Topics in Undergraduate Mathematics

Ryan Joo

This is (still!) an incomplete draft. For corrections and comments, please send an email to the author at ryanjooruian18@gmail.com, or create a pull request at `https://github.com/Ryanjoo18/undergrad-maths`.

Last updated May 12, 2025.

Preface

I began writing this book during the mid-year break in 2023, which was when I had a bit of free time to read up on undergraduate mathematics, and also started learning LaTeX to write this book.

The objective of this book is to serve as a compilation of essential topics at the undergraduate level (as well as serving as my personal notes). The book covers (or aims to cover) the following topics:

1. Abstract algebra, which follows [DF04; Lan05].
2. Linear algebra, which follows [Axl24].
3. Real analysis, which follows [Rud76; Apo57]
4. Multivariable analysis, which follows [Spi65; Rud76].
5. Complex analysis, which follows [SS03].
6. General topology, which follows [Mun18].
7. Measure theory, which follows [Fol99].

Prerequisites

This book is written such that it is accessible to high school students for self-study. No formal prerequisites are required, although some experience with proofs may be helpful.

Presentation

This book follows the typical style of "Definition", "Theorem", etc. As far as possible, I will try to make clear what I define, and what results I wish to show.

For ease of reference, important terms are ***coloured*** when first defined, and are included in the glossary; less important terms are *italicised* when first defined, and are not included in the glossary.

Note on Problem Solving

Mathematics is about problem solving. Pólya outlined the following problem solving cycle in [Pól45].

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.

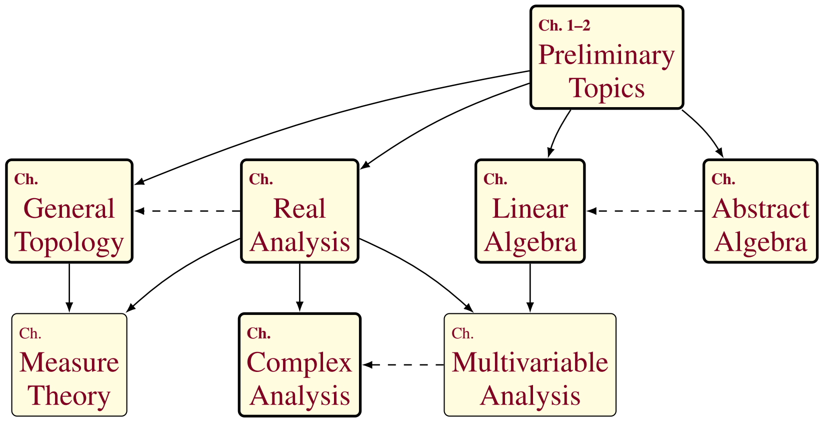
4. **Belief system**: one's perspectives regarding the nature of a discipline and how one goes about working on it.

Acknowledgements

I am greatly thankful to the authors listed in the Bibliography.

Contents

Prerequisite Tree



- A solid arrow ⟶ indicates a required prerequisite, a dotted arrow ····▸ indicates a recommended prerequisite.
- Core topics are in **bold** boxes; other courses (i.e., options or prerequisites) are in light boxes.

I

Preliminary Topics

Chapter 1

Mathematical Reasoning and Logic

This chapter covers basic logic and common methods and proof, which are the bread and butter of mathematics.

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:

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

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

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

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

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

Lemma 1.1.

(i) Double negation law:

$$P \equiv \neg(\neg P)$$

(ii) Commutative laws:

$$P \land Q \equiv Q \land P$$
$$P \lor Q \equiv Q \lor P$$

(iii) Associative laws:

$$(P \wedge Q) \wedge R \equiv P \wedge (Q \wedge R)$$
$$(P \vee Q) \vee R \equiv P \vee (Q \vee R)$$

(iv) Idempotent laws:

$$P \wedge P \equiv P$$
$$P \vee P \equiv P$$

(v) Distributive laws:

$$P \wedge (Q \vee R) \equiv (P \wedge Q) \vee (P \wedge Q)$$
$$P \vee (Q \wedge R) \equiv (P \vee Q) \wedge (P \vee R)$$

(vi) Absorption laws:

$$P \vee (P \wedge Q) \equiv P$$
$$P \wedge (P \vee Q) \equiv P$$

(vii) de Morgan's laws:

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

Remark. Notice that because of the associative laws we can leave out parentheses in statements of the forms $P \wedge Q \wedge R$ and $P \vee Q \vee R$ without ambiguity, because the two possible ways of filling in the parentheses are equivalent.

Statements that are always true are called *tautologies*; for instance $P \vee \neg P$. Similarly, statements that are always false are called *contradictions*; for instance $P \wedge \neg P$.

We can now state a few more useful laws involving tautologies and contradictions.

Lemma 1.2.

(i) Tautology laws: if Q is a tautology, then

$$\begin{aligned} P \wedge Q &\equiv P \\ P \vee Q & \quad \textit{is a tautology} \\ \neg Q & \quad \textit{is a contradiction} \end{aligned}$$

(ii) Contradiction laws: if Q is a contradiction, then

$$\begin{aligned} P \vee Q &\equiv P \\ P \wedge Q &\quad \textit{is a contradiction} \\ \neg Q &\quad \textit{is a tautology} \end{aligned}$$

1.1.1 If, only if

We denote an ***implication*** 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* (or *premise*) 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 $P \implies Q$,

1. assume that P holds,
2. deduce, through some logical steps, that Q holds.

Alternatively, we can prove the contrapositive: assume that Q does not hold, then show that P does not hold.

1.1.2 If and only if, iff

We denote a ***bidirectional implication*** 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:

1. prove $P \implies Q$, and
2. prove $Q \implies P$.

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*.

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

$$\neg\left[\forall x \in X, P(x)\right] \equiv \exists x \in X, \neg P(x)$$
$$\neg\left[\exists x \in X, P(x)\right] \equiv \forall x \in X, \neg P(x)$$

To prove a statement of the form $\forall x \in X, P(x)$,

1. Start with "*let $x \in X$ be given*" to address the quantifier with an arbitrary x (this will prove the statement for all $x \in X$).
2. Show that $P(x)$ is true.

Consider statements of the form $\forall x \in X, P(x) \implies Q(x)$; we say that the statement is *vacuously true* if $P(x)$ is false for all $x \in X$.

To prove a statement of the form $\exists x \in X, P(x)$, there is not such a clear steer about how to continue:

- you can construct such an x with the desired properties (constructive proof);
- you can demonstrate logically that such an x must exist because of some earlier assumption (non-constructive proof);
- you can suppose that such an x does not exist, and consequently arrive at some inconsistency (proof by contradiction).

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

What is a *proof*? Informally, we will define a mathematical proof to be a logical argument that establishes the truth of a mathematical statement. A typical proof proceeds as follows:

1. Start with the given hypotheses.
2. Apply rules of inferences (logical deduction) to get new statements.
3. Repeat Step 2 until we reach the desired conclusion.

We first present some straightforward 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.
$$P \implies \cdots \implies Q$$
- A ***proof by contrapositive*** of $P \implies Q$ is to prove instead $\neg Q \implies \neg P$.
- A ***disproof by counterexample*** is to provide a counterexample to disprove a statement, which makes the negation of the statement true.

  Thus, to disprove $P \implies Q$, the counterexample makes the hypothesis P true, and the conclusion Q false. Likewise, to disprove $\forall x \in X, P(x)$, we prove its negation $\exists x \in X, \neg P(x)$, i.e., find $a \in X$ such that $P(a)$ is false.

  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.

1.3.1 Proof by Contradiction

To ***prove by contradiction***,

1. Assume P is false, i.e., $\neg P$ is true (to prove $P \implies Q$ by contradiction, suppose $P \wedge \neg Q$).
2. 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.

We illustrate this method of proof using a classic example.

Example 1.4 (Irrationality of $\sqrt{2}$)**.** Prove that $\sqrt{2}$ is irrational.

Proof. We prove by contradiction. Suppose otherwise, that $\sqrt{2}$ is rational. Then $\sqrt{2} = \frac{a}{b}$ for some $a, b \in \mathbb{Z}, b \neq 0$, 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 $k \in \mathbb{Z}$. Substituting $a = 2k$ into the above equation and simplifying it gives us

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

This means that b^2 is even, from which follows again that b is even. This contradicts the assumption that a and b coprime, so we are done. □

Example 1.5 (Euclid)**.** Prove that there are infinitely many prime numbers.

Proof. Suppose otherwise, that only finitely many prime numbers exist. List them as $p_1, \ldots, p_n$. Consider the number

$$N = p_1 p_2 \cdots p_n + 1.$$

Note that N 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. □

1.3.2 Proof of Existence and Uniqueness

To prove existential statements, we can adopt two approaches:

1. ***Constructive proof*** (direct proof)

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

To ***prove uniqueness*** (after proving existence), we can either

- assume $\exists x, y \in X$ such that $P(x) \wedge P(y)$ is true, then show $x = y$, or
- assume that $\exists x, y \in X$ are distinct such that $P(x) \wedge P(y)$, then derive a contradiction.

We sometimes use $\exists!$ to mean "there exists a unique".

Example 1.6. Prove that we can find 100 consecutive positive integers which are all composite numbers.

Proof. We proceed by constructive proof; we will construct integers $n, n+1, n+2, \ldots, n+99$, all of which are composite.

Claim. $n = 101! + 2$.

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. □

Example 1.7. Prove that for all $p, q \in \mathbb{Q}$ with $p < q$, there exists $x \in \mathbb{Q}$ such that $p < x < q$.

Proof. We prove by construction; we want to construct x in terms of p and q, which fulfils the required condition.

Claim. $x = \dfrac{p+q}{2}$.

Evidently $x \in \mathbb{Q}$. Since $p < q$,

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

Similarly,

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

Remark. There are two parts to prove: 1) x satisfies the given statement 2) x is within the domain (for this question we do not have to prove x is rational since $\mathbb{Q}$ is closed under addition).

□

Example 1.8. 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.

□

Example 1.9. Prove that every integer greater than 1 is divisible by a prime.

Proof. We proceed by a non-constructive 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 $\mathbb{N}$.

Example 1.10. Prove that the equation $x^2 + y^2 = 3z^2$ has no solutions (x, y, z) in integers where $z \neq 0$.

Proof. Suppose (x, y, z) is a solution. WLOG assume $z > 0$. By the least integer principle, we may also assume that our solution has z minimal. Taking remainders modulo 3, we see that

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

Since perfect squares can only be congruent to 0 or 1 modulo 3, we must have $x \equiv y \equiv 0 \pmod 3$. Writing $x = 3a$ and $y = 3b$ for $a, b \in \mathbb{Z}$ gives

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

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

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

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

1.3.3 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 $\mathbb{N}$ states the following: every non-empty subset $S \subset \mathbb{N}$ has a smallest element; that is, there exists $m \in S$ such that $m \leq k$ for all $k \in S$.

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

Lemma 1.11. *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 $S \neq \mathbb{N}$.

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

Now consider $p-1$. Since p is the least element of $\mathbb{N} \setminus S$, $p-1 \notin \mathbb{N} \setminus S$ 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 $p \in \mathbb{N} \setminus S$.

$\boxed{\impliedby}$ Suppose the principle of induction is true. Then this implies that 1.12 is true, which in turn implies that 1.16 is true. In order to prove the well-ordering of $\mathbb{N}$, we prove the following statement $P(n)$ by strong induction on n: If $S \subset \mathbb{N}$ 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 $\mathbb{N}$.

Now suppose that $P(k)$ is true for $k = 1, \ldots, n$. To show that $P(n+1)$ is true, let $S \subset \mathbb{N}$ 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 $n \in \mathbb{N}$. This implies the well-ordering of $\mathbb{N}$, because if S is a non-empty subset of $\mathbb{N}$, then pick $n \in S$. Since $n \in \mathbb{N}$, $P(n)$ is true, and therefore S has a smallest element. □

Theorem 1.12 (Principle of mathematical induction)**.** *Let $P(n)$ be a family of statements indexed by $\mathbb{N}$. Suppose that*

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

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

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

(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 $S = \{n \in \mathbb{N} \mid P(n) \text{ is true}\}$. □

We illustrate the application of this proving technique using a classic example.

Example 1.13. Prove that for any $n \in \mathbb{N}$,

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

Proof. 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 $k \in \mathbb{N}$, $k \geq 1$; 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. □

Example 1.14 (Bernoulli's inequality)**.** Let $x \in \mathbb{R}$, $x > -1$. Then for all $n \in \mathbb{N}$,

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

Proof. Induct on n. Fix $x > -1$. Let $P(n) : (1+x)^n \geq 1 + nx$.

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

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

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

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

> **Corollary 1.15.** *Let $P(n)$ be a family of statements indexed by integers $n \geq N$ for some $N \in \mathbb{Z}$. Suppose that*
>
> *(i) $P(N)$ is true;*

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

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

Proof. Apply 1.12 to the statement $Q(n) = P(n+N)$ for $n \in \mathbb{N}$. □

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

Theorem 1.16 (Strong induction)**.** *Let* $P(n)$ *be a family of statements indexed by* $\mathbb{N}$. *Suppose that*

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

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

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

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 $n \in \mathbb{N}$, $Q(n) \implies Q(n+1)$. By 1.12, $Q(n)$ holds for all $n \in \mathbb{N}$, and hence $P(n)$ holds for all n. □

Example 1.17 (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 $n \geq 2$ 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 $n \in \mathbb{N}$. □

The following is also another variant on induction.

Theorem 1.18 (Cauchy induction)**.** *Let* $P(n)$ *be a family of statements indexed by* $\mathbb{N}_{\geq 2}$. *Suppose that*

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

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

Then $P(n)$ *is true for all* $n \in \mathbb{N}_{\geq 2}$.

Example 1.19 (AM–GM inequality)**.** Given $n \in \mathbb{N}$, prove that for positive reals $a_1, a_2, \ldots, a_n$,

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

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

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

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

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

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

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

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

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, $\frac{a_1+a_2+\cdots+a_n}{n} \geq \sqrt[n]{a_1 a_2 \cdots a_n}$ Let $a_n = \frac{a_1+a_2+\cdots+a_{n-1}}{n-1}$ Then we have

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

So,

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

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

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

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

By Cauchy induction, this proves the AM–GM inequality for n variables. $\square$

1.3.4 Pigeonhole Principle

Theorem 1.20 (Pigeonhole principle). *If $kn+1$ objects are distributed among n boxes, one of the boxes will contain at least $k+1$ objects.*

Example 1.21 (IMO 1972). Prove that every set of 10 two-digit integer numbers has two disjoint subsets with the same sum of elements.

Proof. 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. □

Example 1.22 (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

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

Proof. Recall that the fractional part of a real number x is $x-\lfloor x \rfloor$. Consider 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, 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

Exercise 1.1. 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 $(\forall x \in \mathbb{R})[x > 2 \implies x^2 > 4]$.

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

□

Exercise 1.2. 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 $f(x) \neq y$.

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

□

Exercise 1.3. 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.]

Solution. Prove by contradiction. Suppose n is not a perfect square and $\sqrt{n}$ is rational. Then $\sqrt{n} = \frac{a}{b}$ for some $a, b \in \mathbb{Z}$. Squaring both sides and clearing denominator gives

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

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

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

$$\begin{aligned}
a = q_1^{e_1} q_2^{e_2} \cdots q_u^{e_u} &\implies a^2 = q_1^{2e_1} q_2^{2e_2} \cdots q_u^{2e_u} \\
b = r_1^{f_1} r_2^{f_2} \cdots r_v^{f_v} &\implies b^2 = r_1^{2f_1} r_2^{2f_2} \cdots r_v^{2f_v}
\end{aligned}$$

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. □

Exercise 1.4. Prove that for every pair of irrational numbers p and q such that $p < q$, there is an irrational x such that $p < x < q$.

Solution. Consider the average of p and q, i.e., $\dfrac{p+q}{2}$. Evidently $p < \dfrac{p+q}{2} < q$.

Since it may not always be the case that $\dfrac{p+q}{2}$ is irrational (so we cannot immediately take $x = \dfrac{p+q}{2}$), we need to consider two cases:

$\dfrac{p+q}{2}$ **is irrational** Take $x = \dfrac{p+q}{2}$ and we are done.

$\dfrac{p+q}{2}$ **is rational** Let $r = \dfrac{p+q}{2}$, and take the average of p and r, i.e., $\dfrac{p+r}{2}$. Evidently $p < \dfrac{p+r}{2} < r < q$. Since p is irrational and r is rational, $\dfrac{p+r}{2}$ is irrational. In this case, take $x = \dfrac{3p+q}{4}$.

□

Exercise 1.5. Given n real numbers $a_1, a_2, \ldots, a_n$. Show that there exists an a_i $(1 \leq i \leq n)$ such that a_i is greater than or equal to the mean of the n numbers.

Solution. 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 $a_i > \bar{a}$. □

Exercise 1.6. Prove that the following statement is false: there is an irrational number a such that for all irrational number b, ab is rational.

Idea. Prove the negation of the statement: for every irrational number a, there is an irrational number b such that ab is irrational. We shall adopt a constructive proof (note that we can consider multiple cases and construct more than one b).

Solution. Given an irrational number a, let us consider $\dfrac{\sqrt{2}}{a}$. We consider cases:

- If $\dfrac{\sqrt{2}}{a}$ is irrational, take $b = \dfrac{\sqrt{2}}{a}$. Then $ab = \sqrt{2}$ which is irrational.
- If $\dfrac{\sqrt{2}}{a}$ is rational, its reciprocal $\dfrac{a}{\sqrt{2}}$ is rational. 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 $ab = a\sqrt{3}$ is irrational.

□

Exercise 1.7. Prove that there are infinitely many prime numbers that are congruent to 3 modulo 4.

Idea. It is not really possible to come up with a direct proof, so we prove by contradiction.

Solution. Suppose, for a contradiction, that 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.

Let $M = (p_1 p_2 \cdots p_m)^2 + 2$.

We have the following observation:

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

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

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

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

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

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

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

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

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

Exercise 1.8. Prove that, for any positive integer n, there exists a perfect square m^2 such that $n \leq m^2 \leq 2n$.

Idea. A direct proof by construction is not quite possible, so we prove by contradiction.

Solution. Suppose, for a contradiction, that $n > m^2$ and $(m+1)^2 > 2n$ for some positive integer n, so that there is no square between n and $2n$. Then

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

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

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

which simplifies to

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

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

Exercise 1.9. Prove that for every positive integer $n \geq 4$,

$$n! > 2^n.$$

Solution. Induct on n. Let $P(n) : n! > 2^n$.

The base case $P(4)$ is clear. Now suppose $P(k)$ is true for some $k \in \mathbb{N}_{\geq 4}$, i.e., $k! > 2^k$. Then

$$(k+1)! = k!(k+1) > 2^k(k+1) > 2^k \cdot 2 = 2^{k+1},$$

so $P(k+1)$ is true. $\square$

Exercise 1.10. Prove by mathematical induction, for $n \geq 2$,

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

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

The base case $P(2)$ is clear. Now assume $P(k)$ is true for $k \geq 2, k \in \mathbb{N}$, i.e., $\sqrt[k]{k} < 2 - \frac{1}{k}$, or

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

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

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

so $P(k+1)$ is true. $\square$

Exercise 1.11. Prove that, for all integers $n \geq 3$,

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

Solution. 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 $k \in \mathbb{N}_{\geq 3}$; 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 $1 \leq i \leq n$, is called the *Fibonacci sequence* if and only if it is defined recursively by $F_1 = 1$, $F_2 = 1$, $F_n = F_{n-1} + F_{n-2}$ for $n > 2$.

Exercise 1.12. 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 $n \in \mathbb{N}$, prove that

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

Idea. Given the recurrence relation, we may need to use *strong induction*: use $P(k), P(k+1), P(k+2)$ to prove $P(k+3)$, for all $k \in \mathbb{N}$.

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

The base cases $P(1), P(2), P(3)$ are clear. Now assume $P(k), P(k+1), P(k+2)$ are true, for some $k \in \mathbb{N}$. We will show that $P(k+3)$ is true.

By the inductive hypothesis, for $k \in \mathbb{N}$ we have

$$a_k \leq 2^k, \quad a_{k+1} \leq 2^{k+1}, \quad a_{k+2} \leq 2^{k+2}.$$

Then

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

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

Exercise 1.13. For $m, n \in \mathbb{N}$, prove that

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

Solution. Induct on n. Let $P(n) : F_{n+m+1} = F_n F_m + F_{n+1} F_{m+1}$ for all $m \in \mathbb{N}$ in the cases $k = n$ and $k = n + 1$.

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 $m \in \mathbb{N}$,

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

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 $m \in \mathbb{N}$. □

Chapter 2

Set Theory

2.1 Basics of Naive Set Theory

2.1.1 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 *set builder notation*:

$$\{x \in S \mid P(x)\}$$

The following sets of numbers are frequently encountered.

- The natural numbers $\mathbb{N} = \{1, 2, 3, \dots\}$ (some people include 0, some do not).
- The integers $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$.
- The rational numbers $\mathbb{Q} = \left\{\frac{p}{q} \mid p, q \in \mathbb{Z}, q \neq 0\right\}$.
- The real numbers $\mathbb{R}$ (the construction of which, using Dedekind cuts, will be discussed in Chapter 13).
- The complex numbers $\mathbb{C} = \{x + yi \mid x, y \in \mathbb{R}\}$.

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

[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.

We say A is a ***subset*** of B if every element of A is in B:

$$A \subset B \quad \text{means} \quad (\forall x)(x \in A \implies x \in B)$$

By construction, we have $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$.

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

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

Proof. Since $A \subset B$, $x \in A \implies x \in B$. Since $B \subset C$, $x \in B \implies x \in C$. Combining these two implications, we have $x \in A \implies x \in C$. Hence $A \subset C$. □

We say two sets A and B are ***equal*** if and only if they contain the same elements:

$$A = B \quad \text{means} \quad x \in A \iff x \in B.$$

Lemma 2.2 (Double inclusion)**.** *Let $A, B \subset S$. 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 $\mathbb{R}$ are known as ***intervals***, which can be visualised as sections of the real line. We define *bounded intervals*

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

and *unbounded intervals*

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

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):

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

Lemma 2.3. *Let $A, B \subset S$. Then*

$$A \subset B \iff \mathcal{P}(A) \subset \mathcal{P}(B).$$

Proof.

$\boxed{\implies}$ Suppose $A \subset B$. Let $X \in \mathcal{P}(A)$. Then $X \subset A$, so $X \subset B$. Thus $X \in \mathcal{P}(B)$. Hence $\mathcal{P}(A) \subset \mathcal{P}(B)$.

$\boxed{\impliedby}$ Suppose A and B are sets such that $\mathcal{P}(A) \subset \mathcal{P}(B)$.

Let $x \in A$. Consider the set $C = \{x\}$. Then $C \subset A$, so $C \in \mathcal{P}(A)$. Thus $C \in \mathcal{P}(B)$, so $C \subset B$.

We know $x \in C$, so we get $x \in B$. Hence $A \subset B$. $\square$

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 is the set of all ordered pairs with the first element of the pair coming from A and the second from B:

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

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 $1 \leq i \leq n$. If all the A_i are the same, we write the product as A^n.

Example 2.4. $\mathbb{R}^2$ is the Euclidean plane, $\mathbb{R}^3$ is the Euclidean space, and $\mathbb{R}^n$ is the n-dimensional Euclidean space.

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

Lemma 2.5. *Let A, B, C, D be sets.*

(i) $A \times \emptyset = \emptyset \times A = \emptyset$.

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

(iii) $A \times (B \cap C) = (A \cap B) \times (A \cap C)$.

(iv) $(A \cap B) \times (C \cap D) = (A \cap C) \times (B \cap D)$.

(v) $(A \cup B) \times (C \cup D) \subset (A \cup C) \times (B \cup D)$.

Proof.

(i) Evidently $\emptyset \subset A \times \emptyset$, which is vacuously true.

To show the opposite containment $A \times \emptyset \subset \emptyset$ is equivalent to showing

$$(\forall x) x \in A \times \emptyset \implies x \in \emptyset;$$

but $x \in \emptyset$ is always false, so this is equivalent to

$$(\forall x) \neg (x \in A \times \emptyset).$$

Note that

$$\begin{aligned} x \in A \times \emptyset &\iff (\exists a \in A)(\exists b \in \emptyset)[x = (a, b)] \\ &\iff (\exists a)(\exists b)[a \in A \wedge b \in \emptyset \wedge x = (a, b)] \end{aligned}$$

which is always false, since $b \in \emptyset$ is always false.

Hence $(\forall x) \neg (x \in A \times \emptyset)$ holds.

(ii)

(iii)

(iv)

(v)

□

2.1.2 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:

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

We say A and B are ***disjoint*** if both sets have no element in common; that is,

$$A \cap B = \emptyset.$$

More generally, we can take unions and intersections of arbitrary numbers of sets (could be finitely or infinitely many). Given an indexed 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 $A \setminus B = A \cap B^c$.

Lemma 2.6 (Distributive laws)**.** *Let $A, B, C \subset S$. Then*

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

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

These properties are straightforward consequences of logic laws.

Proof.

(i) We have

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

(ii) Similar to (i).

$\square$

Lemma 2.7 (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

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

Lemma 2.8. *The following hold:*

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

2.2 Relations

2.2.1 Definition and Examples

Definition 2.9 (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 2.10. 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 $\leq$ on the set of real numbers is
$$\{(x, y) \in \mathbb{R}^2 \mid x \leq y\} \subset \mathbb{R}^2;$$
we write $x \leq y$ if (x, y) is in this set.
- The "divides" relation $\mid$ on $\mathbb{N}$ is
$$\{(m, n) \in \mathbb{N}^2 \mid m \text{ divides } n\} \subset \mathbb{N}^2;$$
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 2.11. Given $R = \{(1, a), (1, b), (2, b), (3, b)\}$, then $\operatorname{dom} R = \{1, 2, 3\}$ and $\operatorname{ran} R = \{a, b\}$.

Definition 2.12 (Binary relation)**.** A ***binary relation*** in A is a relation between A and itself; that is, $R \subset A \times A$.

2.2.2 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 2.13 (Less than or equal to)**.** The relation $\leq$ on R is reflexive, anti-symmetric, and transitive, but not symmetric.

Definition 2.14. 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 $s \leq b$ for all $b \in B$.

Example 2.15.

- 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.

2.2.3 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 2.16 (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 2.17 (Equivalence class)**.** Given an equivalence relation $\sim$ on a set A, and given $x \in A$, the ***equivalence class*** of x is

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

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

Definition 2.18 (Partition)**.** A ***partition*** of a set A is a collection of subsets $\{A_i \mid i \in I\}$ such that

(i) $A_i \neq \emptyset$ for all $i \in I$; (all 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)

Proposition 2.19. *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, 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 2.20 (Quotient set)**.** The ***quotient set*** is the set of all equivalence classes, denoted by $A/\sim$.

Example 2.21 (Modular arithmetic)**.** Fix $n \in \mathbb{Z}^+$. Define a relation on $\mathbb{Z}$:

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

Lemma. *$\sim$ is a equivalence relation on $\mathbb{Z}$.*

Proof.

(i) Since $n \mid a - a = 0$, we have $a \sim a$.

(ii) Suppose $a \sim b$. Then $n \mid a - b$ implies $n \mid b - a$, so $b \sim a$.

(iii) Suppose $a \sim b$ and $b \sim c$. Then $n \mid (a - b)$ and $n \mid (b - c)$, so $n \mid (a - b) + (b - c) = (a - c)$. Thus $a \sim c$.

□

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

We denote the equivalence class of a by $[a]$, called the *congruence class* of a mod n, which consists of the integers which differ from a by an integral multiple of n; that is,

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

There are precisely n distinct congruence classes mod n, namely

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

determined by the possible remainders after division by n; and these congruence classes partition the integers $\mathbb{Z}$. The set of congruence classes is denoted by $\mathbb{Z}/n\mathbb{Z} := \mathbb{Z}/\sim$, called the *integers modulo n*.

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

$$\begin{aligned}[][a]+[b] &= [a+b] \\ [a][b] &= [ab].\end{aligned}$$

Lemma. *Addition and mulltiplication on $\mathbb{Z}/n\mathbb{Z}$ are well-defined.*

Proof. Suppose $[a_1] = [b_1]$ and $[a_2] = [b_2]$.

Then $a_1 \equiv b_1 \pmod n$, or $n \mid (a_1 - b_1)$; let $a_1 = b_1 + sn$ for some integer s. Similarly, let $a_2 = b_2 + tn$ for some integer t.

Then $a_1 + a_2 = (b_1 + b_2) + (s+t)n$, so $a_1 + a_2 \equiv b_1 + b_2 \pmod n$. Hence $[a_1 + a_2] = [b_1 + b_2]$.

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 $a_1 a_2 \equiv b_1 b_2 \pmod n$. Hence $[a_1 a_2] = [b_1 b_2]$. □

Hence we have shown that if $a_1 \equiv b_1 \pmod n$ and $a_2 \equiv b_2 \pmod n$, then

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

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

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

Lemma. $(\mathbb{Z}/n\mathbb{Z})^\times$ *equals the collection of congruence classes whose representatives are relatively prime to n:*

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

Example 2.22 (Rationals)**.** Define a relation on $\mathbb{Z} \times \mathbb{Z}^*$, where $\mathbb{Z}^* = \mathbb{Z} \setminus \{0\}$:

$$(a,b) \sim (c,d) \iff ad = bc.$$

Denote the congruence class of (a,b) by a/b.

Let $\mathbb{Q} := \mathbb{Z} \times \mathbb{Z}^*$, with addition and multiplication defined by

$$\begin{aligned}\frac{a}{b} + \frac{c}{d} &= \frac{ad+bc}{bd} \\ \frac{a}{b}\frac{c}{d} &= \frac{ac}{bd}\end{aligned}$$

Then $\mathbb{Q}$ is a field.

Definition 2.23 (Quotient map)**.** The quotient map is the map

$$\begin{aligned}\pi\colon X &\to X/\sim\\ x &\mapsto [x]\end{aligned}$$

Lemma 2.24. *Quotient maps are surjective.*

Proof. ☐

2.2.4 Axiom of Choice and Its Equivalences

Definition 2.25. Let $(P, \leq)$ be a partially ordered set. Suppose $A \subset P$.

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

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

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

(iv) $C \subset P$ is called a ***chain*** if either $x \leq y$ or $y \leq x$ 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 2.26. 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 2.27. *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 2.28. *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

2.3.1 Definitions

Definition 2.29 (Function)**.** A ***function*** $f\colon X \to Y$ 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 $f\colon \mathbb{R} \to \mathbb{R}$, the assignment $f(x) = \frac{1}{x}$ is not sufficient as it fails at $x = 0$. Similarly, $f(x) = y$ where $y^2 = x$ fails because $f(x)$ is undefined for $x < 0$, and for $x > 0$ it does not return a unique value; in such cases, we say the the function is *ill-defined*. We are interested in the opposite; functions that are *well-defined*.

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

Definition 2.30 (Restriction)**.** Suppose $f\colon X \to Y$. The ***restriction*** of f to $A \subset$ is the map $f|_A\colon A \to Y$.

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

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

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

2.3.2 Injectivity, Surjectivity, Bijectivity

Definition 2.32. Suppose $f\colon X \to Y$.

(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:

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

(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 $f\colon X \to Y$.

2.3.3 Images and Pre-images

Definition 2.33. Suppose $f\colon X \to Y$. The ***image*** of $A \subset X$ under f is

$$f(A) := \{y \in Y \mid \exists x \in A, y = f(x)\}.$$

The ***pre-image*** of $B \subset Y$ under f is

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

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

A useful identity is

$$x \in f^{-1}(A) \iff f(x) \in A.$$

Lemma 2.34. *Let $f\colon X \to Y$.*

(i) If $A \subset Y$, then $f\left(f^{-1}(A)\right) \subset A$.

(ii) If $A \subset X$, then $A \subset f^{-1}\left(f(A)\right)$.

Proof.

(i)

$$\begin{aligned} y \in f\left(f^{-1}(A)\right) &\implies \exists x \in f^{-1}(A), \quad f(x) = y \\ &\iff \exists f(x) \in A, \quad f(x) = y \\ &\implies y \in A. \end{aligned}$$

(ii)

$$\begin{aligned} x \in A &\implies f(x) \in f(A) \\ &\iff x \in f^{-1}\left(f(A)\right). \end{aligned}$$

□

Lemma 2.35 (Algebra of pre-images)**.** *Suppose $f\colon X \to Y$. Then*

(i) $f^{-1}(A^c) = \left[f^{-1}(A)\right]^c$ for every $A \subset Y$;

(ii) $f^{-1}\left(\bigcup_{i\in I} A_i\right) = \bigcup_{i\in I} f^{-1}(A_i)$;

(iii) $f^{-1}\left(\bigcap_{i\in I} A_i\right) = \bigcap_{i\in I} f^{-1}(A_i)$.

Proof.

(i) Suppose $A \subset Y$. Let $x \in X$, then

$$\begin{aligned} x \in f^{-1}(A^c) &\iff f(x) \in A^c \\ &\iff f(x) \notin A \\ &\iff x \notin f^{-1}(A) \\ &\iff x \in f^{-1}(A)^c \end{aligned}$$

Hence $f^{-1}(A^c) = f^{-1}(A)^c$.

(ii) Suppose $\{A_i \mid i \in I\}$ is a collection of subsets of Y. Then

$$\begin{aligned} x \in f^{-1}\left(\bigcup_{i\in I} A_i\right) &\iff f(x) \in \bigcup_{i\in I} A_i \\ &\iff f(x) \in A_i \text{ for some } i \in I \\ &\iff x \in f^{-1}(A_i) \text{ for some } i \in I \\ &\iff x \in \bigcup_{i\in I} f^{-1}(A_i) \end{aligned}$$

Hence $f^{-1}\left(\bigcup_{i\in I} A_i\right) = \bigcup_{i\in I} f^{-1}(A_i)$.

(iii) Suppose $\{A_i \mid i \in I\}$ is a collection of subsets of Y. Then

$$\begin{aligned} x \in f^{-1}\left(\bigcap_{i\in I} A_i\right) &\iff f(x) \in \bigcap_{i\in I} A_i \\ &\iff f(x) \in A_i \text{ for every } i \in I \\ &\iff x \in f^{-1}(A_i) \text{ for every } i \in I \\ &\iff x \in \bigcap_{i\in I} f^{-1}(A_i) \end{aligned}$$

Hence $f^{-1}\left(\bigcap_{i\in I} A_i\right) = \bigcap_{i\in I} f^{-1}(A_i)$.

□

Lemma 2.36 (Algebra of images)**.** *Suppose $f\colon X \to Y$. Then*

(i) $f(A)^c \subset f(A^c)$;

(ii) $f\left(\bigcup_{i\in I} A_i\right) = \bigcup_{i\in I} f(A_i)$;

(iii) $f\left(\bigcap_{i\in I} A_i\right) \subset \bigcap_{i\in I} f(A_i)$.

