)$, which is also the closure of $B_r(p)$.
- Consider $\mathbf{Q} \subset \mathbf{R}$. $\mathbf{Q}' = \overline{\mathbf{Q}} = \mathbf{R}$.

Proposition 9.21. If p is a limit point of E, then every ball of p contains infinitely many points of E.

Proof. We prove by contradiction. Suppose otherwise, for a contradiction, that there exists $B_r(p)$ which contains only a finite number of points of E distinct from p. Then let

$$B_r(p) = \{q_1, \ldots, q_n\},$$

where $p \neq q_i$ for $i = 1, \ldots, n$. Take

$$r = \min\{d(p, q_1), \ldots, d(p, q_n)\},$$

then $B_r(p)$ contains no points of E distinct from p, which is a contradiction. □

Corollary 9.22. A finite point set has no limit points.

Remark. The converse is not true; for example, $\mathbf{N}$ is an infinite set with no limit points. In a later section we will show that infinite sets contained in some open ball always have an limit point; this result is known as the Bolzano–Weierstrass theorem (Theorem 9.42).

A closed set was defined to be the complement of an open set. The next result describes closed sets in another way.

Proposition 9.23. Suppose $E \subset X$. Then E is closed if and only if it contains all its limit points.

Proof.

$\boxed{\Longrightarrow}$ Suppose E is closed. Let p be a limit point of E. We want to prove that $p \in E$.

Suppose otherwise, for a contradiction, that $p \notin E$. Then $p \in E^c$. Since E^c is open, there exists $\varepsilon > 0$ such that $B_\varepsilon(p) \subset E^c$. Thus $B_\varepsilon(p)$ contains no points of E, contradicting the fact that p is a limit point of E.

$\boxed{\Longleftarrow}$ Suppose E contains all its limit points. To show that E is closed, we want to show that E^c is open.

Let $p \in E^c$. Then p is not a limit point of E, so there exists some ball $B_\varepsilon(p)$ which does not intersect E, so $B_\varepsilon(p) \subset E^c$. Hence E^c is open, so E is closed. □

Proposition 9.24. Suppose $E \subset X$. Then E' is a closed subset of X.

Proof. To prove that E' is closed, we need to show that its complement $(E')^c$ is open.

Suppose $p \in (E')^c$. Then p is not a limit point of E, so there exists a ball $B_\varepsilon(p)$ whose intersection with E is either empty or $\{p\}$ (depending on whether $p \in E$ or not).

Claim. $B_{\frac{\varepsilon}{2}}(p) \subset (E')^c$.

Let $q \in B_{\frac{\varepsilon}{2}}(p)$.

- If $q = p$, then clearly $q \in (E')^c$.
- If $q \neq p$, there is some ball about q which is contained in $B_\varepsilon(p)$, but does not contain p: the ball $B_\delta(q)$ where $\delta = \min\left(\frac{\varepsilon}{2}, d(p,q)\right)$ has this property. This ball meets E in the empty set, and so $q \in (E')^c$ in this case too.

□

Proposition 9.25. Suppose $E \subset X$. Then $\overline{E} = E \cup E'$.

Proof. We show double inclusion.

$\boxed{E \cup E' \subset \overline{E}}$ Obviously $E \subset \overline{E}$, so we need only show that $E' \subset \overline{E}$.

We prove by contrapositive. Suppose $p \in \overline{E}^c$. Since $\overline{E}^c$ is open, there is some ball $B_\varepsilon(p)$ which lies in $\overline{E}^c$, and hence also in E^c, and therefore a cannot be a limit point of E.

$\boxed{\overline{E} \subset E \cup E'}$ If $p \in \overline{E}$, we saw in Lemma 5.1.5 that there is a sequence (x_n) of elements of E with $x_n \to p$. If $x_n = p$ for some n then we are done, since this implies that $p \in E$. Suppose, then, that $x_n \neq p$ for all n. Let $\varepsilon > 0$ be given, for sufficiently large n, all the x_n are elements of $B_\varepsilon(p) \setminus \{p\}$, and they all lie in E. It follows that p is a limit point of E, and so we are done in this case also. □

Proposition 9.26. Suppose non-empty $E \subset \mathbf{R}$ is bounded above. Let $y = \sup E$. Then $y \in \overline{E}$. Hence $y \in E$ if E is closed.

Proof. If $y \in E$, since $E \subset \overline{E}$ we have that $y \in \overline{E}$.

For the second part, assume $y \notin E$. For every $h > 0$ there exists then a point $x \in E$ such that $y - h < x < y$, for otherwise $y - h$ would be an upper bound of E. Thus y is a limit point of E. Hence $y \in \overline{E}$. □ review proof

(y is either in E, or a limit point of E)

§9.2 Compactness

Definitions and Properties

Definition 9.27 (Open cover)**.** An ***open cover*** of $K \subset X$ is a collection of open sets $\mathcal{U} = \{U_i \mid i \in I\}$ such that

$$K \subset \bigcup_{i \in I} U_i.$$

A *subcover* of $\mathcal{U}$ is a subcollection $\{U_i \mid i \in I'\}$, where $I' \subset I$, which is an open cover of K. If I' is finite, then it is called a *finite subcover*.

Definition 9.28 (Compactness)**.** $K \subset X$ is ***compact*** if *every* open cover of K has a finite subcover.

That is, if $\mathcal{U} = \{U_i \mid i \in I\}$ is an open cover of K, then there are finitely many indices $i_1, \ldots, i_n \in I$ such that

$$K \subset \bigcup_{k=1}^{n} U_{i_k}.$$

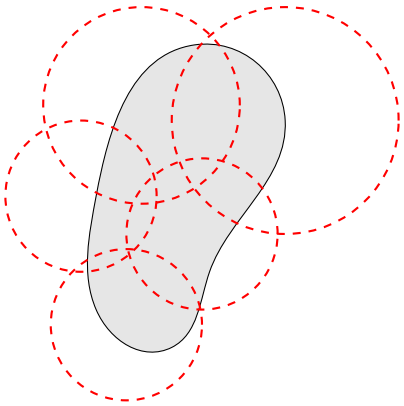


Figure 9.6: Compact set

Example
- The real line $\mathbf{R}$ is not compact. For instance, the open cover $\{(-n, n) \mid n \in \mathbf{N}\}$ has no finite subcover.
- $\mathbf{Z}$ is not compact in $\mathbf{R}$. For instance, the open cover $\left\{\left(n - \frac{1}{2}, n + \frac{1}{2}\right) \;\middle|\; n \in \mathbf{Z}\right\}$ has no finite subcover.
- $[0, 1]$ is compact. (See Proposition 9.33 for the proof.)

Proposition 9.29. Every finite set is compact.

Proof. Let E be finite; let $E = \{p_1, \ldots, p_n\}$.

Let $\mathcal{U} = \{U_i \mid i \in I\}$ be an open cover of E. We need to construct a finite subcover of E.

For each point $p_k \in E$, choose one U_{i_k} such that $p_k \in U_{i_k}$. Then $\{U_{i_k} \mid k = 1, \ldots, n\}$ is a finite subcover of $\mathcal{U}$. □

Notice earlier than if $E \subset Y \subset X$, then E may be open relative to Y, but not open relative to X; this implies that the property of being open depends on the space in which E is embedded. Compactness, however, behaves better, as shown in the next result; it is independent of the metric space.

Proposition 9.30. Suppose Y is a subspace of X, and $K \subset Y$. Then K is compact relative to X if and only if K is compact relative to Y.

Proof.

$\boxed{\Longrightarrow}$ Suppose K is compact relative to X. We will show that K is compact relative to Y. Let $\mathcal{U}$ be an open cover of K in Y; that is, $\mathcal{U} = \{U_i \mid i \in I\}$ is a collection of sets open relative to Y, such that $K \subset \bigcup_{i\in I} U_i$. We want to show that $\mathcal{U}$ has a finite subcover.

Since for all $i \in I$, U_i is open relative to Y, by Proposition 9.12, there exists V_i open relative to X such that $U_i = Y \cap V_i$. Consider $\{V_i \mid i \in I\}$, which is an open cover of K, since it is a collection of open sets. Since K is compact relative to X, there exist finitely many indices $i_1, \ldots, i_n$ such that

$$K \subset \bigcup_{k=1}^{n} V_{i_k}.$$

Since $K \subset \bigcup_{k=1}^{n} V_{i_k}$ and $K \subset Y$, we have that

$$K \subset \left(\bigcup_{k=1}^{n} V_{i_k}\right) \cap Y = \bigcup_{k=1}^{n} (Y \cap V_{i_k}) = \bigcup_{k=1}^{n} U_{i_k},$$

where $\{U_{i_k} \mid k = 1, \ldots, n\}$ forms a finite subcover of $\mathcal{U}$. Hence K is compact relative to Y.

$\boxed{\Longleftarrow}$ Suppose K is compact relative to Y. Let $\mathcal{V}$ be an open cover of K in X; that is, $\mathcal{V} = \{V_i \mid i \in I\}$ is a collection of open subsets of X which covers K. We want to show that $\mathcal{V}$ has a finite subcover.

For $i \in I$, let $U_i = Y \cap V_i$. Then $\{U_i \mid i \in I\}$ cover K in Y. By compactness of K in Y, there exist finitely many indices $i_1, \ldots, i_n$ such that

$$K \subset \bigcup_{k=1}^{n} U_{i_k} \subset \bigcup_{k=1}^{n} V_{i_k}$$

since $U_i \subset V_i$. □

Proposition 9.31. Compact subsets of metric spaces are bounded.

Proof. Suppose $K \subset X$ is compact. To prove that K is bounded, we want to construct some open ball that contains the entirety of K.

Fix $p \in K$. For $n \in \mathbf{N}$, let $U_n = B_n(p)$. Then $\{U_n \mid n \in \mathbf{N}\}$ is an open cover of K. By compactness of K, there exists a finite subcover

$$\{U_{n_i} \mid i = 1, \ldots, m\}.$$

But note that $U_{n_1} \subset \cdots \subset U_{n_m}$, so U_{n_m} contains K. Hence K is bounded. □

Proposition 9.32. Compact subsets of metric spaces are closed.

Proof. Let $K \subset X$ be compact. To prove that K is closed, we need to show that K^c is open. Let $p \in K^c$; our goal is to show that there exists $\varepsilon > 0$ such that $B_\varepsilon(p) \subset K^c$, or $B_\varepsilon(p) \cap K = \emptyset$.

For all $q_i \in K$, consider the pair of open balls $B_{r_i}(p)$ and $B_{r_i}(q_i)$, where $r_i < \frac{1}{2}d(p, q_i)$. Since K is compact, there exists finite many points $q_{i_1}, \ldots, q_{i_n} \in K$ such that

$$K \subset \bigcup_{k=1}^{n} B_{r_{i_k}}(q_{i_k}) = W.$$

Consider the intersection

$$\bigcap_{k=1}^{n} B_{r_{i_k}}(p),$$

which is an open ball at p of radius $\min\{d(p, q_{i_k}) \mid k = 1, \ldots, n\}$.

Claim. $\varepsilon = \min\{d(p, q_{i_k}) \mid k = 1, \ldots, n\}$.

Note that $B_\varepsilon(p) \subset B_{r_{i_k}}(p)$ for all $k = 1, \ldots, n$. By construction, for all $q_i \in K$, the open balls $B_{r_i}(p)$ and $B_{r_i}(q_i)$ are disjoint. In particular,

$$B_\varepsilon(p) \cap B_{r_{i_k}}(q_{i_k}) = \emptyset \quad (k = 1, \ldots, n)$$

Then taking the union,

$$\begin{aligned} \bigcup_{k=1}^{n} \left(B_\varepsilon(p) \cap B_{r_{i_k}}(q_{i_k}) \right) &= \emptyset \\ B_\varepsilon(p) \cap \left(\bigcup_{k=1}^{n} B_{r_{i_k}}(q_{i_k}) \right) &= \emptyset \\ B_\varepsilon(p) \cap W &= \emptyset \end{aligned}$$

as desired. □

Proposition 9.33. Closed subsets of compact sets are compact.

Proof. Suppose $K \subset X$ is compact, $F \subset K$ is closed (relative to X). We will show that F is compact. Let $\mathcal{U} = \{U_i \mid i \in I\}$ be an open cover of F. We want to show that there exists a finite subcover of $\mathcal{U}$.

Since F is closed, its complement F^c is open. Consider the union

$$\Omega = \mathcal{U} \cup \{F^c\},$$

which is an open cover of K.

Since K is compact, there exists a finite subcover of Ω, given by

$$\Phi = \{U_{i_1}, \ldots, U_{i_n}, F^c\}$$

which covers K, and hence F. Now remove F^c from Φ to obtain

$$\Phi' = \{U_{i_1}, \ldots, U_{i_n}\},$$

which is an open cover of F, since $F^c \cap F = \emptyset$. Hence Φ' is a finite subcover of $\mathcal{U}$, so F is compact. □

Remark. Caution: this does *not* say "closed sets are compact"! In fact, closed sets are not necessarily compact. For instance, $\mathbf{R}$ is closed in $\mathbf{R}$, but it is not compact because it is not bounded.