2.3.4 Composition

Definition 2.37 (Composition)**.** Given $f\colon X \to Y$ and $g\colon Y \to Z$, the ***composition*** $g \circ f\colon X \to Z$ 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 2.38 (Associativity of composition)**.** *Suppose* $f : X \to Y$, $g : Y \to Z$, $h : Z \to W$. *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 2.39 (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 2.40. $f : X \to Y$ *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

$$\begin{aligned} g_1 \colon Z \to X, &\quad g_1(1) = x, \\ g_2 \colon Z \to Y, &\quad g_2(1) = y. \end{aligned}$$

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 2.41.** *$f \colon X \to Y$ is surjective if and only if for any set Z and any functions $g_1, g_2 \colon Y \to Z$,*
>
> $$g_1 \circ f = g_2 \circ f \implies g_1 = g_2.$$

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 $g_1, g_2 \colon Y \to Z$ 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

$$\begin{aligned} g_1 \colon Y \to Z, g_1(y) = 1, g_1(y') = y' \\ g_2 \colon Y \to Z, g_2(y) = 2, g_2(y') = y' \end{aligned}$$

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 $g_i\colon Y \to Z$ 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. □

Lemma 2.42 (Inverse image of composition)**.** *Suppose $f\colon X \to Y$, $g\colon Y \to Z$. Then*

$$(g \circ f)^{-1}(A) = f^{-1}\left(g^{-1}(A)\right)$$

for every $A \subset Z$.

Proof. Suppose $A \subset Z$. Let $x \in X$, then we have

$$\begin{aligned} x \in (g \circ f)^{-1}(A) &\iff (g \circ f)(x) \in A \\ &\iff g\left(f(x)\right) \in A \\ &\iff f(x) \in g^{-1}(A) \\ &\iff x \in f^{-1}\left(g^{-1}(A)\right) \end{aligned}$$

Hence $(g \circ f)^{-1}(A) = f^{-1}\left(g^{-1}(A)\right)$. □

2.3.5 Invertibility

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

Definition 2.43 (Invertibility)**.** Suppose $f\colon X \to Y$. We say that

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

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

(iii) f is ***invertible*** if there exists $k\colon Y \to X$ which is a left and right inverse of f; we call k 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 2.44. Let

$$\begin{aligned} f&\colon \mathbb{R} \to [0,\infty), & f(x) &= x^2 \\ g&\colon [0,\infty) \to \mathbb{R}, & g(x) &= \sqrt{x} \end{aligned}$$

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

Lemma 2.45 (Uniqueness of inverse)**.** *If $f\colon X \to Y$ 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. Immediately 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 2.46 (Invertibility criterion)**.** *Suppose $f\colon X \to Y$. 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) $\boxed{\implies}$ 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 $g(f(a)) = g(f(b))$, so $a = b$.

$\boxed{\impliedby}$ Let f be injective. Choose any x_0 in the domain of f. Define $g\colon Y \to X$ 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) $\boxed{\implies}$ 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 $f(g(y)) = \mathrm{id}_Y(y) = y$ so $y \in f(X)$. Thus $f(X) = Y$ so f is surjective.

$\boxed{\impliedby}$ 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 $f(g(y)) = y$. This defines a function $g\colon Y \to X$ which is evidently a right-inverse of f.

(iii) $\boxed{\implies}$ 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.

$\boxed{\impliedby}$ Suppose f is bijective. Then by (i) and (ii), f has a left-inverse $g\colon Y \to X$ and a right-inverse $h\colon Y \to X$. 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 2.47 (Inverse of composition)**.** *Suppose $f\colon X \to Y$, $g\colon Y \to Z$. 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 2.48. *If $f_1, \dots, f_n$ 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 2.49. *$\sim$ 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 $f\colon X \to Y$ is a bijection. Then f is invertible, with inverse $f^{-1}\colon Y \to X$. Since f^{-1} is invertible (with inverse f), it is bijective.

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

□

Theorem 2.50 (Cantor–Schröder–Bernstein). *If $f\colon X \to Y$ and $g\colon Y \to X$ are injective, then $A \sim B$.*

2.4 Cardinality

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

Definition 2.51. A and B said to be ***equivalent*** (or have the same *cardinality*), denoted by $A \sim B$, if there exists a bijection $f\colon A \to B$.

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

$$\begin{aligned} \mathbb{N}_n &= \{k \in \mathbb{N} \mid 1 \leq k \leq n\}, \\ n\mathbb{N} &= \{nk \mid k \in \mathbb{N}\}. \end{aligned}$$

Definition 2.52. For any set A, we say

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

(ii) A is ***countable*** if $A \sim \mathbb{N}$; 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 $a_1, a_2, \dots$ of distinct terms. This technique is particularly useful when there is not possible to deduce an explicit formula for a bijection.

Lemma 2.53. *$\mathbb{N}$ is infinite.*

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

For the base case $n = 1$, if there exists a function $f_1\colon \{1\} \to \mathbb{N}$, consider the set $\mathbb{N} \setminus f_1(\{1\})$. 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 $\mathbb{N}_k$ to $\mathbb{N}$, then there does not exist a bijection from $\mathbb{N}_{k+1}$ to $\mathbb{N}$. We prove the contrapositive: if there exists a bijection from $\mathbb{N}_{k+1} \to \mathbb{N}$, then there exists a bijection from $\mathbb{N}_k$ to $\mathbb{N}$.

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

Corollary 2.54. *Any countable set is infinite.*

Example 2.55. $\mathbb{N}$ is countable since the identity map from $\mathbb{N}$ to $\mathbb{N}$ is a bijection.

Example 2.56. $n\mathbb{N}$ is countable.

Proof. Let $f\colon \mathbb{N} \to n\mathbb{N}$ which sends $k \mapsto nk$. We need to show that f is bijective:

- For any $k_1, k_2 \in \mathbb{N}$, $nk_1 = nk_2$ implies $k_1 = k_2$ so f is injective.
- For any $x \in n\mathbb{N}$, $x = nk$ for some $k \in \mathbb{N}$, thus $\frac{x}{n} = k \in \mathbb{N}$ so f is surjective.

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

Example 2.57. $\mathbb{Z}$ is countable.

Proof. Consider the following arrangement of the elements of $\mathbb{Z}$ and $\mathbb{N}$:

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

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

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

□

Proposition 2.58. *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 $E \sim \mathbb{N}$, as desired. □

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

Proposition 2.59. *The countable union of countable sets is countable.*

Proof. Let $\{A_n \mid n \in \mathbb{N}\}$ be a fanily of countable sets; clearly this is a countable collection of sets (indexed by $\mathbb{N}$). 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.

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

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 2.58, T is countable. $\square$

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 2.60. *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 2.61. *Let A be a countable set. For $n \in \mathbb{N}$, 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 2.59, B_n is countable. $\square$

Corollary 2.62. *$\mathbb{Q}$ is countable.*

Proof. Note that every $x \in \mathbb{Q}$ is of the form $\frac{b}{a}$, where $a, b \in \mathbb{Z}$. 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 2.63. *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 2.64. *$\mathbb{R}$ is uncountable.*

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

Theorem 2.65 (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 $f \colon A \to \mathcal{P}(A)$. 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 $A \to \mathcal{P}(A)$. $\square$

Exercises

Exercise 2.1. Prove that the statement $\forall x \in \emptyset, P(x)$ is vacuously true.

Solution. Let S be the embedding set. The statement $\forall x \in \emptyset, P(x)$ means

$$\forall x \in S, \quad x \in \emptyset \implies P(x).$$

But $x \in \emptyset$ is always false, by the definition of empty set. Hence the statement is always true, regardless of x. □

Exercise 2.2. Prove that for any set $A \subset S$, $\emptyset \subset A$ and $A \subset A$.

Solution. Let $A \subset S$. Let $x \in \emptyset$, then $x \in \emptyset \implies x \in A$ is vacuously true, so $\emptyset \subset A$.

Likewise, let $x \in A$, then $x \in A \implies x \in A$ is always true, so $A \subset A$. □

Exercise 2.3. Let A be the set of all complex polynomials in n variables. Given a subset $T \subset A$, define the *zeros* of T as the set

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

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

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

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

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

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

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

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

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

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

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

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

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

Exercise 2.4 (Complex numbers)**.** Let $\mathbb{R}[x]$ denote the set of real polynomials. Define

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

where

$$f(x) \sim g(x) \iff x^2+1 \text{ divides } f(x)-g(x).$$

The complex number $a+bi$ is defined to be the equivalence class of $a+bx$.

(a) Define the sum and product of two complex numbers and show that such definitions are well-defined.

(b) Define the reciprocal of a complex number.

Exercise 2.5 ([Rud76] 2.2)**.** We say $z \in \mathbb{C}$ is *algebraic* if there exist integers $a_0, \dots, a_n$, not all zero, such that

$$a_0 z^n + a_1 z^{n-1} + \cdots + a_{n-1} z + a_n = 0.$$

Prove that the set of all algebraic numbers is countable. *Hint*: For every positive integer N there are only finitely many equations with

$$n + |a_0| + |a_1| + \cdots + |a_n| = N.$$

Solution. Following the hint, let A_N be the set of numbers z that satisfy $a_0 z^n + a_1 z^{n-1} + \cdots + a_{n-1} z + a_n = 0$, for some coefficients $a_0, \dots, a_n$ which satisfy

$$n + |a_0| + |a_1| + \cdots + |a_n| = N.$$

By the fundamental theorem of algebra, $a_0 z^n + a_1 z^{n-1} + \cdots + a_{n-1} z + a_n = 0$ has at most n solutions, so each A_N is finite. Hence the set of algebraic numbers, which is the union

$$\bigcup_{N=2}^{\infty} A_N$$

is at most countable. Since all rational numbers are algebraic, it follows that the set of algebraic numbers is exactly countable. □

Exercise 2.6 ([Rud76] 2.3)**.** Prove that there exist real numbers which are not algebraic.

Solution. By the previous exercise, the set of real algebraic numbers is countable. If every real number were algebraic, the entire set of real numbers would be countable, a contradiction. □

Exercise 2.7 ([Rud76] 2.4)**.** Is the set of irrational real numbers countable?

Solution. No. If $\mathbb{R} \setminus \mathbb{Q}$ were countable, $\mathbb{R} = \mathbb{Q} \cup (\mathbb{R} \setminus \mathbb{Q})$ would be countable, which is clearly false. □

II
Linear Algebra

Chapter 3

Vector Spaces

3.1 Definition of Vector Space

Let $\mathbf{F}$ denote a field, which can mean either $\mathbb{R}$ or $\mathbb{C}$.

Definition 3.1 (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 $\mathbf{F}$.

Elements of a vector space are called *vectors* or *points*.

Remark. 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* $\mathbf{F}$.

A vector space over $\mathbb{R}$ is called a *real vector space*; a vector space over $\mathbb{C}$ is called a *complex vector space*.

Lemma 3.2 (Uniqueness of additive identity)**.** *A vector space has a unique additive identity.*

Proof. Suppose that $\mathbf{0}$ 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. □

Lemma 3.3 (Uniqueness of additive inverse)**.** *Every element in a vector space has a unique additive inverse.*

Proof. Let $v \in V$. Suppose w and w' are additive inverses of v. Then

$$w = w + \mathbf{0} = w + (v + w') = (w + v) + w' = \mathbf{0} + w' = w'.$$

□

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, and define $w - v$ to be $w + (-v)$.

We now prove some seemingly trivial facts.

Lemma 3.4.

(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) Let $v \in V$,
$$0v = (0 + 0)v = 0v + 0v.$$
Adding the additive inverse of $0v$ to both sides of the equation gives $\mathbf{0} = 0v$.

(ii) Let $a \in \mathbf{F}$,
$$a\mathbf{0} = a(\mathbf{0} + \mathbf{0}) = a\mathbf{0} + a\mathbf{0}.$$
Adding the additive inverse of $a\mathbf{0}$ to both sides of the equation gives $\mathbf{0} = a\mathbf{0}$.

(iii) Let $v \in V$,
$$v + (-1)v = 1v + (-1)v = (1 + (-1))v = 0v = \mathbf{0}.$$
Since $v + (-1)v = \mathbf{0}$, $(-1)v$ is the additive inverse of v.

□

Example 3.5 (n-tuple space)**.** Let $\mathbf{F}^n$ be the set of n-tuples whose elements belong to $\mathbf{F}$:

$$\mathbf{F}^n := \{(x_1, \dots, x_n) \mid x_i \in \mathbf{F}\}$$

For $x = (x_1, \dots, x_n) \in \mathbf{F}^n$ and $i = 1, \dots, n$, we say that x_i is the i-th *coordinate* of x.

Define addition and scalar multiplication on $\mathbf{F}^n$ as

$$(x_1,\dots,x_n)+(y_1,\dots,y_n)=(x_1+y_1,\dots,x_n+y_n)$$
$$\lambda(x_1,\dots,x_n)=(\lambda x_1,\dots,\lambda x_n)$$

Then $\mathbf{F}^n$ is a vector space over $\mathbf{F}$.

Example 3.6. Let $\mathbf{F}^\infty$ be the set of all sequences of elements of $\mathbf{F}$:

$$\mathbf{F}^\infty := \{(x_1,x_2,\dots) \mid x_i \in \mathbf{F}\}$$

Define addition and scalar multiplication on $\mathbf{F}^\infty$ as

$$(x_1,x_2,\dots)+(y_1,y_2,\dots)=(x_1+y_1,x_2+y_2,\dots)$$
$$\lambda(x_1,x_2,\dots)=(\lambda x_1,\lambda x_2,\dots)$$

Then $\mathbf{F}^\infty$ is a vector space over $\mathbf{F}$, where the additive identity is $\mathbf{0}=(0,0,\dots)$.

Example 3.7 (Space of functions from a set to a field)**.** If S is a set, $\mathbf{F}^S := \{f \mid f\colon S \to \mathbf{F}\}$. Define addition and scalar multiplication on $\mathbf{F}^S$ as

$$(f+g)(x)=f(x)+g(x) \quad (x \in S)$$
$$(\lambda f)(x)=\lambda f(x) \quad (x \in S)$$

for all $f,g \in \mathbf{F}^S$, $\lambda \in \mathbf{F}$. Then $\mathbf{F}^S$ is a vector space over $\mathbf{F}$ (if S is a non-empty set), where the additive identity of $\mathbf{F}^S$ is the function $0 : S \to \mathbf{F}$ defined as

$$0(x)=0 \quad (\forall x \in S)$$

and for $f \in \mathbf{F}^S$, additive inverse of f is the function $-f : S \to \mathbf{F}$ defined as

$$(-f)(x)=-f(x) \quad (\forall x \in S)$$

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 3.8 (Complexification)**.** Suppose V is a real vector space. The *complexifcation* of V, denoted by $V_{\mathbb{C}}$, equals $V \times V$. An element of $V_{\mathbb{C}}$ is an ordered pair (u,v), where $u,v \in V$, which we write as $u+iv$.

- Addition on $V_{\mathbb{C}}$ is defined as

  $$(u_1+iv_1)+(u_2+iv_2)=(u_1+u_2)+i(v_1+v_2)$$

  for all $u_1,v_2,u_2,v_2 \in V$.

- Complex scalar multiplication on $V_{\mathbb{C}}$ is defined as

  $$(a+bi)(u+iv)=(au-bv)+i(av+bu)$$

for all $a, b \in \mathbb{R}$ and all $u, v \in V$.

Then $V_{\mathbb{C}}$ is a (complex) vector space.

3.2 Subspaces

Whenever we have a mathematical object with some structure, we want to consider subsets that also have the same structure.

Definition 3.9 (Subspace)**.** We say $U \subset V$ is a ***subspace*** of V, denoted as $U \leq V$, if U is also a vector space (with the same addition and scalar multiplication as on 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 3.10 (Subspace test)**.** *Suppose $U \subset V$. Then $U \leq V$ if and only if U satisfies the following conditions:*

(i) $\mathbf{0} \in U$*;* *(additive identity)*

(ii) $u + w \in U$ *for all* $u, w \in U$*;* *(closed under addition)*

(iii) $\lambda u \in U$ *for all* $\lambda \in \mathbf{F}$, $u \in U$. *(closed under scalar multiplication)*

Proof.

$\boxed{\Longrightarrow}$ If $U \leq V$, 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. □

Lemma 3.11 (A subspace of a subspace is a subspace)**.** *If $U \leq V$ and $W \leq U$, then $W \leq V$.*

Proof. This is immediate from the definition of a subspace. □

Lemma 3.12 (Intersection of subspaces is a subspace)**.** *Let $\{U_i \mid i \in I\}$ be a collection of subspaces of V. Then $\bigcap_{i \in I} U_i \leq V$.*

Proof. Let $U = \bigcap_{i \in I} U_i$.

(i) Since each $U_i \leq V$, $\mathbf{0} \in U_i$ so $\mathbf{0} \in U$.

(ii) Let $u, w \in U$, then $u, w \in U_i$. Since $U_i \leq V$, $u + w \in U_i$ so $u + w \in U$.

(iii) Let $\lambda \in \mathbf{F}$, $u \in U$, then $u \in U_i$. Since $U_i \leq V$, $\lambda u \in U_i$ so $\lambda u \in U$.

□

Definition 3.13 (Sum of subsets)**.** Suppose $U_1, \ldots, U_n \subset V$. The ***sum*** of $U_1, \ldots, U_n$ is the set of all possible sums of elements of $U_1, \ldots, U_n$:

$$U_1 + \cdots + U_n := \{u_1 + \cdots + u_n \mid u_i \in U_i\}.$$

Example 3.14.

- Let $U = \{(x, 0, 0) \in \mathbf{F}^3 \mid x \in F\}$ and $W = \{(0, y, 0) \in \mathbf{F}^3 \mid y \in \mathbf{F}\}$. Then

$$U + W = \{(x, y, 0) \mid x, y \in \mathbf{F}\}.$$

- Let $U = \{(x, x, y, y) \in \mathbf{F}^4 \mid x, y \in \mathbf{F}\}$ 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 3.15. *Suppose $U_1, \ldots, U_n \leq V$. Then $U_1 + \cdots + U_n$ is the smallest subspace of V containing $U_1, \ldots, U_n$.*

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, $U_1 + \cdots + U_n \leq V$.

Let M be the smallest subspace of V containing $U_1, \ldots, U_n$. We want to show that $U_1 + \cdots + U_n = M$. To do so, we show double inclusion.

$\boxed{\supset}$ For all $u_i \in U_i$ $(1 \leq i \leq n)$,

$$u_i = \mathbf{0} + \cdots + \mathbf{0} + u_i + \mathbf{0} + \cdots + \mathbf{0} \in U_1 + \cdots + U_n,$$

where all except one of the u's are $\mathbf{0}$. Thus $U_i \subset U_1 + \cdots + U_n$ for $1 \leq i \leq n$. Hence $M \subset U_1 + \cdots + U_n$.

$\boxed{\subset}$ Conversely, every subspace of V containing $U_1, \ldots, U_n$ 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 3.16 (Direct sum)**.** Suppose $U_1, \ldots, U_n \leq V$. We say $U_1 + \cdots + U_n$ is a ***direct sum*** if each element of $U_1 + \cdots + U_n$ can be written in only one way as a sum $u_1 + \cdots + u_n$, $u_i \in U_i$. In this case, we denote the sum as

$$U_1 \oplus \cdots \oplus U_n.$$

Example 3.17.

- 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 $\mathbf{F}^n = U_1 \oplus \cdots \oplus U_n$.

The definition of direct sum requires every vector in the sum to have a unique representation as an appropriate sum. The next result shows that when deciding whether a sum of subspaces is a direct sum, we only need to consider whether $\mathbf{0}$ can be uniquely written as an appropriate sum.

Lemma 3.18 (Condition for direct sum)**.** *Suppose $V_1, \dots, V_n \leq V$. Then $V_1 \oplus \dots \oplus V_n$ if and only if $v_1 + \dots + v_n = \mathbf{0}$ implies $v_1 = \dots = v_n = \mathbf{0}$.*

Proof.

$\boxed{\text{(i)} \implies \text{(ii)}}$ Suppose $V_1 + \dots + V_n$ is a direct sum. Then by the definition of direct sum, the only way to write $\mathbf{0}$ as a sum $u_1 + \dots + u_n$ is by taking $u_i = \mathbf{0}$.

$\boxed{\text{(ii)} \implies \text{(i)}}$ Suppose that the only way to write $\mathbf{0}$ as a sum $v_1 + \dots + v_n$ by taking $v_1 = \dots = v_n = \mathbf{0}$. To show that $v \in V_1 + \dots + V_n$ is a direct sum, let $v \in V_1 + \dots + V_n$. Then

$$v = v_1 + \dots + v_n \tag{I}$$

for some $v_i \in V_i$. To show that this representation is unique, suppose

$$v = v_1' + \dots + v_n' \tag{II}$$

for some $v_i' \in V_i$. Substracting (II) from (I) gives

$$\mathbf{0} = (v_1 - v_1') + \dots + (v_n - v_n').$$

Since $v_i - v_i' \in V_i$, the equation above implies $v_i - v_i' = \mathbf{0}$, so $v_i = v_i'$. Hence there is only one unique way to represent $v_1 + \dots + v_n$. $\square$

The next result provides a characterisation for direct sum.

Lemma 3.19. *Suppose $U, W \leq V$. Then $U + W$ is a direct sum if and only if $U \cap W = \{\mathbf{0}\}$.*

Proof.

$\boxed{\implies}$ Suppose that $U + W$ is a direct sum. Let $v \in U \cap W$, we will show that $v = \mathbf{0}$.

Note that $\mathbf{0} = v + (-v)$, 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 must have $v = \mathbf{0}$. Hence $U \cap W = \{\mathbf{0}\}$.

$\boxed{\impliedby}$ 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 3.18, $U + W$ is a direct sum. $\square$

3.3 Span and Linear Independence

Definition 3.20 (Linear combination)**.** We say $v \in V$ is a ***linear combination*** of $v_1, \dots, v_n \in V$ if there exists $a_1, \dots, a_n \in \mathbf{F}$ such that

$$\begin{aligned} v &= a_1 v_1 + \cdots + a_n v_n \\ &= \sum_{i=1}^{n} a_i v_i. \end{aligned}$$

Definition 3.21 (Span)**.** The ***span*** of $\{v_1, \dots, v_n\}$ is the set of all linear combinations of $v_1, \dots, v_n$:

$$\operatorname{span}(v_1, \dots, v_n) := \{a_1 v_1 + \cdots + a_n v_n \mid a_i \in \mathbf{F}\}.$$

We say $v_1, \dots, v_n$ *spans* V if $\operatorname{span}(v_1, \dots, v_n) = V$.
If $S \subset V$ is such that $\operatorname{span}(S) = V$, we say S *spans* V, and S is a *spanning set* for V:

$$\operatorname{span}(S) := \{a_1 v_1 + \cdots + a_n v_n \mid v_i \in S, a_i \in \mathbf{F}\}.$$

Proposition 3.22. $\operatorname{span}(v_1, \dots, v_n)$ *in* V *is the smallest subspace of* V *containing* $v_1, \dots, v_n$*.*

Proof. First we show that $\operatorname{span}(v_1, \dots, v_n) \leq V$, using the subspace test.

(i) $\mathbf{0} = 0v_1 + \cdots + 0v_n \in \operatorname{span}(v_1, \dots, v_n)$

(ii) $(a_1v_1 + \cdots + a_nv_n) + (c_1v_1 + \cdots + c_nv_n) = (a_1 + c_1)v_1 + \cdots + (a_n + c_n)v_n \in \operatorname{span}(v_1, \dots, v_n)$, so $\operatorname{span}(v_1, \dots, v_n)$ is closed under addition.

(iii) $\lambda(a_1v_1 + a_nv_n) = (\lambda a_1)v_1 + \cdots + (\lambda a_n)v_n \in \operatorname{span}(v_1, \dots, v_n)$, so $\operatorname{span}(v_1, \dots, v_n)$ is closed under scalar multiplication.

Let M be the smallest vector subspace of V containing $v_1, \dots, v_n$. We claim that $M = \operatorname{span}(v_1, \dots, v_n)$.

$\boxed{\subset}$ Each v_i is a linear combination of $v_1, \dots, v_n$, 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 $v_1, \dots, v_n$, we have that $v_i \in \operatorname{span}(v_1, \dots, v_n)$. But M is the smallest vector subspace containing $v_1, \dots, v_n$, so

$$M \subset \operatorname{span}(v_1, \dots, v_n).$$

$\boxed{\supset}$ Since $v_i \in M$ $(1 \leq i \leq n)$ 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 $v_1, \dots, v_n$). Thus

$$\operatorname{span}(v_1, \dots, v_n) \subset M.$$

□

Definition 3.23 (Finite-dimensional)**.** We say V is ***finite-dimensional*** if it has a finite spanning set; otherwise, it is *infinite-dimensional*.

Example 3.24. For positive integer n, $\mathbf{F}^n$ is finite-dimensional.

Proof. Suppose $(x_1, x_2, \dots, x_n) \in \mathbf{F}^n$, then

$$(x_1, x_2, \dots, x_n) = x_1(1, 0, \dots, 0) + x_2(0, 1, \dots, 0) + \cdots + x_n(0, 0, \dots, 1)$$

so

$$(x_1, \dots, x_n) \in \operatorname{span}\left((1, 0, \dots, 0), (0, 1, \dots, 0), \dots, (0, \dots, 0, 1)\right).$$

The vectors $(1, 0, \dots, 0), (0, 1, \dots, 0), \dots, (0, \dots, 0, 1)$ spans $\mathbf{F}^n$, so $\mathbf{F}^n$ is finite-dimensional. □

Definition 3.25 (Linear independence)**.** We say $v_1, \dots, v_n$ are ***linearly independent*** in V if

$$a_1 v_1 + \cdots + a_n v_n = \mathbf{0} \implies a_1 = \cdots = a_n = 0.$$

Otherwise, the vectors are *linearly dependent*.
We say $S \subset V$ is linearly independent if every finite subset of S is linearly independent.

Lemma 3.26 (Compare coefficients)**.** *Let $v_1, \dots, v_n$ 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 $a_i = b_i$ ($1 \le i \le n$).

Proof. Exercise. □

The following are easy consequences of the definition.

1. Any set which contains a linearly dependent set is linearly dependent.
2. Any subset of a linearly independent set is linearly independent.
3. Any set which contains **0** is linearly dependent.
4. A set S of vectors is linearly independent if and only if each finite subset of S is linearly independent.

The following result will often be useful; (i) states that given a linearly dependent set of vectors, one of the vectors is in the span of the previous ones; furthermore (ii) states that we can throw out that vector without changing the span of the original set.

Lemma 3.27 (Linear dependence lemma)**.** *Suppose $v_1, \dots, v_n$ are linearly dependent in V. Then there exists v_i such that the following hold:*

> *(i)* $v_i \in \text{span}(v_1, \ldots, v_{i-1})$
>
> *(ii)* $\text{span}(v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n) = \text{span}(v_1, \ldots, v_n)$

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 $i = \max\{1, \ldots, n\}$ such that $a_i \neq 0$. Then

$$v_i = -\frac{a_1}{a_i} v_1 - \cdots - \frac{a_{i-1}}{a_i} v_{i-1},$$

since $a_{i-1}, \ldots, a_n = 0$. Thus v_i can be written as a linear combination of $v_1, \ldots, v_{i-1}$, so $v_i \in \text{span}(v_1, \ldots, v_{i-1})$.

(ii) Now suppose i is such that $v_i \in \text{span}(v_1, \ldots, v_{i-1})$. Then there exists $b_1, \ldots, b_{i-1} \in \mathbf{F}$ be such that

$$v_i = b_1 v_1 + \cdots + b_{i-1} v_{i-1}. \tag{I}$$

Suppose $u \in \text{span}(v_1, \ldots, v_n)$. Then there exists $c_1, \ldots, c_n \in \mathbf{F}$ such that

$$u = c_1 v_1 + \cdots + c_n v_n. \tag{II}$$

Substituting (I) into (II) gives

$$\begin{aligned} u &= c_1 v_1 + \cdots + c_{i-1} v_{i-1} + c_i v_i + c_{i+1} v_{i+1} + \cdots + c_n v_n \\ &= c_1 v_1 + \cdots + c_{i-1} v_{i-1} + c_i (b_1 v_1 + \cdots + b_{i-1} v_{i-1}) + c_{i+1} v_{i+1} + \cdots + c_n v_n \\ &= c_1 v_1 + \cdots + c_{i-1} v_{i-1} + c_i b_1 v_1 + \cdots + c_i b_{i-1} v_{i-1} + c_{i+1} v_{i+1} + \cdots + c_n v_n \\ &= (c_1 + b c_i) v_1 + \cdots + (c_{i-1} + b_{i-1} c_i) v_{i-1} + c_{i+1} v_{i+1} + \cdots + c_n v_n. \end{aligned}$$

Thus $u \in \text{span}(v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n)$. This shows that removing v_i from $v_1, \ldots, v_n$ does not change the span of the set.

□

The next result says that no linearly independent set in V is longer than a spanning set in V.

> **Proposition 3.28.** *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.*

That is,

$$\{\text{linearly independent}\} \leq \{\text{spanning set}\}. \tag{3.1}$$

Proof. Suppose $A = \{u_1, \ldots, u_m\}$ is linearly independent in V, $B = \{w_1, \ldots, w_n\}$ spans V. We want to show that $m \leq n$.

We do so through the process described below with m steps; in each step, we add one of the u's and remove one of the w's.

Step 1 Since B spans V, if we add any other vector to B, we will get a linearly dependent set, 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 to B, then the new set

$$\{u_1, w_1, \dots, w_n\}$$

is linearly dependent.

By the linear independence lemma, one of the vectors in the above set is a linear combination of the previous vectors. Since $\{u_1, \dots, u_m\}$ is linearly independent, $u_1 \neq \mathbf{0}$ so $u_1 \notin \operatorname{span}\{\,\} = \{\mathbf{0}\}$. Hence the linear dependence lemma implies we can remove one of the w's, so that the new set B (of length n) consisting of u_1 and the remaining w's spans V.

Step i **(**$2 \leq i \leq m$**)** The set B (of length n) from step $i-1$ spans V. In particular, u_i is in the span of B. If we add u_i to B, placing it just after $u_1, \dots, u_{i-1}$, then the new set (of length $n+1$)

$$\{u_1, \dots, u_{i-1}, u_i, w\text{'s}\}$$

is linearly dependent.

By the linear dependence lemma, one of the vectors in this set is in the span of the previous ones. Since $u_1, \dots, u_i$ are linearly independent, this vector cannot be one of the u's. Hence there still must be at least one remaining w at this step. We can remove from our new set (after adjoining u_i in the proper place) a w that is a linear combination of the previous vectors in the set, so that the new set B (of length n) consisting of $u_1, \dots, u_i$ and the remaining w's spans V.

After step m, we have added *all* the u's and the process stops. At each step as we add a u to B, the linear dependence lemma implies that there is some w to remove. Hence there must be at least as many w's as u's. □

We can use this result to show, without any computations, that certain sets are not linearly independent and that certain sets do not span a given vector space.

Example 3.29.

- $(1,0,0), (0,1,0), (0,0,1)$ spans $\mathbb{R}^3$. Thus no set of length larger than three is linearly independent in $\mathbb{R}^3$.
- $(1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1)$ is linearly independent in $\mathbb{R}^4$. Thus no set of length less than four spans $\mathbb{R}^4$.

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 3.30. *Every subspace of a finite-dimensional vector space is finite-dimensional.*

Proof. Suppose V is finite-dimensional, $U \leq V$. To show that U is finite-dimensional, we shall construct a spanning set of vectors in U, via the following steps.

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 set of vectors.

Step i Our set so far is $\{v_1, \dots, v_{i-1}\}$.

If $U = \operatorname{span}(v_1, \dots, v_{i-1})$, then U is finite-dimensional and we are done.

Otherwise, choose $v_i \in U$ such that $v_i \notin \operatorname{span}(v_1, \dots, v_{i-1})$ and add it to our set.

After each step, we have constructed a set of vectors such that no vector in this set is in the span of the previous vectors; by the linear dependence lemma, our constructed set is linearly independent.

By 3.28, 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. □

3.4 Bases

Definition 3.31 (Basis)**.** We say $B = \{v_1 \ldots, v_n\}$ is a ***basis*** of V if

(i) B is linearly independent in V, and

(ii) B is a spanning set of V.

Example 3.32 (Standard basis)**.** Let $\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, where the i-th coordinate is 1. Then $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is a basis of $\mathbf{F}^n$, known as the *standard basis* of $\mathbf{F}^n$.

The next result helps explain why bases are useful.

Lemma 3.33 (Criterion for basis)**.** *Let $B = \{v_1, \ldots, v_n\}$ be a set of vectors in V. Then B is a basis of V if and only if every $v \in V$ can be uniquely expressed as a linear combination of $v_1, \ldots, v_n$.*

Proof.

$\boxed{\implies}$ Let $v \in V$. Since B is a basis of V, there exist $a_1, \ldots, a_n \in \mathbf{F}$ such that

$$v = a_1 v_1 + \cdots + a_n v_n. \tag{I}$$

To show that the representation is unique, suppose that $c_1, \ldots, c_n \in \mathbf{F}$ also satisfy

$$v = c_1 v_1 + \cdots + c_n v_n. \tag{II}$$

Subtracting (II) from (I) 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.

$\boxed{\impliedby}$ Suppose 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 $a_1, \ldots, a_n \in \mathbf{F}$ are such that

$$a_1 v_1 + \cdots + a_n v_n = \mathbf{0}.$$

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, we conclude that B is a basis of V. $\square$

A spanning set in a vector space may not be a basis because it is not linearly independent. The next result says that given any spanning set, we can *remove* some vectors so that the remaining set is linearly independent and still spans the vector space.

Lemma 3.34. *In a vector space, every spanning set can be reduced to a basis.*

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 i $(2 \leq i \leq n)$ If $v_i \in \operatorname{span}(v_1, \ldots, v_{i-1})$, delete v_i from B. If $v_i \notin \operatorname{span}(v_1, \ldots, v_{i-1})$, leave B unchanged.

Since we only delete vectors from B that are in the span of the previous vectors, by the linear dependence lemma, the set 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, we conclude that B is a basis of V. □

Corollary 3.35. *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 3.34, the spanning set can be reduced to a basis. □

We now show that given any linearly independent set, we can *add* some vectors so that the extended set is still linearly independent but also spans the space.

Lemma 3.36. *In a finite-dimensional vector space, every linearly independent set can be extended to a basis.*

Proof. Suppose $\{u_1, \ldots, u_m\}$ is linearly independent in V, and $\{w_1, \ldots, w_n\}$ spans V. Then the set

$$\{u_1, \ldots, u_m, w_1, \ldots, w_n\}$$

spans V. By 3.34, we can reduce this set 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 3.37. *Suppose V is finite-dimensional, $U \leq V$. Then there exists $W \leq V$ such that $V = U \oplus W$.*

Proof. Since V is finite-dimensional and $U \leq V$, by 3.30, U is finite-dimensional. By 3.35, U has a basis, say $B = \{u_1, \ldots, u_n\}$.

Since B is linearly independent, by 3.36, B can be extended to a basis of V, say

$$\{u_1, \ldots, u_n, w_1, \ldots, w_n\}.$$

Claim. $W = \operatorname{span}(w_1, \dots, w_n)$.

We need to show that $V = U \oplus W$; by 3.18, we need to show (i) $V = U + W$, and (ii) $U \cap W = \{\mathbf{0}\}$.

(i) Let $v \in V$. Since $\{u_1, \dots, u_n, w_1, \dots, w_n\}$ spans V, there exists $a_1, \dots, a_n, b_1, \dots, b_n \in \mathbf{F}$ such that

$$v = a_1 u_1 + \cdots + a_n u_n + b_1 w_1 + \cdots + b_n w_n.$$

Take $u = a_1 u_1 + \cdots + a_n u_n \in U$, $w = b_1 w_1 + \cdots + b_n w_n \in W$. Then $v = u + w \in U + W$, so $V = U + W$.

(ii) Let $v \in U \cap W$. Since $v \in U$, v can be written as a linear combination of $u_1, \dots, u_n$:

$$v = a_1 u_1 + \cdots + a_n u_n. \tag{I}$$

Since $v \in W$, v can be written as a linear combination of $w_1, \dots, w_n$:

$$v = b_1 w_1 + \cdots + b_n w_n. \tag{II}$$

Subtracting (II) from (I) gives

$$\mathbf{0} = a_1 u_1 + \cdots + a_n u_n - b_1 w_1 - \cdots - b_n w_n.$$

Since $u_1, \dots, u_n, w_1, \dots, w_n$ are linearly independent, we have $a_i = b_i = 0$ for all i. Thus $v = \mathbf{0}$, so $U \cap W = \{\mathbf{0}\}$.

□

3.5 Dimension

Lemma 3.38. *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 3.28, $|B_1| \leq |B_2|$.

Similarly, by definition, B_2 is linearly independent in V and B_1 spans V, so $|B_2| \leq |B_1|$.

Since $|B_1| \leq |B_2|$ and $|B_2| \leq |B_1|$, 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 3.39 (Dimension). The ***dimension*** of V is the length of any basis of V, denoted by $\dim V$.

Lemma 3.40 (Dimension of subspace). *Suppose V is finite-dimensional, $U \leq V$. Then $\dim U \leq \dim V$.*

Proof. Since V is finite-dimensional and $U \leq V$, 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 3.28, $|B_U| \leq |B_V|$, so

$$\dim U = |B_U| \leq |B_V| = \dim V.$$

□

To check that a set 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 set has the *right length*, then we only need to check that it satisfies one of the two required properties.

Proposition 3.41. *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 $\dim V = n$, and $\{v_1, \dots, v_n\}$ is linearly independent in V.

By 3.36, $\{v_1, \dots, v_n\}$ can be extended to a basis of V. However, every basis of V has length n, which means no elements are added to $\{v_1, \dots, v_n\}$. Hence $\{v_1, \dots, v_n\}$ is a basis of V.

(ii) Suppose $\dim V = n$, and $\{v_1, \dots, v_n\}$ spans V.

By 3.34, $\{v_1, \dots, v_n\}$ can be reduced to a basis of V. However, every basis of V has length n, which means no elements are removed from $\{v_1, \dots, v_n\}$. Hence $\{v_1, \dots, v_n\}$ is a basis of V.

□

Corollary 3.42. *Suppose V is finite-dimensional, $U \leq V$. If $\dim U = \dim V$, then $U = V$.*

Proof. Let $\dim U = \dim V = n$, let $\{u_1, \dots, u_n\}$ be a basis of U.

Then $\{u_1, \dots, u_n\}$ is linearly independent in V (because it is a basis of U) of length $\dim V$. By 3.41, $\{u_1, \dots, u_n\}$ is a basis of V. In particular every vector in V is a linear combination of $u_1, \dots, u_n$. Thus $U = V$. □

The next result gives a formula for the dimension of the sum of two subspaces of a finite-dimensional vector space.

Lemma 3.43 (Dimension of sum)**.** *Suppose V is finite-dimensional, $U_1, U_2 \leq V$. Then*

$$\dim(U_1 + U_2) = \dim U_1 + \dim U_2 - \dim(U_1 \cap U_2). \tag{3.2}$$

Proof. Let $\{u_1, \dots, u_m\}$ be a basis of $U_1 \cap U_2$; thus $\dim(U_1 \cap U_2) = m$.

Since $\{u_1, \dots, u_m\}$ is a basis of $U_1 \cap U_2$, it is linearly independent in U_1. By 3.36, $\{u_1, \dots, u_m\}$ can be extended to a basis $\{u_1, \dots, u_m, v_1, \dots, v_j\}$ of U_1; thus $\dim U_1 = m + j$. Similarly, extend $\{u_1, \dots, u_m\}$ to a basis $\{u_1, \dots, u_m, v_1, \dots, v_k\}$ of U_2; thus $\dim U_2 = m + k$.

We will show that

$$\{u_1, \dots, u_m, v_1, \dots, v_j, w_1, \dots, w_k\}$$

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 $\{u_1, \dots, u_m, v_1, \dots, v_j, w_1, \dots, w_k\}$ is linearly independent. To prove this, suppose

$$a_1u_1 + \cdots + a_mu_m + b_1v_1 + \cdots + b_jv_j + c_1w_1 + \cdots + c_kw_k = \mathbf{0}, \tag{I}$$

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. (I) can be rewritten as

$$c_1w_1 + \cdots + c_kw_k = -a_1u_1 - \cdots - a_mu_m - b_1v_1 - \cdots - b_jv_j,$$

which shows that $c_1w_1 + \cdots + c_kw_k \in U_1$. But actually all the w_i's are in U_2, so $c_1w_1 + \cdots + c_kw_k \in U_2$. Thus $c_1w_1 + \cdots + c_kw_k \in U_1 \cap U_2$. Then we can write

$$c_1w_1 + \cdots + c_kw_k = d_1u_1 + \cdots + d_mu_m$$

for some $d_i \in \mathbf{F}$. But $u_1, \dots, u_m, w_1, \dots, w_k$ are linearly independent, so $c_i = d_i = 0$ for all i. Thus our original equation (I) becomes

$$a_1u_1 + \cdots + a_mu_m + b_1v_1 + \cdots + b_jv_j = \mathbf{0}.$$

Since $u_1, \ldots, u_m, v_1, \ldots, v_j$ are linearly independent, $a_i = b_i = 0$ for all i, as desired. $\square$

Exercises

Exercise 3.1 ([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. □

Exercise 3.2 ([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. □

Exercise 3.3 ([Axl24] 2A Q12)**.** Suppose $\{v_1, \dots, v_n\}$ is linearly independent in V, $w \in V$. Prove that if $\{v_1 + w, \dots, v_n + w\}$ is linearly dependent, then $w \in \operatorname{span}(v_1, \dots, v_n)$.

Solution. If $\{v_1 + w, \dots, v_n + w\}$ is linearly dependent, then there exists $a_1, \dots, a_n \in \mathbf{F}$, 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 $\{v_1, \dots, v_n\}$ 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 $v_1, \dots, v_n$. Thus by definition of span, $w \in \operatorname{span}(v_1, \dots, v_n)$. □

Exercise 3.4 ([Axl24] 2A Q14)**.** Suppose $\{v_1, \dots, v_n\} \subset V$. Let

$$w_i = v_1 + \cdots + v_i \quad (i = 1, \dots, n)$$

Show that $\{v_1, \dots, v_n\}$ is linearly independent if and only if $\{w_1, \dots, w_n\}$ 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. □

Exercise 3.5 ([Axl24] 2A Q18)**.** Prove that $\mathbf{F}^\infty$ is infinite-dimensional.

Solution. Suppose, for a contradiction, that $\mathbf{F}^\infty$ is finite-dimensional, i.e., there exists a finite spanning set $\{v_1, \dots, v_n\}$. 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 3.28, we have reached a contradiction. □

Exercise 3.6 ([Axl24] 2B Q5)**.** Suppose V is finite-dimensional, $U, W \le V$ 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. □

Exercise 3.7 ([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 $v_1 + v_2, v_2 + v_3, v_3 + v_4, v_4$. $\square$

Exercise 3.8 ([Axl24] 2B Q10)**.** Suppose $U, W \leq V$ 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

$$v = u + w = \sum_{i=1}^{m} a_i u_i + \sum_{i=1}^{n} b_j w_j.$$

$\square$

Exercise 3.9 ([Axl24] 2C Q8)**.**

Exercise 3.10 ([Axl24] 2C Q16)**.**

Exercise 3.11 ([Axl24] 2C Q17)**.** Suppose that $V_1, \ldots, V_n \leq V$ are finite-dimensional. Prove that $V_1 + \cdots + V_n$ is finite-dimensional, and

$$\dim(V_1 + \cdots + V_n) \leq \dim V_1 + \cdots + \dim V_n.$$

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

$$\dim(M_k + V_{k+1}) \leq \dim V_1 + \cdots + \dim V_{k+1},$$

which is finite. $\square$

Chapter 4

Linear Maps

4.1 Vector Space of Linear Maps

Definition 4.1 (Linear map)**.** A ***linear map*** from V to W is a function $T\colon V \to W$ satisfying the following properties:

(i) $T(v+w) = Tv + Tw$ for all $v, w \in V$; (additivity)

(ii) $T(\lambda v) = \lambda T(v)$ for all $\lambda \in \mathbf{F}$, $v \in V$. (homogeneity)

Notation. If there is no ambiguity, we omit the parantheses and write Tv instead of $T(v)$.

Notation. Let $\mathcal{L}(V, W)$ denote the set of linear maps from V to W, and let $\mathcal{L}(V)$ denote the set of linear maps on V (from V to V).

Example 4.2. If V is any vector space, the *identity map* I, defined by $Iv = v$, is a linear map on V. The *zero map* 0, defined by $0v = v$, is a linear map on 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 4.3 (Linear map lemma)**.** *Suppose* $\{v_1, \dots, v_n\}$ *is a basis of* V*, and* $w_1, \dots, w_n \in W$*. Then there exists a unique linear map* $T \in \mathcal{L}(V, W)$ *such that*

$$Tv_i = w_i \quad (i = 1, \dots, n)$$

Proof.

$\boxed{\text{Existence}}$ Define $T\colon V \to W$ as

$$T(c_1 v_1 + \cdots + c_n v_n) = c_1 w_1 + \cdots + c_n w_n,$$

for some $c_i \in \mathbf{F}$. Since $\{v_1, \dots, v_n\}$ is a basis of V, by 3.33, each $v \in V$ can be uniquely expressed as a linear combination of $v_1, \dots, v_n$, thus the equation above does indeed define a function $T\colon V \to W$. For

i ($1 \le i \le n$), take $c_i = 1$ and the other c's equal to 0, then

$$T\left(0v_1 + \cdots + 1v_i + \cdots + 0v_n\right) = 0w_1 + \cdots + 1w_i + \cdots + 0w_n$$

which shows that $Tv_i = w_i$.

We now show that $T\colon V \to W$ 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 $v = c_1v_1 + \cdots + c_nv_n$,

$$\begin{aligned}
T(\lambda v) &= T(\lambda c_1v_1 + \cdots + \lambda c_nv_n) \\
&= \lambda c_1w_1 + \cdots + \lambda c_nw_n \\
&= \lambda(c_1w_1 + \cdots + c_nw_n) \\
&= \lambda Tv.
\end{aligned}$$

$\boxed{\text{Uniqueness}}$ Suppose that $T \in \mathcal{L}(V,W)$ and $Tv_i = w_i$ for $i = 1, \dots, n$. 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 $\operatorname{span}\{v_1, \dots, v_n\}$. Since $\{v_1, \dots, v_n\}$ is a basis of V, this implies that T is uniquely determined on V. ☐

Lemma 4.4. *$\mathcal{L}(V,W)$ is a vector space, with addition and scalar multiplication defined as follows: for all $S, T \in \mathcal{L}(V,W)$, $\lambda \in \mathbf{F}$,*

$$\begin{aligned}
(S+T)(v) &= Sv + Tv \\
(\lambda T)(v) &= \lambda(Tv)
\end{aligned}$$

for all $v \in V$.

Proof. Exercise. ☐

Definition 4.5 (Product of linear maps). Suppose $T \in \mathcal{L}(U,V)$, $S \in \mathcal{L}(V,W)$. Define the ***product*** $ST \in \mathcal{L}(U,W)$ by

$$(ST)(u) := S(Tu) \quad (u \in U).$$

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.

Lemma 4.6 (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) Identity: $TI = IT = T$ *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. □

Lemma 4.7. *Suppose* $T \in \mathcal{L}(V, W)$*. Then* $T(\mathbf{0}) = \mathbf{0}$*.*

Proof. By additivity,

$$T(\mathbf{0}) = T(\mathbf{0} + \mathbf{0}) = T(\mathbf{0}) + T(\mathbf{0}).$$

Add the additive inverse of $T(\mathbf{0})$ to each side of the equation to obtain $T(\mathbf{0}) = \mathbf{0}$. □

4.2 Kernel and Image

Definition 4.8 (Kernel)**.** Suppose $T \in \mathcal{L}(V, W)$. The ***kernel*** of T is

$$\ker T := \{v \in V \mid Tv = \mathbf{0}\}.$$

That is, $\ker T$ is the subset of V consisting of those vectors that T maps to $\mathbf{0}$.

We check that if $T \in \mathcal{L}(V, W)$, then $\ker T \leq V$.

(i) By 4.7, $T(\mathbf{0}) = \mathbf{0}$, 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 4.9 (Injectivity)**.** Suppose $T \in \mathcal{L}(V, W)$. We say T is ***injective*** if

$$Tu = Tv \implies u = v.$$

The next result provides a useful characterisation of injective linear maps.

Lemma 4.10. *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 4.11 (Image)**.** Suppose $T \in \mathcal{L}(V, W)$. The ***image*** of T is

$$\operatorname{im} T := \{Tv \mid v \in V\}.$$

That is, $\operatorname{im} T$ is the subset of W consisting of those vectors that are of the form Tv for some $v \in V$.

We check that if $T \in \mathcal{L}(V, W)$, then $\operatorname{im} T \leq W$.

(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

$$\lambda w = \lambda Tv = T(\lambda v) \in \operatorname{im} T \implies \lambda w \in \operatorname{im} T.$$

Definition 4.12 (Surjectivity)**.** Suppose $T \in \mathcal{L}(V, W)$. T is ***surjective*** if $\operatorname{im} T = W$.

4.2.1 Fundamental Theorem of Linear Maps

Theorem 4.13 (Fundamental theorem of linear maps)**.** *Suppose V is finite-dimensional, $T \in \mathcal{L}(V, W)$. Then $\operatorname{im} T$ is finite-dimensional, and*

$$\dim V = \dim \ker T + \dim \operatorname{im} T. \tag{4.1}$$

Proof. Let $\{u_1, \dots, u_m\}$ be basis of $\ker T$, then $\dim \ker T = m$. The linearly independent list $u_1, \dots, u_m$ can be extended to a basis

$$\{u_1, \dots, u_m, v_1, \dots, v_n\}$$

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 $\{Tv_1, \dots, Tv_n\}$ 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 $\{u_1, \dots, u_m, v_1, \dots, v_n\}$ 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 $u_i \in \ker T$,

$$\begin{aligned} Tv &= T\left(a_1 u_1 + \cdots + a_m u_m + b_1 v_1 + \cdots + b_n v_n\right) \\ &= a_1 \underbrace{Tu_1}_{\mathbf{0}} + \cdots + a_m \underbrace{Tu_m}_{\mathbf{0}} + b_1 T v_1 + \cdots + b_n v_n \\ &= b_1 T v_1 + \cdots + b_n T v_n \in \operatorname{im} T. \end{aligned}$$

Since every element of $\operatorname{im} T$ can be expressed as a linear combination of $Tv_1, \dots, Tv_n$, we have that $\{Tv_1, \dots, Tv_n\}$ 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 $c_1, \dots, c_n \in \mathbf{F}$ such that

$$c_1 T v_1 + \cdots + c_n T v_n = \mathbf{0}.$$

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 $\{u_1, \dots, u_m\}$ 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 $u_1, \dots, u_m, v_1, \dots, v_n$ are linearly independent, $c_i = d_i = 0$. Since $c_i = 0$, $\{Tv_1, \dots, Tv_n\}$ 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 4.14. *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 $\operatorname{im} T \leq W$, by 3.40, we have that $\dim \operatorname{im} T \leq \dim W$. Let $T \in \mathcal{L}(V, W)$. Then

$$\begin{aligned} \dim \ker T &= \dim V - \dim \operatorname{im} T && \text{[by fundamental theorem of linear maps]} \\ &\geq \dim V - \dim W > 0. \end{aligned}$$

Since $\dim \ker T > 0$, this means that $\ker T$ contains some $v \in V \setminus \{\mathbf{0}\}$, so $\ker T \neq \{\mathbf{0}\}$; hence 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 4.15. *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 surjective.*

Proof. Let $T \in \mathcal{L}(V, W)$. Then

$$\begin{aligned} \dim \operatorname{im} T &= \dim V - \dim \ker T && \text{[by fundamental theorem of linear maps]} \\ &\leq \dim V && [\because \dim \ker T \geq 0] \\ &< \dim W. \end{aligned}$$

Since $\dim \operatorname{im} T < \dim W$, $\operatorname{im} T \neq W$ so T is not surjective. □

Example 4.16 (Homogeneous system of linear equations)**.** Consider the homogeneous system of linear equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= 0 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= 0 \end{aligned} \tag{$*$}$$

where $a_{ij} \in \mathbf{F}$.

Define $T \colon \mathbf{F}^n \to \mathbf{F}^m$ by

$$T(x_1, \dots, x_n) = \left(\sum_{i=1}^n a_{1i}x_i, \dots, \sum_{i=1}^n a_{mi}x_i \right).$$

The solution set of $(*)$ is given by

$$\ker T = \left\{ (x_1, \dots, x_n) \in \mathbf{F}^n \;\middle|\; \sum_{i=1}^{n} a_{1i}x_i = 0, \dots, \sum_{i=1}^{n} a_{mi}x_i = 0 \right\}.$$

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 $\dim \mathbf{F}^n < \dim \mathbf{F}^m$, so T is not surjective. Hence there exists $(c_1, \dots, c_m) \in \mathbf{F}^m$ such that

$$\forall (x_1, \dots, x_n) \in \mathbf{F}^n, \quad T(x_1, \dots, x_n) \neq (c_1, \dots, c_m).$$

Thus the choice of constant terms $(c_1, \dots, c_m)$ 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 $(x_1, \dots, x_n)$. □

4.3 Matrices

4.3.1 Representing a Linear Map by a Matrix

Definition 4.17 (Matrix)**.** Suppose $m, n \in \mathbb{N}$. 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.

Notation. We use i for indexing across the m rows, and j for indexing across the n columns.

Let $\mathcal{M}_{m\times n}(\mathbf{F})$ 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 4.18 (Matrix of linear map)**.** Suppose $T \in \mathcal{L}(V, W)$, $\{v_1, \dots, v_n\}$ is a basis of V, $\{w_1, \dots, w_m\}$ 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 $A_{1j}, \dots, A_{mj}$ 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

$$\mathcal{M}(T; \{v_1, \dots, v_n\}, \{w_1, \dots, w_m\}).$$

4.3.2 Addition and Scalar Multiplication of Matrices

Define addition and scalar multiplication on $\mathcal{M}_{m\times n}(\mathbf{F})$ as

$$\begin{aligned}(A+B)_{ij} &= A_{ij}+B_{ij}\\ (\lambda A)_{ij} &= \lambda A_{ij}\end{aligned}$$

Lemma 4.19. *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 $S,T\in\mathcal{L}(V,W)$, $\{v_1,\dots,v_n\}$ is a basis of V, $\{w_1,\dots,w_m\}$ is a basis of W.

(i) Let $\mathcal{M}(S)=A$, $\mathcal{M}(T)=B$. Then

$$Sv_j=\sum_{i=1}^m A_{ij}w_i,\quad Tv_j=\sum_{i=1}^m B_{ij}w_i.$$

Let $\mathcal{M}(S+T)=C$. Then

$$\begin{aligned}(S+T)v_j &= \sum_{i=1}^m C_{ij}w_i\\ Sv_j+Tv_j &= \sum_{i=1}^m C_{ij}w_i\\ \sum_{i=1}^m A_{ij}w_i+\sum_{i=1}^m B_{ij}w_i &= \sum_{i=1}^m C_{ij}w_i\\ \sum_{i=1}^m (A_{ij}+B_{ij})w_i &= \sum_{i=1}^m C_{ij}w_i\\ A_{ij}+B_{ij} &= C_{ij}\end{aligned}$$

which implies $A+B=C$. Hence $\mathcal{M}(S+T)=\mathcal{M}(S)+\mathcal{M}(T)$.

(ii) Let $\mathcal{M}(T)=A$. Then

$$Tv_j=\sum_{i=1}^m A_{ij}w_i.$$

Let $\lambda\in\mathbf{F}$, $\mathcal{M}(\lambda T)=B$. Then

$$\begin{aligned}\lambda Tv_j &= \sum_{i=1}^m B_{ij}w_i\\ \lambda\sum_{i=1}^m A_{ij}w_i &= \sum_{i=1}^m B_{ij}w_i\\ \lambda A_{ij} &= B_{ij}\end{aligned}$$

which implies $\lambda A = B$. Hence $\mathcal{M}(\lambda T) = \lambda \mathcal{M}(T)$.

$\square$

Lemma 4.20. *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 *zero matrix*; 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. $\square$

4.3.3 Matrix Multiplication

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.

Definition 4.21 (Matrix multiplication)**.** Suppose $A \in \mathcal{M}_{m\times n}(\mathbf{F})$, $B \in \mathcal{M}_{n\times p}(\mathbf{F})$. Then $AB \in \mathcal{M}_{m\times p}$ is defined as

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}.$$

This means that the entry in row i, column j of AB is computed by taking row i of A and column j of B, multiplying together corresponding entries, and then summing.

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 $ST \in \mathcal{L}(U,W)$.

Lemma 4.22 (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. Let $\{v_1,\dots,v_n\}$ be a basis of V, $\{w_1,\dots,w_m\}$ be a basis of W, $\{u_1,\dots,u_p\}$ be a basis of U. Let $\mathcal{M}(S) = A$, $\mathcal{M}(T) = B$. For $j = 1,\dots,p$,

$$\begin{aligned}
(ST)u_j &= S(Tu_j) \\
&= S\left(\sum_{k=1}^{n} B_{kj}v_k\right) \\
&= \sum_{k=1}^{n} B_{kj}Sv_k \\
&= \sum_{k=1}^{n} B_{kj}\left(\sum_{i=1}^{m} A_{ik}w_i\right) \\
&= \sum_{i=1}^{m}\left(\sum_{k=1}^{n} A_{ik}B_{kj}\right) w_i.
\end{aligned}$$

□

Notation. Let $A_{i,\cdot}$ denote the row vector corresponding to the i-th row of A, and let $A_{\cdot,j}$ denote the column vector corresponding to the j-th column of A.

Lemma 4.23. *Suppose $A \in \mathcal{M}_{m\times n}(\mathbf{F})$, $B \in \mathcal{M}_{n\times p}(\mathbf{F})$. Then*

$$(AB)_{ij} = A_{i,\cdot}B_{\cdot,j}.$$

That is, the entry in row i, column j of AB equals (row i of A) times (column j of B).

Proof. By definition of matrix multiplication,

$$A_{i,\cdot}B_{\cdot,j} = \begin{pmatrix} A_{i1} & \cdots & A_{in} \end{pmatrix} \begin{pmatrix} B_{1j} \\ \vdots \\ B_{nj} \end{pmatrix} = \sum_{k=1}^{n} a_{ik}b_{kj} = (AB)_{ij}.$$

□

Lemma 4.24. *Suppose $A \in \mathcal{M}_{m\times n}(\mathbf{F})$, $B \in \mathcal{M}_{n\times p}(\mathbf{F})$. Then*

$$(AB)_{\cdot,j} = AB_{\cdot,j} \quad (j = 1, \ldots, p).$$

That is, column j of AB equals A times column j of B.

Proof. Using the previous result,

$$AB_{\cdot,j} = \begin{pmatrix} A_{1,\cdot}B_{\cdot,j} \\ \vdots \\ A_{n,\cdot}B_{\cdot,j} \end{pmatrix} = \begin{pmatrix} (AB)_{1j} \\ \vdots \\ (AB)_{nj} \end{pmatrix} = (AB)_{\cdot,j}$$

□

Lemma 4.25 (Linear combination of columns)**.** *Suppose $A \in \mathcal{M}_{m\times n}(\mathbf{F})$, $b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$. Then*

$$Ab = b_1A_{\cdot,1} + \cdots + b_nA_{\cdot,n}.$$

That is, Ab is a linear combination of the columns of A, with the scalars that multiply the columns coming from b.

Proof. We have

$$\begin{aligned} Ab &= \begin{pmatrix} A_{11}b_1 + \cdots + A_{1n}b_n \\ \vdots \\ A_{m1}b_1 + \cdots + A_{mn}b_n \end{pmatrix} = \begin{pmatrix} A_{11}b_1 \\ \vdots \\ A_{m1}b_1 \end{pmatrix} + \cdots + \begin{pmatrix} A_{1n}b_n \\ \vdots \\ A_{mn}b_n \end{pmatrix} \\ &= b_1 \begin{pmatrix} A_{11} \\ \vdots \\ A_{m1} \end{pmatrix} + \cdots + b_n \begin{pmatrix} A_{1n} \\ \vdots \\ A_{mn} \end{pmatrix} = b_1A_{\cdot,1} + \cdots + b_nA_{\cdot,n}. \end{aligned}$$

□

The next result is the main tool used to prove the column–row factorisation 4.30 and to prove that the column rank of a matrix equals the row rank. To be consistent with the notation often used with the column–row factorisation, we denote the matrices as C and R instead of A and B.

Lemma 4.26. *Suppose $C \in \mathcal{M}_{m\times c}(\mathbf{F})$, $R \in \mathcal{M}_{c\times n}(\mathbf{F})$. Then*

(i) Columns: for $j = 1, \ldots, n$, $(CR)_{\cdot,j}$ is a linear combination of $C_{\cdot,1}, \ldots, C_{\cdot,c}$, with coefficients coming from $R_{\cdot,j}$.

(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) Suppose $j \in \{1, \ldots, n\}$.
$$(CR)_{\cdot,j} = CR_{\cdot,j}$$
and then apply the previous result.

(ii) Similar.

□

4.3.4 Rank of a Matrix

We begin by defining two non-negative integers associated with each matrix.

Definition 4.27. Suppose $A \in \mathcal{M}_{m\times n}(\mathbf{F})$. The *row space* of A is the span of its rows, and the *column space* of A is the span of its columns:

$$\begin{aligned}\operatorname{Row}(A) &:= \operatorname{span}\left(A_{i,\cdot} \mid 1 \leq i \leq m\right), \\ \operatorname{Col}(A) &:= \operatorname{span}\left(A_{\cdot,j} \mid 1 \leq j \leq n\right).\end{aligned}$$

The ***row rank*** and ***column rank*** of A are defined as

$$\begin{aligned}r(A) &:= \dim \operatorname{Row}(A), \\ c(A) &:= \dim \operatorname{Col}(A).\end{aligned}$$

If A is an $m \times n$ matrix, then the column rank of A is at most n (because A has n columns) and the column rank of A is also at most m (because $\dim \mathcal{M}_{m\times 1} = m$). Similar remarks hold for the row rank of A.

We now define the *transpose* of a matrix.

Definition 4.28 (Transpose)**.** Suppose $A \in \mathcal{M}_{m\times n}(\mathbf{F})$. The ***transpose*** of A is the $n \times m$ matrix A^T whose entries are defined by

$$(A^T)_{ij} = A_{ji}.$$

Lemma 4.29 (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) $(AC)^T = C^T A^T$.

Proof.

(i) ${(A+B)^T}_{ij} = (A+B)_{ji} = A_{ji} + B_{ji} = (A^T)_{ij} + (B^T)_{ij}$

(ii) ${(\lambda A)^T}_{ij} = (\lambda A)_{ji} = \lambda A_{ji} = \lambda (A^T)_{ij}$

(iii) ${(AC)^T}_{ij} = (AC)_{ji} = \sum_{k=1}^n A_{jk}C_{ji} = \sum_{k=1}^n C_{ji}A_{jk} = \sum_{k=1}^n (C^T)_{ik}(A^T)_{kj} = (C^T A^T)_{ij}$

□

The next result will be the main tool used to prove that the column rank equals the row rank.

Proposition 4.30 (Column-row factorisation)**.** *Suppose $A \in \mathcal{M}_{m\times n}(\mathbf{F})$, $c(A) \geq 1$. 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

$$\{A_{\cdot,1}, \ldots, A_{\cdot,n}\}$$

is a spanning set of $\mathrm{Col}(A)$, so it can be reduced to a basis of $\mathrm{Col}(A)$, by 3.34. This basis 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 $j \in \{1, \ldots, n\}$, the j-th column of A is a linear combination of the columns of C. Make the coefficients of this linear combination into column j of a $c(A) \times n$ matrix, which we call R. By 4.26(i), it follows that $A = CR$. □

Theorem 4.31. *The column rank of a matrix equals its row rank.*

Proof. Suppose $A \in \mathcal{M}_{m\times n}(\mathbf{F})$. Let $A = CR$ be the column-row factorisation of A given by 4.30, where $C \in \mathcal{M}_{m\times c(A)}(\mathbf{F})$, $R \in \mathcal{M}_{c(A)\times n}(\mathbf{F})$.

Then 4.26(ii) tells us that every row of A is a linear combination of the rows of R. Because R has $c(A)$ rows, this implies that the row rank of A is less than or equal to the column rank $c(A)$ of A.

To prove the inequality in the other direction, apply the result in the previous paragraph to A^T, getting

$$\begin{aligned} c(A) &= r(A^T) \\ &\leq c(A^T) \\ &= r(A). \end{aligned}$$

Thus the column rank of A equals the row rank of A. □

Since the column rank equals row rank, we can dispense with the terms “column rank” and “row rank”, and just use the simpler term “rank”.

Definition 4.32 (Rank)**.** The ***rank*** of a matrix A is defined as

$$\operatorname{rank} A := r(A) = c(A).$$

4.4 Invertibility and Isomorphism

4.4.1 Invertibility

We begin this section by defining the notions of invertible and inverse in the context of linear maps.

Definition 4.33 (Invertibility)**.** We say $T \in \mathcal{L}(V, W)$ is ***invertible*** if there exists $S \in \mathcal{L}(W, V)$ such that $ST = I_V$, $TS = I_W$; we call S an *inverse* of T.

Lemma 4.34. *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

$$S_1 = S_1 I_W = S_1(TS_2) = (S_1 T)S_2 = I_V S_2 = S_2.$$

□

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

Notation. If T is invertible, we denote its inverse by T^{-1}.

The following result is useful in determing if a linear map is invertible.

Lemma 4.35 (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 with 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 $\ker T = \{\mathbf{0}\}$, so $\dim \ker T = 0$. By the fundamental theorem of linear maps,

$$\dim \operatorname{im} T = \dim V - \dim \ker T = \dim V = \dim W$$

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,

$$\dim \ker T = \dim V - \dim \operatorname{im} T = \dim V - \dim W = 0$$

which implies that T is injective.

□

Corollary 4.36. *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 4.35, 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 the above; reverse the roles of S and T (and V and W) to show that if $TS = I$ then $ST = I$. □

4.4.2 Isomorphism

The next definition captures the idea of two vector spaces that are essentially the same, except for the names of their elements.

> **Definition 4.37** (Isomorphism)**.** An ***isomorphism*** is an invertible linear map. We say V is ***isomorphic*** to W, and denote $V \cong W$, 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 4.38.** *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. By 4.35, 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 $\{v_1, \dots, v_n\}$ be a basis of V, $\{w_1, \dots, w_n\}$ 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

$$Tv_i = w_i \quad (i = 1, \dots, n)$$

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

$$\begin{aligned}T(a_1 v_1 + \cdots + a_n v_n) = w &\implies w \in \operatorname{im} T \\ &\implies W = \operatorname{im} T \\ &\implies T \text{ is surjective} \\ &\implies T \text{ is invertible.}\end{aligned}$$

$\square$

> **Proposition 4.39.** *Suppose $\{v_1, \dots, v_n\}$ is a basis of V, $\{w_1, \dots, w_m\}$ 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

$$Tv_j = 0 \quad (j = 1, \dots, n)$$

Since $v_1, \dots, v_n$ 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

$$Tv_j = \sum_{i=1}^{m} A_{ij} w_i \quad (j = 1, \dots, n)$$

Since $\mathcal{M}(T) = A$, $\operatorname{im} \mathcal{M} = \mathcal{M}_{m \times n}(\mathbf{F})$ so $\mathcal{M}$ is surjective.

$\square$

Now we can determine the dimension of the vector space of linear maps from one finite-dimensional vector space to another.

Corollary 4.40. *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 $\mathcal{L}(V, W) \cong \mathcal{M}_{m \times n}(\mathbf{F})$,

$$\dim \mathcal{L}(V, W) = \dim \mathcal{M}_{m \times n}(\mathbf{F}) = mn = (\dim V)(\dim W).$$

$\square$

4.4.3 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 4.41 (Matrix of a vector)**.** Suppose $v \in V$, $\{v_1, \dots, v_n\}$ 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 $b_1, \dots, b_n \in \mathbf{F}$ are such that

$$v = b_1 v_1 + \cdots + b_n v_n.$$

Example 4.42. If $x = (x_1, \dots, x_n) \in \mathbf{F}^n$, then the matrix of the vector x with respect to the standard basis of $\mathbf{F}^n$ is

$$\mathcal{M}(x) = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

Lemma 4.43. *Suppose $T \in \mathcal{L}(V, W)$. Let $\{v_1, \dots, v_n\}$ be a basis of V, $\{w_1, \dots, w_m\}$ be a basis of W. Then*

$$\mathcal{M}(T)_{\cdot,j} = \mathcal{M}(Tv_j) \quad (j = 1, \dots, n)$$

Proof. By definition, the entries of $\mathcal{M}(T)$ are defined such that

$$Tv_j = \sum_{i=1}^{m} A_{ij} w_i \quad (j = 1, \dots, n)$$

Then since $Tv_j \in W$, by definition, the matrix of Tv_j with respect to the basis $\{w_1, \dots, w_m\}$ is

$$\mathcal{M}(Tv_j) = \begin{pmatrix} A_1 j \\ \vdots \\ A_{mj} \end{pmatrix}$$

which is precisely the j-th column of $\mathcal{M}(T)_{\cdot,j}$. □

The following result shows that linear maps act like matrix multiplication.

Lemma 4.44. *Suppose $T \in \mathcal{L}(V, W)$. Let $\{v_1, \dots, v_n\}$ be a basis of V, $\{w_1, \dots, w_m\}$ 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 4.45. *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. □

4.4.4 Change of Basis

For $n \in \mathbb{N}$, 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.

The next result justifies the name "identity matrix".

Lemma 4.46. *Suppose $A \in \mathcal{M}_{n\times n}(\mathbf{F})$. Then $AI_n = I_nA = A$.*

Proof. Exercise. □

Definition 4.47 (Invertible matrix). We say $A \in \mathcal{M}_{n\times n}(\mathbf{F})$ is ***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.

Lemma 4.48 (Uniqueness of inverse). *The inverse of an invertible square matrix is unique.*

Proof. Let A be an invertible square matrix, let B and C be inverses of A. 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}.

Lemma 4.49.

(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) We have

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

so the inverse of A^{-1} is A.

(ii) We have

$$\begin{aligned}(AC)(C^{-1}A^{-1}) &= A(CC^{-1})A^{-1} \\ &= AIA^{-1} \\ &= AA^{-1} \\ &= I,\end{aligned}$$

and similarly $(C^{-1}A^{-1})(AC) = I$.

□

Lemma 4.50 (Matrix of product of linear maps)**.** *Suppose $T \in \mathcal{L}(U, V)$, $S \in \mathcal{L}(V, W)$. Let $\{u_1, \dots, u_m\}$ be a basis of U, $\{v_1, \dots, v_n\}$ be a basis of V, $\{w_1, \dots, w_p\}$ be a basis of W. Then*

$$\mathcal{M}\left(ST; \{u_1, \dots, u_m\}, \{w_1, \dots, w_p\}\right) = \\ \mathcal{M}\left(S; \{v_1, \dots, v_n\}, \{w_1, \dots, w_p\}\right) \mathcal{M}\left(T; \{u_1, \dots, u_m\}, \{v_1, \dots, v_n\}\right).$$

Proof. Refer to previous section. Now we are just being more explicit about the bases involved. □

Corollary 4.51. *Let $\{u_1, \dots, u_n\}$ and $\{v_1, \dots, v_n\}$ be bases of V. Then the matrices*

$$\mathcal{M}\left(I; \{u_1, \dots, u_n\}, \{v_1, \dots, v_n\}\right) \quad \text{and} \quad \mathcal{M}\left(I; \{v_1, \dots, v_n\}, \{u_1, \dots, u_n\}\right)$$

are invertible, and each is the inverse of the other.