Note that closed and bounded sets are not necessarily compact for general metric spaces, but they are compact in $\mathbf{R}^n$ (by Theorem 9.41).

Corollary 9.34. If F is closed and K is compact, then $F \cap K$ is compact.

Proof. Suppose F is closed, K is compact. By Proposition 9.32, K is closed. By Proposition 9.15, the intersection of two closed sets is closed, so $F \cap K$ is closed.

Since $F \cap K \subset K$ is closed, and K is compact, by Proposition 9.33, $F \cap K$ is compact. □

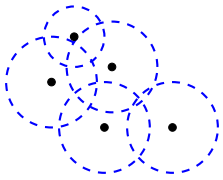
Heine–Borel Theorem

Proposition 9.35. K is compact if and only if every infinite subset of K has a limit point in K.

Proof.

$\boxed{\Longrightarrow}$ Suppose K is compact. Let E be an infinite subset of K. Suppose otherwise, for a contradiction, that E has no limit point in K.

For all $p \in K$, p is not a limit point of E, so there exists $r_p > 0$ such that $B_{r_p}(p) \cap E \setminus \{p\} = \emptyset$.



Consider the open cover of K given by the collection of open balls at each $p \in K$:

$$\mathcal{U} = \left\{ B_{r_p}(p) \mid p \in E \right\}.$$

It is clear that $\mathcal{U}$ has no finite subcover, since E is infinite, and each $B_{r_p}(p)$ contains at most one point of E.

Since $E \subset K$, the above is also true for K. This contradicts the compactness of K.

$\boxed{\Longleftarrow}$ Suppose every infinite subset of K that has a limit point in K. Fix an arbitrary open cover $\mathcal{U} = \{U_i \mid i \in I\}$ of K. We will show that $\mathcal{U}$ has a finite subcover, by construction.

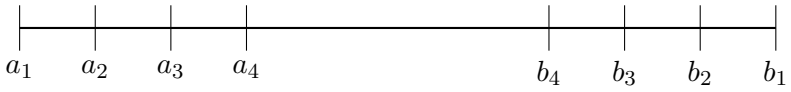
Before that, we will reindex $\mathcal{U}$ to make it more convenient, as follows. By the definition of a cover, every $p \in K$ is contained in some U_i. Pick *one* such U_i for each $p \in K$, and call it U_p. Then our open cover is now $\mathcal{U} = \{U_p \mid p \in K\}$, and for all $p \in K$ we have $p \in U_p$.

To complete proof

□

Proposition 9.36 (Nested interval theorem)**.** Suppose (I_n) is a decreasing sequence of closed and bounded intervals in $\mathbf{R}$; that is, $I_1 \supset I_2 \supset \cdots$. Then

$$\bigcap_{n=1}^{\infty} I_n \neq \emptyset.$$



Proof. For $n \in \mathbf{N}$, let $I_n = [a_n, b_n]$. Let $E = \{a_n \mid n \in \mathbf{N}\}$. Since E is non-empty and bounded above (by b_1), it has a supremum in $\mathbf{R}$; let $x = \sup E$.

Claim. $x \in \bigcap_{n=1}^{\infty} I_n$.

Since x is the supremum, we have that $a_n \leqslant x$ for all $n \in \mathbf{N}$. Note that for $m > n$, $I_n \supset I_m$ implies $a_n \leqslant a_m \leqslant b_m \leqslant b_n$. This means b_n is an upper bound for all a_n; hence $x \leqslant b_n$ for all $n \in \mathbf{N}$.

Therefore $x \in I_n$ for $n = 1, 2, \ldots$ □

To generalise the notion of intervals, we define a *k-cell* as

$$\{(x_1, \ldots, x_k) \in \mathbf{R}^k \mid a_i \leqslant x_i \leqslant b_i,\ 1 \leqslant i \leqslant k\}.$$

Example

A 1-cell is an interval, a 2-cell is a rectangle, and a 3-cell is a rectangular solid. In this regard, we can think of a k-cell as a higher-dimensional version of a rectangle or rectangular solid; it is the Cartesian product of k closed intervals.

The previous result can be generalised to k-cells, which we will now prove.

Proposition 9.37. Suppose (I_n) is a decreasing sequence of k-cells; that is, $I_1 \supset I_2 \supset \cdots$. Then $\bigcap_{n=1}^{\infty} I_n \neq \emptyset$.

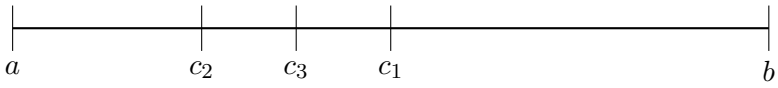
Proof. For $n \in \mathbf{N}$, let

$$I_n = \{(x_1, \ldots, x_k) \mid a_{n_i} \leqslant x_i \leqslant b_{n_i},\ 1 \leqslant i \leqslant k\},$$

and let $I_{n_i} = [a_{n_i}, b_{n_i}]$, where $I_n = I_{n_1} \times \cdots \times I_{n_k}$ is the Cartesian product of k closed intervals.

For each i $(i = 1, \ldots, k)$, □

Lemma 9.38. Every closed interval is compact (in $\mathbf{R}$).



Proof. Suppose otherwise, for a contradiction, that a closed interval $[a, b] \subset \mathbf{R}$ is not compact. Then there exists an open cover $\mathcal{U} = \{U_i \mid i \in I\}$ with no finite subcover.

Let $c_1 = \frac{1}{2}(a, b)$. Subdivide $[a, b]$ into subintervals $[a, c_1]$ and $[c_1, b]$. Then $\mathcal{U}$ covers $[a, c_1]$ and $[c_1, b]$, but at least one of these subintervals has no finite subcover (if not, then both subintervals have finite subcovers, so we can take the union of the two finite subcovers to obtain a larger subcover of the entire interval). WLOG, assume $[a, c_1]$ has no finite subcover; let $I_1 = [a, c_1]$.

Again subdivide I_1 in half to get $[a, c_2]$ and $[c_2, c_1]$. At least one of these subintervals has no finite subcover.

Repeat the above process of subdividing intervals into half. Then we obtain a decreasing sequence of closed intervals

$$I_1 \supset I_2 \supset I_3 \supset \cdots$$

where all of them have no finite subcover of $\mathcal{U}$.

By the nested interval theorem (Proposition 9.36), there exists $x' \in I_n$ for all $n \in \mathbf{N}$. Notice x' is in some U_i, which is open. Then there exists $\varepsilon > 0$ such that $B_\varepsilon(x') \subset U_i$.

Since the length of the subintervals is decreasing and tends to zero, there exists some subinterval I_n so small such that $I_n \subset B_\varepsilon(x')$. This means $I_n \subset U_i$, so U_i itself is an open cover of I_n, which contradicts the fact that I_n has no finite subcover of $\mathcal{U}$. □

We now show a more general result.

Lemma 9.39. Every k-cell is compact (in $\mathbf{R}^k$).

Proof. We proceed in a similar manner to the proof the previous result.

Suppose I is a k-cell; that is,

$$I = \{(x_1, \ldots, x_k) \mid a_i \leqslant x_i \leqslant b_i,\ 1 \leqslant i \leqslant k\}.$$

Write $\mathbf{x} = (x_1, \ldots, x_k) \in \mathbf{R}^k$. Let

$$\delta = \left(\sum_{i=1}^{k} (b_i - a_i)^2 \right)^{\frac{1}{2}}$$

that is, δ is the distance between the points $(a_1, \ldots, a_k)$ and $(b_1, \ldots, b_k)$, which is the maximum distance between two points in I. Then

$$|\mathbf{x} - \mathbf{y}| \leqslant \delta \quad (\forall \mathbf{x}, \mathbf{y} \in I)$$

Suppose otherwise, for a contradiction, that I is not compact; that is, there exists an open cover $\mathcal{U} = \{U_i\}$ of I which contains no finite subcover of I.

For $1 \leqslant i \leqslant k$, let $c_i = \frac{1}{2}(a_i + b_i)$. The intervals $[a_i, c_i]$ and $[c_i, b_i]$ then determine 2^k k-cells Q_i whose union is I. At least one of these sets Q_i, call it I_1, cannot be covered by any finite subcollection of $\mathcal{U}$ (otherwise I could be so covered). We next subdivide I_1 and continue the process. We obtain a sequence (I_n) with the following properties:

(i) $I \supset I_1 \supset I_2 \supset \cdots$

(ii) I_n is not covered by any finite subcollection of $\mathcal{U}$

(iii) $|\mathbf{x} - \mathbf{y}| \leqslant 2^{-n}\delta$ for all $\mathbf{x}, \mathbf{y} \in I_n$

By (i) and Theorem 2.39, there is a point $\mathbf{x}'$ which lies in every I_n. For some i, $\mathbf{x}' \in U_i$. Since U_i is open, there exists $r > 0$ such that $|\mathbf{y} - \mathbf{x}'| < r$ implies that $y \in U_i$. If n is so large that $2^{-n}\delta < r$ (there is such an n, for otherwise $2^n \leqslant \frac{\delta}{r}$ for all positive integers n, which is absurd since $\mathbf{R}$ is archimedean), then (iii) implies that $I_n \subset U_i$, which contradicts (ii). □

We have now come to an important result, which will be crucial in proving the Heine–Borel theorem and Bolzano–Weierstrass theorem.

Proposition 9.40. If $E \subset \mathbf{R}^k$ has one of the following three properties, then it has the other two:

(i) E is closed and bounded.

(ii) E is compact.

(iii) Every infinite subset of E has a limit point in E.

Proof.

$\boxed{\text{(i)} \implies \text{(ii)}}$ Suppose E is closed and bounded. Since E is bounded, then $E \subset I$ for some k-cell I.

From Lemma 9.39, we have that I is compact. Since E is a closed subset of a compact set, by Proposition 9.33, E is compact.

$\boxed{\text{(ii)} \implies \text{(iii)}}$ This directly follows from Proposition 9.35.

$\boxed{\text{(iii)} \implies \text{(i)}}$ If E is not bounded, then E contains points $\mathbf{x}_n$ with

$$|\mathbf{x}_n| > n \quad (n = 1, 2, 3, \dots)$$

The set S consisting of these points $\mathbf{x}_n$ is infinite and clearly has no limit point in $\mathbf{R}^k$, hence has none in E. Thus (iii) implies that E is bounded.

If E is not closed, then there is a point $\mathbf{x}_0 \in \mathbf{R}^k$ which is a limit point of E but not a point of E. For $n = 1, 2, 3, \dots$, there are points $\mathbf{x}_n \in E$ such that $|\mathbf{x}_n - \mathbf{x}_0| < \frac{1}{n}$. Let S be the set of these points $\mathbf{x}_n$.

Then S is infinite (otherwise $|\mathbf{x}_n - \mathbf{x}_0|$ would have a constant positive value, for infinitely many n), S has $\mathbf{x}_0$ as a limit point, and S has no other limit point in $\mathbf{R}^k$. For if $\mathbf{y} \in \mathbf{R}^k$, $\mathbf{y} \neq \mathbf{x}_0$, then

$$\begin{aligned} |\mathbf{x}_n - \mathbf{y}| &\geqslant |\mathbf{x}_0 - \mathbf{y}| - |\mathbf{x}_n - \mathbf{x}_0| \\ &\geqslant |\mathbf{x}_0 - \mathbf{y}| - \frac{1}{n} \\ &\geqslant \frac{1}{2}|\mathbf{x}_0 - \mathbf{y}| \end{aligned}$$

for all but finitely many n; this shows that $\mathbf{y}$ is not a limit point of S (Theorem 2.20).

Thus S has no limit point in E; hence E must be closed if (iii) holds. □

review proof

Theorem 9.41 (Heine–Borel theorem)**.** $E \subset \mathbf{R}^n$ is compact if and only if E is closed and bounded.

Proof. This is simply (i) $\iff$ (ii) in the previous result. □

Bolzano–Weierstrass Theorem

Theorem 9.42 (Bolzano–Weierstrass theorem)**.** Every bounded infinite subset of $\mathbf{R}^n$ has a limit point in $\mathbf{R}^n$.

Proof. Suppose E is a bounded infinite subset of $\mathbf{R}^n$.

Since E is bounded, there exists an n-cell $I \subset \mathbf{R}^n$ such that $E \subset I$. Since I is compact, by Proposition 9.35, E has a limit point in I and thus $\mathbf{R}^n$. □

Cantor Intersection Theorem

A collection $\mathcal{A} = \{A_i \mid i \in I\}$ of subsets of X is said to have the *finite intersection property*, if the intersection of every finite subcollection of $\mathcal{A}$ is non-empty.