Proof. In the previous result, replace w_i with u_i, and replace S and T with I, to obtain

$$I = \mathcal{M}(I; \{v_1, \dots, v_n\}, \{u_1, \dots, u_n\})\mathcal{M}(I; \{u_1, \dots, u_n\}, \{v_1, \dots, v_n\}).$$

Now interchange the roles of u's and v's, which gives

$$I = \mathcal{M}(I; \{u_1, \dots, u_n\}, \{v_1, \dots, v_n\})\mathcal{M}(I; \{v_1, \dots, v_n\}, \{u_1, \dots, u_n\}).$$

These two equations above give the desired result. □

Theorem 4.52 (Change-of-basis formula)**.** *Suppose $T \in \mathcal{L}(V)$. Let $\{u_1, \dots, u_n\}$ and $\{v_1, \dots, v_n\}$ be bases of V. Let*

$$A = \mathcal{M}(T; \{u_1, \dots, u_n\}), \quad B = \mathcal{M}(T; \{v_1, \dots, v_n\}),$$

and $C = \mathcal{M}(I; \{u_1, \dots, u_n\}, \{v_1, \dots, v_n\})$. Then

$$A = C^{-1}BC. \tag{4.2}$$

Proof. Note that

$$\begin{aligned}\mathcal{M}(T;\{u_1,\ldots,u_n\},\{v_1,\ldots,v_n\}) &= \underbrace{\mathcal{M}(T;\{v_1,\ldots,v_n\})}_{B}\underbrace{\mathcal{M}(I;\{u_1,\ldots,u_n\},\{v_1,\ldots,v_n\})}_{C} \\ &= \underbrace{\mathcal{M}(I;\{u_1,\ldots,u_n\},\{v_1,\ldots,v_n\})}_{C}\underbrace{\mathcal{M}(T;\{u_1,\ldots,u_n\})}_{A}\end{aligned}$$

Hence $BC = CA$, and the desired result follows. □

THe next result states that the matrix of inverse equals the inverse of matrix.

Lemma 4.53. *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

$$\mathcal{M}\left(T^{-1}\right)\mathcal{M}(T) = \mathcal{M}\left(T^{-1}T\right) = \mathcal{M}(I) = I.$$

□

4.5 Products and Quotients of Vector Spaces

4.5.1 Products of Vector Spaces

As usual when dealing with more than one vector space, all vector spaces in use should be over the same field.

Definition 4.54 (Product). Suppose $V_1, \dots, V_n$ are vector spaces over $\mathbf{F}$. The ***product*** $V_1 \times \cdots \times V_n$ is defined by

$$V_1 \times \cdots \times V_n := \{(v_1, \dots, v_n) \mid v_i \in V_i\}.$$

Remark. This is analagous to the Cartesian product of sets.

Lemma 4.55. *$V_1 \times \cdots \times V_n$ is a vector space over $\mathbf{F}$, with addition and scalar multiplication defined by*

$$\begin{aligned}(u_1, \dots, u_n) + (v_1, \dots, v_n) &= (u_1 + v_1, \dots, u_n + v_n) \\ \lambda(v_1, \dots, v_n) &= (\lambda v_1, \dots, \lambda v_n)\end{aligned}$$

The next result shows that the dimension of a product is the sum of dimensions.

Lemma 4.56 (Dimension of product). *Suppose $V_1, \dots, V_n$ 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. Choose a basis of each V_i. For each basis vector of each V_i, consider the element of $V_1 \times \cdots \times V_n$ that equals the basis vector in the i-th slot and 0 in the other slots. The set of all such vectors is linearly independent and spans $V_1 \times \cdots \times V_n$. Thus it is a basis of $V_1 \times \cdots \times V_n$. The length of this basis is $\dim V_1 + \cdots + \dim V_n$, as desired.

□

Products are also related to direct sums, by the following result.

Proposition 4.57. *Suppose that $V_1, \dots, V_n \leq V$. Define a linear map*

$$\begin{aligned}\Gamma : V_1 \times \cdots \times V_n &\to V_1 + \cdots + V_n \\ (v_1, \dots, v_n) &\mapsto v_1 + \cdots + v_n\end{aligned}$$

Then $V_1 + \cdots + V_n$ is a direct sum if and only if Γ is injective.

Proof.

$$\begin{aligned}
\Gamma \text{ is injective} &\iff \ker\Gamma = \{\mathbf{0}\} \\
&\iff (v_1, \ldots, v_n) = \mathbf{0} \\
&\iff v_1 = \cdots = v_n = 0 \\
&\iff V_1 \oplus \cdots \oplus V_n && \text{[by 3.18]}
\end{aligned}$$

□

The next result says that a sum is a direct sum if and only if dimensions add up.

Proposition 4.58. *Suppose V is finite-dimensional, $V_1, \ldots, V_n \leq V$. 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

$$\dim(V_1 + \cdots + V_n) = \dim(V_1 \times \cdots \times V_n).$$

Then use the previous two results above. □

4.5.2 Quotient Spaces

We begin our approach to quotient spaces by defining a *coset*.

Definition 4.59 (Coset)**.** Suppose $v \in V$, $U \subset V$. We call $v + U$ a ***coset*** of U, defined by

$$v + U := \{v + u \mid u \in U\}.$$

Definition 4.60 (Quotient space)**.** Suppose $U \leq V$. Then the ***quotient space*** V/U is the set of cosets of U:

$$V/U := \{v + U \mid v \in V\}.$$

Example 4.61. If $U = \{(x, 2x) \in \mathbb{R}^2 \mid x \in \mathbb{R}\}$, then $\mathbb{R}^2/U$ is the set of lines in $\mathbb{R}^2$ that have gradient of 2.

The next result shows that two cosets of a subspace are equal or disjoint.

Lemma 4.62. *Suppose $U \leq V$, 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

$$v + u = w + ((v - w) + u) \in w + U.$$

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. $\square$

We can define a vector space structure on V/U.

Lemma 4.63. *Suppose $U \leq V$. 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. We first need to show that addition and scalar multiplication are well-defined.

Addition Suppose $v_1, v_2, w_1, w_2 \in V$ are such that

$$v_1 + U = v_2 + U, \quad w_1 + U = w_2 + U.$$

By 4.62,

$$v_1 - v_2 \in U, \quad w_1 - w_2 \in U.$$

Since $U \leq V$, U is closed under addition, so $(v_1 - v_2) + (w_1 - w_2) \in U$. Thus $(v_1 + w_1) - (v_2 + w_2) \in U$. Using 4.62 again, we see that

$$(v_1 + w_1) + U = (v_2 + w_2) + U,$$

as desired. Hence addition on V/U is well-defined.

Scalar multiplication Suppose $v_1, v_2 \in V$ are such that $v_1 + U = v_2 + U$, suppose $\lambda \in \mathbf{F}$.

Since $U \leq V$, U is closed under scalar multiplication, so $\lambda(v_1 - v_2) \in U$. Thus $\lambda v_1 - \lambda v_2 \in U$. By 4.62,

$$(\lambda v_1) + U = (\lambda v_2) + U.$$

Hence scalar multiplication on V/U is well-defined.

The verification that addition and scalar multiplication make V/U into a vector space is straightforward and is left to the reader. Note that the additive identity of V/U is $0 + U$ (which equals U) and that the additive inverse of $v + U$ is $(-v) + U$. □

Definition 4.64 (Quotient map)**.** Suppose $U \leq V$. The ***quotient map*** is the map

$$\begin{aligned} \pi : V &\to V/U \\ v &\mapsto v + U \end{aligned}$$

for all $v \in V$.

Notation. Although π depends on U as well as V, these spaces are left out of the notation because they should be clear from the context.

We check that the quotient map is a linear map: let $v, w \in V$, $\lambda \in \mathbf{F}$,

(i) $\pi(v) + \pi(w) = (v + U) + (w + U) = (v + w) + U = \pi(v + w)$.

(ii) $\pi(\lambda v) = (\lambda v) + U = \lambda(v + U) = \lambda(\pi v)$.

Lemma 4.65 (Dimension of quotient space)**.** *Suppose V is finite-dimensional, $U \leq V$. Then*

$$\dim V/U = \dim V - \dim U.$$

Idea. Since dimensions are involved, think of the fundamental theorem of linear maps.

Proof. Let the quotient map $\pi : V \to V/U$.

- Let $v \in V$. Then

$$\begin{aligned} v \in \ker \pi &\iff \pi(v) = \mathbf{0} + U = U \\ &\iff v + U = U \quad \text{[by 4.62]} \\ &\iff v \in U \end{aligned}$$

  so $\ker \pi = U$.

- The definition of π implies $\operatorname{im} \pi = V/U$.

By the fundamental theorem of linear maps,

$$\begin{aligned} \dim V &= \dim \ker \pi + \dim \operatorname{im} \pi \\ &= \dim U + \dim V/U \end{aligned}$$

which gives the desired result. □

Each linear map T on V induces a linear map $\tilde{T}$ on $V/\ker T$, as defined below.

Definition 4.66. Suppose $T \in \mathcal{L}(V, W)$. Define

$$\begin{aligned} \tilde{T} : V/\ker T &\to W \\ v + \ker T &\mapsto Tv \end{aligned}$$

$$\begin{tikzcd} V \arrow[r, "\tilde{T}"] \arrow[dr, "T"'] & V/\ker T \arrow[d, "\pi"] \\ & W \end{tikzcd}$$

We first show that $\tilde{T}$ is well-defined.

Proof. Suppose $u, v \in V$ are such that

$$u + \ker T = v + \ker T.$$

By 4.62, $u - v \in \ker T$. Thus $T(u - v) = \mathbf{0}$, so $Tu = Tv$. □

We then check that $\tilde{T}$ is a linear map from $V/\ker T$ to W.

The next result shows that we can think of $\tilde{T}$ as a modified version of T, with a domain that produces an injective map.

Lemma 4.67. *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) $\operatorname{im} \tilde{T} = \operatorname{im} T$.

Proof.

(i) Let $v \in V$. Then
$$(\tilde{T} \circ \pi)(v) = \tilde{T}\,(\pi(v)) = \tilde{T}(v + \ker T) = Tv.$$

(ii) Let $v + \ker T \in \ker \tilde{T}$. Then

$$\begin{aligned} \tilde{T}(v + \ker T) = \mathbf{0} &\implies Tv = \mathbf{0} \\ &\implies v \in \ker T \\ &\implies v + \ker T = \ker T \end{aligned}$$

so $\ker \tilde{T} = \{\mathbf{0} + \ker T\}$.

(iii) The definition of $\tilde{T}$ shows that $\operatorname{im} \tilde{T} = \operatorname{im} T$.

□

Theorem 4.68 (First isomorphism theorem)**.** *Suppose $T \in \mathcal{L}(V, W)$ is an isomorphism. Then*

$$V/\ker T \cong \operatorname{im} T. \tag{4.3}$$

Proof. (ii) and (iii) imply that if we think of $\tilde{T}$ as mapping into $\operatorname{im} T$, then $\tilde{T}$ is an isomorphism from $V/\ker T$ onto $\operatorname{im} T$. □

Theorem 4.69 (Second isomorphism theorem)**.** *Suppose $U, W \leq V$. Then*

$$(U + W)/W \cong U/(U \cap W). \tag{4.4}$$

Theorem 4.70 (Third isomorphism theorem)**.** *$U \subset V \subset W$, then*

$$W/V \cong (W/U)/(V/U). \tag{4.5}$$

4.6 Duality

4.6.1 Dual Space and Dual Map

Linear maps into the scalar field $\mathbf{F}$ play a special role in linear algebra, so they get a special name.

Definition 4.71 (Linear functional). A ***linear functional*** on V is a linear map from V to $\mathbf{F}$.

That is, a linear functional is an element of $\mathcal{L}(V, \mathbf{F})$.

Example 4.72. $\phi : \mathbb{R}^3 \to \mathbb{R}$ defined by $\phi(x, y, z) = x + y + z$ is a linear functional on $\mathbb{R}^3$.

Definition 4.73 (Dual space). The ***dual space*** V' of V is the vector space of linear functionals on V.

That is, $V' := \mathcal{L}(V, \mathbf{F})$.

Lemma 4.74 (Dimension of dual space). *Suppose V is finite-dimensional. Then V' is finite-dimensional, and*

$$\dim V' = \dim V.$$

Proof. By 4.40,

$$\dim V' := \dim \mathcal{L}(V, \mathbf{F}) = (\dim V)(\dim \mathbf{F}) = \dim V.$$

□

Definition 4.75 (Dual basis). Let $\{v_1, \ldots, v_n\}$ be a basis of V. Then the ***dual basis*** of $\{v_1, \ldots, v_n\}$ is

$$\{\phi_1, \ldots, \phi_n\} \subset V',$$

where each ϕ_i is the linear functional on V such that

$$\phi_i(v_j) = \delta_{ij} = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j) \end{cases}$$

Example 4.76 (Dual basis of the standard basis of $\mathbf{F}^n$). Fix a positive integer n. For $i = 1, \ldots, n$, define ϕ_i to be the linear functional on $\mathbf{F}^n$ that selects the i-th coordinate of a vector in $\mathbf{F}^n$:

$$\phi_i(x_1, \ldots, x_n) = x_i$$

for each $(x_1, \ldots, x_n) \in \mathbf{F}^n$.

Let $\{e_1, \ldots, e_n\}$ be the standard basis of $\mathbf{F}^n$. Then

$$\phi_i(e_j) = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j) \end{cases}$$

Thus $\phi_1, \ldots, \phi_n$ is the dual basis of the standard basis $e_1, \ldots, e_n$ of $\mathbf{F}^n$.

The next result shows that the dual basis of a basis of V consists of the linear functionals on V that give the coefficients for expressing a vector in V as a linear combination of the basis vectors.

Proposition 4.77. *Suppose $\{v_1, \dots, v_n\}$ is a basis of V, and $\{\phi_1, \dots, \phi_n\}$ is the dual basis. Then for each $v \in V$,*

$$v = \phi_1(v)v_1 + \cdots + \phi_n(v)v_n.$$

Proof. Let $v \in V$. Since $\{v_1, \dots, v_n\}$ is a basis of V, there exist $c_1, \dots, c_n \in \mathbf{F}$ such that

$$v = c_1 v_1 + \cdots + c_n v_n.$$

For $i = 1, \dots, n$, applying ϕ_i to both sides of the equation above gives

$$\phi_i(v) = c_i.$$

□

The next result shows that the dual basis is a basis of the dual space. Thus the terminology "dual basis" is justified.

Lemma 4.78. *Suppose V is finite-dimensional. Then the dual basis of a basis of V is a basis of V'.*

Proof. Suppose $\{v_1, \dots, v_n\}$ is a basis of V. Let $\{\phi_1, \dots, \phi_n\}$ denote the dual basis.

Since $\{\phi_1, \dots, \phi_n\}$ has length $\dim V$, in order to show that it is a basis of V', it suffices to show that it is linearly independent in V'.

Suppose $a_1, \dots, a_n \in \mathbf{F}$ are such that

$$a_1\phi_1 + \cdots + a_n\phi_n = 0. \tag{I}$$

Now for each $i = 1, \dots, n$,

$$(a_1\phi_1 + \cdots + a_n\phi_n)(v_i) = a_i.$$

Thus (I) shows that $a_1 = \cdots = a_n = 0$. Hence $\{\phi_1, \dots, \phi_n\}$ is linearly independent. □

Definition 4.79 (Dual map)**.** Suppose $T \in \mathcal{L}(V, W)$. The ***dual map*** of T is the linear map

$$\begin{aligned} T' : W' &\to V' \\ \phi &\mapsto \phi \circ T \end{aligned}$$

If $T \in \mathcal{L}(V, W)$ and $\phi \in W'$, then $T'(\phi)$ is defined above to be the composition of the linear maps ϕ and T. Thus $T'(\phi)$ is indeed a linear map from V to $\mathbf{F}$, i.e., $T'(\phi) \in V'$.

We check that $T' \in \mathcal{L}(W', V')$: let $\phi, \psi \in W'$, $\lambda \in \mathbf{F}$,

(i) $T'(\phi + \psi) = (\phi + \psi) \circ T = \phi \circ T + \psi \circ T = T'(\phi) + T'(\psi)$

(ii) $T'(\lambda\phi) = (\lambda\phi) \circ T = \lambda(\phi \circ T) = \lambda(T'(\phi))$

Lemma 4.80 (Algebraic properties of dual map)**.** *Suppose* $T \in \mathcal{L}(V, W)$. *Then*

(i) $(S+T)' = S' + T'$ *for all* $S \in \mathcal{L}(V, W)$

(ii) $(\lambda T)' = \lambda T'$ *for all* $\lambda \in \mathbf{F}$

(iii) $(ST)' = T'S'$ *for all* $S \in \mathcal{L}(W, U)$

Proof.

(i)

(ii)

(iii) Let $\phi \in U'$. Then

$$(ST)'(\phi) = \phi \circ (ST) = (\phi \circ S) \circ T = T'(\phi \circ S) = T'(S'(\phi)) = (T'S')(\phi).$$

□

(i) and (ii) imply that the function that takes T to T' is a linear map from $\mathcal{L}(V, W)$ to $\mathcal{L}(W', V')$.

4.6.2 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 $\ker T$. To do this, we will need the next definition.

Definition 4.81 (Annihilator)**.** For $U \subset V$, the ***annihilator*** of U is defined by

$$U^0 := \{\phi \in V' \mid \phi(u) = \mathbf{0},\ \forall u \in U\}.$$

Example 4.82. $\{\mathbf{0}\}^0 = V'$ and $V^0 = \{\mathbf{0}\}$.

We check that $U^0 \leq V$:

(i) Note that $0 \in U^0$ (here 0 is the zero linear functional on V) because the zero linear functional applied to every vector in U equals $\mathbf{0} \in \mathbf{F}$.

(ii) Suppose $\phi, \psi \in U^0$. Thus $\phi, \psi \in V'$ and $\phi(u) = \psi(u) = \mathbf{0}$ for every $u \in U$.

Let $u \in U$, then

$$(\phi + \psi)(u) = \phi(u) + \psi(u) = \mathbf{0} + \mathbf{0} = \mathbf{0}.$$

Thus $\phi + \psi \in U^0$, so U^0 is closed under addition.

(iii) Suppose $\phi \in U^0$, $\lambda \in \mathbf{F}$, let $u \in U$, then

$$\phi(\lambda u) = \lambda\phi(u) = \mathbf{0}$$

so $\lambda\phi \in U^0$, so U^0 is closed under scalar multiplication.

Lemma 4.83 (Dimension of annihilator)**.** *Suppose V is finite-dimensional, and $U \leq V$. Then*

$$\dim U^0 = \dim V - \dim U.$$

Proof. Let $i \in \mathcal{L}(U, V)$ be the inclusion map defined by $i(u) = u$ for each $u \in U$. Thus the dual map i' is a linear map from V' to U'. The fundamental theorem of linear maps applied to i' shows that

$$\dim \ker i' + \dim \operatorname{im} i' = \dim V'. \tag{I}$$

However, $\ker i' = U^0$ (as can be seen by thinking about the definitions) and $\dim V' = \dim V$ (by 4.74), so we can rewrite (I) as

$$\dim U^0 + \dim \operatorname{im} i' = \dim V. \tag{II}$$

If $\phi \in U'$, then ϕ can be extended to a linear functional ψ on V (see, for example, Exercise 13 in Section 3A). The definition of i' shows that $i'(\psi) = \phi$. Thus $\phi \in \operatorname{im} i'$, which implies that $\operatorname{im} i' = U'$. Hence

$$\dim \ker i' = \dim U' = \dim U,$$

and then (II) becomes the equation $\dim U + \dim U^0 = \dim V$, as desired. □

The next result provides conditions for the annihilator to equal $\{\mathbf{0}\}$ or the whole space.

Lemma 4.84. *Suppose V is finite-dimensional, and $U \leq V$. Then*

(i) $U^0 = \{\mathbf{0}\} \iff U = V$

(ii) $U^0 = V' \iff U = \{\mathbf{0}\}$

Proof.

(i)

$$\begin{aligned} U^0 = \{\mathbf{0}\} &\iff \dim U^0 = 0 \\ &\iff \dim U = \dim V && \text{[by 4.83]} \\ &\iff U = V \end{aligned}$$

(ii)

$$\begin{aligned} U^0 = V' &\iff \dim U^0 = \dim V' \\ &\iff \dim U^0 = \dim V && \text{[by 4.74]} \\ &\iff \dim U = 0 && \text{[by 4.83]} \\ &\iff U = \{\mathbf{0}\} \end{aligned}$$

□

The next result concerns $\ker T'$.

Lemma 4.85. *Suppose V and W are finite-dimensional, $T \in \mathcal{L}(V, W)$. Then*

(i) $\ker T' = (\operatorname{im} T)^0$

(ii) $\dim \ker T' = \dim \ker T + \dim W - \dim V$

Proof.

(i) $\boxed{\subset}$ Let $\phi \in \ker T'$. Then $0 = T'(\phi) = \phi \circ T$. Hence

$$0 = (\phi \circ T)(v) = \phi(Tv) \quad (\forall v \in V).$$

Thus $\phi \in (\operatorname{im} T)^0$. This implies that $\ker T' \subset (\operatorname{im} T)^0$.

$\boxed{\supset}$ Let $\phi \in (\operatorname{im} T)^0$. Then $\phi(Tv) = 0$ for every $v \in V$. Hence $0 = \phi \circ T = T'(\phi)$, i.e., $\phi \in \ker T'$. Thus $\phi \in \ker T'$, which shows that $(\operatorname{im} T)^0 \subset \ker T'$.

(ii) We have

$$\begin{aligned}
\dim\ker T' &= \dim(\operatorname{im} T)^0 && \text{[by (i)]} \\
&= \dim W - \dim\operatorname{im} T && \text{[by 4.83]} \\
&= \dim W - (\dim W - \dim\ker T) && \text{[by fundamental theorem of linear maps]} \\
&= \dim\ker T + \dim W - \dim V.
\end{aligned}$$

□

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.

Lemma 4.86. *Suppose V and W are finite-dimensional, $T \in \mathcal{L}(V, W)$. Then*

$$T \text{ is surjective} \iff T' \text{ is injective.}$$

Proof. Let $T \in \mathcal{L}(V, W)$. We have

$$\begin{aligned}
T \text{ is surjective} &\iff \operatorname{im} T = W \\
&\iff (\operatorname{im} T)^0 = \{\mathbf{0}\} && \text{[by 4.84]} \\
&\iff \operatorname{im} T' = \{\mathbf{0}\} && \text{[by 4.85]} \\
&\iff T' \text{ is injective}
\end{aligned}$$

□

The following result concerns $\operatorname{im} T'$.

Lemma 4.87. *Suppose V and W finite-dimensional, $T \in \mathcal{L}(V, W)$. Then*

(i) $\dim\operatorname{im} T' = \dim\operatorname{im} T$

(ii) $\operatorname{im} T' = (\ker T)^0$

Proof.

(i) We have

$$\begin{aligned}
\dim\operatorname{im} T' &= \dim W' - \dim\ker T' && \text{[by fundamental theorem of linear maps]} \\
&= \dim W - \dim(\ker T)^0 && \text{[by 4.74 and 4.85]} \\
&= \dim\operatorname{im} T && \text{[by 4.83]}
\end{aligned}$$

(ii) We first show that $\operatorname{im} T' \subset (\ker T)^0$.

Let $\phi \in \ker T'$. Then there exists $\psi \in W'$ such that $\phi = T'(\psi)$.

If $v \in \ker T$, then

$$\phi(v) = \big(T'(\psi)\big)\, v = (\psi \circ T)(v) = \psi(Tv) = \psi(\mathbf{0}) = \mathbf{0}.$$

Hence $\phi \in (\ker T)^0$. This implies that $\operatorname{im} T' \in (\ker T)^0$.

We will complete the proof by showing that $\operatorname{im} T'$ and $(\ker T)^0$ have the same dimension. To do this, note that

$$\begin{aligned}
\dim \operatorname{im} T' &= \dim \operatorname{im} T && \text{[by 4.74]} \\
&= \dim V - \ker T && \text{[by fundamental theorem of linear maps]} \\
&= \dim (\ker T)^0 && \text{[by 4.83]}
\end{aligned}$$

□

Lemma 4.88. *Suppose V and W are finite-dimensional, $T \in \mathcal{L}(V, W)$. Then*

$$T \text{ is injective} \iff T' \text{ is surjective.}$$

Proof. Let $T \in \mathcal{L}(V, W)$. We have

$$\begin{aligned}
T \text{ is injective} &\iff \ker T = \{\mathbf{0}\} \\
&\iff (\ker T)^0 = V' && \text{[by 4.84]} \\
&\iff \operatorname{im} T' = V' && \text{[by 4.87]}
\end{aligned}$$

□

4.6.3 Matrix of Dual of Linear Map

The setting for the next result is the assumption that we have a basis $\{v_1, \ldots, v_n\}$ of V, along with its dual basis $\{\phi_1, \ldots, \phi_n\}$ of V'. We also have a basis $\{w_1, \ldots, w_m\}$ of W, along with its dual basis $\{\psi_1, \ldots, \psi_m\}$ of W'.

Thus $\mathcal{M}(T)$ is computed with respect to the aforementioned bases of V and W, and $\mathcal{M}(T')$ is computed with respect to the aforementioned dual bases of W' and V'. Using these bases gives the following result.

Lemma 4.89. *Suppose V and W are finite-dimensional, $T \in \mathcal{L}(V, W)$. Then*

$$\mathcal{M}(T') = \mathcal{M}(T)^T.$$

Proof. Let $\mathcal{M}(T) = A$, $\mathcal{M}(T') = C$. From the definition of $\mathcal{M}(T')$ we have

$$T'(\psi_i) = \sum_{k=1}^{n} C_{ki}\phi_k.$$

The left side of the equation above equals $\psi_i \circ T$. Thus applying both sides of the equation above to v_j gives

$$\begin{aligned}(\psi_i \circ T)(v_j) &= \sum_{k=1}^{n} C_{ki}\phi_k(v_j)\\ &= C_{ji}.\end{aligned}$$

We also have

$$\begin{aligned}(\psi_i \circ T)(v_j) &= \psi_i(Tv_j)\\ &= \psi_i\left(\sum_{k=1}^{m} A_{kj}w_k\right)\\ &= \sum_{k=1}^{m} A_{kj}\psi_i(w_k)\\ &= A_{ij}.\end{aligned}$$

Comparing the last line of the last two sets of equations, we have $C_{ji} = A_{ij}$. Thus $C = A^T$, so $\mathcal{M}(T') = \mathcal{M}(T)^T$ as desired. $\square$

Exercises

Exercise 4.1 ([Axl24] 3A)**.** Suppose $b, c \in \mathbb{R}$. Define $T\colon \mathbb{R}^3 \to \mathbb{R}^2$ by

$$T(x, y, z) = (2x - 4y + 3z + b, 6x + cxyz).$$

Show that T is linear if and only if $b = c = 0$.

Exercise 4.2 ([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 $S \in \mathcal{L}(V)$.

Exercise 4.3 ([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

$$\begin{aligned} & a_1 Tv_1 + \cdots + a_n Tv_n = \mathbf{0} \\ \Longrightarrow\ & T(a_1 v_1 + \cdots + a_n v_n) = 0 \\ \Longrightarrow\ & a_1 v_1 + \cdots + a_n v_n \in \ker T \end{aligned}$$

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. □

Exercise 4.4 ([Axl24] 3B Q11)**.** Suppose that V is finite-dimensional, $T \in \mathcal{L}(V, W)$. Prove that there exists $U \leq V$ such that

$$U \cap \ker T = \{\mathbf{0}\} \quad \text{and} \quad \operatorname{im} T = T(U).$$

Solution. □

Exercise 4.5 ([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. □

Exercise 4.6 ([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.

Exercise 4.7 ([Axl24] 3B 22)**.** Suppose U, V are finite-dimensional, $S \in \mathcal{L}(V, W)$, $T \in \mathcal{L}(U, V)$. Prove that

$$\dim \ker ST \leq \dim \ker S + \dim \ker T.$$

Solution. □

Exercise 4.8 ([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 $\left(T^{-1}\right)^{-1} = T$. □

3C Q15,16,17

3D Q11,12,17,22,23,24

Exercise 4.9 ([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.

$$(ST)(T^{-1}S^{-1}) = S(TT^{-1})S^{-1} = I = T^{-1}S^{-1}ST.$$

□

Exercise 4.10 ([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) $\{Tv_1, \dots, Tv_n\}$ is a basis of V for every basis $\{v_1, \dots, v_n\}$ of V;

(iii) $\{Tv_1, \dots, Tv_n\}$ is a basis of V for some basis $\{v_1, \dots, v_n\}$ 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. □

Exercise 4.11 ([Axl24] 3E Q3)**.** Suppose $V_1, \dots, V_m$ are vector spaces. Prove that

$$\mathcal{L}(V_1 \times \cdots \times V_m, W) \cong \mathcal{L}(V_1, W) \times \cdots \times \mathcal{L}(V_m, W).$$

Exercise 4.12 ([Axl24] 3E Q4)**.** Suppose $V_1, \dots, V_m$ are vector spaces. Prove that

$$\mathcal{L}(V, W_1 \times \cdots \times W_m) \cong \mathcal{L}(V, W_1) \times \cdots \times \mathcal{L}(V, W_m).$$

Exercise 4.13 ([Axl24] 3E Q5)**.** For a positive integer m, define V^m by

$$V^m = \underbrace{V \times \cdots \times V}_{m \text{ times}}.$$

Prove that $V^m \cong \mathcal{L}(\mathbf{F}^m, V)$.

Exercise 4.14 ([Axl24] 3E Q6)**.** Suppose that $v, x \in V$ and $U, W \leq V$ are such that $v + U = x + W$. Prove that $U = W$.

Exercise 4.15 ([Axl24] 3E Q12, Barycentric coordinates)**.** Suppose $v_1, \ldots, v_m \in V$. Let

$$A = \{\lambda_1 v_1 + \cdots + \lambda_m v_m \mid \lambda_i \in \mathbf{F}, \lambda_1 + \cdots + \lambda_m = 1\}.$$

(i) Prove that A is a coset of some subspace of V.

(ii) Prove that if B is a coset of some subspace of V, and $\{v_1, \ldots, v_m\} \subset B$, then $A \subset B$.

(iii) Prove that A is a coset of some subspace of V, where $\dim V < m$.

Exercise 4.16 ([Axl24] 3E Q13)**.** Suppose $U \leq V$, and V/U is finite-dimensional. Prove that $V \cong U \times (V/U)$.

Solution.

$$\dim V = \dim U + (\dim V - \dim U) = \dim U + \dim(V/U).$$

□

Exercise 4.17 ([Axl24] 3E Q14)**.** Suppose $U, W \leq V$ such that $V = U \oplus W$. Suppose $w_1, \ldots, w_m$ is a basis of W. Prove that $w_1 + U, \ldots, w_m + U$ is a basis of V/U.

Exercise 4.18 ([Axl24] 3E Q15)**.**

Exercise 4.19 ([Axl24] 3E Q16)**.** Suppose $\phi \in \mathcal{L}(V, \mathbf{F})$ and $\phi \neq 0$. Prove that $\dim V/\ker\phi = 1$.

Exercise 4.20 ([Axl24] 3E Q18)**.**

Exercise 4.21 ([Axl24] 3E Q19)**.** Suppose $T \in \mathcal{L}(V, W)$ and $U \leq V$. Let π denote the quotient map from V to V/U. Prove that there exists $S \in \mathcal{L}(V/U, W)$ such that

$$T = S \circ \pi \iff U \subset \ker T.$$

Chapter 5

Polynomials

5.1 Definitions

Definition 5.1 (Polynomial)**.** We say $p : \mathbf{F} \to \mathbf{F}$ 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. Let $\mathbf{F}[z]$ denote the set of polynomials with coefficients in $\mathbf{F}$.

Lemma 5.2. *With the usual operations of addition and scalar multiplication,* $\mathbf{F}[z]$ *is a vector space over* $\mathbf{F}$.
Hence $\mathbf{F}[z]$ *is a subspace of* $\mathbf{F}^{\mathbf{F}}$ *(vector space of functions from* $\mathbf{F}$ *to* $\mathbf{F}$*).*

Definition 5.3 (Degree)**.** A polynomial $p \in \mathbf{F}[z]$ is has ***degree*** n, denoted by $\deg p = n$, if there exist scalars $a_0, a_1, \dots, a_n \in \mathbf{F}$ 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.

Lemma 5.4. *For non-negative integer* n, $\mathbf{F}_n[z]$ *is finite-dimensional.*

Proof. $\mathbf{F}_n[z] = \operatorname{span}(1, z, z^2, \dots, z^n)$ [here we slightly abuse notation by letting z^k denote a function]. □

Lemma 5.5. $\mathbf{F}[z]$ *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. □

5.2 Zeros of Polynomials

Definition 5.6 (Zero of polynomial)**.** We call $\lambda \in \mathbf{F}$ a ***zero*** of a polynomial $p \in \mathbf{F}[z]$ if

$$p(\lambda) = 0.$$

Lemma 5.7 (Factor theorem)**.** *Suppose $n \in \mathbb{N}$, $p \in \mathbf{F}_n[z]$. Suppose $\lambda \in \mathbf{F}$, then $p(\lambda) = 0$ if and only if there exists $q \in \mathbf{F}_{n-1}[z]$ such that*

$$p(z) = (z - \lambda)q(z) \quad (z \in \mathbf{F}).$$

Proof.

$\boxed{\implies}$ Suppose $p(\lambda) = 0$. Let $a_0, a_1, \dots, a_n \in \mathbf{F}$ be such that

$$p(z) = a_n z^n + \dots + a_1 z + a_0 \quad (z \in \mathbf{F}).$$

Then for all $z \in \mathbf{F}$,

$$\begin{aligned} p(z) &= p(z) - p(\lambda) \\ &= (a_n z^n + \dots + a_1 z + a_0) - (a_n \lambda^n + \dots + a_1 \lambda + a_0) \\ &= a_n (z^n - \lambda^n) + \dots + a_1 (z - \lambda). \end{aligned}$$

Note that for each $k = 1, \dots, n$, we can factorise

$$z^k - \lambda^k = (z - \lambda)\left(z^{k-1} + z^{k-2}\lambda + \dots + \lambda^{k-1}\right).$$

Thus p equals $z - \lambda$ times some polynomial of degree $n - 1$, as desired.

$\boxed{\impliedby}$ Now suppose that there exists a polynomial $q \in \mathbf{F}[z]$ such that

$$p(z) = (z - \lambda)q(z) \quad (z \in \mathbf{F}).$$

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 5.8. *Suppose $n \in \mathbb{N}$, $p \in \mathbf{F}_n[z]$. 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 5.7, 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 $\mathbf{F}$. $\square$

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.

5.3 Division Algorithm for Polynomials

Proposition 5.9 (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 $n \geq m$. The set

$$S = \{1, z, \ldots, z^{m-1}, s, zs, \ldots, z^{n-m}s\}$$

is linearly independent in $\mathbf{F}[z]$ because each polynomial in S has a different degree. Also, S has length $n + 1$, which equals $\dim \mathbf{F}[z]$. Hence S is a basis of $\mathbf{F}[z]$.