Proposition 9.43. Suppose $\mathcal{K} = \{K_i \mid i \in I\}$ is a collection of compact subsets of a metric space X, which satisfies the finite intersection property. Then $\bigcap_{i \in I} K_i \neq \emptyset$.

Proof. We fix a member $K_1 \subset \mathcal{K}$. Suppose otherwise, for a contradiction, that $\bigcap_{i \in I} K_i = \emptyset$; that is, no point of K_1 belongs to every $K_i \in \mathcal{K}$.

For $i \in I$, let $U_i = K_i^c$. Then the sets $\{U_i \mid i \in I\}$ form an open cover of K_1. Since K_1 is compact by assumption, there exist finitely many indices $i_1, \ldots, i_n$ such that

$$K_1 \subset \bigcup_{k=1}^{n} U_{i_k}.$$

By de Morgan's laws, we have that

$$\bigcup_{k=1}^{n} U_{i_k} = \bigcup_{k=1}^{n} {K_{i_k}}^c = \left(\bigcap_{k=1}^{n} K_{i_k} \right)^c .$$

Thus

$$K_1 \subset \left(\bigcap_{k=1}^{n} K_{i_k} \right)^c ,$$

which means that

$$K_1 \cap \bigcap_{k=1}^{n} K_{i_k} = \emptyset.$$

Thus $K_1, K_{i_1}, \ldots, K_{i_n}$ is a finite subcollection of $\mathcal{K}$ which has an empty intersection; this contradicts the finite intersection property of $\mathcal{K}$. □

Theorem 9.44 (Cantor's intersection theorem)**.** Suppose (K_n) is a decreasing sequence of non-empty compact sets; that is, $K_1 \supset K_2 \supset \cdots$. Then $\bigcap_{n=1}^{\infty} K_n \neq \emptyset$.

Proof. This follows from the previous result; it is obvious that the intersection of every finite subcollection of a decreasing sequence of sets must be non-empty. □

The following result is a characterisation of compact sets.

Proposition 9.45. K is compact if and only if every collection of closed subsets of K satisfies the finite intersection property.

Proof.

$\boxed{\Longrightarrow}$ Suppose K is compact.

If $\mathcal{U}$ is an open covering of K, then the collection $\mathcal{F}$ of complements of sets in $\mathcal{U}$ is a collection of closed sets whose intersection is empty (why?); and

conversely, if $\mathcal{F}$ is a collection of closed sets whose intersection is empty, then the collection $\mathcal{U}$ of complements of sets in $\mathcal{F}$ is an open covering.

To complete proof

□

Sequential Compactness

Definition 9.46 (Sequential compactness)**.** $K \subset X$ is ***sequentially compact*** if every sequence in K has a convergent subsequence in K.

We now show that compactness and sequential compactness are equivalent.

Proposition 9.47. $K \subset X$ is compact if and only if it is sequentially compact.

Proof.

$\boxed{\Longrightarrow}$ Suppose $K \subset X$ is compact. Take any sequence (y_n) from K. Suppose, for a contradiction, that every point $x \in K$ is not a limit of any subsequence of (y_n). Then for all $x \in K$, there exists $r_x > 0$ such that $B_{r_x}(x)$ contains at most one point in (y_n), which is x.

Consider the collection of open balls at each $x \in K$:

$$\{B_{r_x}(x) \mid x \in K\}.$$

This is an open cover of K. By the compactness of K, there exists a finite subcover of K:

$$\left\{B_{r_{x_1}}(x_1), \ldots, B_{r_{x_N}}(x_N)\right\}.$$

In particular, these open balls cover $\{y_n\}$. Hence there must be some x_i $(1 \leqslant i \leqslant N)$ such that there are infinitely many $y_j = x_i$. Consider the sequence (y_j) where each term in this sequence is equal to x_i; this is a subsequence of (y_n) that converges to $x_i \in K$. This contradicts the assumption.

$\boxed{\Longleftarrow}$ Suppose, for a contradiction, that K is not compact. Then there exists an open cover $\{U_\alpha \mid \alpha \in \Lambda_\alpha\}$ which has no finite subcover. Then Λ must be an infinite set.

If Λ is countable, WLOG, assume $\Lambda = \mathbf{N}$. Since any finite union

$$\bigcup_{i=1}^{n} U_i$$

cannot cover K, we can take some $x_n \in K \setminus \bigcup_{i=1}^{n} U_i$ for every $n \in \mathbf{N}$. Then we obtain a sequence (x_n) in K and so must have a convergent subsequence (x_{n_k}) that converges to some $x_0 \in K$. It follows that there must be some U_N such that $x_0 \in U_N$. Since U_N is open, there exists $r > 0$ such that

$$B_r(x_0) \subset U_N.$$

On the other hand, since $x_{n_k} \to x_0$, there exists $N' \in \mathbf{N}$ such that if $n_k \geqslant N'$ then

$$x_{n_k} \in B_r(x_0).$$

However, by our way of choosing x_n, whenever $n_k > \max\{N', N\}$, $x_{n_k} \notin U_N$. This leads to a contradiction. $\square$

§9.3 Perfect Sets

Definition 9.48 (Perfect set)**.** E is ***perfect*** if

(i) E is closed;

(ii) every point of E is a limit point of E.

Proposition 9.49. Let non-empty $P \subset \mathbf{R}^k$ be perfect. Then P is uncountable.

Proof. Since P has limit points, P must be infinite. Suppose, for a contradiction, that P is countable. This means we can list the points of P in a sequence:

$$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots$$

We now construct a sequence (B_n) of open balls, where B_n is any open ball centred at $\mathbf{x}_n$:

$$B_n = \{\mathbf{y} \in \mathbf{R}^k \mid |\mathbf{y} - \mathbf{x}_n| < r\}.$$

Then its closure $\overline{B}_n$ is the closed ball

$$\overline{B}_n = \{\mathbf{y} \in \mathbf{R}^k \mid |\mathbf{y} - \mathbf{x}_n| \leqslant r\}.$$

Suppose B_n has been constructed. Note that $B_n \cap P$ is not empty. Since P is perfect, every point of P is a limit point of P, so there exists is a neighborhood Vn+i such that (i) Yn + 1 C: vn , (ii) Xn ¢ Yn + 1, (iii) Vn + 1 n p is not empty. By (iii), Vn+i satisfies our induction hypothesis, and the construction can proceed. Put Kn = Yn n P. Since Yn is closed and bounded, Yn is compact. Since Xn ¢ Kn+l, no point of plies in n'? Kn. Since Kn C: P, this implies that nf Kn is empty. But each Kn is nonempty, by (iii), and Kn=> Kn+t, by (i); this contradicts the Corollary to Theorem 2.36. □

Corollary 9.50. Every interval $[a, b]$ is uncountable. In particular, $\mathbf{R}$ is uncountable.

Cantor Set

We now construct the Cantor set. Consider the interval

$$C_0 = [0, 1].$$

Remove the middle third $\left(\frac{1}{3}, \frac{2}{3}\right)$ to give

$$C_1 = \left[0, \frac{1}{3}\right] \cup \left[\frac{2}{3}, 1\right].$$

Remove the middle thirds of these intervals to give

$$C_2 = \left[0, \frac{1}{9}\right] \cup \left[\frac{2}{9}, \frac{3}{9}\right] \cup \left[\frac{6}{9}, \frac{7}{9}\right] \cup \left[\frac{8}{9}, 1\right].$$

Figure 9.7: Cantor set

Repeating this process, we obtain a monotonically decreasing sequence of compact sets (C_n), where C_n is the union of 2^n intervals, each of length 3^{-n}. Recursively, we have that $C_{n+1} = \frac{1}{3}C_n \cup \left(\frac{1}{3}C_n + \frac{2}{3}\right)$.

Note that each C_n has the following properties:

(i) closed (since each C_n is a finite union of closed sets, which is closed)

(ii) compact (since each C_n is a closed subset of a compact set $[a, b]$)

(iii) non-empty (since the endpoints 0 and 1 are in each C_n)

The ***Cantor set*** is defined to be the union

$$C \coloneqq \bigcap_{n=1}^{\infty} C_n.$$

Lemma 9.51 (Properties of the Cantor set)**.**

(i) C is closed.

(ii) C is compact.

(iii) C is not empty.

(iv) C has no interior points.

Proof.

(i) C is the intersection of arbitrarily many closed sets, so C is closed.

(ii) C is bounded in $[0, 1]$, by definition. Since C is closed and bounded, by the Heine–Borel theorem, C is compact.

(iii) Since (C_n) is a decreasing sequence of non-empty compact sets, by Cantor's intersection theorem, $\bigcap_{n=1}^{\infty} C_n = C \neq \emptyset$.

(iv) Suppose, for a contradiction, that there exists $p \in C$ which is an interior point. Then there exists some open interval around p, i.e. $p \in (a, b)$.

However in C_n, each interval has length $\frac{1}{3^n}$. Hence for any (a, b) we can find some $n \in \mathbf{N}$ such that (a, b) is not contained in C_n and hence not contained in C.

□

Proposition 9.52. C is a perfect set in $\mathbf{R}$ which contains no segment.

Proof. We prove (i) C contains no segment, and (ii) C is perfect.

(i) No segment of the form

$$\left(\frac{3k+1}{3^m}, \frac{3k+2}{3^m}\right),$$

where $k, m \in \mathbf{Z}^+$, has a point in common with C. Since every segment (α, β) contains a segment of the above form, if

$$3^{-m} < \frac{\beta - \alpha}{6},$$

C contains no segment.

(ii) Sincw we have shown that C is closed, it suffices to show that every point of C is a limit point of C.

Let $x \in C$, and let S be any segment containing x. Let I_n be that interval of C_n which contains x. Choose n large enough, so that $I_n \subset S$. Let x_n be an endpoint of I_n, such that $x_n \neq x$.

It follows from the construction of C that $x_n \in P$. Hence x is a limit point of C, and C is perfect.

□

Corollary 9.53. C is uncountable.

The following are a few more interesting properties of the Cantor set.

Proposition 9.54. C is precisely the set of all real numbers in $[0, 1]$ whose ternary expansion contain only 0's or and 2's.

Remark. Finite decimal expansions, as always, are not formally well-defined; for instance, $\frac{1}{3} = 0.1 = 0.0222\cdots$ so that $\frac{1}{3} \in C$ because it can be expressed with only zeros and twos in at least one of its ternary expansions.

Proof. □

Proposition 9.55. C has measure zero; that is, for all $\varepsilon > 0$, C can be covered by intervals of total length less than ε.

§9.4 Connectedness

Definition 9.56 (Connectedness)**.** A and B are ***separated*** if

(i) $A \cap \overline{B} = \emptyset$, and

(ii) $\overline{A} \cap B = \emptyset$;

that is, no point of A lies in the closure of B, and no point of B lies in the closure of A. (Equivalently, no point of one set is a limit point of the other set.)

$E \subset X$ is ***connected*** if E is not the union of two non-empty separated sets.

Remark. Separated sets are of course disjoint, but disjoint sets need not be separated. For example, the interval $[0, 1]$ and the segment $(1, 2)$ are not separated, since 1 is a limit point of $(1, 2)$. However, the segments $(0, 1)$ and $(1, 2)$ are separated.

Example

In $\mathbf{R}^2$, consider the set

$$E = \{(x, y) \mid x, y \in \mathbf{Q}\}.$$

Then E is not connected; if we let

$$A = \left\{(x, y) \mid x, y \in \mathbf{Q}, x < \sqrt{2}\right\},$$
$$B = \left\{(x, y) \mid x, y \in \mathbf{Q}, x > \sqrt{2}\right\},$$

then note that $A \cup B = E$, as well as $A \cap \overline{B} = \emptyset$ and $\overline{A} \cap B = \emptyset$.

Proposition 9.57. Closed intervals in $\mathbf{R}$ are connected.

Proof. Suppose otherwise, for a contradiction, that a closed interval $[a, b]$ is not connected. Then by definition, there exists non-empty sets A and B, with $A \cap \overline{B} = \emptyset$ and $\overline{A} \cap B = \emptyset$. WLOG let $a \in A$.

Let $s = \sup A$. Then by Proposition 9.26, $s \in \overline{A}$. Then $\overline{A} \cap B = \emptyset$ implies $s \notin B$, so $s \in A$. Thus $A \cap \overline{B} = \emptyset$ implies $s \notin \overline{B}$. Hence there exists an open interval $(s - \varepsilon, s + \varepsilon)$ around s that is disjoint from B. But since $A \cup B = [a, b]$, we must have $(s - \varepsilon, s + \varepsilon) \subset A$. This contradicts the fact that s is the supremum of A. □

Proposition 9.58. $E \subset \mathbf{R}$ is connected if and only if it has the following property: if $x, y \in E$ and $x < z < y$, then $z \in E$.

Proof.

$\boxed{\Longleftarrow}$ If there exists $x, y \in E$ and some $z \in (x, y)$ such that $z \notin E$, then $E = A_z \cup B_z$ where

$$A_z = E \cap (-\infty, z), \quad B_z = E \cap (z, \infty).$$

Since $x \in A_z$ and $y \in B_z$, A and B are non-empty. Since $A_z \subset (-\infty, z)$ and $B_z \subset (z, \infty)$, they are separated. Hence E is not connected.

$\boxed{\Longrightarrow}$ Suppose E is not connected. Then there are non-empty separated sets A and B such that $A \cup B = E$. Pick $x \in A$, $y \in B$, and WLOG assume that $x < y$. Define

$$z := \sup(A \cap [x, y].)$$

By □

Definition 9.59. We say that a metric space is ***disconnected*** if we can write it as the disjoint union of two nonempty open sets. We say that a space is ***connected*** if it is not disconnected.

If X is written as a disjoint union of two nonempty open sets U and V then we say that these sets ***disconnect*** X.

Example

If $X = [0, 1] \cup [2, 3] \subset \mathbf{R}$ then we have seen that both $[0, 1]$ and $[2, 3]$ are open in X. Since X is their disjoint union, X is disconnected.

The following lemma gives some equivalent ways to formulate the concept of connected space.

Lemma 9.60. The following are equivalent:

(i) X is connected.

(ii) If $f : X \to \{0, 1\}$ is a continuous function then f is constant.

(iii) The only subsets of X which are both open and closed are X and $\emptyset$.

(Here the set $\{0, 1\}$ is viewed as a metric space via its embedding in $\mathbf{R}$, or equivalently with the discrete metric.)

Proof. □

Frequently one has a metric space X and a subset E of it whose connectedness or otherwise one wishes to ascertain. To this end, it is useful to record the following lemma.

Lemma 9.61. Let $E \subset X$, considered as a metric space with the metric induced from X. Then E is connected if and only if the following is true: if U, V are open subsets of X, and $U \cap V \cap E = \emptyset$, then $E \subset U \cup V$ implies either $E \subset U$ or $E \subset V$.

Proof. □

We now turn to some basic properties of the notion of connectedness. These broadly conform with one's intuition about how connected sets should behave.

Lemma 9.62 (Sunflower lemma)**.** Let $\{E_i \mid i \in I\}$ be a collection of connected subsets of X such that $\bigcap_{i\in I} E_i \neq \emptyset$. Then $\bigcup_{i\in I} E_i$ is connected.

Proof. □

Proposition 9.63. The Cantor set C is totally disconnected.

Exercises

Problem 9.1. Prove that the following are metrics.

(i) On an arbitrary set X, define

$$d(x,y) = \begin{cases} 1 & (x \neq y) \\ 0 & (x = y) \end{cases}$$

(This is called the *discrete metric*.)

(ii) On $\mathbf{Z}$, define $d(x,y)$ to be 2^{-m}, where 2^m is the largest power of two dividing $x-y$. The triangle inequality holds in the following stronger form, known as the ultrametric property:

$$d(x,z) \leqslant \max\{d(x,y), d(y,z)\}.$$

Indeed, this is just a rephrasing of the statement that if 2^m divides both $x-y$ and $y-z$, then 2^m divides $x-z$.

(This is called the *2-adic metric*. The role of 2 can be replaced by any other prime p, and the metric may also be extended in a natural way to the rationals $\mathbf{Q}$.)

(iii) Let $\mathcal{G} = (V,E)$ be a connected graph. Define d on V as follows: $d(v,v) = 0$, and $d(v,w)$ is the length of the shortest path from v to w.

(This is known as the *path metric*.)

(iv) Let G be a group generated by elements a, b and their inverses. Define a distance on G as follows: $d(v,w)$ is the minimal k such that $v = wg_1 \cdots g_k$, where $g_i \in \{a, b, a^{-1}, b^{-1}\}$ for all i.

(This is known as the *word metric*.)

(v) Let $X = \{0,1\}^n$ (the boolean cube), the set of all strings of n zeroes and ones. Define $d(x,y)$ to be the number of coordinates in which x and y differ.

(This is known as the *Hamming distance*.)

(vi) Consider the set $P(\mathbf{R}^n)$ of one-dimensional subspaces of $\mathbf{R}^n$, that is to say lines through the origin. One way to define a distance on this set is to take, for lines L_1, L_2, the distance between L_1 and L_2 to be

$$d(L_1, L_2) = \sqrt{1 - \frac{|\langle v, w\rangle|^2}{\|v\|^2 \|w\|^2}},$$

where v and w are any non-zero vectors in L_1 and L_2 respectively.

When $n = 2$, the distance between two lines is $\sin\theta$ where θ is the angle between those lines.

(This is known as the *projective space*.)

Problem 9.2 (Product space)**.** If (X, d_X) and (Y, d_Y) are metric spaces, set

$$d_{X\times Y}\left((x_1,y_1),(x_2,y_2)\right) = \sqrt{d_X(x_1,x_2)^2 + d_Y(y_1,y_2)^2}.$$

for $x_1, x_2 \in X$, $y_1, y_2 \in Y$.

Prove that $d_{X\times Y}$ gives a metric on $X \times Y$; we call $X \times Y$ the *product space*.

Solution. Reflexivity and symmetry are obvious. Less clear is the triangle inequality. We need to prove that

$$\begin{aligned}&\sqrt{d_X(x_1,x_3)^2 + d_Y(y_1,y_3)^2} + \sqrt{d_X(x_3,x_2)^2 + d_Y(y_3,y_2)^2}\\ &\geqslant \sqrt{d_X(x_1,x_2)^2 + d_Y(y_1,y_2)^2}\end{aligned} \tag{1}$$

Write $a_1 = d_X(x_2,x_3)$, $a_2 = d_X(x_1,x_3)$, $a_3 = d_X(x_1,x_2)$ and similarly $b_1 = d_Y(y_2,y_3)$, $b_2 = d_Y(y_1,y_3)$ and $b_3 = d_Y(y_1,y_2)$. Thus we want to show

$$\sqrt{{a_2}^2 + {b_2}^2} + \sqrt{{a_1}^2 + {b_1}^2} \geqslant \sqrt{{a_3}^2 + {b_3}^2}. \tag{2}$$

To prove this, note that from the triangle inequality we have $a_1 + a_2 \geqslant a_3$, $b_1 + b_2 \geqslant b_3$. Squaring and adding gives

$${a_1}^2 + {b_1}^2 + {a_2}^2 + {b_2}^2 + 2(a_1a_2 + b_1b_2) \geqslant {a_3}^2 + {b_3}^2.$$

By Cauchy–Schwarz,

$$a_1a_2 + b_1b_2 \leqslant \sqrt{{a_1}^2 + {b_1}^2}\sqrt{{a_2}^2 + {b_2}^2}.$$

Substituting this into the previous line gives precisely the square of (2), and (1) follows. □

10 Numerical Sequences and Series

Tthis chapter will deal primarily with sequences and series in $\mathbf{R}$ (and also $\mathbf{C}$). The basic facts about convergence, however, are just as easily explained in a more general setting (metric spaces).

As usual, let (X, d) be a metric space.

§10.1 Sequences

Convergence

Definition 10.1 (Sequence)**.** A ***sequence*** (x_n) in X is a function $f : \mathbf{N} \to X$ which maps $n \mapsto x_n$.

Definition 10.2. The ***range*** of a sequence (x_n) is the set

$$\{a \in X \mid \exists n \in \mathbf{N}, a = x_n\}.$$

Note that the range of a sequence may be a finite set or it may be infinite. (x_n) is ***bounded*** if its range is bounded.

Definition 10.3 (Convergence)**.** A sequence (x_n) ***converges*** to $x \in X$, denoted by $x_n \to x$, if

$$\forall \varepsilon > 0, \quad \exists N \in \mathbf{N}, \quad \forall n \geqslant N, \quad d(x_n, x) < \varepsilon.$$

We call x a *limit* of (x_n). If (x_n) does not converge, it is said to *diverge*.

Remark. This limit process conveys the intuitive idea that x_n can be made arbitrarily close to x, provided that n is sufficiently large.

Remark. If $x_n \not\to x$, simply negate the definition for convergence:

$$\exists \varepsilon > 0, \quad \forall N \in \mathbf{N}, \quad \exists n \geqslant N, \quad d(x_n, x) \geqslant \varepsilon.$$

Remark. From the definition, the convergence of a sequence depends not only on the sequence itself, but also on the metric space X. For instance, the sequence given by $a_n = \frac{1}{n}$ converges in $\mathbf{R}$ (to 0), but fails to converge in $\mathbf{R}^+$. In cases of possible ambiguity, we shall specify "convergent in X" rather than "convergent".

> **Example**
>
> Show that $\frac{1}{n} \to 0$.

Solution. Fix $\varepsilon > 0$. By the Archimedian property, there exists $N \in \mathbf{N}$ such that $\frac{1}{N} < \varepsilon$. Take $N = \left\lfloor \frac{1}{\varepsilon} \right\rfloor + 1$. Then for all $n \geqslant N$,

$$\left| \frac{1}{n} - 0 \right| = \frac{1}{n} \leqslant \frac{1}{N} = \frac{1}{\left\lfloor \frac{1}{\varepsilon} \right\rfloor + 1} < \frac{1}{\frac{1}{\varepsilon}} = \varepsilon$$

as desired. Therefore $\frac{1}{n} \to 0$. □

A useful tip for finding the required N (in terms of ε) is to work backwards from the result we wish to show, as illustrated in the following example.

Example

Let $a_n = 1 + (-1)^n \frac{1}{\sqrt{n}}$. Show that $a_n \to 1$.

Before our proof, we aim to find some $N \in \mathbf{N}$ such that if $n \geqslant N$ then

$$\begin{aligned} |a_n - 1| &< \varepsilon \\ \frac{1}{\sqrt{n}} = \left| (-1)^n \frac{1}{\sqrt{n}} \right| &< \varepsilon \\ \frac{1}{n} &< \varepsilon^2 \\ n &> \frac{1}{\varepsilon^2} \end{aligned}$$

Hence take $N = \left\lfloor \frac{1}{\varepsilon^2} \right\rfloor + 1$.

Solution. Let $\varepsilon > 0$ be given. Take $N = \left\lfloor \frac{1}{\varepsilon^2} \right\rfloor + 1$. If $n \geqslant N$, then

$$\begin{aligned} |a_n - 1| = \left| (-1)^n \frac{1}{\sqrt{n}} \right| &= \frac{1}{\sqrt{n}} \\ &\leqslant \frac{1}{\sqrt{N}} = \frac{1}{\sqrt{\left\lfloor \frac{1}{\varepsilon^2} \right\rfloor + 1}} \\ &< \frac{1}{\sqrt{\frac{1}{\varepsilon^2}}} = \varepsilon \end{aligned}$$

as desired. Therefore $a_n \to 1$. □

Lemma 10.4 (Uniqueness of limit)**.** If a sequence converges, then its limit is unique.

Proof. Let (x_n) be a sequence in X. Suppose that $x_n \to x$ and $x_n \to x'$ for $x, x' \in X$. We will show that $x' = x$.

Let $\varepsilon > 0$ be given. Then there exists $N, N' \in \mathbf{N}$ such that

$$n \geqslant N \implies d(x_n, x) < \frac{\varepsilon}{2}$$

and

$$n \geqslant N' \implies d(x_n, x') < \frac{\varepsilon}{2}.$$

Take $N_1 \coloneqq \max\{N, N'\}$. If $n \geqslant N_1$, then both hold. By the triangle inequality,

$$\begin{aligned} d(x, x') &\leqslant d(x, x_n) + d(x_n, x') \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \end{aligned}$$

for all $\varepsilon > 0$. Hence $d(x, x') = 0$ and thus $x = x'$. □

Since the limit is unique, we can give it a notation.

Notation. If (x_n) converges to x, we denote $\lim_{n\to\infty} x_n = x$.

We now outline some important properties of convergent sequences in metric spaces.

Proposition 10.5. Let (x_n) be a sequence in X.

(i) $x_n \to x$ if and only if every open ball of x contains x_n for all but finitely many n.

(ii) If (x_n) converges, then (x_n) is bounded.

(iii) Suppose $E \subset X$. Then x is a limit point of E if and only if there exists a sequence (x_n) in $E \setminus \{x\}$ such that $x_n \to x$.

Proof.

(i) $\boxed{\implies}$ Suppose $x_n \to x$. Let $\varepsilon > 0$ be given, then there exists $N \in \mathbf{N}$ such that

$$n \geqslant N \implies d(x_n, x) < \varepsilon.$$

Corresponding to this ε, consider the open ball $B_\varepsilon(x)$. Then by definition, for $y \in X$,

$$d(y, x) < \varepsilon \implies y \in B_\varepsilon(x).$$

Hence $n \geqslant N$ implies $x_n \in B_\varepsilon(x)$.

$\boxed{\impliedby}$ Suppose every open ball of x contains all but finitely many of the x_n.