Since $p \in \mathbf{F}[z]$ and S is a basis of $\mathbf{F}[z]$, there exist unique constants $a_0, a_1, \ldots, a_{m-1} \in \mathbf{F}$ and $b_0, b_1, \ldots, b_{n-m} \in \mathbf{F}$ such that

$$\begin{aligned} p &= a_0 + a_1 z + \cdots + a_{m-1} z^{m-1} + b_0 s + b_1 zs + \cdots + b_{n-m} z^{n-m} s \\ &= \underbrace{a_0 + a_1 z + \cdots + a_{m-1} z^{m-1}}_{r} + s \left(\underbrace{b_0 + b_1 z + \cdots + b_{n-m} z^{n-m}}_{q} \right). \end{aligned}$$

With r and q as defined above, we see that p can be written as $p = sq + r$ with $\deg r < \deg s$, as desired.

The uniqueness of $q, r \in \mathbf{F}[z]$ satisfying these conditions follows from the uniqueness of the constants $a_0, a_1, \ldots, a_{m-1} \in \mathbf{F}$ and $b_0, b_1, \ldots, b_{n-m} \in \mathbf{F}$. ∎

5.4 Factorisation of Polynomials over $\mathbb{C}$

Theorem 5.10 (Fundamental theorem of algebra, first version)**.** *Every non-constant polynomial with complex coefficients has a zero in* $\mathbb{C}$.

Remark. The fundamental theorem of algebra is an existence theorem. Its proof does not lead to a method for finding zeros.

The quadratic formula gives the zeros explicitly for polynomials of degree 2. Similar but more complicated formulas exist for polynomials of degree 3 and 4. However no such formulas exist for polynomials of degree 5 and above.

Theorem 5.11 (Fundamental theorem of algebra, second version)**.** *If* $p \in \mathbb{C}[z]$ *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 $c, \lambda_1, \ldots, \lambda_n \in \mathbb{C}$.

5.5 Factorisation of Polynomials over $\mathbb{R}$

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 $\mathbb{R}$, we will use our factorisation theorem over $\mathbb{C}$. We begin with the next result.

> **Lemma 5.12.** *Suppose $p \in \mathbb{C}[z]$ is a polynomial with real coefficients. If $\lambda \in \mathbb{C}$ is a zero of p, then so is its conjugate $\overline{\lambda}$.*

Proof. Let
$$p(z) = a_0 + a_1 z + \cdots + a_n z^n,$$
where $a_0, \ldots, a_n \in \mathbb{R}$. Suppose $\lambda \in \mathbb{C}$ is a zero of p, then

$$a_0 + a_1\lambda + \cdots + a_n\lambda^n = 0.$$

Taking the complex conjugate on both sides of the equation gives

$$a_0 + a_1\overline{\lambda} + \cdots + a_n\overline{\lambda}^n = 0.$$

Hence $\overline{\lambda}$ is a zero of p. □

We want a factorisation theorem for polynomials with real coefficients. We begin with the following result.

Remark. Think about the quadratic formula in connection with the result below.

> **Lemma 5.13** (Factorisation of quadratic polynomial)**.** *Suppose $b, c \in \mathbb{R}$. Then there is a polynomial factorisation of the form*
>
> $$x^2 + bx + c = (x - \lambda_1)(x - \lambda_2)$$
>
> *with $\lambda_1, \lambda_2 \in \mathbb{R}$ if and only if $b^2 \geq 4c$.*

Proof. Completing the square gives

$$x^2 + bx + c = \left(x + \frac{b}{2}\right)^2 + \left(c - \frac{b^2}{4}\right). \tag{I}$$

$\boxed{\Longrightarrow}$ We prove the contrapositive. Suppose $b^2 < 4c$, then the RHS of (I) is positive for every $x \in \mathbb{R}$. Hence the polynomial x^2+bx+c has no real zeros and thus cannot be factored in the form $(x-\lambda_1)(x-\lambda_2)$ with $\lambda_1, \lambda_2 \in \mathbb{R}$.

$\boxed{\Longleftarrow}$ Suppose $b^2 \geq 4c$. Then there is a real number d such that $d^2 = \frac{b^2}{4} - c$. We can rewrite (I) as

$$\begin{aligned} x^2 + bx + c &= \left(x + \frac{b}{2}\right)^2 - d^2 \\ &= \left(x + \frac{b}{2} + d\right)\left(x + \frac{b}{2} - d\right), \end{aligned}$$

which gives the desired factorisation. □

Theorem 5.14 (Factorisation of polynomial over $\mathbb{R}$)**.** *Suppose $p \in \mathbb{R}[x]$ 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 $c, \lambda_1, \ldots, \lambda_n, b_1, \ldots, b_N, c_1, \ldots, c_N \in \mathbb{R}$, with ${b_i}^2 < 4c_i$ for each i.

Chapter 6

Eigenvalues and Eigenvectors

6.1 Invariant Subspaces

6.1.1 Eigenvalues

Definition 6.1 (Operator)**.** An ***operator*** is a linear map from a vector space to itself.

Definition 6.2 (Invariant subspace)**.** Suppose $T \in \mathcal{L}(V)$. $U \leq V$ is ***invariant*** under T if $Tu \in U$ for all $u \in U$.

Example 6.3. 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: if $u \in \{\mathbf{0}\}$, then $u = \mathbf{0}$ so $Tu = \mathbf{0} \in \{\mathbf{0}\}$.

(ii) The subspace V is invariant under T: if $u \in V$, then $Tu \in V$.

(iii) The subspace $\ker T$ is invariant under T: 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: if $u \in \operatorname{im} T$, then $Tu \in \operatorname{im} T$ by definition.

Definition 6.4 (Eigenvalue and eigenvector)**.** Suppose $T \in \mathcal{L}(V)$. $\lambda \in \mathbf{F}$ is an ***eigenvalue*** of T if there exists $v \in V \setminus \{\mathbf{0}\}$ such that $Tv = \lambda v$; we say v is an ***eigenvector*** of T corresponding to λ.

Lemma 6.5 (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 4.35. ☐

Proposition 6.6 (Linearly independent eigenvectors)**.** *Suppose $T \in \mathcal{L}(V)$. Then every set of eigenvectors of T corresponding to distinct eigenvalues of T is linearly independent.*

Proof. 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 6.7. *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 6.6, 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 $m \leq \dim V$, as desired. ☐

6.1.2 Polynomials Applied to Operators

Notation. Suppose $T \in \mathcal{L}(V)$, $n \in \mathbb{Z}^+$. $T^n \in \mathcal{L}(V)$ 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 6.8. 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

$$p(T) := a_n T^n + \cdots + a_1 T + a_0.$$

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 6.9 (Product of polynomials)**.** Suppose $p, q \in \mathbf{F}[z]$. Then $pq \in \mathbf{F}[z]$ is the polynomial defined by

$$(pq)(z) = p(z)q(z) \quad (z \in \mathbf{F})$$

Lemma 6.10. *Suppose $p, q \in \mathbf{F}[z]$, $T \in \mathcal{L}(V)$. Then*

(i) $(pq)(T) = p(T)q(T)$; *(multiplicativity)*

(ii) $p(T)q(T) = q(T)p(T)$. *(commutativity)*

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

$$p(T)q(T) = (pq)(T) = (qp)(T) = q(T)p(T)$$

since the multiplication of polynomials is commutative.

□

Lemma 6.11. *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) Let $u \in \ker p(T)$. Then $p(T)u = \mathbf{0}$. Thus

$$(p(T))\,(Tu) = (p(T)T)\,(u) = (Tp(T))\,(u) = T\,(p(T)u) = T(\mathbf{0}) = \mathbf{0}.$$

Hence $Tu \in \ker p(T)$, so $\ker p(T)$ is invariant under T.

(ii) Let $u \in \operatorname{im} p(T)$. Then there exists $v \in V$ such that $u = p(T)v$. Thus

$$Tu = T\,(p(T)v) = p(T)(Tv).$$

Hence $Tu \in \operatorname{im} p(T)$, so $\operatorname{im} p(T)$ is invariant under T.

□

6.2 The Minimal Polynomial

6.2.1 Existence of Eigenvalues on Complex Vector Spaces

The following is one of the most important results in linear algebra.

Theorem 6.12 (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

$$S = \{v, Tv, T^2v, \dots, T^nv\}.$$

Since $\dim V = n$ and S has length $n+1$, S is not linearly independent. Thus there exist $a_0, \dots, a_n \in \mathbb{C}$, not all 0, such that

$$a_0v + a_1Tv + a_2T^2v + \cdots + a_nT^nv = \mathbf{0},$$

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 (5.10), there exists a root of p in $\mathbb{C}$; let $\lambda \in \mathbb{C}$ be a root of p. By the factor theorem,

$$p(z) = (z-\lambda)q(z) \quad (z \in \mathbb{C}).$$

Thus

$$\begin{aligned} p(T) &= (T-\lambda I)q(T) \\ \mathbf{0} = p(T)v &= (T-\lambda I)q(T)v \\ Tq(T)v &= \lambda q(T)v \end{aligned}$$

Since p is the minimal polynomial and $\deg q < \deg p$, we must have that $q(T)v \neq \mathbf{0}$. Therefore λ is an eigenvalue of T, with corresponding eigenvector $q(T)v$. □

Example 6.13. Note that the hypothesis in 6.12 that $\mathbf{F} = \mathbb{C}$ cannot be replaced with the hypothesis that $\mathbf{F} = \mathbb{R}$. For instance, consider $T \in \mathcal{L}(\mathbb{R}^2)$ defined by

$$Tv = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} v. \tag{$*$}$$

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 $T \in \mathcal{L}(\mathbb{C}^2)$ 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 $\begin{pmatrix} i \\ 1 \end{pmatrix}$.

6.2.2 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*.

> **Theorem 6.14.** *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 $p(T) = 0$. Furthermore, $\deg p \leq \dim V$.*

Proof.

$\boxed{\text{Existence}}$ 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 0 operator.

Let $u \in V \setminus \{\mathbf{0}\}$, consider the set
$$\{u, Tu, T^2u, \dots, T^nu\}.$$

This set has length $n+1$, so it is linearly dependent. By the linear dependence lemma, there exists a smallest positive integer $m \leq n$ such that T^mu is a linear combination of $u, Tu, \dots, T^{m-1}u$; 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,

$$q(T)\left(T^ku\right) = T^k\left(q(T)u\right) = T^k(\mathbf{0}) = \mathbf{0}.$$

By the linear dependence lemma, $\{u, Tu, \dots, T^{m-1}u\}$ is linearly independent. Thus the above equation implies that $\dim\ker q(T) \geq m$. Hence by the fundamental theorem of linear maps,

$$\begin{aligned}\dim\operatorname{im} q(T) &= \dim V - \dim\ker q(T) \\ &\leq \dim V - m.\end{aligned}$$

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 $\deg s \leq \dim V - m$ such that

$$s\left(T|_{\operatorname{im} q(T)}\right) = 0.$$

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 $s(T)|_{\operatorname{im} q(T)} = s(T|_{\operatorname{im} q(T)}) = 0$. Thus sq is a monic polynomial such that $\deg sq \leq \dim V$ and $(sq)(T) = 0$, as desired.

$\boxed{\text{Uniqueness}}$ Let $p \in \mathbf{F}[z]$ be a monic polynomial of smallest degree such that $p(T) = 0$; let $r \in \mathbf{F}[z]$ be a monic polynomial of same degree and $r(T) = 0$. Then $(p - r)(T) = 0$ and also $\deg(p - r) < \deg p$.

We claim that $p-r=0$. Suppose, for a contradiction, that $p-r\neq 0$. 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 $s(T)=0$ and also $\deg s=\deg(p-r)<\deg p$, a contradiction. □

Definition 6.15 (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 $p(T)=0$.

Theorem 6.16. *Suppose V is finite-dimensional, $T\in\mathcal{L}(V)$. Then the zeros of the minimal polynomial of T are eigenvalues of T.*

Proof. Let p be the minimal polynomial of T.

$\boxed{\Longrightarrow}$ First suppose $\lambda\in\mathbf{F}$ is a zero of p. Then p can be written in the form

$$p(z)=(z-\lambda)q(z)$$

where q is a monic polynomial with coefficients in $\mathbf{F}$. Since $p(T)=0$, we have

$$\mathbf{0}=(T-\lambda I)(q(T)v)\quad(v\in V).$$

Since $\deg p<\deg p$ and p is the minimal polynomial of T, there exists at least one $v\in V$ such that $q(T)v\neq\mathbf{0}$. The equation above thus implies that λ is an eigenvalue of T, as desired.

$\boxed{\Longleftarrow}$ To prove that every eigenvalue of T is a zero of p, now suppose $\lambda\in\mathbf{F}$ is an eigenvalue of T. Thus there exists $v\in V\setminus\{\mathbf{0}\}$ such that $Tv=\lambda v$. Repeated applications of T to both sides of this equation show that $T^kv=\lambda^kv$ for every nonnegative integer k. Thus

$$p(T)v=p(\lambda)v.$$

Since p is the minimal polynomial of T, we have $p(T)v=\mathbf{0}$. Hence the equation above implies that $p(\lambda)=0$. Thus λ is a zero of p, as desired. □

If V is a complex vector space, by the fundamental theorem of algebra (5.11), the minimal polynomial of T has the factorisation

$$(z-\lambda_1)\cdots(z-\lambda_m),\tag{6.1}$$

where $\lambda_1,\dots,\lambda_m$ are eigenvalues of T.

The next result completely characterises the polynomials that when applied to an operator give the 0 operator.

Proposition 6.17. *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.*

Proof. Let p denote the minimal polynomial of T.

$\boxed{\implies}$ Suppose $q(T) = 0$. By the division algorithm, there exist polynomials $s, r \in \mathbf{F}[z]$ such that

$$q + ps + r \tag{$*$}$$

and $\deg r < \deg p$. We have

$$0 = q(T) = p(T)s(T) + r(T) = r(T).$$

The equation above implies that $r = 0$ (otherwise, dividing r by its highest-degree coefficient would produce a monic polynomial that when applied to T gives 0; this polynomial would have a smaller degree than the minimal polynomial, which would be a contradiction).

Thus $(*)$ becomes $q = ps$, so q is a polynomial multiple of p.

$\boxed{\impliedby}$ Suppose q is a polynomial multiple of p. Thus $q = ps$ for some polynomial $s \in \mathbf{F}[z]$, so

$$q(T) = p(T)s(T) = 0s(T) = 0$$

as desired. □

The following corollary concerns the minimal polynomial of a restriction operator.

Corollary 6.18. *Suppose V is finite-dimensional, $T \in \mathcal{L}(V)$, $U \leq V$ is invariant under T. Then the minimal polynomial of T is a polynomial multiple of the minimal polynomial of $T|_U$.*

Proof. Let p be the minimal polynomial of T. Then $p(T)v = \mathbf{0}$ for all $v \in V$. In particular,

$$p(T)u = \mathbf{0} \quad (u \in U).$$

Thus $p\left(T|_U\right) = 0$. By 6.17 (applied to $T|_U$ in place of T), p is a polynomial multiple of the minimal polynomial of $T|_U$. □

The next result shows that the constant term of the minimal polynomial of an operator determines whether the operator is invertible.

Corollary 6.19. *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.*

Proof. Suppose $T \in \mathcal{L}(V)$, let p be the minimal polynomial of T. Then

$$\begin{aligned}
T \text{ is not invertible} &\iff 0 \text{ is an eigenvalue of } T && \text{[by 6.5]} \\
&\iff 0 \text{ is a zero of } p && \text{[by 6.16]} \\
&\iff \text{constant term of } p \text{ is } 0.
\end{aligned}$$

□

6.2.3 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.

Lemma 6.20. *Suppose V is a finite-dimensional, real vector space. Suppose $T \in \mathcal{L}(V)$ and $b, c \in \mathbb{R}$ with $b^2 < 4c$. Then $\dim \ker(T^2 + bT + cI)$ is even.*

Proof. By 6.11, $\ker(T^2 + bT + cI)$ is invariant under T. By replacing V with $\ker(T^2 + bT + cI)$ and replacing T with T restricted to $\ker(T^2 + bT + cI)$, we can assume that $T^2 + bT + cI = 0$; we now need to prove that $\dim V$ is even.

Suppose $\lambda \in \mathbb{R}$ and $v \in V$ are such that $Tv = \lambda v$. Then

$$\mathbf{0} = (T^2 + bT + cI)v = (\lambda^2 + b\lambda + c)v = \underbrace{\left(\left(\lambda + \frac{b}{2}\right)^2 + c - \frac{b^2}{4}\right)}_{>0} v$$

implies $v = \mathbf{0}$. Hence we have shown that T has no eigenvectors.

Let $U \leq V$ be invariant under T, and has the largest dimension among all subspaces of V that are invariant under T and have even dimension. If $U = V$, then we are done; otherwise assume there exists $w \in V$ such that $w \notin U$.

Let $W = \operatorname{span}(w, Tw)$. Then W is invariant under T, since $T(Tw) = -bTw - cw$. Furthermore, $\dim W = 2$, since otherwise w would be an eigenvector of T. Now

$$\dim(U + W) = \dim U + \dim W - \dim(U \cap W) = \dim U + 2,$$

where $U \cap W = \{\mathbf{0}\}$ because otherwise $U \cap W$ would be a one-dimensional subspace of V that is invariant under T (impossible because T has no eigenvectors).

Because $U + W$ is invariant under T, the equation above shows that there exists a subspace of V invariant under T of even dimension larger than $\dim U$. Thus the assumption that $U \neq V$ was incorrect. Hence V has even dimension. □

The next result states that on odd-dimensional vector spaces, every operator has an eigenvalue. We already know this result for finite-dimensional complex vectors spaces (without the odd hypothesis). Thus in the proof below, we will assume that $\mathbf{F} = \mathbb{R}$.

Theorem 6.21. *Every operator on an odd-dimensional vector space has an eigenvalue.*

Proof. Suppose V is a finite-dimensional real vector space, $\dim V = n$ is odd. Let $T \in \mathcal{L}(V)$. We will induct on n in steps of size two to show that T has an eigenvalue.

The base case where $\dim V = 1$ holds, because then every non-zero vector in V is an eigenvector of T.

Now suppose that $n \geq 3$ and the desired result holds for all operators on all odd-dimensional vector spaces of dimension less than n.

Let p be the minimal polynomial of T. If p is a polynomial multiple of $x - \lambda$ for some $\lambda \in \mathbb{R}$, by 6.16, λ is an eigenvalue of T and we are done. Thus we can assume that there exist $b, c \in \mathbb{R}$ such that $b^2 < 4c$ and p is a polynomial multiple of $x^2 + bx + c$ (see 5.14). There exists a monic polynomial $q \in \mathbb{R}[x]$ such that $p(x) = q(x)(x^2 + bx + c)$ for all $x \in \mathbb{R}$. Then

$$0 = p(T) = (q(T))(T^2 + bT + cI),$$

which means that $q(T)$ equals 0 on $\operatorname{im}(T^2 + bT + cI)$. Since $\deg q < \deg p$ and p is the minimal polynomial of T, this implies that $\operatorname{im}(T^2 + bT + cI) \neq V$.

By the fundamental theorem of linear maps,

$$\underbrace{\dim V}_{\text{odd}} = \underbrace{\dim \ker(T^2 + bT + cI)}_{\text{even}} + \dim \operatorname{im}(T^2 + bT + cI)$$

implies that $\dim \operatorname{im}(T^2 + bT + cI)$ is odd.

Hence $\operatorname{im}(T^2 + bT + cI)$ is a subspace of V that is invariant under T (by 6.11) and that has odd dimension less than $\dim V$. Our induction hypothesis now implies that T restricted to $\operatorname{im}(T^2 + bT + cI)$ has an eigenvalue, which means that T has an eigenvalue. $\square$

6.3 Upper-Triangular Matrices

Suppose $T \in \mathcal{L}(V)$. Recall that the matrix of T with respect to a basis $\{v_1, \dots, v_n\}$ of V is the $n \times n$ matrix whose entries A_{ij} are defined by

$$Tv_j = \sum_{i=1}^{n} A_{ij} v_i \quad (j = 1, \dots, n).$$

Notation. If the basis is not clear from context, we denote the matrix of T as $\mathcal{M}(T; \{v_1, \dots, v_n\})$.

Remark. The matrices of operators are square matrices.

The *diagonal* of a square matrix consists of the entries on the line from the upper left corner to the bottom right corner.

Definition 6.22 (Upper-triangular matrix)**.** A square matrix is called ***upper triangular*** if all the entries below the diagonal are 0.

We represent an upper-triangular matrix in the form

$$\begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

where the 0 indicates that all entries below the diagonal equal 0, and $*$ denotes entries that we do not know or that are irrelevant to the questions being discussed.

The next result provides a useful connection between upper-triangular matrices and invariant subspaces.

Lemma 6.23 (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) $\mathcal{M}(T; \{v_1, \dots, v_n\})$ *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 $k = 1, \dots, n$.*

Proof.

$\boxed{\text{(i)} \implies \text{(ii)}}$ Suppose $k \in \{1, \dots, n\}$. Since the matrix of T with respect to $\{v_1, \dots, v_n\}$ is upper triangular, if $j \in \{1, \dots, n\}$, then

$$Tv_j \in \operatorname{span}(v_1, \dots, v_j).$$

If $j \le k$, then $\operatorname{span}(v_1, \dots, v_j) \subset \operatorname{span}(v_1, \dots, v_k)$, so

$$Tv_j \in \operatorname{span}(v_1, \dots, v_k)$$

for each $j \in \{1, \dots, k\}$. Thus $\operatorname{span}(v_1, \dots, v_k)$ is invariant under T.

$\boxed{\text{(ii)} \implies \text{(iii)}}$ Suppose (ii) holds, so $\operatorname{span}(v_1, \dots,)$ is invariant under T for each $k = 1, \dots, n$. In particular, $Tv_k \in \operatorname{span}(v_1, \dots, v_k)$ for each $k = 1, \dots, n$.

$\boxed{\text{(iii)} \implies \text{(i)}}$ Suppose $Tv_k \in \operatorname{span}(v_1, \dots, v_k)$ for each $k = 1, \dots, n$.

Then when writing each Tv_k as a linear combination of basis vectors $v_1, \dots, v_n$, we need to use only $v_1, \dots, v_k$. Hence all entries under the diagonal of $\mathcal{M}(T)$ are 0, so $\mathcal{M}(T)$ is an upper-triangular matrix. □

The next result tells us that if $\mathcal{M}(T)$ is upper-triangular with respect to some basis of V, then T satisfies a simple equation depending on the diagonal entries.

Proposition 6.24. *Suppose $T \in \mathcal{L}(V)$ has an upper-triangular matrix with respect to some basis of V, with diagonal entries $\lambda_1, \dots, \lambda_n$. Then*

$$(T - \lambda_1 I) \cdots (T - \lambda_n I) = 0. \tag{6.2}$$

Proof. Let $\{v_1, \dots, v_n\}$ be a basis of V, with respect to which T has an upper-triangular matrix with diagonal entries $\lambda_1, \dots, \lambda_n$:

$$\mathcal{M}(T) = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

- Considering the first column of $\mathcal{M}(T)$, we have

$$\begin{aligned} & Tv_1 = \lambda_1 v_1 \\ \implies & (T - \lambda_1 I)v_1 = \mathbf{0} \\ \implies & (T - \lambda_1 I) \cdots (T - \lambda_m I)v_1 = \mathbf{0} \quad (m = 1, \dots, n). \end{aligned}$$

- Considering the second column of $\mathcal{M}(T)$, we have

$$\begin{aligned} & (T - \lambda_2 I)v_2 \in \operatorname{span}(v_1) \\ \implies & (T - \lambda_1 I)(T - \lambda_2 I)v_2 = \mathbf{0} \\ \implies & (T - \lambda_1 I) \cdots (T - \lambda_m I)v_2 = \mathbf{0} \quad (m = 2, \dots, n). \end{aligned}$$

- Considering the third column of $\mathcal{M}(T)$, we have

$$\begin{aligned} & (T - \lambda_3 I)v_3 \in \operatorname{span}(v_1, v_2) \\ \implies & (T - \lambda_1 I)(T - \lambda_2 I)(T - \lambda_3 I)v_3 = \mathbf{0} \\ \implies & (T - \lambda_1 I) \cdots (T - \lambda_m I)v_3 = \mathbf{0} \quad (m = 3, \dots, n). \end{aligned}$$

Continuing this pattern, we see that

$$(T - \lambda_1 I) \cdots (T - \lambda_n I)v_k = \mathbf{0} \quad (k = 1, \dots, n).$$

Hence $(T-\lambda_1 I)\cdots(T-\lambda_n I)$ is the 0 operator, because it is $\mathbf{0}$ on each vector in a basis of V. □

The next result tells us that the eigenvalues of an operator can be determined from the upper-triangular matrix.

Proposition 6.25. *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.*

Proof. Let $\{v_1, \ldots, v_n\}$ be a basis of V with respect to which T has an upper-triangular matrix:

$$\mathcal{M}(T) = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

Since $Tv_1 = \lambda_1 v_1$, we see that λ_1 is an eigenvalue of T.

Let $k \in \{2, \ldots, n\}$, then $(T - \lambda_k I)v_k \in \operatorname{span}(v_1, \ldots, v_{k-1})$, so $T - \lambda_k I$ maps $\operatorname{span}(v_1, \ldots, v_k)$ into $\operatorname{span}(v_1, \ldots, v_{k-1})$. Since

$$\dim \operatorname{span}(v_1, \ldots, v_k) = k, \quad \dim \operatorname{span}(v_1, \ldots, v_{k-1}) = k-1,$$

this implies that $T - \lambda_k I$ restricted to $\operatorname{span}(v_1, \ldots, v_k)$ is not injective (by 4.14). Thus there exists $v \in \operatorname{span}(v_1, \ldots, v_k)$ such that $v \neq \mathbf{0}$ and $(T - \lambda_k I)v = \mathbf{0}$. Thus λ_k is an eigenvalue of T. Hence every entry on the diagonal of $\mathcal{M}(T)$ is an eigenvalue of T.

We now prove T has no other eigenvalues. Let q be the polynomial defined by

$$q(z) = (z - \lambda_1)\cdots(z - \lambda_n).$$

By 6.24, $q(T) = 0$. By 6.17, q is a polynomial multiple of the minimal polynomial of T. Thus every zero of the minimal polynomial of T is a zero of q.

By 6.16, the zeros of the minimal polynomial of T are the eigenvalues of T. This implies that every eigenvalue of T is a zero of q.

Hence the eigenvalues of T are all contained in the set $\{\lambda_1, \ldots, \lambda_n\}$. □

Example 6.26. Define $T \in \mathcal{L}(\mathbf{F}^3)$ by $T(x, y, z) = (2x + y, 5y + 3z, 8z)$. The matrix of T with respect to the standard basis is

$$\mathcal{M}(T) = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 5 & 3 \\ 0 & 0 & 8 \end{pmatrix}.$$

Thus the eigenvalues of T are 2, 5, and 8.

The following result gives a *necessary and sufficient condition* to have an upper-triangular matrix.

Proposition 6.27. *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 $\lambda_1,\ldots,\lambda_m \in \mathbf{F}$.

Proof.

$\boxed{\Longrightarrow}$ Suppose T has an upper-triangular matrix with respect to some basis of V. Let $\alpha_1,\ldots,\alpha_n$ denote the diagonal entries of that matrix. Define a polynomial $q \in \mathbf{F}[z]$ by

$$q(z) = (z-\alpha_1)\cdots(z-\alpha_n).$$

By 6.24, $q(T) = 0$. By 6.17, q is a polynomial multiple of the minimal polynomial. Thus the minimal polynomial of T equals $(z-\lambda_1)\cdots(z-\lambda_m)$ for some $\lambda_1,\ldots,\lambda_m \in \mathbf{F}$ with $\{\lambda_1,\ldots,\lambda_m\} \subset \{\alpha_1,\ldots,\alpha_n\}$.

$\boxed{\Longleftarrow}$ Suppose the minimal polynomial of T equals $(z-\lambda_1)\cdots(z-\lambda_m)$ for some $\lambda_1,\ldots,\lambda_m \in \mathbf{F}$. We induct on m.

For the base case $m = 1$, $z - \lambda_1$ is the minimal polynomial of T, which implies that $T = \lambda_1 I$, so the matrix of T (with respect to any basis of V) is upper-triangular.

Now suppose $m > 1$ and the desired result holds for all smaller positive integers. Let

$$U = \operatorname{im}(T - \lambda_m I).$$

Then U is invariant under T (by 6.11), so $T|_U$ is an operator on U.

If $u \in U$, then $u = (T-\lambda_m I)v$ for some $v \in V$ and

$$(T-\lambda_1 I)\cdots(T-\lambda_{m-1}I)u = (T-\lambda_1 I)\cdots(T-\lambda_m I)v = \mathbf{0}.$$

Hence $(z-\lambda_1)\cdots(z-\lambda_{m-1})$ is a polynomial multiple of the minimal polynomial of $T|_U$, by 6.17. Thus the minimal polynomial of $T|_U$ is the product of at most $m-1$ terms of the form $z - \lambda_k$.

By our induction hypothesis, there is a basis $\{u_1,\ldots,u_M\}$ of U, with respect to which $T|_U$ has an upper-triangular matrix. Thus for each $k \in \{1,\ldots,M\}$, we have (using 6.23)

$$Tu_k = (T|_U)(u_k) \in \operatorname{span}(u_1,\ldots,u_k). \tag{I}$$

Extend $\{u_1,\ldots,u_M\}$ to a basis $\{u_1,\ldots,u_M,v_1,\ldots,v_N\}$ of V. If $k \in \{1,\ldots,N\}$, then

$$Tv_k = (T-\lambda_m I)v_k + \lambda_m v_k.$$

The definition of U shows that $(T-\lambda_m I)v_k \in U = \operatorname{span}(u_1,\ldots,u_M)$. Thus the equation above shows that

$$Tv_k \in \operatorname{span}(u_1,\ldots,u_M,v_1,\ldots,v_k). \tag{II}$$

From (I) and (II), we conclude (using 6.23) that T has an upper-triangular matrix with respect to the basis $\{u_1,\ldots,u_M,v_1,\ldots,v_N\}$ of V, as desired. $\square$

We conclude with an important result: every operator on a finite-dimensional complex vector space has an upper-triangular matrix.

Theorem 6.28. *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.*

Proof. The desired result follows immediately from 6.27 and the second version of the fundamental theorem of algebra (5.11). $\square$

6.4 Diagonalisable Operators

6.4.1 Diagonal Matrices

Definition 6.29 (Diagonal matrix)**.** A ***diagonal matrix*** is a square matrix that is 0 everywhere except possibly on the diagonal.

That is, a diagonal matrix has the form

$$\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

for some $\lambda_1, \dots, \lambda_n \in \mathbf{F}$.

Remark. If an operator has a diagonal matrix with respect to some basis, then the entries on the diagonal are precisely the eigenvalues of the operator, by 6.25.

Definition 6.30 (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 6.31 (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

$$E(\lambda, T) := \ker(T - \lambda I) = \{v \in V \mid Tv = \lambda v\}.$$

Hence $E(\lambda, T)$ is the set of all eigenvectors of T corresponding to λ, along with the $\mathbf{0}$ vector. The definitions imply that λ is an eigenvalue of T if and only if $E(\lambda, T) \neq \{\mathbf{0}\}$.

By 6.11, $E(\lambda, T)$ is a subspace of V.

The next result states that the sum of eigenspaces is a direct sum.

Proposition 6.32. *Suppose $T \in \mathcal{L}(V)$, $\lambda_1, \dots, \lambda_m$ are distinct eigenvalues of T. Then*

$$E(\lambda_1, T) \oplus \cdots \oplus E(\lambda_m, T).$$

Furthermore, if V is finite-dimensional, then

$$\dim E(\lambda_1, T) + \cdots + \dim E(\lambda_m, T) \leq \dim V.$$

Proof. To show that $E(\lambda_1, T) + \cdots + E(\lambda_m, T)$ is a direct sum, suppose

$$v_1 + \cdots + v_m = \mathbf{0},$$

where each $v_i \in E(\lambda_i, T)$. By 6.6, eigenvectors corresponding to distinct eigenvalues are linearly independent, so each $v_i = \mathbf{0}$. Thus by 3.18, $E(\lambda_1, T) + \cdots + E(\lambda_m, T)$ is a direct sum.

If V is finite-dimensional, then

$$\begin{aligned} \dim E(\lambda_1, T) + \cdots + \dim E(\lambda_m, T) &= \dim \left(E(\lambda_1, T) \oplus \cdots \oplus E(\lambda_m, T)\right) && \text{[by 4.58]} \\ &\leq \dim V && \text{[by 3.40]} \end{aligned}$$

□

6.4.2 Conditions for Diagonalisability

The following characterisations of diagonalisable operators will be useful.

Lemma 6.33 (Conditions equivalent to diagonalisability). *Suppose V is finite-dimensional, $T \in \mathcal{L}(V)$, $\lambda_1, \dots, \lambda_m$ 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) $V = E(\lambda_1, T) \oplus \cdots \oplus E(\lambda_m, T)$.

(iv) $\dim V = \dim E(\lambda_1, T) + \cdots + \dim E(\lambda_m, T)$.

Proof.

$\boxed{\text{(i)} \iff \text{(ii)}}$ By definition, $T \in \mathcal{L}(V)$ is diagonalisable if and only if T has a diagonal matrix

$$\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

with respect to a basis $\{v_1, \dots, v_n\}$ of V, if and only if $Tv_i = \lambda_i v_i$ for each $i = 1, \dots, n$.

$\boxed{\text{(ii)} \implies \text{(iii)}}$ Suppose V has a basis consisting of eigenvectors of T. Hence every vector in V can be written as a linear combination of T, which implies that

$$V = E(\lambda_1, T) + \cdots + E(\lambda_m, T).$$

By 6.32, this is a direct sum.

$\boxed{\text{(iii)} \implies \text{(iv)}}$ This follows from 4.58.

$\boxed{\text{(iv)} \implies \text{(ii)}}$ Suppose

$$\dim V = \dim E(\lambda_1, T) + \cdots + \dim E(\lambda_m, T).$$

Choose a basis of each $E(\lambda_i, T)$; put all these bases together to form a set $\{v_1, \dots, v_n\}$ of eigenvectors of T, where $n = \dim V$.

Claim. $\{v_1, \dots, v_n\}$ is a basis of T.

By 3.41, it suffices to show that $\{v_1, \dots, v_n\}$ is linearly independent. Suppose $a_1, \dots, a_n \in \mathbf{F}$ are such that

$$a_1 v_1 + \cdots + a_n v_n = \mathbf{0}.$$

For each $i = 1, \dots, m$, let u_i denote the sum of all the terms $a_j v_j$ such that $v_j \in E(\lambda_i, T)$. Thus each u_i is in $E(\lambda_i, T)$, and

$$u_1 + \cdots + u_m = \mathbf{0}.$$

By 6.6, since eigenvectors corresponding to distinct eigenvalues are linearly independent, this implies that each u_i equals $\mathbf{0}$. Because each u_i is a sum of terms $a_j v_j$, where the v_j's were chosen to be a basis of $E(\lambda_i, T)$, this implies that all a_j's equal 0. Thus $\{v_1, \ldots, v_n\}$ is linearly independent and hence is a basis of V. $\square$

The next result shows that if an operator has as many distinct eigenvalues as the dimension of its domain, then the operator is diagonalisable.

Corollary 6.34. *Suppose V is finite-dimensional, $T \in \mathcal{L}(V)$ has $\dim V$ distinct eigenvalues. Then T is diagonalisable.*

Proof. Suppose T has distinct eigenvalues $\lambda_1, \ldots, \lambda_{\dim V}$, with corresponding eigenvectors $v_1, \ldots, v_{\dim V}$. By 6.6, eigenvectors corresponding to distinct eigenvalues are linearly independent, so $v_1, \ldots, v_{\dim V}$ are linearly independent, and thus forms a basis of V.

With respect to this basis consisting of eigenvectors, T has a diagonal matrix. $\square$

In later chapters, we will find additional conditions that imply that certain operators are diagonalisable. For example, see the real spectral theorem (8.18) and the complex spectral theorem (8.19).

The next result provides a necessary and sufficient condition for diagonalisability.

Theorem 6.35. *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 $\lambda_1, \ldots, \lambda_m \in \mathbf{F}$.*

Proof.

$\boxed{\implies}$ Suppose T is diagonalisable. Thus there is a basis $\{v_1, \ldots, v_n\}$ of V consisting of eigenvectors of T. Let $\lambda_1, \ldots, \lambda_m$ be distinct eigenvalues of T. Then for each v_j, there exists λ_k with $(T - \lambda_k I)v_j = \mathbf{0}$. Thus

$$(T - \lambda_1 I) \cdots (T - \lambda_m I)v_j = \mathbf{0},$$

which implies that the minimal polynomial of T equals $(z - \lambda_1) \cdots (z - \lambda_m)$.

$\boxed{\impliedby}$ Suppose the minimal polynomial of T equals $(z - \lambda_1) \cdots (z - \lambda_m)$ for distinct $\lambda_1, \ldots, \lambda_m \in \mathbf{F}$. Thus

$$(T - \lambda_1 I) \cdots (T - \lambda_m I) = 0. \tag{I}$$

We will prove that T is diagonalisable by inducting on m. For the base case $m = 1$, $T - \lambda_1 I = 0$, which means that T is a scalar multiple of the identity operator, so T is diagonalisable.

Now suppose that $m > 1$ and the desired result holds for all smaller values of m. The subspace $\operatorname{im}(T - \lambda_m I)$ is invariant under T (by 6.11). Thus $T|_{\operatorname{im}(T - \lambda_m I)} \in \mathcal{L}(\operatorname{im}(T - \lambda_m I))$.

If $u \in \operatorname{im}(T - \lambda_m I)$, then $u = (T - \lambda_m I)v$ for some $v \in V$, and (I) implies

$$(T - \lambda_1 I) \cdots (T - \lambda_{m-1} I)u = (T - \lambda_1) \cdots (T - \lambda_m I)v = \mathbf{0}. \tag{II}$$

Hence $(z - \lambda_1) \cdots (z - \lambda_{m-1})$ is a polynomial multiple of the minimal polynomial of $T|_{\operatorname{im}(T - \lambda_m I)}$, by 6.17. Thus by induction hypothesis, there is a basis of $\operatorname{im}(T - \lambda_m I)$ consisting of eigenvectors of T.

Let $u \in \operatorname{im}(T - \lambda_m I) \cap \ker(T - \lambda_m I)$. Then $Tu = \lambda_m u$. Now (II) implies that

$$\begin{aligned} \mathbf{0} &= (T - \lambda_1 I) \cdots (T - \lambda_{m-1} I)u \\ &= (\lambda_m - \lambda_1) \cdots (\lambda_m - \lambda_{m-1})u. \end{aligned}$$

Since $\lambda_1, \dots, \lambda_m$ are distinct, the equation above implies that $u = \mathbf{0}$. Hence $\operatorname{im}(T - \lambda_m I) \cap \ker(T - \lambda_m I) = \{\mathbf{0}\}$.

By 3.19, this implies that $\operatorname{im}(T - \lambda_m I) + \ker(T - \lambda_m I)$ is a direct sum, whose dimension is $\dim V$ (by 4.58 and the fundamental theorem of linear maps). Hence $\operatorname{im}(T - \lambda_m I) \oplus \ker(T - \lambda_m I) = V$. Every nonzero vector in $\ker(T - \lambda_m I)$ is an eigenvector of T with eigenvalue λ_m. Earlier in this proof we saw that there is a basis of $\operatorname{im}(T - \lambda_m I)$ consisting of eigenvectors of T. Adjoining to that basis a basis of $\ker(T - \lambda_m I)$ gives a basis of V consisting of eigenvectors of T. The matrix of T with respect to this basis is a diagonal matrix, as desired. □

The next result states that the restriction of a diagonalisable operator to an invariant subspace is diagonalisable.

Corollary 6.36. *Suppose $T \in \mathcal{L}(V)$ is diagonalisable, $U \leq V$ is invariant under T. Then $T|_U$ is a diagonalisable operator on U.*

Proof. Since T is diagonalisable, by 6.35, the minimal polynomial of T equals $(z - \lambda_1) \cdots (z - \lambda_m)$ for some distinct $\lambda_1, \dots, \lambda_m \in \mathbf{F}$.

By 6.18, the minimal polynomial of T is a polynomial multiple of the minimal polynomial of $T|_U$. Hence the minimal polynomial of $T|_U$ has the form required by 6.35, which shows that $T|_U$ is diagonalisable. □

6.5 Commuting Operators

Definition 6.37 (Commute)**.** Two operators S and T on the same vector space ***commute*** if $ST = TS$.
Two square matrices A and B of the same size *commute* if $AB = BA$.

The next result shows that two operators commute if and only if their matrices (with respect to the same basis) commute.

Lemma 6.38. *Suppose $S, T \in \mathcal{L}(V)$ and $\{v_1, \dots, v_n\}$ is a basis of V. Then S and T commute if and only if $\mathcal{M}(S; \{v_1, \dots, v_n\})$ and $\mathcal{M}(T; \{v_1, \dots, v_n\})$ commute.*

Proof. We have

$$\begin{aligned} S \text{ and } T \text{ commute} &\iff ST = TS \\ &\iff \mathcal{M}(ST) = \mathcal{M}(TS) \\ &\iff \mathcal{M}(S)\mathcal{M}(T) = \mathcal{M}(T)\mathcal{M}(S) \\ &\iff \mathcal{M}(S) \text{ and } \mathcal{M}(T) \text{ commute} \end{aligned}$$

as desired. □

The next result shows that if two operators commute, then every eigenspace for one operator is invariant under the other operator.

Lemma 6.39. *Suppose $S, T \in \mathcal{L}(V)$ commute, $\lambda \in \mathbf{F}$. Then $E(\lambda, S)$ is invariant under T.*

Proof. Let $v \in E(\lambda, S)$. Then

$$S(Tv) = (ST)v = (TS)v = T(Sv) = T(\lambda v) = \lambda Tv$$

so $Tv \in E(\lambda, S)$. Hence $E(\lambda, S)$ is invariant under T. □

Suppose we have two operators, each of which is diagonalisable. If we want to do computations involving both operators, then we want the two operators to be diagonalisable by the same basis, which according to the next result is possible when the two operators commute.

Proposition 6.40. *Two diagonalisable operators on the same vector space have diagonal matrices with respect to the same basis if and only if the two operators commute.*

Proof.

$\boxed{\implies}$ Suppose $S, T \in \mathcal{L}(V)$ have diagonal matrices with respect to the same basis.

Since any two diagonal matrices of the same size commute, by 6.38, S and T commute.

$\boxed{\impliedby}$ Suppose $S, T \in \mathcal{L}(V)$ are diagonalisable operators that commute, so $ST = TS$. Let $\lambda_1, \dots, \lambda_m$ denote the distinct eigenvalues of S.

Since S is diagonalisable, by 6.33,

$$V = E(\lambda_1, S) \oplus \cdots \oplus E(\lambda_m, S). \tag{$*$}$$

For each $i = 1, \dots, m$, the subspace $E(\lambda_i, S)$ is invariant under T (by 6.39). Since T is diagonalisable, by 6.36, the restriction $T|_{E(\lambda_i,S)}$ is diagonalisable for each i.

Hence for each $i = 1, \dots, m$, there is a basis of $E(\lambda_i, S)$ consisting of eigenvectors of T. Putting these bases together gives a basis of V (because of $(*)$), with each vector in this basis being an eigenvector of both S and T. Thus S and T both have diagonal matrices with respect to this basis, as desired. □

Suppose V is a finite-dimensional nonzero complex vector space. Then every operator on V has an eigenvector (by 6.12). The next result shows that if two operators on V commute, then there is a vector in V that is an eigenvector for both operators (but the two commuting operators might not have a common eigenvalue).

Lemma 6.41. *Every pairs of commuting operators on a finite-dimensional nonzero complex vector space has a common eigenvector.*

Proof. Suppose V is a finite-dimensional nonzero complex vector space. Suppose $S, T \in \mathcal{L}(V)$ commute.

Let λ be an eigenvalue of S (6.12 tells us that S does indeed have an eigenvalue). Thus $E(\lambda, S) \neq \{\mathbf{0}\}$. Also, $E(\lambda, S)$ is invariant under T (by 6.39).

Thus $T|_{E(\lambda,S)}$ has an eigenvector (again using 6.12), which is an eigenvector for both S and T, completing the proof. □

Remark. The hypothesis $\mathbb{C}$ is needed, since all vector spaces over $\mathbb{C}$ have eigenvalues, by 6.12.

Recall that 6.28 states that for every operator, there exists a basis that gives an upper-triangular matrix. We now extend this result to two commuting operators.

Proposition 6.42. *Suppose V is a finite-dimensional complex vector space, $S, T \in \mathcal{L}(V)$ commute. Then there exists a basis of V, with respect to which both S and T have upper-triangular matrices.*

Proof. Let $n = \dim V$. Induct on n.

The desired result holds if $n = 1$, since all 1×1 matrices are upper triangular.

Now suppose $n > 1$ and the desired result holds for all complex vector spaces whose dimension is $n - 1$.

Since S and T commute, by 6.41, let v_1 be any common eigenvalue of S and T. Hence $Sv_1 \in \operatorname{span}(v_1)$ and $Tv_1 \in \operatorname{span}(v_1)$. Let W be a subspace of V such that

$$V = \operatorname{span}(v_1) \oplus W;$$

see 2.33 for the existence of W. Define a linear map $P \colon V \to W$ by

$$P(av_1 + w) = w \quad (a \in \mathbb{C},\ w \in W).$$

Define $\tilde{S}, \tilde{T} \in \mathcal{L}(W)$ by

$$\tilde{S}w = P(Sw), \quad \tilde{T}w = P(Tw)$$

for each $w \in W$. To apply the induction hypothesis to $\tilde{S}$ and $\tilde{T}$, we must first show that they commute. Let $w \in W$, then there exists $a \in \mathbb{C}$ such that

$$(\tilde{S}\tilde{T})w = \tilde{S}(P(Tw)) = \tilde{S}(Tw - av_1) = P(S(Tw - av_1)) = P((ST)w),$$

where the last equality holds because v_1 is an eigenvector of S and $Pv_1 = 0$. Similarly,

$$(\tilde{T}\tilde{S})w = P((TS)w).$$

Since S and T commute, the last two displayed equations show that $(\tilde{S}T)w = (\tilde{T}S)w$. Hence $\tilde{S}$ and $\tilde{T}$ commute.

Thus we can use our induction hypothesis to state that there exists a basis $\{v_2, \dots, v_n\}$ of W such that $\tilde{S}$ and $\tilde{T}$ both have upper-triangular matrices with respect to this basis. The list $\{v_1, \dots, v_n\}$ is a basis of V.

If $k \in \{2, \dots, n\}$, then there exist $a_k, b_k \in \mathbb{C}$ such that

$$\begin{aligned} Sv_k &= a_k v_1 + \tilde{S}v_k \\ Tv_k &= b_k v_1 + \tilde{T}v_k \end{aligned}$$

Since $\tilde{S}$ and $\tilde{T}$ have upper-triangular matrices with respect to $\{v_2, \dots, v_n\}$, we know that $\tilde{S}v_k \in \operatorname{span}(v_2, \dots, v_k)$ and $\tilde{T}v_k \in \operatorname{span}(v_2, \dots, v_k)$. Hence the equations above imply that

$$Sv_k \in \operatorname{span}(v_1, \dots, v_k), \quad Tv_k \in \operatorname{span}(v_1, \dots, v_k).$$

Hence S and T have upper-triangular matrices with respect to $\{v_1, \dots, v_n\}$, as desired. □

In general, it is not possible to determine the eigenvalues of the sum or product of two operators from the eigenvalues of the two operators. However, the next result shows that something nice happens when the two operators commute.

Proposition 6.43 (Eigenvalues of sum and product of commuting operators)**.** *Suppose V is a finite-dimensional complex vector space, S and T are commuting operators on V. Then*

(i) every eigenvalue of $S + T$ is an eigenvalue of S plus an eigenvalue of T;

(ii) every eigenvalue of ST is an eigenvalue of S times an eigenvalue of T.

Proof.

(i) By 6.42, there exists a basis of V, with respect to which both S and T have upper-triangular matrices. With respect to that basis,

$$\mathcal{M}(S + T) = \mathcal{M}(S) + \mathcal{M}(T).$$

By definition of matrix addition, each entry on the diagonal of $\mathcal{M}(S + T)$ equals the sum of the corresponding entries on the diagonals of $\mathcal{M}(S)$ and $\mathcal{M}(T)$. Furthermore, $\mathcal{M}(S+T)$ is upper-triangular (as you should verify).

By 6.25,

- every entry on the diagonal of $\mathcal{M}(S)$ is an eigenvalue of S,
- every entry on the diagonal of $\mathcal{M}(T)$ is an eigenvalue of T, and
- every eigenvalue of $S+T$ is on the diagonal of $\mathcal{M}(S+T)$.

Hence every eigenvalue of $S+T$ is an eigenvalue of S plus an eigenvalue of T.

(ii) Similar to above.

$\square$

Exercises

Exercise 6.1 ([Axl24] 5A Q1)**.** Suppose $T \in \mathcal{L}(V)$, $U \leq V$. 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 $Tu \in U$.

□

Exercise 6.2 ([Axl24] 5A Q4)**.**

Exercise 6.3 ([Axl24] 5A Q8)**.**

Exercise 6.4 ([Axl24] 5A Q11)**.**

Exercise 6.5 ([Axl24] 5A Q13)**.**

Exercise 6.6 ([Axl24] 5A Q28)**.**

Exercise 6.7 ([Axl24] 5A Q32)**.**

5B 2 7 10 11 13 17 18 22

Exercise 6.8 ([Axl24] 5D Q1)**.** Suppose V is a finite-dimensional complex vector space and $T \in \mathcal{L}(V)$.

(i) Prove that if $T^4 = I$, then T is diagonalisable.

(ii) Prove that if $T^4 = T$, then T is diagonalisable.

(iii) Give an example of an operator $T \in \mathcal{L}(\mathbb{C}^2)$ such that $T^4 = T^2$ and T is not diagonalisable.

Solution.

(i) If $T^4 = I$, then $T^4 - I = 0$. Let

$$\begin{aligned} p(x) &= x^4 - 1 \\ &= (x+1)(x-1)(x+i)(x-i). \end{aligned}$$

Let $m(x)$ be the minimal polynomial of T. Then m divides p, which implies m only has simple roots (no repeated roots), so T is diagonalisable, by 5.62.

(ii) Similar to the above, consider $p(x) = x^4 - x = x(x-1)(x+i)(x-i)$.

(iii) Consider

$$T = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Then we have that

$$T^2 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = T^4.$$

□

Exercise 6.9 ([Axl24] 5D Q2)**.** Suppose $T \in \mathcal{L}(V)$ has a diagonal matrix A with respect to some basis of V. Prove that if $\lambda \in \mathbf{F}$, then λ appears on the diagonal of A precisely $\dim E(\lambda, T)$ times.

Exercise 6.10 ([Axl24] 5D Q3)**.** Suppose V is finite-dimensional and $T \in \mathcal{L}(V)$. Prove that if the operator T is diagonalisable, then $V = \ker T \oplus \operatorname{im} T$.

Exercise 6.11 ([Axl24] 5D Q4)**.** Suppose V is finite-dimensional and $T \in \mathcal{L}(V)$. Prove that the following are equivalent.

(i) $V = \ker T \oplus \operatorname{im} T$.

(ii) $V = \ker T + \operatorname{im} T$.

(iii) $\ker T \cap \operatorname{im} T = \{\mathbf{0}\}$.

Exercise 6.12 ([Axl24] 5D Q5)**.** Suppose V is a finite-dimensional complex vector space and $T \in \mathcal{L}(V)$. Prove that T is diagonalisable if and only if

$$V = \ker(T - \lambda I) \oplus \operatorname{im}(T - \lambda I)$$

for every $\lambda \in \mathbb{C}$.

Exercise 6.13 ([Axl24] 5D Q9)**.** Suppose $R, T \in \mathcal{L}(\mathbf{F}^3)$ each have 2, 6, 7 as eigenvalues. Prove that there exists an invertible operator $S \in \mathcal{L}(\mathbf{F}^3)$ such that $R = S^{-1}TS$.

Exercise 6.14 ([Axl24] 5D Q14)**.** Suppose $\mathbf{F} = \mathbb{C}$, k is a positive integer, and $T \in \mathcal{L}(V)$ is invertible. Prove that T is diagonalisable if and only if T^k is diagonalisable.

Exercise 6.15 ([Axl24] 5D Q20)**.** Suppose V is finite-dimensional and $T \in \mathcal{L}(V)$. Prove that T is diagonalisable if and only if the dual operator T^* is diagonalisable.

Exercise 6.16 ([Axl24] 5E Q2)**.** Suppose $\mathcal{E}$ is a subset of $\mathcal{L}(V)$ and every element of $\mathcal{E}$ is diagonalisable. Prove that there exists a basis of V with respect to which every element of $\mathcal{E}$ has a diagonal matrix if and only if every pair of elements of $\mathcal{E}$ commutes.

Exercise 6.17 ([Axl24] 5E Q3)**.** Suppose $S, T \in \mathcal{L}(V)$ are such that $ST = TS$. Suppose $p \in \mathbf{F}[z]$.

(i) Prove that $\ker p(S)$ is invariant under T.

(ii) Prove that $\operatorname{im} p(S)$ is invariant under T.

Exercise 6.18 ([Axl24] 5E Q4)**.** Prove or give a counterexample: If A is a diagonal matrix and B is an upper-triangular matrix of the same size as A, then A and B commute.

Solution. Counterexample:

$$\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$
$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix}$$

□

Exercise 6.19 ([Axl24] 5E Q5)**.** Prove that a pair of operators on a finite-dimensional vector space commute if and only if their dual operators commute.

Exercise 6.20 ([Axl24] 5E Q7)**.** Suppose V is a complex vector space, $S \in \mathcal{L}(V)$ is diagonalisable, and $T \in \mathcal{L}(V)$ commutes with S. Prove that there is a basis of V such that S has a diagonal matrix with respect to this basis and T has an upper-triangular matrix with respect to this basis.

Chapter 7

Inner Product Spaces

7.1 Inner Products and Norms

7.1.1 Inner Products

Recall that we can define a dot product on the Euclidean space $\mathbb{R}^n$ as

$$x \cdot y = x_1 y_1 + \cdots + x_n y_n,$$

where $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$.

We now generalise this notion.

Definition 7.1 (Inner product space)**.** An ***inner product*** on V is a map $\langle \cdot, \cdot \rangle : V \times V \to \mathbf{F}$ such that for all $u, v, w \in V$, $\lambda \in \mathbf{F}$,

(i) $\langle v, v \rangle \geq 0$, where equality holds if and only if $v = \mathbf{0}$ (positive definite)

(ii) $\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle$ (sesquilinear)
$\langle \lambda u, v \rangle = \lambda \langle u, v \rangle$

(iii) $\langle u, v \rangle = \overline{\langle v, u \rangle}$ (conjugate symmetric)

An ***inner product space*** $(V, \langle \cdot, \cdot \rangle)$ is a vector space V along with an inner product $\langle \cdot, \cdot \rangle$ on V.

Notation. If the inner product on V is clear from context, we omit it and simply denote the inner product space as V.

Remark. Every real number equals its complex conjugate. Thus if we are dealing with a real vector space, then in (iii) we can dispense with the complex conjugate, so $\langle u, v \rangle = \langle v, u \rangle$ for all $u, v \in V$.

Example 7.2.

- The *Euclidean inner product* on $\mathbf{F}^n$ is defined by

$$\langle (w_1, \dots, w_n), (z_1, \dots, z_n) \rangle = w_1 \overline{z_1} + \cdots + w_n \overline{z_n}$$

  for all $(w_1, \dots, w_n), (z_1, \dots, z_n) \in \mathbf{F}^n$.

- An inner product can be defined on the vector space $\mathcal{C}([-1,1],\mathbb{R})$ by

$$\langle f,g\rangle = \int_{-1}^{1} fg$$

for all $f,g \in \mathcal{C}([-1,1],\mathbb{R})$.

Lemma 7.3 (Basic properties of inner product)**.**

(i) For each fixed $u \in V$, the function that sends $u \mapsto \langle u,v\rangle$ is a linear map from V to $\mathbf{F}$.

(ii) $\langle 0,v\rangle = 0$ for every $v \in V$.

(iii) $\langle v,0\rangle = 0$ for every $v \in V$.

(iv) $\langle u,v+w\rangle = \langle u,v\rangle + \langle u,w\rangle$ for all $u,v,w \in V$.

(v) $\langle u,\lambda v\rangle = \overline{\lambda}\langle u,v\rangle$ for all $\lambda \in \mathbf{F}$, $u,v \in V$.

Proof.

(i) For $v \in V$, the linearity of $u \mapsto \langle u,v\rangle$ follows from the sesquilinearity of the inner product.

(ii) Every linear map takes $\mathbf{0}$ to 0. Thus (ii) follows from (i).

(iii) If $v \in V$, by conjugate symmetry and (ii),

$$\langle v,0\rangle = \overline{\langle 0,v\rangle} = \overline{0} = 0.$$

(iv) Suppose $u,v,w \in V$. Then

$$\begin{aligned}\langle u,v+w\rangle &= \overline{\langle v+w,u\rangle}\\ &= \overline{\langle v,u\rangle + \langle w,u\rangle}\\ &= \overline{\langle v,u\rangle} + \overline{\langle w,u\rangle}\\ &= \langle u,v\rangle + \langle u,w\rangle.\end{aligned}$$

(v) Suppose $\lambda \in \mathbf{F}$, $u,v \in V$. Then

$$\begin{aligned}\langle u,\lambda v\rangle &= \overline{\lambda v,u}\\ &= \overline{\lambda\langle v,u\rangle}\\ &= \overline{\lambda}\overline{\langle v,u\rangle}\\ &= \overline{\lambda}\langle u,v\rangle.\end{aligned}$$

□

7.1.2 Norms

Each inner product determines a norm.

Definition 7.4 (Norm)**.** For $v \in V$, the ***norm*** of v is

$$\|v\| := \sqrt{\langle v, v\rangle}.$$

Lemma 7.5 (Basic properties of norm)**.** *Suppose* $v \in V$.

(i) $\|v\| = 0$ *if and only if* $v = \mathbf{0}$.

(ii) $\|\lambda v\| = |\lambda|\|v\|$ *for all* $\lambda \in \mathbf{F}$.

Proof.

(i) By positive definiteness of the inner product, $\langle v, v\rangle = 0$ if and only if $v = \mathbf{0}$. Take square root to get $\|v\| = 0$.

(ii) Suppose $\lambda \in \mathbf{F}$. Then

$$\|\lambda v\|^2 = \langle \lambda v, \lambda v\rangle = \lambda \langle v, \lambda v\rangle = \lambda\bar{\lambda} \langle v, v\rangle = |\lambda|^2\|v\|^2.$$

Taking square roots yields the desired equality.

□

Remark. Working with norms squared is usually easier than working directly with norms.

Now we come to a crucial definition.

Definition 7.6 (Orthogonal vectors)**.** We say $u, v \in V$ are ***orthogonal*** if $\langle u, v\rangle = 0$.

Lemma 7.7 (Orthogonality and $\mathbf{0}$)**.**

(i) $\mathbf{0}$ *is orthogonal to every vector in* V.

(ii) $\mathbf{0}$ *is the only vector in* V *that is orthogonal to itself.*

Proof.

(i) Recall that $\langle \mathbf{0}, v\rangle = 0$ for every $v \in V$.

(ii) If $v \in V$ and $\langle v, v\rangle = 0$, then $v = \mathbf{0}$, by positive definiteness.

□

Lemma 7.8 (Pythagoras' theorem)**.** *Suppose* $u, v \in V$. *If* u *and* v *are orthogonal, then*

$$\|u + v\|^2 = \|u\|^2 + \|v\|^2. \tag{7.1}$$

Proof. Suppose $\langle u, v\rangle = 0$. Then

$$\begin{aligned}\|u+v\|^2 &= \langle u+v, u+v\rangle \\ &= \langle u, u+v\rangle + \langle v, u+v\rangle \\ &= \langle u,u\rangle + \langle u,v\rangle + \langle v,u\rangle + \langle v,v\rangle \\ &= \|u\|^2 + 0 + \bar{0} + \|v\|^2 \\ &= \|u\|^2 + \|v\|^2\end{aligned}$$

as desired. □

We now introdoce a process known as *orthogonal decomposition*. Suppose $u, v \in V$, $u \neq \mathbf{0}$. Then the ***orthogonal projection*** of v onto u is

$$\operatorname{proj}_u(v) := \frac{\langle v,u\rangle}{\langle u,u\rangle}u, \tag{7.2}$$

which is parallel to u. We check that $v - \operatorname{proj}_u(v)$ and u are orthogonal:

$$\begin{aligned}\langle v - \operatorname{proj}_u(v), u\rangle &= \langle v,u\rangle - \left\langle \frac{\langle v,u\rangle}{\langle u,u\rangle}u, u\right\rangle \\ &= \langle v,u\rangle - \frac{\langle v,u\rangle}{\langle u,u\rangle}\langle u,u\rangle = 0.\end{aligned}$$

Lemma 7.9 (Cauchy–Schwarz inequality)**.** *Suppose $u, v \in V$. Then*

$$|\langle u,v\rangle| \leq \|u\|\|v\|, \tag{7.3}$$

where equality holds if and only if $u = \lambda v$ for some scalar λ.

Proof. If $u = \mathbf{0}$, then both sides of the desired inequality equal 0. Thus assume $u \neq \mathbf{0}$. Consider the orthogonal decomposition of v:

$$v = (v - \operatorname{proj}_u(v)) + \operatorname{proj}_u(v).$$

By the Pythagoras' theorem,

$$\|v\|^2 = \underbrace{\|v - \operatorname{proj}_u(v)\|^2}_{\geq 0} + \|\operatorname{proj}_u(v)\|^2,$$

so

$$\|v\| \geq \|\operatorname{proj}_u(v)\| = \left|\frac{\langle v,u\rangle}{\langle u,u\rangle}\right| \|u\| = \frac{|\langle v,u\rangle|}{\|u\|}$$

and rearranging gives the desired inequality. Equality holds if and only if $v = \operatorname{proj}_u(v)$, i.e.,

$$\frac{\langle v,u\rangle}{\langle u,u\rangle}u = v.$$

□

Lemma 7.10 (Triangle inequality). *Suppose $u, v \in V$. Then*

$$\|u+v\| \leq \|u\| + \|v\|, \tag{7.4}$$

where equality holds if and only if $u = \lambda v$ for some $\lambda \in \mathbb{R}_{\geq 0}$.

Proof. We have

$$\begin{aligned}
\|u+v\|^2 &= \langle u+v, u+v\rangle \\
&= \langle u,u\rangle + \langle v,v\rangle + \langle u,v\rangle + \langle v,u\rangle \\
&= \langle u,u\rangle + \langle v,v\rangle + \langle u,v\rangle + \overline{\langle u,v\rangle} \\
&= \|u\|^2 + \|v\|^2 + 2\operatorname{Re}\langle u,v\rangle \\
&\leq \|u\|^2 + \|v\|^2 + 2|\langle u,v\rangle| && (*) \\
&\leq \|u\|^2 + \|v\|^2 + 2\|u\|\|v\| && (**) \\
&= (\|u\| + \|v\|)^2,
\end{aligned}$$

where $(**)$ follows from the Cauchy–Schwarz inequality. Taking square roots of both sides of the above inequality gives the desired inequality.

Equality holds if and only if equality holds in $(*)$ and $(**)$, i.e.,

$$\langle u,v\rangle = \|u\|\|v\|.$$

If $u = \lambda v$ for $\lambda \in \mathbb{R}_{\geq 0}$, then the above equation holds.

Conversely, suppose the above equation holds. Then equality in the Cauchy–Schwarz inequality implies that $u = \lambda v$ for some scalar λ. By the above equation, λ must be a non-negative real number, completing the proof. □

Corollary 7.11 (Reverse triangle inequality). *Suppose $u, v \in V$. Then*

$$\big|\|u\| - \|v\|\big| \leq \|u - v\|.$$

Proof. We have

$$\begin{aligned}
\|u-v\|^2 &= \langle u-v, u-v\rangle \\
&= \|u\|^2 + \|v\|^2 - (\langle u,v\rangle + \langle v,u\rangle) \\
&\geq \|u\|^2 + \|v\|^2 - 2\|u\|\|v\| \\
&= (\|u\| - \|v\|)^2.
\end{aligned}$$

Taking square roots yields the desired result. □

Lemma 7.12 (Parallelogram equality). *Suppose $u, v \in V$. Then*

$$\|u+v\|^2 + \|u-v\|^2 = 2\left(\|u\|^2 + \|v\|^2\right). \tag{7.5}$$

Proof. We have

$$\begin{aligned}
\|u+v\|^2 + \|u-v\|^2 &= \langle u+v, u+v\rangle + \langle u-v, u-v\rangle \\
&= \left(\|u\|^2 + \|v\|^2 + \langle u,v\rangle + \langle v,u\rangle\right) + \left(\|u\|^2 + \|v\|^2 - \langle u,v\rangle - \langle v,u\rangle\right) \\
&= 2\left(\|u\|^2 + \|v\|^2\right)
\end{aligned}$$

as desired. □

7.2 Orthonormal Bases

7.2.1 Orthonormal Bases

Definition 7.13 (Orthonormal basis)**.** We say $\{e_1, \dots, e_n\} \subset V \setminus \{\mathbf{0}\}$ is ***orthonormal*** if

(i) $\|e_i\| = 1$;

(ii) the vectors are pairwise orthogonal.

If additionally $\{e_1, \dots, e_n\}$ is a basis of V, then $\{e_1, \dots, e_n\}$ is a ***orthonormal basis*** of V.

Lemma 7.14. *Suppose $\{e_1, \dots, e_n\}$ is a orthonormal set of vectors in V. Then*

$$\|a_1 e_1 + \cdots + a_n e_n\|^2 = |a_1|^2 + \cdots + |a_n|^2$$

for all $a_1, \dots, a_n \in \mathbf{F}$.

Proof. By the Pythagoras' theorem,

$$\begin{aligned}\|a_1 e_1 + \cdots + a_n e_n\|^2 &= \|a_1 e_1\|^2 + \cdots + \|a_n e_n\|^2 \\ &= |a_1|^2 \|e_1\|^2 + \cdots + |a_n|^2 \|e_n\|^2 \\ &= |a_1|^2 + \cdots + |a_n|^2\end{aligned}$$

since each $\|e_i\| = 1$. □

The result above has the following important corollary.

Corollary 7.15. *Every orthonormal set of vectors is linearly independent.*

Proof. Suppose $\{e_1, \dots, e_n\}$ is an orthonormal set of vectors in V. Suppose $a_1, \dots, a_n \in \mathbf{F}$ are such that

$$a_1 e_1 + \cdots + a_n e_n = \mathbf{0}.$$

By the previous result,

$$|a_1|^2 + \cdots + |a_n|^2 = 0,$$

so $a_1 = \cdots = a_n = 0$. Hence $e_1, \dots, e_n$ are linearly independent. □

Corollary 7.16. *Suppose V is finite-dimensional. Then every orthonormal set of vectors in V of length $\dim V$ is an orthonormal basis of V.*

Proof. By 7.15, every orthonormal set of vectors in V is linearly independent. Thus by 3.41, every such set of length $\dim V$ is a basis. □

Now we come to an important inequality.

Lemma 7.17 (Bessel's inequality). *Suppose $\{e_1, \dots, e_n\}$ is an orthonormal set of vectors in V. If $v \in V$ then*

$$|\langle v, e_1\rangle|^2 + \cdots + |\langle v, e_n\rangle|^2 \leq \|v\|^2. \tag{7.6}$$

Proof. Let $v \in V$. For $i = 1, \dots, n$, consider the orthogonal projection of v onto e_i:

$$\begin{aligned} v &= (v - \operatorname{proj}_{e_i}(v)) + \operatorname{proj}_{e_i}(v) \\ &= \left(v - \frac{\langle v, e_i\rangle}{\langle e_i, e_i\rangle} e_i\right) + \frac{\langle v, e_i\rangle}{\langle e_i, e_i\rangle} e_i \\ &= (v - \langle v, e_i\rangle e_i) + \langle v, e_i\rangle e_i. \end{aligned}$$

Then by Pythagoras' theorem,

$$\begin{aligned} \|v\|^2 &= \|v - \langle v, e_i\rangle e_i\|^2 + \|\langle v, e_i\rangle e_i\| \\ &= \|v - \langle v, e_i\rangle e_i\|^2 + |\langle v, e_i\rangle|^2. \end{aligned}$$

Write

$$\begin{aligned} v &= \operatorname{proj}_{e_1}(v) + \cdots + \operatorname{proj}_{e_n}(v) + w \\ &= \langle v, e_1\rangle e_1 + \cdots + \langle v, e_n\rangle e_n + w \end{aligned}$$

for some $w \in V$. Note that for $i = 1, \dots, n$,

$$\begin{aligned} \langle v, e_i\rangle &= \langle\langle v, e_i\rangle e_i, e_i\rangle + \langle w, e_i\rangle \\ &= \langle v, e_i\rangle + \langle w, e_i\rangle \end{aligned}$$

which implies $\langle w, e_i\rangle = 0$, so w is orthogonal to $e_1, \dots, e_n$. Thus $e_1, \dots, e_n, w$ are pairwise orthogonal. By Pythagoras' theorem,

$$\begin{aligned} \|v\|^2 &= |\langle v, e_1\rangle|^2 + \cdots + |\langle v, e_n\rangle|^2 + \underbrace{\|w\|^2}_{\geq 0} \\ &\geq |\langle v, e_1\rangle|^2 + \cdots + |\langle v, e_n\rangle|^2 \end{aligned}$$

as desired. Equality holds for orthonormal bases (as we will see later). □

The next result states that a vector can be expressed as a linear combination of an orthonormal basis. Usually we write $v = \sum_{i=1}^n a_i v_i$, but with orthonormal basis we can just take $a_i = \langle v, e_i\rangle$.