Let $\varepsilon > 0$ be given. Consider the open ball $B_\varepsilon(x)$. Since $B_\varepsilon(x)$ is a open ball of x, it will also eventually contain all x_n; that is, there exists $N \in \mathbf{N}$ such that if $n \geqslant N$, then $x_n \in B_\varepsilon(x)$, i.e. $d(x_n, x) < \varepsilon$. Hence $x_n \to x$.

(ii) Suppose $x_n \to x$. Let $\varepsilon > 0$ be given. Then there exists $N \in \mathbf{N}$ such that $n \geqslant N$ implies $d(x_n, x) < 1$. Now let

$$r = \max\{1, d(x_1, x), \ldots, d(x_N, x)\}.$$

Then $d(x_n, x) \leqslant r$ for $n = 1, 2, \ldots, N$, so the range of x_n is bounded by $B_r(x)$. Hence (x_n) is bounded.

(iii) $\boxed{\Longrightarrow}$ Suppose x is a limit point of E.

Consider a sequence of open balls $\left(B_{\frac{1}{n}}(x)\right)$, for $n \in \mathbf{N}$. Since x is a limit point, each open ball intersects with E at some point which is not x. We pick one such point x_n from each $B_{\frac{1}{n}}(x) \cap E$. Then

$$d(x_n, x) < \frac{1}{n}.$$

Let $\varepsilon > 0$ be given. Then by the Archimedian property, there exists $N \in \mathbf{N}$ such that $\frac{1}{N} < \varepsilon$. If $n \geqslant N$,

$$d(x_n, x) \leqslant \frac{1}{n} \leqslant \frac{1}{N} < \varepsilon,$$

which shows that $x_n \to x$.

$\boxed{\Longleftarrow}$ Suppose that there exists a sequence (x_n) in $E \setminus \{x\}$ such that $x_n \to x$. Then for each open ball $B_\varepsilon(x)$, we can find some $N \in \mathbf{N}$ such that if $n \in \mathbf{N}$ then

$$x_n \in B_\varepsilon(x).$$

Since $x_n \in E \setminus \{x\}$, this shows that x is a limit point of E.

□

Proposition 10.6 (Ordering)**.** Suppose (a_n) and (b_n) are convergent sequences, and $a_n \leqslant b_n$. Then

$$\lim_{n\to\infty} a_n \leqslant \lim_{n\to\infty} b_n.$$

Proof. Let $a = \lim_{n\to\infty} a_n$, $b = \lim_{n\to\infty} b_n$. Suppose, for a contradiction, that $a > b$.

Let $\varepsilon = a - b > 0$ be given. There exists $N_1, N_2 \in \mathbf{N}$ such that

$$n \geqslant N_1 \implies |a_n - a| < \frac{\varepsilon}{2},$$
$$n \geqslant N_2 \implies |b_n - b| < \frac{\varepsilon}{2}.$$

Let $N = \max\{N_1, N_2\}$, then $n \geqslant N$ implies

$$a_n > a - \frac{\varepsilon}{2}, \quad b_n < b + \frac{\varepsilon}{2}$$

and thus

$$a_n - b_n > a - b - \varepsilon = 0$$

so $a_n > b_n$, which is a contradiction. □

Remark. If $a_n < b_n$, we may not necessarily have $\lim_{n\to\infty} a_n < \lim_{n\to\infty} b_n$. For instance, $-\frac{1}{n} < \frac{1}{n}$ but their limits are both 0.

Proposition 10.7 (Arithmetic properties)**.** Suppose (a_n) and (b_n) are convergent seqeunces in $\mathbf{C}$; let $a = \lim_{n\to\infty} a_n$, $b = \lim_{n\to\infty} b_n$. Then

(i) $\lim_{n\to\infty} ca_n = ca$, where c is a constant (scalar multiplication)

(ii) $\lim_{n\to\infty}(a_n + b_n) = a + b$ (addition)

(iii) $\lim_{n\to\infty}(a_n b_n) = ab$ (multiplication)

(iv) $\lim_{n\to\infty}\frac{a_n}{b_n} = \frac{a}{b}$ $(b_n \neq 0, b \neq 0)$ (division)

Proof.

(i) The case where $c = 0$ is trivial. Now suppose $c \neq 0$. Let $\varepsilon > 0$ be given. Then there exists $N \in \mathbf{N}$ such that

$$n \geqslant N \implies |a_n - a| < \frac{\varepsilon}{|c|}.$$

Then if $n \geqslant N$,

$$|ca_n - ca| = |c|\,|a_n - a| < \varepsilon.$$

(ii) Let $\varepsilon > 0$ be given. Since $a_n \to a$ and $b_n \to b$, there exists $N_1, N_2 \in \mathbf{N}$ such that

$$n \geqslant N_1 \implies |a_n - a| < \frac{\varepsilon}{2},$$
$$n \geqslant N_2 \implies |b_n - b| < \frac{\varepsilon}{2}.$$

Let $N = \max\{N_1, N_2\}$, then $n \geqslant N$ implies

$$\begin{aligned}|(a_n + b_n) - (a + b)| &\leqslant |a_n - a| + |b_n - b| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.\end{aligned}$$

Hence $\lim_{n\to\infty}(a_n + b_n) = a + b$, as desired.

(iii) Write

$$a_n b_n - ab = (a_n - a)(b_n - b) + a(b_n - b) + b(a_n - a).$$

Let $\varepsilon > 0$ be given. Since $a_n \to a$ and $b_n \to b$, there exist $N_1, N_2 \in \mathbf{N}$ such that

$$n \geqslant N_1 \implies |a_n - a| < \sqrt{\varepsilon},$$
$$n \geqslant N_2 \implies |b_n - b| < \sqrt{\varepsilon}.$$

Let $N = \max\{N_1, N_2\}$. Then $n \geqslant N$ implies

$$|(a_n - a)(b_n - b)| < \varepsilon,$$

and thus $\lim_{n\to\infty}(a_n - a)(b_n - b) = 0$.

Note that $\lim_{n\to\infty} a(b_n - b) = \lim_{n\to\infty} b(a_n - a) = 0$. Hence

$$\lim_{n\to\infty}(a_n b_n - ab) = 0.$$

(iv) Since we have proven multiplication, it suffices to show that $\lim_{n\to\infty}\frac{1}{b_n} = \frac{1}{b}$.

Since $b_n \to b$, there exists $m \in \mathbf{N}$ such that

$$n \geqslant m \implies |b_n - b| < \frac{1}{2}|b|.$$

Let $\varepsilon > 0$ be given. There exists $N \in \mathbf{N}$, $N > m$ such that

$$n \geqslant N \implies |b_n - b| < \frac{1}{2}|b|^2\varepsilon.$$

Hence for $n \geqslant N$,

$$\left|\frac{1}{b_n} - \frac{1}{b}\right| = \left|\frac{b - b_n}{b_n b}\right| < \frac{2}{|b|^2}|b_n - b| < \varepsilon.$$

□

Proposition 10.8 (Squeeze theorem)**.** Let $a_n \leqslant c_n \leqslant b_n$ where (a_n) and (b_n) are convergent sequences such that $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = L$. Then (c_n) is also a convergent sequence, and

$$\lim_{n\to\infty} c_n = L.$$

Proof. Let $\varepsilon > 0$ be given. There exist $N_1, N_2 \in \mathbf{N}$ such that

$$n \geqslant N_1 \implies |a_n - L| < \varepsilon,$$
$$n \geqslant N_2 \implies |b_n - L| < \varepsilon.$$

In particular, we have

$$a_n > L - \varepsilon, \quad b_n < L + \varepsilon.$$

Let $N = \max\{N_1, N_2\}$. Then $n \geqslant N$ implies

$$L - \varepsilon < a_n \leqslant c_n \leqslant b_n < L + \varepsilon$$

or

$$|c_n - L| < \varepsilon.$$

Hence (c_n) is convergent, and $c_n \to L$. □

Subsequences

Definition 10.9 (Subsequence)**.** Given a sequence (x_n), consider a sequence (n_k) of positive integers such that $n_1 < n_2 < \cdots$. Then (x_{n_k}) is called a ***subsequence*** of (x_n).

If (x_{n_k}) converges, its limit is called a *subsequential limit* of (x_n).

Proposition 10.10. (x_n) converges to x if and only if every subsequence of (x_n) converges to x.

Proof.

$\boxed{\implies}$ Suppose $x_n \to x$. Let $\varepsilon > 0$ be give. Then there exists $N \in \mathbf{N}$ such that

$$n \geqslant N \implies d(x_n, x) < \varepsilon.$$

Every subsequence of (x_n) can be written in the form (x_{n_k}) where $n_1 < n_2 < \cdots$ is a strictly increasing sequence of positive integers. Pick M such that $n_M \geqslant N$. Then

$$k > M \implies n_k > n_M \geqslant N \implies d(x_{n_k}, x) < \varepsilon.$$

Hence every subsequence of (x_n) converges to x.

$\boxed{\impliedby}$ Suppose every subsequence of (x_n) converges to x. Since (x_n) is a subsequence of itself, we must have $x_n \to x$. □

Proposition 10.11. In a compact metric space, any sequence has a convergent subsequence.

Proof. Suppose (x_n) is a sequence in a compact metric space X.

Let E be the range of (x_n). We have to consider two cases: (i) E is finite, (ii) E is infinite.

(i) We prove by directly constructing the desired convergent subsequence.

Notice that there are infinitely many terms in the sequence (x_n), but only finitely many distinct terms in E. Hence by the pigeonhole principle, at least one term of E appears infinitely many times in the sequence. That is, there exists $x \in E$ and a sequence (n_k) with $n_1 < n_2 < \cdots$ such that

$$x_{n_1} = x_{n_2} = \cdots = x.$$

This subsequence (x_{n_k}) that we have constructed evidently converges to x.

(ii) If E is infinite, then E is an infinite subset of a compact set. By Proposition 9.35, E has a limit point $x \in X$.

We now construct a subsequence (x_{n_k}) of (x_n) such that $x_{n_k} \to x$. Choose n_1 so that $d(x, x_{n_1}) < 1$. Having chosen $n_1, \ldots, n_{k-1}$, choose n_k where $n_k > n_{k-1}$ such that $d(x, x_{n_k}) < \frac{1}{k}$ (such n_k exists due to Proposition 9.21). Then $x_{n_k} \to x$.

□

Corollary 10.12 (Bolzano–Weierstrass)**.** Every bounded sequence in $\mathbf{R}^k$ contains a convergent subsequence.

Proof. By Proposition 9.40, every bounded sequence in $\mathbf{R}^k$ lives in a compact subset of $\mathbf{R}^k$, and therefore it lives in a compact metric space. Hence by the previous result, it contains a convergent subsequence converging to a point in $\mathbf{R}^k$. □

Lemma 10.13. Suppose (x_n) is a sequence in X. Then the subsequential limits of (x_n) form a closed subset of X.

Proof. Let E be the set of all subsequential limits of (x_n), let q be a limit point of E. We want to show that $q \in E$.

Choose n_1 so that $x_{n_1} \neq q$. (If no such n_1 exists, then E has only one point, and there is nothing to prove.) Put $\delta = d(q, x_{n_1})$. Suppose $n_1, \ldots, n_{i-1}$ are chosen. Since q is a limit point of E, there is an $x \in E$ with $d(x, q) < 2^{-1}\delta$. Since $x \in E$, there is an $n_i > n_{i-1}$ such that $d(x, x_{n_k}) < 2^{-i}\delta$. Thus

$$d(q, x_{n_k}) < 2^{1-i}\delta$$

for $i = 1, 2, 3, \ldots$. This says that (x_{n_k}) converges to q. Hence $q \in E$. □

Cauchy Sequences

This is a 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*.

Definition 10.14 (Cauchy sequence)**.** A sequence (x_n) in X is a ***Cauchy sequence*** if

$$\forall \varepsilon > 0, \quad \exists N \in \mathbf{N}, \quad \forall n, m \geqslant N, \quad d(x_n, x_m) < \varepsilon.$$

Remark. Intuitively, we see that the distances between any two terms is sufficiently small after a certain point.

A natural question is regarding the relationship between convergent sequences and Cauchy sequences. We now address this.

Proposition 10.15.

(i) In any metric space, every convergent sequence is a Cauchy sequence.

(ii) If X is a compact metric space and if (x_n) is a Cauchy sequence in X, then (x_n) converges to some point of X.

(iii) In $\mathbf{R}^k$, every Cauchy sequence converges.

Remark. The converse of (i) is not true. For instance, the sequence $\{3, 3.1, 3.14, 3.141, 3.1415, \ldots\}$ is a Cauchy sequence but does not converge in $\mathbf{Q}$.

Proof.

(i) Suppose $x_n \to x$. Let $\varepsilon > 0$. There exists $N \in \mathbf{N}$ such that for all $n \geqslant N$,

$$d(x_n, x) < \frac{\varepsilon}{2}.$$

Then for all $n, m \geqslant N$,

$$d(x_n, x_m) \leqslant d(x_n, x) + d(x_m, x) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

as desired. Hence (x_n) is a Cauchy sequence.