Lemma 7.18. *Suppose $\{e_1, \dots, e_n\}$ is an orthonormal basis of V, and $u, v \in V$. Then*

$$v = \langle v, e_1\rangle e_1 + \cdots + \langle v, e_n\rangle e_n. \tag{7.7}$$

Proof. Since $\{e_1, \dots, e_n\}$ is a basis of V, there exist $a_1, \dots, a_n \in \mathbf{F}$ such that

$$v = a_1 e_1 + \cdots + a_n e_n.$$

Since $e_1, \ldots, e_n$ are orthonormal, taking the inner product of both sides with e_i gives

$$\langle v, e_i \rangle = a_i \quad (i = 1, \ldots, n).$$

Hence we are done. $\square$

Applying Pythagoras' theorem to Eq. (7.7), we obtain *Parseval's identity*:

$$\|v\|^2 = |\langle v, e_1 \rangle|^2 + \cdots + |\langle v, e_n \rangle|^2. \tag{7.8}$$

Let $u, v \in V$. Taking the inner product of u on both sides of Eq. (7.7) gives

$$\begin{aligned}
\langle u, v \rangle &= \langle u, \langle v, e_1 \rangle e_1 + \cdots + \langle v, e_n \rangle e_n \rangle \\
&= \langle u, \langle v, e_1 \rangle e_1 \rangle + \cdot + \langle u, \langle v, e_n \rangle e_n \rangle \\
&= \overline{\langle v, e_1 \rangle} \langle u, e_1 \rangle + \cdots + \overline{\langle v, e_n \rangle} \langle u, e_n \rangle
\end{aligned}$$

that is,

$$\langle u, v \rangle = \langle u, e_1 \rangle \overline{\langle v, e_1 \rangle} + \cdots + \langle u, e_n \rangle \overline{\langle v, e_n \rangle}. \tag{7.9}$$

7.2.2 Gram–Schmidt Procedure

The *Gram–Schmidt procedure* is a method for constructing orthonormal basis, by turning a linearly independent set into an orthonormal set with the same span as the original set. It guarantees the existence of orthonormal bases.

Theorem 7.19 (Gram–Schmidt procedure)**.** *Suppose $v_1, \dots, v_n$ are linearly independent in V. Define*

$$u_k = \begin{cases} v_1 & (k=1) \\ v_k - \operatorname{proj}_{u_1}(v_k) - \cdots - \operatorname{proj}_{u_{k-1}}(v_k) & (k=2,\dots,n) \end{cases}$$

and let

$$e_k = \frac{u_k}{\|u_k\|}.$$

Then $\{e_1, \dots, e_n\}$ is an orthonormal set of vectors in V such that

$$\operatorname{span}(v_1, \dots, v_k) = \operatorname{span}(e_1, \dots, e_k)$$

for $k = 1, \dots, n$.

Proof. Induct on k.

For the base case $k = 1$, since $e_1 = \frac{u_1}{\|u_1\|}$ we have $\|e_1\| = 1$, and $\operatorname{span}(v_1) = \operatorname{span}(e_1)$ because e_1 is a non-zero multiple of v_1.

Suppose the desired result holds for $k-1$; that is, the set $\{e_1, \dots, e_{k-1}\}$ generated by the above procedure is an orthonormal set, and

$$\operatorname{span}(v_1, \dots, v_{k-1}) = \operatorname{span}(e_1, \dots, e_{k-1}). \tag{I}$$

Since $v_1, \dots, v_n$ are linearly independent, we have $v_k \notin \operatorname{span}(v_1, \dots, v_{k-1})$. Thus $v_k \notin \operatorname{span}(e_1, \dots, e_{k-1}) = \operatorname{span}(u_1, \dots, u_{k-1})$, which implies that $u_k \neq \mathbf{0}$ (so we are not dividing by 0); thus $\|e_k\| = 1$.

We now check that $e_1, \dots, e_k$ is an orthonormal set. For $j \in \{1, \dots, k-1\}$,

$$\begin{aligned}
\langle e_k, e_j \rangle &= \left\langle \frac{u_k}{\|u_k\|}, \frac{u_j}{\|u_j\|} \right\rangle \\
&= \frac{1}{\|u_k\|\|u_j\|} \langle u_k, u_j \rangle \\
&= \frac{1}{\|u_k\|\|u_j\|} \left\langle v_k - \operatorname{proj}_{u_1}(v_k) - \cdots - \operatorname{proj}_{u_j}(v_k) - \cdots - \operatorname{proj}_{u_{k-1}}(v_k), u_j \right\rangle \\
&= \frac{1}{\|u_k\|\|u_j\|} \left\langle v_k - \frac{\langle v_k, u_1 \rangle}{\langle u_1, u_1 \rangle} u_1 - \cdots - \frac{\langle v_k, u_j \rangle}{\langle u_j, u_j \rangle} u_j - \cdots - \frac{\langle v_k, u_{k-1} \rangle}{\langle u_{k-1}, u_{k-1} \rangle} u_{k-1}, u_j \right\rangle \\
&= \frac{1}{\|u_k\|\|u_j\|} \left(\langle v_k, u_j \rangle - \left\langle \frac{\langle v_k, u_j \rangle}{\langle u_j, u_j \rangle} u_j, u_j \right\rangle \right) \\
&= \frac{1}{\|u_k\|\|u_j\|} \left(\langle v_k, u_j \rangle - \langle v_k, u_j \rangle \right) = 0
\end{aligned}$$

so e_k is orthogonal to $e_1, \dots, e_{k-1}$. Hence $e_1, \dots, e_k$ is an orthonormal set of vectors.

From the definition of e_k, we see that $v_k \in \operatorname{span}(e_1, \ldots, e_k)$. Combining this information with (I) shows that

$$\operatorname{span}(v_1, \ldots, v_k) \subset \operatorname{span}(e_1, \ldots, e_k).$$

Both sets above are linearly independent (the v's by hypothesis, and the e's by orthonormality and 7.15). Thus both subspaces above have dimension k, and hence they are equal, completing the induction step and thus completing the proof. □

Now we can answer the question about the *existence* of orthonormal bases.

Corollary 7.20. *Every finite-dimensional inner product space has an orthonormal basis.*

Proof. Suppose V is finite-dimensional. Choose a basis of V.

Apply the Gram–Schmidt procedure (7.19) to it, producing an orthonormal set of length $\dim V$. By 7.16, this orthonormal set is an orthonormal basis of V. □

Sometimes we need to know not only that an orthonormal basis exists, but also that every orthonormal set can be extended to an orthonormal basis. In the next corollary, the Gram–Schmidt procedure shows that such an extension is always possible.

Corollary 7.21. *Suppose V is finite-dimensional. Then every orthonormal set of vectors in V can be extended to an orthonormal basis of V.*

Proof. Suppose $\{e_1, \ldots, e_m\}$ is an orthonormal set of vectors in V. By 7.15, $\{e_1, \ldots, e_m\}$ is linearly independent, and thus can be extended to a basis $\{e_1, \ldots, e_m, v_1, \ldots, v_n\}$ of V.

Now apply the Gram–Schmidt procedure to $\{e_1, \ldots, e_m, v_1, \ldots, v_n\}$, producing an orthonormal set

$$\{e_1, \ldots, e_m, u_1, \ldots, u_n\}$$

where the first m vectors are unchanged because they are already orthonormal. The set above is an orthonormal basis of V by 7.16. □

The next result shows that the condition for an operator to have an upper-triangular matrix with respect to some *orthonormal basis* is the same as the condition to have an upper-triangular matrix with respect to an *arbitrary basis* (recall 6.27).

Proposition 7.22. *Suppose V is finite-dimensional, $T \in \mathcal{L}(V)$. Then T has an upper-triangular matrix with respect to some orthonormal basis of V if and only if the minimal polynomial of T equals*

$$(z - \lambda_1) \cdots (z - \lambda_m)$$

for some $\lambda_1, \ldots, \lambda_m \in \mathbf{F}$.

Proof. Suppose T has an upper-triangular matrix with respect to some basis $\{v_1, \ldots, v_n\}$ of V. Thus by 6.23, $\operatorname{span}(v_1, \ldots, v_k)$ is invariant under T for each $k = 1, \ldots, n$.

Apply the Gram–Schmidt procedure to $\{v_1, \ldots, v_n\}$, producing an orthonormal basis $\{e_1, \ldots, e_n\}$ of V. Since

$$\operatorname{span}(e_1, \ldots, e_k) = \operatorname{span}(v_1, \ldots, v_k)$$

for each k, we conclude that $\operatorname{span}(e_1, \ldots, e_k)$ is invariant under T for each $k = 1, \ldots, n$. Thus by 6.23, T has an upper-triangular matrix with respect to the orthonormal basis $\{e_1, \ldots, e_n\}$. Now use 6.27 to complete the proof. □

Theorem 7.23 (Schur's theorem)**.** *Every operator on a finite-dimensional complex inner product space has an upper-triangular matrix with respect to some orthonormal basis.*

Proof. The desired result follows from the second version of the fundamental theorem of algebra (5.11) and 7.22. □

7.2.3 Linear Functionals on Inner Product Spaces

Theorem 7.24 (Riesz representation theorem)**.** *Suppose V is finite-dimensional, and ϕ is a linear functional on V. Then for every $u \in V$, there exists a unique $v \in V$ such that*

$$\phi(u) = \langle u, v \rangle .$$

Proof.

$\boxed{\text{Existence}}$ Pick an orthonormal basis $\{e_1, \ldots, e_n\}$ of V. Let $u \in V$. By 7.18,

$$u = \langle u, e_1 \rangle e_1 + \cdots + \langle u, e_n \rangle e_n.$$

Applying ϕ on u gives

$$\begin{aligned}\phi(u) &= \phi\left(\langle u, e_1 \rangle e_1 + \cdots + \langle u, e_n \rangle e_n\right) \\ &= \langle u, e_1 \rangle \phi(e_1) + \cdots + \langle u, e_n \rangle \phi(e_n) \\ &= \left\langle u, \overline{\phi(e_1)} e_1 \right\rangle + \cdots + \left\langle u, \overline{\phi(e_n)} e_n \right\rangle \\ &= \left\langle u, \overline{\phi(e_1)} e_1 + \cdots + \overline{\phi(e_n)} e_n \right\rangle .\end{aligned}$$

Pick

$$v = \overline{\phi(e_1)} e_1 + \cdots + \overline{\phi(e_n)} e_n.$$

Then we have $\phi(u) = \langle u, v \rangle$ for every $u \in V$, as desired.

$\boxed{\text{Uniqueness}}$ Suppose $v, v' \in V$ satisfy

$$\phi(u) = \langle u, v \rangle = \langle u, v' \rangle$$

for every $u \in V$. Then

$$0 = \langle u, v \rangle - \langle u, v' \rangle = \langle u, v - v' \rangle$$

for every $u \in V$. Taking $u = v - v'$ shows that $v - v' = \mathbf{0}$, so $v = v'$. $\square$

7.3 Orthogonal Complements and Minimisation Problems

7.3.1 Orthogonal Complements

Definition 7.25 (Orthogonal complement)**.** The ***orthogonal complement*** of $U \subset V$ is

$$U^\perp := \{v \in V \mid \langle u, v\rangle = 0, \forall u \in U\}.$$

That is, $U^\perp$ is the set of vectors in V that are orthogonal to every vector in U.

We check that if $U \subset V$, then $U^\perp \leq V$:

(i) $\langle u, \mathbf{0}\rangle = 0$ for every $u \in U$, so $\mathbf{0} \in U^\perp$.

(ii) Let $v, w \in U^\perp$. For every $u \in U$,

$$\langle u, v + w\rangle = \langle u, v\rangle + \langle u, w\rangle = 0 + 0 = 0 \implies v + w \in U^\perp$$

so $U^\perp$ is closed under addition.

(iii) Let $v \in U^\perp$, $\lambda \in \mathbf{F}$. For every $u \in U$,

$$\langle u, \lambda v\rangle = \overline{\lambda}\,\langle u, v\rangle = \overline{\lambda} \cdot 0 = 0 \implies \lambda v \in U^\perp$$

so $U^\perp$ is closed under scalar multiplication.

Example 7.26.

- Let U be a plane in $\mathbb{R}^3$ containing the origin. Then $U^\perp$ is the line containing the origin that is perpendicular to U.
- Let U be a line in $\mathbb{R}^3$ containing the origin. Then $U^\perp$ is the plane containing the origin that is perpendicular to U.

We begin with some straightforward consequences of the definition.

Lemma 7.27 (Properties of orthogonal complement)**.**

(i) $\{\mathbf{0}\}^\perp = V$, $V^\perp = \{\mathbf{0}\}$.

(ii) If $U \subset V$, *then* $U \cap U^\perp \subset \{\mathbf{0}\}$.

(iii) If $G \subset H \subset V$, *then* $H^\perp \subset G^\perp$.

Proof.

(i)

$$v \in \{\mathbf{0}\}^\perp \iff \langle \mathbf{0}, v\rangle = 0 \iff v \in V$$

$$v \in V^\perp \iff \langle v, v\rangle = 0 \iff v = \mathbf{0}$$

(ii) Suppose $U \subset V$. Let $u \in U \cap U^\perp$, then $\langle u, u\rangle = 0$ so $u = \mathbf{0}$. Hence $U \cap U^\perp \subset \{\mathbf{0}\}$.

(iii) Suppose $G \subset H \subset V$. Let $v \in H^\perp$, then

$$\langle u, v\rangle = 0 \quad (\forall u \in H)$$

which implies

$$\langle u, v\rangle = 0 \quad (\forall u \in G).$$

Hence $v \in G^\perp$, so $H^\perp \subset G^\perp$.

□

The next result shows that every *finite-dimensional* subspace of V leads to a natural direct sum decomposition of V.

Lemma 7.28. *Suppose $U \leq V$ is finite-dimensional. Then*

$$V = U \oplus U^\perp.$$

Proof. We first show that $V = U + U^\perp$. Let $v \in V$, pick an orthonormal basis $\{e_1, \dots, e_m\}$ of U. We can write

$$v = \underbrace{\langle v, e_1\rangle e_1 + \cdots + \langle v, e_m\rangle e_m}_{u} + \underbrace{v - \langle v, e_1\rangle e_1 - \cdots - \langle v, e_m\rangle e_m}_{w}. \tag{I}$$

We are left to check that $u \in U$ and $w \in U^\perp$.

- Since each $u_i \in U$, we see that $u \in U$.
- Since $\{e_1, \dots, e_m\}$ is an orthonormal set of vectors, for each $i = 1, \dots, m$,

$$\begin{aligned}\langle w, e_i\rangle &= \langle v - \langle v, e_1\rangle e_1 - \cdots - \langle v, e_m\rangle e_m, e_i\rangle \\ &= \langle v, e_i\rangle - \langle v, e_i\rangle = 0.\end{aligned}$$

  Thus w is orthogonal to every vector in $\operatorname{span}(e_1, \dots, e_m)$, which shows that $w \in U^\perp$.

Since $U \cap U^\perp = \{\mathbf{0}\}$, by 3.19, $U + U^\perp$ is a direct sum. □

Corollary 7.29. *Suppose V is finite-dimensional and $U \leq V$. Then*

$$\dim U^\perp = \dim V - \dim U.$$

Proof. Since $U + U^\perp$ is a direct sum, by 4.58, we have that $\dim V = \dim U + \dim U^\perp$, or

$$\dim U^\perp = \dim V - \dim U.$$

□

Corollary 7.30. *Suppose $U \le V$ is finite-dimensional. Then*

$$U = (U^\perp)^\perp.$$

Proof.

$\boxed{\subset}$ Let $u \in U$. Then $\langle u, w\rangle = 0$ for every $w \in U^\perp$ (by definition of $U^\perp$). Since u is orthogonal to every vector in $U^\perp$, we have $u \in (U^\perp)^\perp$.

Hence $U \subset (U^\perp)^\perp$.

$\boxed{\supset}$ Let $v \in (U^\perp)^\perp$. Since $U + U^\perp$ is a direct product, we write $v = u + w$ for some $u \in U$, $w \in U^\perp$. Then $v - u = w \in U^\perp$. Since $v \in (U^\perp)^\perp$ and $u \in (U^\perp)^\perp$ (as $U \subset (U^\perp)^\perp$), we have $v - u \in (U^\perp)^\perp$. Thus $v - u \in U^\perp \cap (U^\perp)^\perp$, which implies that $v - u = \mathbf{0}$, so $v = u$, and thus $v \in U$.

Hence $(U^\perp)^\perp \in U$. □

Suppose U is a subspace of V and we want to show that $U = V$. Sometimes the easiest way to do so is to show that the only vector orthogonal to U is $\mathbf{0}$, and then use the result below.

Corollary 7.31. *Suppose $U \le V$ is finite-dimensional. Then*

$$U^\perp = \{\mathbf{0}\} \iff U = V.$$

Proof.

$\boxed{\implies}$ Suppose $U^\perp = \{\mathbf{0}\}$. Then $U = (U^\perp)^\perp = \{\mathbf{0}\}^\perp = V$, as desired.

$\boxed{\impliedby}$ If $U = V$, then $U^\perp = V^\perp = \{\mathbf{0}\}$. □

We now define an operator P_U for each finite-dimensional subspace U of V.

Definition 7.32 (Orthogonal projection)**.** Suppose $U \le V$ is finite-dimensional. The ***orthogonal projection*** is the operator $P_U \in \mathcal{L}(V)$ defined as follows: For each $v \in V$, write $v = u + w$ for some $u \in U$, $w \in U^\perp$. Then let $P_U v = u$.

Remark. The direct sum decomposition $V = U \oplus U^\perp$ shows that each $v \in V$ can be uniquely written in the form $v = u + w$ with $u \in U$, $w \in U^\perp$. Thus $P_U v$ is well defined.

Suppose $u \in V$ with $u \neq \mathbf{0}$ and $U = \operatorname{span}(u)$. If $v \in V$ then

$$v = \frac{\langle v, u\rangle}{\|u\|^2}u + \left(v - \frac{\langle v, u\rangle}{\|u\|^2}u\right).$$

Then this implies that

$$P_U v := \frac{\langle v, u\rangle}{\|u\|^2}u.$$

We now check that $P_U \in \mathcal{L}(V)$.

(i) Let $v_1, v_2 \in V$. Write

$$v_1 = u_1 + w_1, \quad v_2 = u_2 + w_2$$

for some $u_1, u_2 \in U$, $w_1, w_2 \in U^\perp$. Thus $P_U v_1 = u_1$ and $P_U v_2 = u_2$. Since

$$v_1 + v_2 = \underbrace{(u_1 + u_2)}_{\in U} + \underbrace{(w_1 + w_2)}_{\in U^\perp},$$

we have

$$P_U(v_1 + v_2) = u_1 + u_2 = P_U v_1 + P_U v_2.$$

(ii) Let $\lambda \in \mathbf{F}$, $v \in V$. Write $v = u + w$, where $u \in U$, $w \in U^\perp$. Then

$$\lambda v = \underbrace{\lambda u}_{\in U} + \underbrace{\lambda w}_{\in U^\perp},$$

so

$$P_U(\lambda v) = \lambda u = \lambda P_U v.$$

Lemma 7.33 (Properties of orthogonal projection)**.** *Suppose $U \leq V$ is finite-dimensional.*

(i) $P_U u = u$ *for every* $u \in U$, $P_U w = \mathbf{0}$ *for every* $w \in U^\perp$.

(ii) $\operatorname{im} P_U = U$, $\ker P_U = U^\perp$.

(iii) $v - P_U v \in U^\perp$ *for every* $v \in V$.

(iv) ${P_U}^2 = P_U$.

(v) $\|P_U v\| \leq \|v\|$ *for every* $v \in V$.

(vi) If $\{e_1, \dots, e_n\}$ *is an orthonormal basis of* U*, and* $v \in V$*, then*

$$P_U v = \langle v, e_1 \rangle e_1 + \cdots + \langle v, e_n \rangle e_n.$$

Proof.

(i) Let $u \in U$. We can write $u = u + \mathbf{0}$, where $u \in U$, $\mathbf{0} \in U^\perp$. Thus $P_U u = u$.

Let $w \in U^\perp$. We can write $w = \mathbf{0} + w$, where $\mathbf{0} \in U$, $w \in U^\perp$. Thus $P_U w = \mathbf{0}$.

(ii) The definition of P_U implies that $\operatorname{im} P_U \subset U$. Furthermore, (i) implies that $U \subset \operatorname{im} P_U$. Hence $\operatorname{im} P_U = U$.

The inclusion $U^\perp \subset \ker P_U$ follows from (i). To prove the inclusion in the other direction, if $v \in \ker P_U$, then the decomposition given by 7.28 must be $v = \mathbf{0} + v$, where $\mathbf{0} \in U$ and $v \in U^\perp$. Thus $\ker P_U \subset U^\perp$.

(iii) If $v \in V$ and $v = u + w$ with $u \in U$, $w \in U^\perp$, then

$$v - P_U v = v - u = w \in U^\perp.$$

(iv) If $v \in V$ and $v = u + w$ with $u \in U$, $w \in U^\perp$, then

$$({P_U}^2)v = P_U(P_U v) = P_U u = u = P_U v.$$

(v) If $v \in V$ and $v = u + w$ with $u \in U$, $w \in U^{\perp}$, then

$$\|P_U v\|^2 \leq \|u\|^2 \leq \|u\|^2 + \|w\|^2 = \|v\|^2,$$

where the last equality comes from the Pythagorean theorem.

(vi) The formula for $P_U v$ follows from equation (I) in the proof of 7.28.

□

7.3.2 Minimisation Problems

Given a subspace U of V and a point $v \in V$, we want to find a point $u \in U$ that minimises $\|v - u\|$. The next result shows that $u = P_U v$ is the unique solution of this minimisation problem.

Theorem 7.34 (Minimising distance to a subspace)**.** *Suppose $U \leq V$ is finite-dimensional. Fix $v \in V$. Then for all $u \in U$*

$$\|v - P_U v\| \leq \|v - u\|, \tag{7.10}$$

where equality holds if and only if $u = P_U v$.

Proof. We have

$$\begin{aligned}
\|v - P_U v\|^2 &\leq \|v - P_U v\|^2 + \|P_U v - u\|^2 && [\because \|P_U v - u\|^2 \geq 0] \\
&= \|(v - P_U v) + (P_U v - u)\|^2 && [\text{by Pythagoras' theorem}] \\
&= \|v - u\|^2.
\end{aligned}$$

Taking square roots gives the desired inequality. Equality holds if and only if $\|P_U v - u\| = 0$, which holds if and only if $u = P_U v$. □

insert figure

7.3.3 Pseudoinverse

Suppose $T \in \mathcal{L}(V, W)$ and $w \in W$. Consider the problem of finding $v \in V$ such that

$$Tv = w.$$

If T is invertible, then evidently we are done. The pseudoinverse will provide the tool to solve the equation above as well as possible, even when T is not invertible.

We need the next result to define the pseudoinverse; it states that we can restrict a linear map to obtain a bijective map.

Lemma 7.35. *Suppose V is finite-dimensional, $T \in \mathcal{L}(V)$. Then $T|_{(\ker T)^\perp}$ is an bijective map from $(\ker T)^\perp$ to $\operatorname{im} T$.*

Proof. To prove bijectivity, we need to show injectivity and surjectivity.

Injectivity Let $v \in (\ker T)^\perp$ be such that $v \in \ker T|_{(\ker T)^\perp}$. Then

$$\begin{aligned} T|_{(\ker T)^\perp} v = \mathbf{0} &\implies Tv = \mathbf{0} \\ &\implies v \in (\ker T) \cap (\ker T)^\perp \\ &\implies v = \mathbf{0} && \text{[by 7.27]} \end{aligned}$$

Hence $\ker T|_{(\ker T)^\perp} = \{\mathbf{0}\}$, so $T|_{(\ker T)^\perp}$ is injective.

Surjectivity Clearly $\operatorname{im} T|_{(\ker T)^\perp} \subset \operatorname{im} T$. To prove the inclusion in the other direction, let $w \in \operatorname{im} T$, so there exists $v \in V$ such that $w = Tv$.

By 7.28, $V = \ker T \oplus (\ker T)^\perp$. Thus $v = u + x$ for some $u \in \ker T$, $x \in (\ker T)^\perp$. Now

$$T|_{(\ker T)^\perp} x = Tx = Tv - Tu = w - \mathbf{0} = w,$$

which shows that $w \in \operatorname{im} T|_{(\ker T)^\perp}$, so $\operatorname{im} T \subset \operatorname{im} T|_{(\ker T)^\perp}$. Hence $\operatorname{im} T|_{(\ker T)^\perp} = \operatorname{im} T$.

□

Now we can define the *pseudoinverse* of a linear map T. In the next definition (and from now on), we can think of $T|_{(\ker T)^\perp}$ as an invertible linear map from $(\ker T)^\perp$ to $\operatorname{im} T$, as is justified by the result above.

Definition 7.36 (Pseudoinverse)**.** Suppose V is finite-dimensional, $T \in \mathcal{L}(V)$. The ***pseudoinverse*** (or *Moore–Penrose inverse*) $T^+ \in \mathcal{L}(W, V)$ of T is defined by

$$T^+ w := \left(T|_{(\ker T)^\perp} \right)^{-1} P_{\operatorname{im} T} w \quad (w \in W).$$

The pseudoinverse behaves much like an inverse, as we will see.

Lemma 7.37 (Properties of pseudoinverse)**.** *Suppose V is finite-dimensional, and $T \in \mathcal{L}(V)$.*

(i) If T is invertible, then $T^+ = T^{-1}$.

(ii) $TT^+ = P_{\operatorname{im} T}$ *(orthogonal projection of* W *onto* $\operatorname{im} T$*).*

(iii) $T^+T = P_{(\ker T)^\perp}$ *(orthogonal projection of* V *onto* $(\ker T)^\perp$*).*

Proof.

(i) Suppose T is invertible. Then injectivity implies $(\ker T)^\perp = V$, and surjectivity implies $\operatorname{im} T = W$.

Thus $T|_{(\ker T)^\perp} = T$, and $P_{\operatorname{im} T} = I_W$. Hence $T^+ = T$.

(ii) Let $w \in \operatorname{im} T$. Thus

$$TT^+w = T\left(T|_{(\ker T)^\perp}\right)^{-1} w = w = P_{\operatorname{im} T} w.$$

Let $w \in (\operatorname{im} T)^\perp$, then $T^+w = \mathbf{0}$. Hence $TT^+w = \mathbf{0} = P_{\operatorname{im} T} w$.

Thus TT^+ and $P_{\operatorname{im} T}$ agree on $\operatorname{im} T$ and on $(\operatorname{im} T)^\perp$. Hence these two linear maps are equal (by 7.28).

(iii) Let $v \in (\ker T)^\perp$. Since $Tv \in \operatorname{im} T$, the definition of T^+ shows that

$$T^+(Tv) = \left(T|_{(\ker T)^\perp}\right)^{-1}(Tv) = v = P_{(\ker T)^\perp} v.$$

Thus T^+T and $P_{(\ker T)^\perp}$ agree on $(\ker T)^\perp$ and on $\ker T$. Hence these two linear maps are equal (by 7.28).

□

For $T \in \mathcal{L}(V, W)$ and $w \in W$, we now return to the problem of finding $v \in V$ that solves the equation

$$Tv = w.$$

As we noted earlier, if T is invertible, then $v = T^{-1}w$ is the unique solution, but if T is not invertible, then T^{-1} is not defined. However, the pseudoinverse T^+ is defined.

In the next result, (i) shows that taking $v = T^+w$ makes Tv as close to w as possible; thus the pseudoinverse provides a *best fit* to the equation above. (ii) shows that among all vectors $v \in V$ that make Tv as close as possible to w, the vector T^+w has the smallest norm.

Theorem 7.38 (Pseudoinverse provides best approximate solution or best solution)**.** *Suppose* V *is finite-dimensional,* $T \in \mathcal{L}(V, W)$*, and* $w \in W$*.*

(i) If $v \in V$*, then*

$$\left\|T\left(T^+w\right) - w\right\| \leq \|Tv - w\|, \tag{7.11}$$

where equality holds if and only if $v \in T^+w + \ker T$*.*

(ii) If $v \in T^+w + \ker T$*, then*

$$\left\|T^+w\right\| \leq \|v\|, \tag{7.12}$$

where equality holds if and only if $v = T^+w$.

Proof.

(i) Let $v \in V$. Then
$$Tv - w = (Tv - TT^+w) + (TT^+w - w).$$
The first term in parentheses above is in $\operatorname{im} T$. Since the operator TT^+ is the orthogonal projection of W onto $\operatorname{im} T$ (by 7.37), the second term in parentheses above is in $(\operatorname{im} T)^\perp$ (by 7.33).

Thus the Pythagorean theorem implies the desired inequality that the norm of the second term in parentheses above is less than or equal to $\|Tv - w\|$, where equality holds if and only if the first term in parentheses above equals 0. Hence equality holds if and only if $v - T^+w + \ker T$, which is equivalent to the statement that $v \in T^+w + \ker T$.

(ii) Let $v \in T^+w + \ker T$. Then $v - T^+w \in \ker T$. Now
$$v = \big(v - T^+w\big) + T^+w.$$
The definition of T^+ implies that $T^+w \in (\ker T)^\perp$. Thus the Pythagorean theorem implies that $\|T^+w\| \leq \|v\|$, where equality holds if and only if $v = T^+w$.

$\square$

Exercises

Exercise 7.1 ([Axl24] 6A Q1). Show that if $v_1, \dots, v_m \in V$, then

$$\sum_{j=1}^{m} \sum_{k=1}^{m} \langle v_j, v_k \rangle \geq 0.$$

Solution.

$$\sum_{j=1}^{m} \left(\sum_{k=1}^{m} \langle v_j, v_k \rangle \right) = \sum_{j=1}^{m} \left\langle v_j, \sum_{k=1}^{m} v_k \right\rangle = \left\langle \sum_{j=1}^{m} v_j, \sum_{k=1}^{m} v_k \right\rangle = \left\| \sum_{k=1}^{m} v_k \right\|^2 \geq 0.$$

□

Exercise 7.2 ([Axl24] 6A Q2). Suppose $S \in \mathcal{L}(V)$. Define $\langle \cdot, \cdot \rangle_1$ by

$$\langle u, v \rangle_1 = \langle Su, Sv \rangle$$

for all $u, v \in V$. Show that $\langle \cdot, \cdot \rangle_1$ is an inner product on V if and only if S is injective.

Solution.

$\boxed{\Longrightarrow}$ Suppose $\langle \cdot, \cdot \rangle_1$ is an inner product on V. Let $u \in \ker S$, then

$$Su = \mathbf{0} \implies \langle Su, Su \rangle = \langle u, u \rangle_1 = 0 \implies u = \mathbf{0}.$$

Hence $\ker S = \{\mathbf{0}\}$, so S is injective.

$\boxed{\Longleftarrow}$ Check conditions for inner product:

(i) $\langle u, u \rangle_1 = \langle Su, Su \rangle \geq 0$, and $\langle u, u \rangle_1 = 0 \iff \langle Su, Su \rangle = 0 \iff Su = \mathbf{0} \iff u = \mathbf{0}$ by injectivity of S.

(ii)

(iii)

□

Exercise 7.3 ([Axl24] 6A Q4, modified). Suppose $T \in \mathcal{L}(V)$ is a *contraction*; that is, $\|Tv\| \leq \|v\|$ for every $v \in V$. Prove that if $|\lambda| > 1$, then $T - \lambda I$ is injective.

Solution. Let $v \in \ker(T - \lambda I)$, then $Tv = \lambda v$, so

$$\begin{aligned}
&\|Tv\| = \|\lambda v\| = |\lambda|\|v\| \\
\implies &|\lambda|\|v\| \leq \|v\| \\
\implies &\underbrace{(|\lambda| - 1)}_{>0}\|v\| \leq 0 \\
\implies &\|v\| \leq 0
\end{aligned}$$

and hence $v = \mathbf{0}$. □

Exercise 7.4 ([Axl24] 6A Q5)**.** Suppose V is a real inner product space.

(i) Show that $\langle u+v, u-v\rangle = \|u\|^2 - \|v\|^2$ for every $u, v \in V$.

(ii) Show that if $u, v \in V$ have the same norm, then $u+v$ is orthogonal to $u-v$.

(iii) Use (ii) to show that the diagonals of a rhombus are perpendicular to each other.

Solution.

(i) We have that

$$\begin{aligned}\langle u+v, u-v\rangle &= \langle u, v\rangle - \langle v, v\rangle - \langle u, v\rangle + \langle v, u\rangle \\ &= \|u\|^2 - \|v\|^2\end{aligned}$$

(ii) We know $\|u\| = \|v\|$, then

$$\langle u+v, u-v\rangle = \|u\|^2 - \|v\|^2 = 0$$

which shows that they are orthogonal.

(iii)

$\square$

Exercise 7.5 ([Axl24] 6A Q6)**.** Suppose $u, v \in V$. Prove that $\langle u, v\rangle = 0 \iff \|u\| \leq \|u+av\|$ for all $a \in \mathbf{F}$.

Solution.

$\boxed{\implies}$ We have

$$\begin{aligned}\|u+av\|^2 &= \langle u+av, u+av\rangle \\ &= \langle u, u\rangle + a\underbrace{\langle v, u\rangle}_{0} + \overline{a}\underbrace{\langle u, v\rangle}_{0} + |a|^2\langle v, v\rangle \\ &= \|u\|^2 + |a|^2\|v\|^2 \geq \|u\|^2.\end{aligned}$$

$\boxed{\impliedby}$ If $v \neq \mathbf{0}$, then it is trivial. Assume $v \neq \mathbf{0}$. Let $a = \frac{\langle u,v\rangle}{\|v\|^2}$. Then we have

$$\begin{aligned}\left\|u - \frac{\langle u, v\rangle}{\|v\|^2}v\right\|^2 &= \left\langle u - \frac{\langle u, v\rangle}{\|v\|^2}v, u - \frac{\langle u, v\rangle}{\|v\|^2}v\right\rangle \\ &= \|u\|^2 - \frac{\overline{\langle u, v\rangle}}{\|v\|^2}\langle u, v\rangle - \frac{\langle u, v\rangle}{\|v\|^2}\langle v, u\rangle + \left|\frac{\langle u, v\rangle}{\|v\|^2}\right|^2\|v\|^2 \\ &= \|u\|^2 - 2\frac{|\langle u, v\rangle|^2}{\|v\|^2} + \frac{|\langle u, v\rangle|^2}{\|v\|^2} \\ &= \|u\|^2 - \frac{|\langle u, v\rangle|^2}{\|v\|^2} \geq \|u\|^2.\end{aligned}$$

$\square$

Exercise 7.6 ([Axl24] 6A Q9)**.** Suppose $u, v \in V$ and $\|u\| = \|v\| = 1$ and $\langle u, v \rangle = 1$. Prove that $u = v$.

Solution. Cauchy–Schwarz inequality. □

Exercise 7.7 ([Axl24] 6A Q14)**.** Suppose $v \in V \setminus \{\mathbf{0}\}$. Prove that $v/\|v\|$ is the unique closest element on the unit sphere of V to v. More precisely, prove that if $u \in V$ and $\|u\| = 1$, then

$$\left\| v - \frac{v}{\|v\|} \right\| \leq \|v - u\|,$$

where equality holds if and only if $u = v/\|v\|$.