(ii) Let (x_n) be a Cauchy sequence in X. Since X is compact, it is sequentially compact. Then there exists a subsequence (x_{n_k}) such that $x_{n_k} \to x$.

Claim. $x_n \to x$.

Let $\varepsilon > 0$. Since (x_n) is a Cauchy sequence, there exists $N_1 \in \mathbf{N}$ such that

$$n, m \geqslant N_1 \implies d(x_n - x_m) < \frac{\varepsilon}{2}.$$

$x_{n_k} \to x$ implies there exists $N_2 \in \mathbf{N}$ such that

$$n_k \geqslant N_2 \implies d(x_{n_k}, x) < \frac{\varepsilon}{2}.$$

Let $N = \max\{N_1, N_2\}$, fix some $n_k \geqslant N$. Then $n \geqslant N$ implies

$$d(x_n, x) \leqslant d(x_n, x_{n_k}) + d(x_{n_k}, x) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

(iii) Suppose (x_n) is a Cauchy sequence.

We perform three steps:

- We first show that (x_n) is bounded:
  Pick $N \in \mathbf{N}$ such that $|x_n - x_N| \leqslant 1$ for all $n \geqslant N$. Then

  $$|x_n| \leqslant \max\{1 + |x_N|, |x_1|, \ldots, |x_{N-1}|\}.$$

- Since (x_n) is bounded, by Bolzano–Weierstrass, (x_n) contains a subsequence (x_{n_k}) which converges to x.
- We now show that $x_n \to x$.
  Let $\varepsilon > 0$ be given. Since (x_n) is a Cauchy sequence, there exists $N_1 \in \mathbf{N}$ such that

  $$n, m \geqslant N_1 \implies |x_n - x_m| < \frac{\varepsilon}{2}.$$

  Since $x_{n_k} \to x$, there exists $M \in \mathbf{N}$ such that for all $k > M$,

  $$n_k > n_M \implies |x_{n_k} - x| < \frac{\varepsilon}{2}.$$

  Now since $n_1 < n_2 < \cdots$ is a sequence of strictly increasing positive integers, we can pick $i > M$ such that $n_k > N_1$. Then for all $n \geqslant N_1$, by setting $m = n_k$ we obtain

  $$|x_n - x_{n_k}| < \frac{\varepsilon}{2}, \quad |x_{n_k} - x| < \frac{\varepsilon}{2}.$$

Hence

$$|x_n - x| \leqslant |x_n - x_{n_k}| + |x_{n_k} - x| < \varepsilon.$$

Therefore (x_n) is convergent, and $x_n \to x$.

□

Definition 10.16. A metric space X is ***complete*** if every Cauchy sequence in X converges.

Remark. The above result shows that that all compact metric spaces and all Euclidean spaces are complete. It also implies that every closed subset E of a complete metric space X is complete. (Every Cauchy sequence in E is a Cauchy sequence in X, hence it converges to some $x \in X$, and actually $x \in E$ since E is closed.)

Example

The sequence (x_n) is defined as follows:

$$x_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}.$$

(x_n) does not converge in $\mathbf{R}$.

Proof. We claim that (x_n) is not a Cauchy sequence. WLOG assume $n > m$. Consider

$$|x_n - x_m| = \frac{1}{m+1} + \frac{1}{m+2} + \cdots + \frac{1}{n} \geqslant \frac{n-m}{n} = 1 - \frac{m}{n}.$$

Let $n = 2m$, then

$$|x_n - x_m| = |x_{2m} - x_m| > \frac{1}{2}.$$

Hence (x_n) is not a Cauchy sequence, so it does not converge. □

Monotonic Sequences

Definition 10.17 (Monotonic sequence)**.** A sequence (x_n) in $\mathbf{R}$ is

(i) *monotonically increasing* if $x_n \leqslant x_{n+1}$ for $n \in \mathbf{N}$;

(ii) *monotonically decreasing* if $x_n \geqslant x_{n+1}$ for $n \in \mathbf{N}$;

(iii) ***monotonic*** if it is either monotonically increasing or monotonically decreasing.

Lemma 10.18 (Monotone convergence theorem)**.** A monotonic sequence in $\mathbf{R}$ converges if and only if it is bounded.

Proof. We show the case for monotically increasing sequences; the case for monotonically decreasing sequences is similar.

$\boxed{\Longleftarrow}$ Suppose (x_n) is a monotonically increasing sequence bounded above. Let E be the range of x_n. By lub property of $\mathbf{R}$, E has a supremum in $\mathbf{R}$; let $x = \sup E$.

Claim. $x_n \to x$.

By definition of supremum, $x_n \leqslant x$ for all $n \in \mathbf{N}$. For every $\varepsilon > 0$, there exists $N \in \mathbf{N}$ such that

$$x - \varepsilon < x_N \leqslant x,$$

otherwise $x - \varepsilon$ would be an upper bound of E. Since (x_n) is monotically increasing, $n \geqslant N$ implies $x_N \leqslant x_n \leqslant x$, so

$$x - \varepsilon < x_n \leqslant x,$$

which implies $|x_n - x| < \varepsilon$. Hence $x_n \to x$. □

Limit Superior and Inferior

For divergent sequences, we have the following definition.

Definition 10.19. Suppose (x_n) is a sequence in $\mathbf{R}$. We write $x_n \to \infty$ if

$$\forall M \in \mathbf{R}, \quad \exists N \in \mathbf{N}, \quad \forall n \geqslant N, \quad x_n \geqslant M.$$

Similarly, we write $x_n \to -\infty$ if

$$\forall M \in \mathbf{R}, \quad \exists N \in \mathbf{N}, \quad \forall n \geqslant N, \quad x_n \leqslant M.$$

Definition 10.20. Suppose (x_n) is a sequence in $\mathbf{R}$. Let $E \subset \overline{\mathbf{R}}$ be the set of all subsequential limits of (x_n) (possibly including $+\infty$ and $-\infty$). Define

$$\limsup_{n\to\infty} x_n \coloneqq \sup E,$$
$$\liminf_{n\to\infty} x_n \coloneqq \inf E,$$

known as the ***limit superior*** and ***limit infimum*** of (x_n) respectively.

Remark. That is, limit superior is the "largest" subsequential limit; limit infimum is the "smallest" subsequential limit.

Remark. The limit superior and limit infimum exist due to the existence of supremum and infimum in $\overline{\mathbf{R}}$.

Lemma 10.21. Equivalently, we can define the limit superior (limit inferior) as the limit of supremum (infimum) of tails:

$$\limsup_{n\to\infty} x_n = \lim_{n\to\infty} \left(\sup_{k\geqslant n} x_k \right),$$
$$\liminf_{n\to\infty} x_n = \lim_{n\to\infty} \left(\inf_{k\geqslant n} x_k \right).$$

Proposition 10.22. Suppose (x_n) is a sequence in $\mathbf{R}$. Then

(i) $\limsup_{n\to\infty} x_n \in E$;

(ii) if $x > \limsup_{n\to\infty} x_n$, there exists $N \in \mathbf{N}$ such that $x_n < x$ for all $n \geqslant N$.

Moreover, $\limsup_{n\to\infty} x_n$ is the only number that satisfies (i) and (ii).

Proof.

(i) We consider three cases for the value of $\limsup_{n\to\infty} x_n$:

- If $\limsup_{n\to\infty} x_n = +\infty$, then $\sup E = +\infty$, so E is not bounded above. Hence (x_n) is not bounded above, so (x_n) has a subsequence (x_{n_k}) such that $x_{n_k} \to \infty$
- If $\limsup_{n\to\infty} x_n \in \mathbf{R}$, then $\sup E \in \mathbf{R}$, so E is bounded above. Hence at least one subsequential limit exists, so that (i) follows from Theorems 3.7 and 2.28.
- If $\limsup_{n\to\infty} x_n = -\infty$, then $\sup E = -\infty$, so E contains only one element, namely $-\infty$. Hence (x_n) has no subsequential limit. Thus for any $M \in \mathbf{R}$, $x_n > M$ for at most a finite number of values of n, so that $x_n \to -\infty$.

(ii) We prove by contradiction.

Suppose there is a number $x > \limsup_{n\to\infty} x_n$ such that $x_n \geqslant x$ for infinitely many values of n. In that case, there is a number $y \in E$ such that $y \geqslant x > \limsup_{n\to\infty} x_n$, contradicting the definition of $\limsup_{n\to\infty} x_n$.

We now show uniqueness. Suppose, for a contradiction, that two numbers p and q satisfy (i) and (ii). WLOG assume $p < q$. Then choose x such that $p < x < q$. Since p satisfies (i), we have $x_n < x$ for all $n \geqslant N$. But then q cannot satisfy (i). □

Of course, an analogous result is true for $\liminf_{n\to\infty} x_n$.

Example

- Let (x_n) be a sequence containing all rationals. Then every real number is a subsequential limit, and
$$\limsup_{n\to\infty} x_n = +\infty, \quad \liminf_{n\to\infty} = -\infty.$$
- Let $x_n = \dfrac{(-1)^n}{1+\frac{1}{n}}$. Then
$$\limsup_{n\to\infty} x_n = 1, \quad \liminf_{n\to\infty} x_n = -1.$$
- For a seqeunce (x_n) in $\mathbf{R}$, $x_n \to x$ if and only if
$$\limsup_{n\to\infty} x_n = \liminf_{n\to\infty} x_n = x.$$

Proposition 10.23. If $a_n \leqslant b_n$ for $n \geqslant N$ where N is fixed, then

$$\liminf_{n\to\infty} a_n \leqslant \liminf_{n\to\infty} b_n,$$
$$\limsup_{n\to\infty} a_n \leqslant \limsup_{n\to\infty} b_n.$$

Proposition 10.24 (Arithmetic properties)**.**

(i) If $k > 0$, $\limsup_{n\to\infty} ka_n = k \limsup_{n\to\infty} a_n$.
If $k < 0$, $\limsup_{n\to\infty} ka_n = k \liminf_{n\to\infty} a_n$.

(ii) $\limsup(a_n + b_n) \leqslant \limsup a_n + \limsup b_n$

Moreover, $\limsup_{n\to\infty}(a_n + b_n)$ may be bounded from below as follows:

$$\limsup_{n\to\infty}(a_n + b_n) \geqslant \limsup_{n\to\infty} a_n + \liminf_{n\to\infty} b_n.$$

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

§10.2 Series

Definition 10.25 (Series)**.** Given a sequence (a_n), we associate a sequence (s_n), where

$$s_n = \sum_{k=1}^{n} a_k = a_1 + a_2 + \cdots + a_n,$$

where the term s_n is called the *n-th partial sum*. The sequence (s_n) is often written as

$$\sum_{n=1}^{\infty} a_n,$$

which we call a ***series***.

Definition 10.26 (Convergence of series)**.** We say that the series *converges* if $s_n \to s$ (the sequence of partial sums converges), and write $\sum_{n=1}^{\infty} a_n = s$; that is,

$$\forall \varepsilon > 0, \quad \exists N \in \mathbf{N}, \quad \forall n \geqslant N, \quad \left| \sum_{k=1}^{n} a_k - s \right| < \varepsilon.$$

The number s is called the *sum* of the series. If (s_n) diverges, the series is said to *diverge*.

Notation. When there is no possible ambiguity, we write $\sum_{n=1}^{\infty} a_n$ simply as $\sum a_n$.

The Cauchy criterion can be restated in the following form:

Proposition 10.27 (Cauchy criterion)**.** $\sum a_n$ converges if and only if

$$\forall \varepsilon > 0, \quad \exists N \in \mathbf{N}, \quad \forall n \geqslant m \geqslant N, \quad \left| \sum_{k=m}^{n} a_k \right| \leqslant \varepsilon.$$

Convergence Tests

To determine the convergence of a series, apart from using the definition and the Cauchy criterion, we also have the following methods:

- Divergence test (Lemma 10.28)
- Boundedness of partial sums (Lemma 10.29, for series of non-negative terms)
- Comparison test (Lemma 10.30)
- Root test (Lemma 10.31)
- Ratio test (Lemma 10.32)
- Absolute convergence (Lemma 10.33)

Lemma 10.28 (Divergence test)**.** If $a_n \not\to 0$, then $\sum a_n$ diverges.

Proof. We prove the contrapositive: if $\sum a_n$ converges, then $a_n \to 0$.

In the Cauchy criterion, take $m = n$, then $|a_n| \leqslant \varepsilon$ for all $n \geqslant N$. □

Remark. The converse is not true; a counterexample of the harmonic series.

Lemma 10.29. A series of non-negative terms converges if and only if its partial sums form a bounded sequence.

Proof. Partial sums are monotonically increasing. But bounded monotonic sequences converge. □

Lemma 10.30 (Comparison test)**.** Consider two sequences (a_n) and (b_n).

(i) Suppose $|a_n| \leqslant b_n$ for all $n \geqslant N_0$ (where N_0 is some fixed integer). If $\sum b_n$ converges, then $\sum a_n$ converges.

(ii) Suppose $a_n \geqslant b_n \geqslant 0$ for all $n \geqslant N_0$. If $\sum b_n$ diverges, then $\sum a_n$ diverges.

Proof.

(i) Since $\sum b_n$ converges, by the Cauchy criterion, fix $\varepsilon > 0$, there exists $N \in \mathbf{N}$, $N \geqslant N_0$ such that for $n \geqslant m \geqslant N$,

$$\sum_{k=m}^{n} b_k \leqslant \varepsilon.$$

By the triangle inequality,

$$\left|\sum_{k=m}^{n} a_k\right| \leqslant \sum_{k=m}^{n} |a_k| \leqslant \sum_{k=m}^{n} b_k \leqslant \varepsilon,$$

so $\sum a_n$ converges, by the Cauchy criterion.

(ii) We prove the contrapositive. If $\sum a_n$ converges, and since $|b_n| \leqslant a_n$ for all $n \geqslant N_0$, then by (i), $\sum b_n$ converges.

□

To employ the comparison test, we need to be familiar with several series whose convergence or divergence is known.

Example (Geometric series)

A geometric series takes the form

$$\sum_{n=0}^{\infty} x^n.$$

Proposition.

(i) If $|x| < 1$, then $\sum x^n$ converges;

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}.$$

(ii) If $|x| \geqslant 1$, then $\sum x^n$ diverges.

Proof.

(i) For $|x| < 1$, the n-th partial sum is given by

$$\sum_{k=0}^{n} x^k = 1 + x + x^2 + \cdots + x^n. \tag{1}$$

Multiplying both sides of (1) by x gives

$$x \sum_{k=0}^{n} x^k = x + x^2 + x^3 \cdots + x^{n+1}. \tag{2}$$

Taking the difference of (1) and (2),

$$(1-x) \sum_{k=0}^{n} x^k = 1 - x^{n+1}$$

a nd so

$$\sum_{k=0}^{n} x^k = \frac{1 - x^{n+1}}{1-x}.$$

Taking limits $n \to \infty$, the result follows.

(ii) For $|x| \geqslant 1$, $x^n \not\to 0$. By the divergence test, $\sum x^n$ diverges.

□

Example (p-series)

A p-series takes the form

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

To determine the convergence of p-series, we first prove the following lemma, which states that a rather "thin" subsequence of (a_n) determines the convergence of $\sum a_n$.

Lemma (Cauchy condensation test)**.** Suppose $a_1 \geqslant a_2 \geqslant \cdots \geqslant 0$. Then $\sum a_n$ converges if and only if the series

$$\sum_{k=0}^{\infty} 2^k a_{2^k} = a_1 + 2a_2 + 4a_4 + \cdots$$

converges.

Proof. Let s_n and t_k denote the n-th partial sum of (a_n) and the k-th partial sum of $(2^k a_{2^k})$ re-

spectively; that is,

$$\begin{aligned} s_n &= a_1 + a_2 + \cdots + a_n, \\ t_k &= a_1 + 2a_2 + \cdots + 2^k a_{2^k}. \end{aligned}$$

We consider two cases:

- For $n < 2^k$, group terms to give

$$\begin{aligned} s_n &= a_1 + a_2 + \cdots + a_n \\ &\leqslant a_1 + (a_2 + a_3) + \cdots + (a_{2^k} + \cdots + a_{2^{k+1}-1}) \\ &\leqslant a_1 + 2a_2 + \cdots + 2^k a_{2^k} \\ &= t_k. \end{aligned}$$

  By comparison test, if (t_k) converges, then (s_n) converges.

- For $n > 2^k$,

$$\begin{aligned} s_n &\geqslant a_1 + a_2 + (a_3 + a_4) + \cdots + (a_{2^{k-1}+1} + \cdots + a_{2^k}) \\ &\geqslant \frac{1}{2} a_1 + a_2 + 2a_4 + \cdots + 2^{k-1} a_{2^k} \\ &= \frac{1}{2} t_k. \end{aligned}$$

  By comparison test, if (s_n) converges, then (t_k) converges.

□

Proposition (p-test)**.**

(i) If $p > 1$, $\sum \frac{1}{n^p}$ converges.

(ii) If $p \leqslant 1$, $\sum \frac{1}{n^p}$ diverges.

Proof. Note that if $p \leqslant 0$, then $\frac{1}{n^p} \not\to 0$. By the divergence test, $\sum \frac{1}{n^p}$ diverges.

If $p > 0$, we want to apply the above lemma. Consider the series

$$\sum_{k=0}^{\infty} 2^k \cdot \frac{1}{(2^k)^p} = \sum_{k=0}^{\infty} 2^{(1-p)k} = \sum_{k=0}^{\infty} \left(2^{1-p}\right)^k,$$

which is a geometric series. Hence the above series converges if and only if $|2^{1-p}| < 1$, which holds if and only if $1 - p < 0$. Then apply the above lemma to conclude the convergence of $\frac{1}{n^p}$. □

Remark. If $p = 1$, the resulting series is known as the *harmonic series* (which diverges). If $p = 2$, the resulting series converges, and the sum of this series is $\frac{\pi^2}{6}$ (Basel problem).

Example (The number e)

Consider the series

$$\sum_{n=0}^{\infty} \frac{1}{n!}.$$

Claim. The above series converges.

Consider the n-th partial sum:

$$\begin{aligned}\sum_{k=0}^{n} \frac{1}{k!} &= \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{n!} \\ &\leqslant 1 + 1 + \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^{n-1}} \\ &< 1 + 1 + \frac{1}{2} + \frac{1}{2^2} + \cdots = 3.\end{aligned}$$

Since the partial sums are bounded (by 3), and the terms are non-negative, the series converges. Then we can make the following definition for the sum of the series:

$$e \coloneqq \sum_{n=0}^{\infty} \frac{1}{n!}$$

Proposition. e is irrational.

Proof. Suppose, for a contradiction, that e is rational. Then $e = \frac{p}{q}$, where p and q are positive integers. Let s_n denote the n-th partial sum:

$$s_n = \sum_{k=0}^{n} \frac{1}{k!}.$$

Then

$$\begin{aligned}e - s_n &= \frac{1}{(n+1)!} + \frac{1}{(n+2)!} + \frac{1}{(n+3)!} + \cdots \\ &< \frac{1}{(n+1)!}\left(1 + \frac{1}{n+1} + \frac{1}{(n+1)^2} + \cdots\right) \\ &= \frac{1}{(n+1)!} \cdot \frac{n+1}{n} = \frac{1}{n!n}\end{aligned}$$

and thus

$$0 < e - s_n < \frac{1}{n!n}.$$

Taking $n = q$ and multiplying both sides by $q!$ gives

$$0 < q!(e - s_q) < \frac{1}{q}.$$

Note that $q!e$ is an integer (by assumption), and

$$q!s_q = q!\left(1 + 1 + \frac{1}{2!} + \cdots + \frac{1}{q!}\right)$$

is an integer, so $q!(e - s_n)$ is an integer. Since $q \geqslant 1$, this implies the existence of an integer between 0 and 1, which is absurd. Hence we have reached a contradiction. □

Lemma. e is equivalent to the following:

$$\lim_{n\to\infty}\left(1 + \frac{1}{n}\right)^n = e.$$

Proof. Let

$$s_n = \sum_{k=0}^{n} \frac{1}{k!}, \quad t_n = \left(1 + \frac{1}{n}\right)^n.$$

By the binomial theorem,

$$t_n = 1 + 1 + \frac{1}{2!}\left(1 - \frac{1}{n}\right) + \frac{1}{3!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) + \cdots + \frac{1}{n!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{n-1}{n}\right).$$

Comparing term by term, we see that $t_n \leqslant s_n$. By Proposition 10.23, we have that

$$\limsup_{n\to\infty} t_n \leqslant \limsup_{n\to\infty} s_n = e.$$

Next, if $n \geqslant m$,

$$t_n \geqslant 1 + 1 + \frac{1}{2!}\left(1 - \frac{1}{n}\right) + \cdots + \frac{1}{m!}\left(1 - \frac{1}{n}\right)\cdots\left(1 - \frac{m-1}{n}\right).$$

Let $n \to \infty$, keeping m fixed. We get

$$\liminf_{n\to\infty} t_n \geqslant 1 + 1 + \frac{1}{2!} + \cdots + \frac{1}{m!},$$

so that

$$s_m \leqslant \liminf_{n\to\infty} t_n.$$

Letting $m \to \infty$, we finally get

$$e \leqslant \liminf_{n\to\infty} t_n.$$

□

Lemma 10.31 (Root test)**.** Given $\sum a_n$, put $\alpha = \limsup_{n\to\infty} \sqrt[n]{|a_n|}$. Then

(i) if $\alpha < 1$, $\sum a_n$ converges;

(ii) if $\alpha > 1$, $\sum a_n$ diverges;

(iii) if $\alpha = 1$, the test gives no information.

Proof.

(i) If $\alpha > 1$, we can choose β so that $\alpha < \beta < 1$, and $n \in \mathbf{N}$ such that for all $n \geqslant N$,

$$\sqrt[n]{|a_n|} < \beta.$$

by Theorem 3.17(b). Since $0 < \beta < 1$, $\sum \beta^n$ converges. Hence by the comparison test, $\sum a_n$ converges.

(ii) If $\alpha > 1$, by Theorem 3.17, there is a sequence (n_k) such that

$$\sqrt[n_k]{|a_{n_k}|} \to \alpha.$$

Hence $|a_n| > 1$ for infinitely many values of n so that the condition $a_n \to 0$, necessary for convergence of $\sum a_n$, does not hold (Theorem 3.23).

(iii) Consider the series $\sum \frac{1}{n}$ and $\sum \frac{1}{n^2}$. For each of these series $\alpha = 1$, but the first diverges, the second converges. Hence the condition that $\alpha = 1$ does not give us information on the convergence of a series.

□

Lemma 10.32 (Ratio test)**.** The series $\sum a_n$

(i) converges if $\limsup_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| < 1$;

(ii) diverges if $\left|\frac{a_{n+1}}{a_n}\right| \geqslant 1$ for all $n \geqslant n_0$, where n_0 is some fixed integer.

Proof.

(i) If $\limsup_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| < 1$, there exists $\beta < 1$ and $N \in \mathbf{N}$ such tht for all $n \geqslant N$,

$$\left|\frac{a_{n+1}}{a_n}\right| < \beta.$$

In particular, from $n = N$ to $n = N + p$,

$$\begin{aligned} |a_{N+1}| &< \beta |a_N| \\ |a_{N+2}| &< \beta |a_{N+1}| < \beta^2 |a_N| \\ &\vdots \\ |a_{N+p}| &< \beta^p |a_N| \end{aligned}$$

Hence for all $n \geqslant N$,

$$|a_n| < |a_N| \beta^{-N} \cdot \beta^n.$$

Since $\sum \beta^n$ converges, by the comparison test, $\sum a_n$ converges.

(ii) Suppose $\left|\frac{a_{n+1}}{a_n}\right| \geqslant 1$ for all $n \geqslant n_0$, where n_0 is some fixed integer. Then $|a_{n+1}| \geqslant |a_n|$ for $n \geqslant n_0$, and it is easily seen that $a_n \not\to 0$, so $\sum a_n$ diverges.

□

The series $\sum a_n$ is said to *converge absolutely* if the series $\sum |a_n|$ converges.

Lemma 10.33 (Absolute convergence)**.** If $\sum a_n$ converges absolutely, then $\sum a_n$ converges.

Proof. □

Example (Power series)

Given a sequence (c_n) of complex numbers, the series

$$\sum_{n=0}^{\infty} c_n z^n$$

is called a ***power series***. The numbers c_n are called the *coefficients* of the series.

In general, the series will converge or diverge, depending on the choice of z. More specifically, with every power series there is associated a circle, the circle of convergence, such that $\sum c_n z^n$ converges if z is in the interior of the circle and diverges if z is in the exterior.

Proposition. Given the power series $\sum c_n z^n$, let

$$\alpha = \limsup_{n\to\infty} \sqrt[n]{|c_n|}, \quad R = \frac{1}{\alpha}.$$

(If $\alpha = 0$, $R = +\infty$; if $\alpha = +\infty$, $R = 0$.) Then $\sum c_n z^n$

(i) converges if $|z| < R$,

(ii) diverges if $|z| > R$.

R is called the *radius of convergence* of $\sum c_n z^n$.

Proof. Put $a_n = c_n z^n$, then apply the root test:

$$\begin{aligned}\limsup_{n\to\infty} \sqrt[n]{|a_n|} &= \limsup_{n\to\infty} \sqrt[n]{|c_n z^n|} \\ &= |z| \limsup_{n\to\infty} \sqrt[n]{|c_n|} \\ &= |z|\alpha \\ &= \frac{|z|}{R}.\end{aligned}$$

(i) If $|z| < R$, then $\limsup_{n\to\infty} \sqrt[n]{|a_n|} < 1$. By the root test, $\sum c_n z^n$ converges.

(ii) If $|z| > R$, then $\limsup_{n\to\infty} \sqrt[n]{|a_n|} > 1$. By the root test, $\sum c_n z^n$ diverges.

□

Further properties of power series will be discussed in **??**.

Summation by Parts

Proposition 10.34 (Partial summation formula)**.** Given two sequences (a_n) and (b_n), put

$$A_n = \sum_{k=0}^{n} a_k$$

if $n \geqslant 0$; put $A_{-1} = 0$. Then, if $0 \leqslant p \leqslant q$, we have

$$\sum_{n=p}^{q} a_n b_n = \sum_{n=p}^{q-1} A_n (b_n - b_{n+1}) + A_q b_q - A_{p-1} b_p.$$

Proof. The RHS can be written as

$$\begin{aligned}
&\sum_{n=p}^{q-1} A_n b_n + A_q b_q - \sum_{n=p}^{q-1} A_n b_{n+1} - A_{p-1} b_p \\
&= \sum_{n=p}^{q} A_n b_n - \sum_{n=p-1}^{q-1} A_n b_{n+1} \\
&= \sum_{n=p}^{q} A_n b_n - \sum_{n=p}^{q} A_{n-1} b_n \\
&= \sum_{n=p}^{q} (A_n - A_{n-1})\, b_n \\
&= \sum_{n=p}^{q} a_n b_n
\end{aligned}$$

which is equal to the LHS. □

Proposition 10.35. Suppose the partial sums A_n of $\sum a_n$ form a bounded sequence, $b_0 \geqslant b_1 \geqslant b_2 \geqslant \cdots$, and $\lim_{n\to\infty} b_n = 0$. Then $\sum a_n b_n = 0$.

Proof. □

Proposition 10.36. Suppose $|c_1| \geqslant |c_2| \geqslant |c_3| \geqslant \cdots$, $c_{2m-1} \geqslant 0, c_{2m} \leqslant 0$ for $m = 1, 2, 3, \ldots$, and $\lim_{n\to\infty} c_n = 0$. Then $\sum c_n$ converges.

Addition and Multiplication of Series

Proposition 10.37. If $\sum a_n = A$ and $\sum b_n = B$, then

(i) $\sum(a_n + b_n) = A + B$,

(ii) $\sum ca_n = cA$ for any fixed c.

Proof.

(i) Let $A_n = \sum_{k=0}^{n} a_k$, $B_n = \sum_{k=0}^{n} b_k$. Then

$$A_n + B_n = \sum_{k=0}^{n}(a_k + b_k).$$

Since $\lim_{n\to\infty} A_n = A$ and $\lim_{n\to\infty} B_n = B$, we see that

$$\lim_{n\to\infty}(A_n + B_n) = A + B.$$

(ii)

□

Thus two convergent series may be added term by term, and the resulting series converges to the sum of the two series. The situation becomes more complicated when we consider multiplication of two series. To begin with, we have to define the product. This can be done in several ways; we shall consider the so-called "Cauchy product".

Definition 10.38 (Cauchy product)**.** Given $\sum a_n$ and $\sum b_n$, let

$$c_n = \sum_{k=0}^{n} a_k b_{n-k} \quad (n = 0, 1, 2, \dots)$$

We call $\sum c_n$ the *product* of the two given series.

This definition may be motivated as follows. If we take two power series $\sum a_n z^n$ and $\sum b_n z^n$, multiply them term by term, and collect terms containing the same power of z, we get

$$\begin{aligned}\sum_{n=0}^{\infty} a_n z^n \cdot \sum_{n=0}^{\infty} b_n z^n &= \left(a_0 + a_1 z + a_2 z^2 + \cdots\right)\left(b_0 + b_1 z + b_2 z^2 + \cdots\right) \\ &= a_0 b_0 + (a_0 b_1 + a_1 b_0) z + (a_0 b_2 + a_1 b_1 + a_2 b_0) z^2 + \cdots \\ &= c_0 + c_1 z + c_2 z^2.\end{aligned}$$

Setting $z = 1$, we arrive at the above definition.

Theorem 10.39 (Mertens)**.** Suppose $\sum a_n = A$, $\sum b_n = B$, $\sum a_n$ converges absolutely. Then their Cauchy product converges to AB.

That is, the product of two convergent series converges, and to the right value, if at least one of the two series converges absolutely.

Proof. Let $A_n = \sum_{k=0}^{n} a_k$, $B_n = \sum_{k=0}^{n} b_k$, $C_n = \sum_{k=0}^{n} c_k$. Also let $\beta_n = B_n - B$. Then

$$\begin{aligned} C_n &= a_0b_0 + (a_0b_1 + a_1b_0) + \cdots + (a_0b_n + a_1b_{n-1} + \cdots + a_nb_0) \\ &= a_0B_n + a_1B_{n-1} + \cdots + a_nB_0 \\ &= a_0(B + \beta_n) + a_1(B + \beta_{n-1}) + \cdots + a_n(B + \beta_0) \\ &= A_nB + a_0\beta_n + a_1\beta_{n-1} + \cdots + a_n\beta_0. \end{aligned}$$

Let

$$\gamma_n = a_0\beta_n + a_1\beta_{n-1} + \cdots + a_n\beta_0.$$

We wish to show that $C_n \to AB$. Since $A_nB \to AB$, it suffices to show that $\lim_{n\to\infty} \gamma_n = 0$.

Let

$$\alpha = \sum_{n=0}^{\infty} |a_n|.$$

Let $\varepsilon > 0$. Since $B_n \to B$, $\beta_n \to 0$. Hence we can choose $N \in \mathbf{N}$ such that for all $n \geqslant N$, $|\beta_n| \leqslant \varepsilon$, in which case

$$\begin{aligned} |\gamma_n| &= |\beta_0a_n + \cdots + \beta_Na_{n-N}| + |\beta_{N+1}a_{n-N}a_{n-N-1} + \cdots + \beta_na_0| \\ &\leqslant |\beta_0a_n + \cdots + \beta_Na_{n-N}| + \varepsilon\alpha. \end{aligned}$$

Keeping N fixed, and letting $n \to \infty$, we get

$$\limsup_{n\to\infty} |\gamma_n| \leqslant \varepsilon\alpha,$$

sine $a_k \to 0$ as $k \to \infty$. Since ε is arbitrary, we have $\lim_{n\to\infty} \gamma_n = 0$, as desired. □

Theorem 10.40 (Abel)**.** Let the series $\sum a_n$, $\sum b_n$, $\sum c_n$ converge to A, B, C respectively, and $\sum c_n$ is the Cauchy product of $\sum a_n$ and $\sum b_n$. Then $C = AB$.

Rearrangements

Definition 10.41 (Rearrangement)**.** Let (k_n) be a sequence in which every positive integer appears once and only once. Putting

$$a'_n = a_{k_n} \quad (\forall n \in \mathbf{N})$$

we say that $\sum a'_n$ is a *rearrangement* of $\sum a_n$.

Proposition 10.42. Let $\sum a_n$ be a series of real numbers which converges, but not absolutely. Suppose $-\infty \leqslant \alpha \leqslant \beta \leqslant \infty$. Then there exists a rearrangement $\sum a'_n$ with partial sums s'_n such that

$$\liminf_{n\to\infty} s'_n = \alpha, \quad \limsup_{n\to\infty} s'_n = \beta.$$

Proposition 10.43. If $\sum a_n$ is a series of complex numbers which converges absolutely, then every rearrangement of $\sum a_n$ converges, and they all converge to the same sum.

Exercises

Problem 10.1. Show the following:

(i) $\lim_{n\to\infty} \dfrac{1}{n^p} = 0 \ (p > 0)$

(ii) $\lim_{n\to\infty} \sqrt[n]{p} = 1 \ (p > 0)$

(iii) $\lim_{n\to\infty} \sqrt[n]{n} = 1$

(iv) $\lim_{n\to\infty} \dfrac{n^\alpha}{(1+p)^n} = 0 \ (p > 0, \alpha \in \mathbf{R})$

(v) $\lim_{n\to\infty} x^n = 0 \ (|x| < 1)$

Solution.

(i) Let $\varepsilon > 0$ be given. Take $N = \left\lfloor \left(\frac{1}{\varepsilon}\right)^{\frac{1}{p}} \right\rfloor + 1$. Then $n \geqslant N$ implies

$$\left|\frac{1}{n^p} - 0\right| = \frac{1}{n^p} \leqslant \frac{1}{N^p} < \frac{1}{\left(\left(\frac{1}{\varepsilon}\right)^{\frac{1}{p}}\right)^p} = \varepsilon.$$

(ii) We need to consider the cases when $p > 1$, $p = 1$, and $0 < p < 1$.

If $p > 1$,

(iii)

(iv)

(v)

□

Problem 10.2. Let (x_n) be a sequence in $\mathbf{R}$, let $\alpha \geqslant 2$ be a constant. Define the sequence (y_n) as follows:

$$y_n = x_n + \alpha x_{n+1} \quad (n = 1, 2, \dots)$$

Show that if (y_n) is convergent, then (x_n) is also convergent.

Problem 10.3 (Contractive sequence)**.** A sequence (x_n) in $\mathbf{R}$ is *contractive* if there exists $k \in [0, 1)$ such that

$$|x_{n+2} - x_{n+1}| \leqslant k|x_{n+1} - x_n| \quad (\forall n \in \mathbf{N})$$

Show that every contractive sequence is convergent.

Solution. By induction on n, we have

$$|a_{n+1} - a_n| \leqslant k^{n-1}|a_2 - a_1| \quad (\forall n \in \mathbf{N})$$

Thus

$$\begin{aligned}
|a_{n+p} - a_n| &\leqslant |a_{n+1} - a_n| + |a_{n+2} - a_{n+1}| + \cdots + |a_{n+p} - a_{n+p-1}| \\
&\leqslant \left(k^{n-1} + k^n + \cdots + k^{n+p-2}\right)|a_2 - a_1| \\
&\leqslant k^{n-1}\left(1 + k + k^2 + \cdots + k^{p-1}\right)|a_2 - a_1| \\
&\leqslant \frac{k^{n-1}}{1-k}|a_2 - a_1|
\end{aligned}$$

for all $n, p \in \mathbf{N}$. Since $k^{n-1} \to 0$ as $n \to \infty$ (independently of p), this implies (a_n) is a Cauchy sequence (in $\mathbf{R}$) and, hence, it is convergent. □

Problem 10.4. The sequence (x_n) is recursively defined by

$$\begin{cases} x_0 = \sqrt{2}, \\ x_{n+1} = \sqrt{2 + x_n} \quad n \geqslant 0. \end{cases}$$

Show that (x_n) converges.

Proof. We first prove by induction that $x_n \leqslant x_{n+1} \leqslant 2$ for all $n \in \mathbf{N}$. For $n = 0$,

$$x_0 = \sqrt{2} \leqslant \sqrt{2 + \sqrt{2}} = x_1 \leqslant \sqrt{2 + \sqrt{4}} = 2.$$

If $x_{n-1} \leqslant x_n \leqslant 2$, then

$$x_n = \sqrt{2 + x_{n-1}} \leqslant \sqrt{2 + x_n} = x_{n+1} \leqslant \sqrt{2 + 2} = 2.$$

Hence (x_n) is monotonically increasing and bounded above by 2. By the monotone convergence theorem, (x_n) converges; let $x_n \to x$. Applying the limit on both sides of $x_{n+1} = \sqrt{2 + x_n}$,

$$\begin{aligned}
\lim_{n\to\infty} x_{n+1} &= \lim_{n\to\infty} \sqrt{2 + x_n} \\
x &= \sqrt{2 + x} \\
x &= 2 \text{ or } 1
\end{aligned}$$

Since all $x_n \geqslant 0$, we must have $x = 2$. □

Bibliography

[Alc14] L. Alcock. *How to Think About Analysis*. Oxford University Press, 2014.

[Apo57] T. M. Apostol. *Mathematical Analysis*. Addison-Wesley, 1957.

[Art11] M. Artin. *Algebra*. Pearson Education, 2011.

[Axl24] S. Axler. *Linear Algebra Done Right, 4th edition*. Springer, 2024.

[BS11] R. G. Bartle and D. R. Sherbert. *Introduction to Real Analysis*. John Wiley & Sons, Inc., 2011.

[DF04] D. S. Dummit and R. M. Foote. *Abstract Algebra*. John Wiley & Sons, 2004.

[HS65] E. Hewitt and K. Stromberg. *Real and Abstract Analysis*. Springer-Verlag Berlin Heiderlberg, 1965.

[Pól45] G. Pólya. *How to Solve It*. Princeton University Press, 1945.

[Rud76] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 1976.

[Sch92] A. H. Schoenfeld. "Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics". In: *Handbook for Research on Mathematics Teaching and Learning*. Macmillan, 1992, pp. 334–370.

Index