Solution. We have

$$\begin{aligned} \left\| v - \frac{v}{\|v\|} \right\| &= \left\| \left(1 - \frac{1}{\|v\|}\right) v \right\| \\ &= \left| 1 - \frac{1}{\|v\|} \right| \|v\| \\ &= \left| \frac{\|v\| - 1}{\|v\|} \right| \|v\| \\ &= |\|v\| - 1| \end{aligned}$$

and

$$\begin{aligned} \|v - u\| &= \langle v - u, v - u \rangle \\ &= \|v\|^2 - \langle v, u \rangle - \langle u, v \rangle + \|u\|^2 \\ &= \|v\|^2 - 2 \operatorname{Re} \langle u, v \rangle + 1 \\ &\geq \|v\|^2 - 2 \left| \langle u, v \rangle \right| + 1 && (*) \\ &\geq \|v\|^2 - 2\|u\|\|v\| + 1 && \text{[by Cauchy–Schwarz inequality]} \quad (**) \\ &= \|v\|^2 - 2\|v\| + 1 = (\|v\| - 1)^2 . \end{aligned}$$

Thus

$$\|v - u\| \geq |\|v\| - 1| = \left\| v - \frac{v}{\|v\|} \right\|.$$

Equality holds if and only if equality in both $(*)$ and $(**)$ hold simultaneously. Equality in $(*)$ holds if and only if

$$\begin{aligned} & \operatorname{Re} \langle u, v \rangle = |\langle u, v \rangle| = \sqrt{\operatorname{Re} \langle u, v \rangle^2 + \operatorname{Im} \langle u, v \rangle^2} \\ \iff & \langle u, v \rangle \in \mathbb{R}_{\geq 0} \\ \iff & k\|v\|^2 \in \mathbb{R}_{\geq 0} \\ \iff & k \in \mathbb{R}_{\geq 0} \end{aligned}$$

and equality in $(**)$ holds if and only if

$$\begin{aligned} & u = kv \\ \iff & |k| = \frac{\|kv\|}{\|v\|} = \frac{\|u\|}{\|v\|} = \frac{1}{\|v\|} \end{aligned}$$

Hence $k = \frac{1}{\|v\|}$, so $u = \frac{v}{\|v\|}$. □

Exercise 7.8 ([Axl24] 6A Q26, polarisation identity)**.** Suppose V is a real inner product space. Prove that

$$\langle u, v \rangle = \frac{\|u+v\|^2 - \|u-v\|^2}{4}.$$

for all $u, v \in V$.

Solution. We have

$$\begin{aligned}
\|u+v\|^2 - \|u-v\|^2 &= \langle u+v, u+v \rangle - \langle u-v, u-v \rangle \\
&= \left(\|u\|^2 + 2\langle u, v \rangle + \|v\|^2\right) - \left(\|u\|^2 - 2\langle u, v \rangle + \|v\|^2\right) \\
&= 4\langle u, v \rangle .
\end{aligned}$$

□

Exercise 7.9 ([Axl24] 6A Q27, polarisation identity)**.** Suppose V is a complex inner product space. Prove that

$$\langle u, v \rangle = \frac{\|u+v\|^2 - \|u-v\|^2 + \|u+iv\|^2 i - \|u-iv\|^2 i}{4}$$

for all $u, v \in V$.

Solution. Similar to previous exercise. □

Exercise 7.10 ([Axl24] 6B Q1)**.** Suppose $\{e_1, \ldots, e_m\}$ is a set of vectors in V such that

$$\|a_1 e_1 + \cdots + a_m e_m\|^2 = |a_1|^2 + \cdots + |a_m|^2$$

for all $a_1, \ldots, a_m \in \mathbf{F}$. Show that $\{e_1, \ldots, e_m\}$ is an orthonormal set of vectors.

Proof. We have

$$\begin{aligned}
\left\|\sum_{i=1}^m a_i e_i\right\|^2 &= \left\langle \sum_{i=1}^m a_i e_i, \sum_{i=1}^m a_i e_i \right\rangle \\
&= \sum_{i=1}^m \sum_{j=1}^m a_i \overline{a_j} \langle e_i, e_j \rangle \\
&= \sum_{i=1}^m |a_i|^2.
\end{aligned}$$

For this holds for arbitrary choices of $a_1, \ldots, a_m \in \mathbf{F}$, we need to have that

$$\langle e_i, e_j \rangle = \delta_{ij},$$

which shows that the vectors are orthogonal to each other. To see that each of them has norm 1, we can set $a_k = 1$ and $a_j = 0$ for all $j \neq k$, which gives that $\|e_k\|^2 = |a_k| = 1$, and thus each of the vectors is normalised, completing the proof. □

Exercise 7.11 ([Axl24] 6B Q3)**.** Suppose $\{e_1, \dots, e_m\}$ is an orthonormal set in V and $v \in V$. Prove that

$$\|v\|^2 = |\langle v, e_1\rangle|^2 + \cdots + |\langle v, e_m\rangle|^2 \iff v \in \operatorname{span}(e_1, \dots, e_m).$$

Solution. $\boxed{\implies}$ We can decompose v into two parts, one is

$$v_{\text{proj}} = \sum_{i=1}^{m} \langle v, e_i\rangle\, e_i,$$

which is the orthogonal projection of v onto the subspace spanned by $e_1, \dots, e_m$. We claim that $v - v_{\text{proj}}$ is orthogonal to v_{proj}. This can be seen as

$$\begin{aligned} \left\langle v_{\text{proj}}, v - v_{\text{proj}}\right\rangle &= \left\langle \sum_{i=1}^{m} \langle v, e_i\rangle\, e_i, v - \sum_{i=1}^{m} \langle v, e_i\rangle\, e_i \right\rangle \\ &= \sum_{i=1}^{m} |\langle v, e_i\rangle|^2 - \sum_{i=1}^{m} |\langle v, e_i\rangle|^2 = 0. \end{aligned}$$

Then by Pythagoras' theorem we have

$$\|v\|^2 = \left\|v_{\text{proj}}\right\|^2 + \left\|v - v_{\text{proj}}\right\|^2$$

where $\|v\|^2 = \left\|v_{\text{proj}}\right\|^2$ and thus $v = v_{\text{proj}}$. Equivalently, $v \in \operatorname{span}(e_1, \dots, e_m)$.

$\boxed{\impliedby}$ This means that $v = \sum_{i=1}^{m} a_i e_i$. However, we know that $a_i = \langle v, e_i\rangle$, so $\|v\|^2 = \sum_{i=1}^{m} |\langle v, e_i\rangle|^2$ by repeatedly applying Pythagoras' theorem. □

Exercise 7.12 ([Axl24] 6B Q6)**.** Suppose $\{e_1, \dots, e_n\}$ is an orthonormal basis of V.

(i) Prove that if $v_1, \dots, v_n$ are vectors in V such that

$$\|e_i - v_i\| < \frac{1}{\sqrt{n}}$$

for each i, then $\{v_1, \dots, v_n\}$ is a basis of V.

(ii) Show that there exist $v_1, \dots, v_n \in V$ such that

$$\|e_i - v_i\| \leq \frac{1}{\sqrt{n}}$$

for each i, but $\{v_1, \dots, v_n\}$ is not linearly independent.

Exercise 7.13 ([Axl24] 6B Q9)**.** Suppose $e_1, \dots, e_m$ is the result of applying the Gram–Schmidt procedure to a linearly independent set $v_1, \dots, v_m$ in V. Prove that $\langle v_i, e_i\rangle > 0$ for each $i = 1, \dots, m$.

Exercise 7.14 ([Axl24] 6B Q10)**.** Suppose $\{v_1, \dots, v_m\}$ is a linearly independent set in V. Explain why the orthonormal set produced by the formulae of the Gram–Schmidt procedure is the only orthonormal set $\{e_1, \dots, e_m\}$ in V such that $\langle v_i, e_i\rangle > 0$ and $\operatorname{span}(v_1, \dots, v_i) = \operatorname{span}(e_1, \dots, e_i)$ for each $i = 1, \dots, m$.

Exercise 7.15 ([Axl24] 6B Q13)**.** Show that a set $v_1, \dots, v_m$ of vectors in V is linearly dependent if and only if the Gram–Schmidt formula produces $u_i = \mathbf{0}$ for some $i \in \{1, \dots, m\}$.

Exercise 7.16 ([Axl24] 6B Q14)**.** Suppose V is a real inner product space and $v_1, \dots, v_m$ is a linearly independent set of vectors in V. Prove that there exist exactly 2^m orthonormal sets $\{e_1, \dots, e_m\}$ of vectors in V such that

$$\operatorname{span}(v_1, \dots, v_i) = \operatorname{span}(e_1, \dots, e_i)$$

for all $i = 1, \dots, m$.

Exercise 7.17 ([Axl24] 6B Q15)**.** Suppose $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$ are inner products on V such that $\langle u, v \rangle_1 = 0$ if and only if $\langle u, v \rangle_2 = 0$. Prove that there exists $c > 0$ such that $\langle u, v \rangle_1 = c \langle u, v \rangle_2$ for every $u, v \in V$.

This exercise shows that if two inner products have the same pairs of orthogonal vectors, then each of the inner products is a scalar multiple of the other inner product.

Exercise 7.18 ([Axl24] 6B Q16)**.** Suppose V is finite-dimensional. Suppose $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$ are inner products on V with corresponding norms $\|\cdot\|_1$ and $\|\cdot\|_2$. Prove that there exists $c > 0$ such that $\|v\|_1 \leq c\|v\|_2$ for every $v \in V$.

Exercise 7.19 ([Axl24] 6B Q17)**.** Suppose V is a complex finite-dimensional vector space. Prove that if T is an operator on V such that 1 is the only eigenvalue of T and $\|Tv\| \leq \|v\|$ for all $v \in V$, then T is the identity operator.

6C 1 4 5 6 7 8 9 10 11 12 14

Exercise 7.20 ([Axl24] 6C Q1)**.**

Exercise 7.21 ([Axl24] 6C Q4)**.**

Exercise 7.22 ([Axl24] 6C Q5)**.**

Exercise 7.23 ([Axl24] 6C Q6)**.**

Exercise 7.24 ([Axl24] 6C Q7)**.**

Exercise 7.25 ([Axl24] 6C Q8)**.** Suppose $U \leq V$ is finite-dimensional, $v \in V$. Define a linear functional $\phi \colon U \to \mathbf{F}$ by

$$\phi(u) = \langle u, v \rangle$$

for all $u \in U$. By the Riesz representation theorem applied to the inner product space U, there exists a unique vector $w \in U$ such that

$$\phi(u) = \langle u, w \rangle$$

for all $u \in U$. Show that $w = P_U v$.

Solution. For each $u \in U$, we have $\langle u, v \rangle = \langle u, w \rangle$. Write

$$\langle u, v \rangle = \langle u, P_U v + (v - P_U v) \rangle .$$

Since $v - P_U v \in U^\perp$, this implies $\langle u, v - P_U v \rangle = 0$. Thus

$$\langle u, v \rangle = \langle u, P_U v + (v - P_U v) \rangle = \langle u, P_U v \rangle .$$

By the uniqueness of w, we must have $w = P_U v$. □

Exercise 7.26 ([Axl24] 6C Q9)**.**

Exercise 7.27 ([Axl24] 6C Q10)**.** Suppose V is finite-dimensional, and $P \in \mathcal{L}(V)$ is such that $P^2 = P$ and

$$\|Pv\| \leq \|v\| \quad (v \in V).$$

Prove that there exists a subspace U of V such that $P = P_U$.

Solution. □

Exercise 7.28 ([Axl24] 6C Q11)**.**

Exercise 7.29 ([Axl24] 6C Q12)**.**

Exercise 7.30 ([Axl24] 6C Q14)**.**

Chapter 8

Operators on Inner Product Spaces

8.1 Self-Adjoint and Normal Operators

8.1.1 Adjoints

Definition 8.1 (Adjoint)**.** The ***adjoint*** of $T \in \mathcal{L}(V, W)$ is the function $T^* \colon W \to V$ such that

$$\langle Tv, w \rangle = \langle v, T^* w \rangle \quad (v \in V,\ w \in W).$$

To see why the definition above makes sense, suppose $T \in \mathcal{L}(V, W)$. Fix $w \in W$. Consider the linear functional on V which maps

$$v \mapsto \langle Tv, w \rangle .$$

(Note that this linear functional depends on T and w.) By the Riesz representation theorem, there exists a unique vector in V such that this linear functional is given by taking the inner product with it. We call this unique vector $T^* w$. In other words, $T^* w$ is the unique vector in V such that

$$\langle Tv, w \rangle = \langle v, T^* w \rangle$$

for every $v \in V$.

Remark. In the equation above, the inner product on the LHS takes place in W, and the inner product on the right takes place in V. However, we use the same notation $\langle \cdot, \cdot \rangle$ for both inner products.

We check that if $T \in \mathcal{L}(V, W)$, then $T^* \in \mathcal{L}(W, V)$; that is, the adjoint of a linear map is a linear map.

(i) Let $v \in V$, $w_1, w_2 \in W$. Then

$$\begin{aligned} \langle Tv, w_1 + w_2 \rangle &= \langle Tv, w_1 \rangle + \langle Tv, w_2 \rangle \\ &= \langle v, T^* w_1 \rangle + \langle v, T^* w_2 \rangle \\ &= \langle v, T^* w_1 + T^* w_2 \rangle . \end{aligned}$$

By the Riesz representation theorem, this implies that $T^*(w_1 + w_2) = T^* w_1 + T^* w_2$.

(ii) Let $v \in V$, $\lambda \in \mathbf{F}$, $w \in W$. Then

$$\begin{aligned}\langle Tv, \lambda w\rangle &= \overline{\lambda}\,\langle Tv, w\rangle \\ &= \overline{\lambda}\,\langle v, T^* w\rangle \\ &= \langle v, \lambda T^* w\rangle .\end{aligned}$$

By the Riesz representation theorem, this implies that $T^*(\lambda w) = \lambda T^* w$.

Remark. To compute T^*, start with a formula for $\langle Tv, w\rangle$, then manipulate it to get *only* v in the first slot; the entry in the second slot will then be $T^* w$.

Lemma 8.2 (Properties of adjoint)**.** *Suppose $T \in \mathcal{L}(V, W)$. Then*

(i) $(S+T)^* = S^* + T^*$ *for all* $S \in \mathcal{L}(V, W)$

(ii) $(\lambda T)^* = \overline{\lambda} T^*$ *for all* $\lambda \in \mathbf{F}$

(iii) $(T^*)^* = T$

(iv) $(ST)^* = T^* S^*$ *for all* $S \in \mathcal{L}(W, U)$*, where* U *is a finite-dimensional inner product space over* $\mathbf{F}$

(v) $I^* = I$*, where* I *is the identity operator on* V

(vi) if T *is invertible, then* T^* *is invertible, and* $(T^*)^{-1} = (T^{-1})^*$

Proof. Let $v \in V$, $w \in V$.

(i) If $S \in \mathcal{L}(V, W)$, then

$$\begin{aligned}\langle (S+T)v, w\rangle &= \langle Sv, w\rangle + \langle Tv, w\rangle \\ &= \langle v, S^* w\rangle + \langle v, T^* w\rangle \\ &= \langle v, (S^* + T^*)w\rangle .\end{aligned}$$

Hence $(S+T)^* w = (S^* + T^*)w$.

(ii) Let $\lambda \in \mathbf{F}$, then

$$\langle (\lambda T)v, w\rangle = \lambda\,\langle Tv, w\rangle = \lambda\,\langle v, T^* w\rangle = \left\langle v, \overline{\lambda} T^* w\right\rangle .$$

Hence $(\lambda T)^* w = \overline{\lambda} T^* w$.

(iii) We have

$$\langle T^* w, v\rangle = \overline{\langle v, T^* w\rangle} = \overline{\langle Tv, w\rangle} = \langle w, Tv\rangle .$$

Hence $(T^*)^* v = Tv$.

(iv) Let $S \in \mathcal{L}(W, U)$, $u \in U$. Then

$$\langle (ST)v, u\rangle = \langle S(Tv), u\rangle = \langle Tv, S^* u\rangle = \langle v, T^*(S^* u)\rangle .$$

Hence $(ST)^* u = T^* S^* u$.

(v) Let $u \in V$. Then
$$\langle Iu, v\rangle = \langle u, v\rangle .$$
Hence $I^*v = v$.

(vi) Suppose T is invertible. Then $T^{-1}T = I$. Taking adjoints of both sides and applying (iv) and (v) gives
$$T^*(T^{-1})^* = I.$$
Similarly, the equation $TT^{-1} = I$ implies
$$(T^{-1})^*T^* = I.$$
Hence $(T^{-1})^*$ is the inverse of T^*.

□

If $\mathbf{F} = \mathbb{R}$, then the map $T \mapsto T^*$ is a linear map from $\mathcal{L}(V, W)$ to $\mathcal{L}(W, V)$, as follows from (i) and (ii). However if $\mathbf{F} = \mathbb{C}$, then this map is not linear due to the complex conjugate in (ii).

The next result shows the relationship between the kernel and image of a linear map and its adjoint.

Lemma 8.3 (Kernel and image of T^*)**.** *Suppose $T \in \mathcal{L}(V, W)$. Then*

(i) $\ker T^* = (\operatorname{im} T)^\perp$

(ii) $\operatorname{im} T^* = (\ker T)^\perp$

(iii) $\ker T = (\operatorname{im} T^*)^\perp$

(iv) $\operatorname{im} T = (\ker T^*)^\perp$

Proof.

(i) Let $w \in W$. Then
$$\begin{aligned} w \in \ker T^* &\iff T^*w = 0 \\ &\iff \langle v, T^*w\rangle = 0 \quad \forall v \in V \\ &\iff \langle Tv, w\rangle = 0 \quad \forall v \in V \\ &\iff w \in (\operatorname{im} T)^\perp. \end{aligned}$$
Hence $\ker T^* = (\operatorname{im} T)^\perp$.

(ii) Replace T with T^* in (iv).

(iii) Replace T with T^* in (i), and use the fact that $(T^*)^* = T$.

(iv) Take the orthogonal complement of both sides of (i), and use the fact that $U = (U^\perp)^\perp$ if $U \leq V$.

□

As we will soon see, the next definition is closely related to the matrix of the adjoint of a linear map.

Definition 8.4 (Conjugate transpose)**.** The ***conjugate transpose*** of a $m \times n$ matrix A is the $n \times m$ matrix A^* obtained by taking the complex conjugate of each entry of A^T.

That is, $(A^*)_{ij} = \overline{a_{ji}}$.

The next result shows how to compute the matrix of T^* from the matrix of T.

Lemma 8.5. *Let $T \in \mathcal{L}(V, W)$. Suppose $\{e_1, \dots, e_n\}$ is an orthonormal basis of V, and $\{f_1, \dots, f_m\}$ is an orthonormal basis of W. Then*

$$\mathcal{M}(T^*) = \mathcal{M}(T)^*.$$

Remark. $\mathcal{M}(T; \{e_1, \dots, e_n\}, \{f_1, \dots, f_m\})$ and $\mathcal{M}(T^*; \{f_1, \dots, f_m\}, \{e_1, \dots, e_n\})$.

Proof. Let $\mathcal{M}(T) = A$, $\mathcal{M}(T^*) = B$.

Since $\{f_1, \dots, f_m\}$ is an orthonormal basis of W, we can write

$$Te_j = \langle Te_j, f_1 \rangle f_1 + \cdots + \langle Te_j, f_m \rangle f_m$$

where $j = 1, \dots, n$. Thus $A_{ij} = \langle Te_j, f_i \rangle$.

Replacing T with T^*, and interchanging $\{e_1, \dots, e_n\}$ and $\{f_1, \dots, f_m\}$ gives

$$B_{ij} = \langle T^* f_j, e_i \rangle = \langle f_j, Te_i \rangle = \overline{\langle Te_i, f_j \rangle} = \overline{A_{ji}}.$$

Hence $\mathcal{M}(T^*) = \mathcal{M}(T)^*$. ◻

8.1.2 Self-Adjoint Operators

Definition 8.6 (Self-adjoint operator)**.** An operator is ***self-adjoint*** if it equals its adjoint.

That is, $T \in \mathcal{L}(V)$ is self-adjoint if $T = T^*$, i.e.,

$$\langle Tv, w\rangle = \langle v, Tw\rangle \quad (v, w \in V).$$

Lemma 8.7. *Every eigenvalue of a self-adjoint operator is real.*

Proof. Suppose T is a self-adjoint operator on V. Let λ be an eigenvalue of T, and let $v \in V \setminus \{\mathbf{0}\}$ be an eigenvector corresponding to λ, i.e., $Tv = \lambda v$. Then

$$\lambda\|v\|^2 = \langle \lambda v, v\rangle = \langle Tv, v\rangle = \langle v, Tv\rangle = \langle v, \lambda v\rangle = \overline{\lambda}\|v\|^2.$$

Since $v \neq \mathbf{0}$, we have $\lambda = \overline{\lambda}$, which means that λ is real. □

Lemma 8.8. *Suppose V is a* complex *inner product space, and $T \in \mathcal{L}(V)$. Then*

$$\langle Tv, v\rangle = 0 \quad \forall v \in V \iff T = 0.$$

Remark. This result does not hold for real inner product spaces. For instance, the operator $T \in \mathcal{L}(\mathbb{R}^2)$ that is a counterclockwise rotation of 90° around the origin; thus $T(x, y) = (-y, x)$. Notice that Tv is orthogonal to v for every $v \in \mathbb{R}^2$, even though $T \neq 0$.

Proof.

$\boxed{\Longleftarrow}$ Suppose $T = 0$. If $u, w \in V$, then

$$\begin{aligned}\langle Tu, w\rangle &= \frac{\langle T(u+w), u+w\rangle - \langle T(u-w), u-w\rangle}{4} \\ &\quad + \frac{\langle T(u+iw), u+iw\rangle - \langle T(u-iw), u-iw\rangle}{4} i.\end{aligned}$$

Note that each term on the RHS is of the form $\langle Tv, v\rangle$ for appropriate $v \in V$.

$\boxed{\Longrightarrow}$ Suppose $\langle Tv, v\rangle = 0$ for every $v \in V$.

Then the equation above implies that $\langle Tu, w\rangle = 0$ for all $u, w \in V$. Taking $w = Tu$ for every $u \in V$, we obtain $Tu = \mathbf{0}$ for every $u \in V$. Hence $T = 0$ as desired. □

Lemma 8.9. *Suppose V is a* complex *inner product space, and $T \in \mathcal{L}(V)$. Then*

$$T \text{ is self-adjoint} \iff \langle Tv, v\rangle \in \mathbb{R} \quad \forall v \in V.$$

Remark. This result does not hold for real inner product spaces, by considering any operator on a real inner product space that is not self-adjoint.

Proof. If $v \in V$, then

$$\langle T^*v, v\rangle = \overline{\langle v, T^*v\rangle} = \overline{\langle Tv, v\rangle}. \tag{I}$$

where the second equality follows since T is self-adjoint. Thus

$$\begin{aligned}
T \text{ is self-adjoint} &\iff T = T^* \\
&\iff T - T^* = 0 \\
&\iff \langle (T - T^*)v, v\rangle = 0 \quad \forall v \in V && \text{[by 8.8]} \\
&\iff \langle Tv, v\rangle = \langle T^*v, v\rangle \quad \forall v \in V \\
&\iff \langle Tv, v\rangle = \overline{\langle Tv, v\rangle} \quad \forall v \in V && \text{[by (I)]} \\
&\iff \langle Tv, v\rangle \in \mathbb{R} \quad \forall v \in V.
\end{aligned}$$

□

On a real inner product space V, a non-zero operator T might satisfy $\langle Tv, v\rangle = 0$ for all $v \in V$. However, the next result shows that this cannot happen for a self-adjoint operator.

Lemma 8.10. *Suppose T is a self-adjoint operator on V. Then*

$$\langle Tv, v\rangle = 0 \quad \forall v \in V \iff T = 0.$$

Proof. We have already proved this (without the hypothesis that T is self-adjoint) when V is a complex inner product space (see 8.8). Thus we can assume that V is a real inner product space.

$\boxed{\Longleftarrow}$ Let $u, v \in V$, then

$$\langle Tu, w\rangle = \frac{\langle T(u + w), u + w\rangle - \langle T(u - w), u - w\rangle}{4} \tag{I}$$

as can be proved by computing the RHS using $\langle Tw, u\rangle = \langle w, Tu\rangle = \langle Tu, w\rangle$, where the first equality holds because T is self-adjoint, and the second equality holds because we are working in a real inner product space.

$\boxed{\Longrightarrow}$ Suppose $\langle Tv, v\rangle = 0$ for every $v \in V$.

Since each term on the RHS of (I) is of the form $\langle Tv, v\rangle$ for appropriate v, this implies that $\langle Tu, w\rangle = 0$ for all $u, w \in V$. Thus taking $w = Tu$, we obtain $Tu = \mathbf{0}$ for every $u \in V$. Hence $T = 0$, as desired. □

8.1.3 Normal Operators

Definition 8.11 (Normal operator)**.** An operator is ***normal*** if it commutes with its adjoint.

That is, $T \in \mathcal{L}(V)$ is normal if $TT^* = T^*T$.

Remark. Every self-adjoint operator is normal, but not vice versa.

The next result provides a useful characterisation of normal operators.

Lemma 8.12 (Characterisation of normal operators)**.** *Suppose $T \in \mathcal{L}(V)$. Then*

$$T \text{ is normal} \iff \|Tv\| = \|T^*v\| \quad \forall v \in V.$$

Proof. We have

$$\begin{aligned} T \text{ is normal} &\iff TT^* = T^*T \\ &\iff T^*T - TT^* = 0 \\ &\iff \langle (T^*T - TT^*)v, v\rangle = 0 \quad \forall v \in V \\ &\iff \langle T^*Tv, v\rangle = \langle TT^*v = v\rangle \quad \forall v \in V \\ &\iff \langle Tv, Tv\rangle = \langle T^*v, T^*v\rangle \quad \forall v \in V \\ &\iff \|Tv\| = \|T^*v\| \quad \forall v \in V. \end{aligned}$$

□

The next result presents several consequences of the result above.

Lemma 8.13. *Suppose $T \in \mathcal{L}(V)$ is normal. Then*

(i) $\ker T = \ker T^*$

(ii) $\operatorname{im} T = \operatorname{im} T^*$

(iii) $V = \ker T \oplus \operatorname{im} T$

(iv) $T - \lambda I$ *is normal for every* $\lambda \in \mathbf{F}$

(v) if $v \in V$ and $\lambda \in \mathbf{F}$, then $Tv = \lambda v \iff T^*v = \overline{\lambda}v$

Proof.

(i) Let $v \in V$. Then

$$\begin{aligned} v \in \ker T &\iff Tv = \mathbf{0} \\ &\iff \|Tv\| = 0 \\ &\iff \|T^*v\| = 0 && \text{[by 8.12]} \\ &\iff T^*v = \mathbf{0} \\ &\iff v \in \ker T^* \end{aligned}$$

(ii) We have

$$\begin{aligned}
\operatorname{im} T &= (\ker T^*)^\perp && \text{[by 8.3]} \\
&= (\ker T)^\perp && \text{[by (i)]} \\
&= \operatorname{im} T^* && \text{[by 8.3]}
\end{aligned}$$

(iii) We have

$$\begin{aligned}
V &= (\ker T) \oplus (\ker T)^\perp \\
&= \ker T \oplus \operatorname{im} T^* && \text{[by 8.3]} \\
&= \ker T \oplus \operatorname{im} T && \text{[by (ii)]}
\end{aligned}$$

(iv) Let $\lambda \in \mathbf{F}$, then

$$\begin{aligned}
(T - \lambda I)(T - \lambda I)^* &= (T - \lambda I)(T^* - \overline{\lambda} I) \\
&= TT^* - \overline{\lambda} T - \lambda T^* + |\lambda|^2 I \\
&= T^*T - \overline{\lambda} T - \lambda T^* + |\lambda|^2 I \\
&= (T^* - \overline{\lambda} I)(T - \lambda I) \\
&= (T - \lambda I)^*(T - \lambda I).
\end{aligned}$$

Thus $T - \lambda I$ commutes with its adjoint. Hence $T - \lambda I$ is normal.

(v) Let $v \in V$, $\lambda \in \mathbf{F}$. Then (iv) and 8.12 imply that

$$\|(T - \lambda I)v\| = \|(T - \lambda I)^* v\| = \left\|(T^* - \overline{\lambda} I)v\right\|.$$

Thus $\|(T - \lambda I)v\| = 0$ if and only if $\left\|(T^* - \overline{\lambda} I)v\right\| = 0$. Hence $Tv = \lambda v$ if and only if $T^*v = \overline{\lambda} v$.

□

Proposition 8.14. *Suppose $T \in \mathcal{L}(V)$ is normal. Then the eigenvectors of T corresponding to distinct eigenvalues are orthogonal.*

Proof. Let α, β be distinct eigenvalues of T, with corresponding eigenvectors u, v. Thus $Tu = \lambda u$ and $Tv = \beta v$. By 8.13, $Tv = \beta v \iff T^*v = \overline{\beta} v$. Thus

$$\begin{aligned}
(\alpha - \beta) \langle u, v \rangle &= \langle \alpha u, v \rangle - \langle \beta u, v \rangle \\
&= \langle \alpha u, v \rangle - \langle u, \overline{\beta} v \rangle \\
&= \langle Tu, v \rangle - \langle u, T^* v \rangle \\
&= 0.
\end{aligned}$$

Since $\alpha \neq \beta$, the equation above implies that $\langle u, v \rangle = 0$. Hence u and v are orthogonal, as desired. □

Proposition 8.15. *Suppose* $\mathbf{F} = \mathbb{C}$ *and* $T \in \mathcal{L}(V)$*. Then* T *is normal if and only if there exist commuting self-adjoint operators* A *and* B *such that* $T = A + iB$*.*

Proof.

$\boxed{\Longrightarrow}$ Suppose T is normal.

Claim. $A = \dfrac{T + T^*}{2}$, $B = \dfrac{T - T^*}{2}$.

Then A and B are self-adjoint, and $T = A + iB$. A quick computation shows that

$$AB - BA = \frac{T^*T - TT^*}{2i}.$$

Since T is normal, the RHS equals 0. Thus $AB = BA$, so A and B commute.

$\boxed{\Longleftarrow}$ Suppose there exist commuting self-adjoint operators A and B such that

$$T = A + iB.$$

Then

$$T^* = A - iB.$$

Solving for A and B gives

$$A = \frac{T + T^*}{2}, \quad B = \frac{T - T^*}{2}.$$

This implies that

$$AB - BA = \frac{T^*T - TT^*}{2i}.$$

Since A and B commute, the LHS equals 0, so T is normal, as desired. □

8.2 Spectral Theorem

The *spectral theorem* characterises the self-adjoint operators when $\mathbf{F} = \mathbb{R}$, and the normal operators when $\mathbf{F} = \mathbb{C}$.

8.2.1 Real Spectral Theorem

To prove the real spectral theorem, we will need two preliminary results. These preliminary results hold on both real and complex inner product spaces, but they are not needed for the proof of the complex spectral theorem.

Lemma 8.16 (Invertible quadratic expressions)**.** *Suppose $T \in \mathcal{L}(V)$ is self-adjoint and $b, c \in \mathbb{R}$ are such that $b^2 < 4c$. Then*

$$T^2 + bT + cI$$

is an invertible operator.

Proof. It suffices to prove that $T^2 + bT + cI$ is injective.

Let $v \in V$ be a non-zero vector. Then

$$\begin{aligned}
&\left\langle (T^2 + bT + cI)v, v \right\rangle \\
&= \left\langle T^2 v, v \right\rangle + b \langle Tv, v \rangle + c \langle v, v \rangle \\
&= \langle Tv, Tv \rangle + b \langle Tv, v \rangle + c\|v\|^2 \\
&\geq \|Tv\|^2 - |b|\|Tv\|\|v\| + c\|v\|^2 && \text{[by Cauchy–Schwarz inequality]} \\
&= \left(\|Tv\| - \frac{|b|\|v\|}{2} \right)^2 + \left(c - \frac{b^2}{4} \right) \|v\|^2 > 0 && \text{[completing the square]}
\end{aligned}$$

This implies that $(T^2 + bT + cI)v \neq \mathbf{0}$ for all $v \neq \mathbf{0}$. Thus $\ker(T^2 + bT + cI) = \{\mathbf{0}\}$, so $T^2 + bT + cI$ is injective. □

Lemma 8.17 (Minimal polynomial of self-adjoint operator)**.** *Suppose $T \in \mathcal{L}(V)$ is self-adjoint. Then the minimal polynomial of T equals*

$$(z - \lambda_1) \cdots (z - \lambda_m)$$

for some $\lambda_1, \dots, \lambda_m \in \mathbb{R}$.

Proof. First suppose $\mathbf{F} = \mathbb{C}$. By 6.16, the zeros of the minimal polynomial of T are the eigenvalues of T. By 8.7, all eigenvalues of T are real. Thus the second version of the fundamental theorem of algebra (5.11) tells us that the minimal polynomial of T has the desired form.

Now suppose $\mathbf{F} = \mathbb{R}$. By the factorisation of a polynomial over $\mathbb{R}$ (5.14), there exist $\lambda_1, \dots, \lambda_m \in \mathbb{R}$ and $b_1, \dots, b_N, c_1, \dots, c_N \in \mathbb{R}$ with ${b_i}^2 < 4c_i$ for each i such that the minimal polynomial of T equals

$$(z - \lambda_1) \cdots (z - \lambda_m)(z^2 + b_1 z + c_1) \cdots (z^2 + b_N z + c_N); \tag{I}$$

here either m or N might equal 0, meaning that there are no terms of the corresponding form. Now

$$(T-\lambda_1 I)\cdots(T-\lambda_m I)(T^2+b_1T+c_1I)\cdots(T^2+b_NT+c_NI)=0.$$

If $N>0$, then we could multiply both sides of the equation above on the right by the inverse of $T^2+b_NT+c_NI$ (which is an invertible operator, by 8.16) to obtain a polynomial expression of T that equals 0. The corresponding polynomial would have degree two less than the degree of (I), violating the minimality of the degree of the polynomial with this property. Thus we must have $N=0$, which means that the minimal polynomial in (I) has the form $(z-\lambda_1)\cdots(z-\lambda_m)$, as desired. □

The next result, which gives a complete description of the self-adjoint operators on a real inner product space, is one of the major theorems in linear algebra.

Theorem 8.18 (Real spectral theorem)**.** *Suppose* $\mathbf{F}=\mathbb{R}$ *and* $T\in\mathcal{L}(V)$*. Then the following are equivalent:*

(i) T *is self-adjoint.*

(ii) T *has a diagonal matrix with respect to some orthonormal basis of* V*.*

(iii) V *has an orthonormal basis consisting of eigenvectors of* T*.*

Proof.

$\boxed{\text{(i)} \implies \text{(ii)}}$ Suppose T is self-adjoint. Our results on minimal polynomials, specifically 7.22 and 8.17, imply that T has an upper-triangular matrix $\mathcal{M}(T)$ with respect to some orthonormal basis of V. With respect to this orthonormal basis, $\mathcal{M}(T^*)=\mathcal{M}(T)^T$.

However, $T^*=T$. Thus $\mathcal{M}(T)^T=\mathcal{M}(T)$. Since $\mathcal{M}(T)$ is upper-triangular, this means that all entries of the matrix above and below the diagonal are 0. Hence the matrix of T is a diagonal matrix with respect to the orthonormal basis.

$\boxed{\text{(ii)} \implies \text{(i)}}$ Suppose T has a diagonal matrix $\mathcal{M}(T)$ with respect to some orthonormal basis of V.

That diagonal matrix equals its transpose. Thus with respect to that basis, $\mathcal{M}(T^*)=\mathcal{M}(T)$. Hence $T^*=T$, so T is self-adjoint.

$\boxed{\text{(ii)} \iff \text{(iii)}}$ This follows from the definitions. □

8.2.2 Complex Spectral Theorem

The next result gives a complete description of the normal operators on a complex inner product space.

Theorem 8.19 (Complex spectral theorem)**.** *Suppose* $\mathbf{F} = \mathbb{C}$ *and* $T \in \mathcal{L}(V)$*. Then the following are equivalent:*

(i) T *is normal.*

(ii) T *has a diagonal matrix with respect to some orthonormal basis of* V*.*

(iii) V *has an orthonormal basis consisting of eigenvectors of* T*.*

Proof.

$\boxed{\text{(i)} \implies \text{(ii)}}$ Suppose T is normal.

By Schur's theorem, there exists an orthonormal basis $\{e_1, \dots, e_n\}$ of V, with respect to which T has an upper-triangular matrix:

$$\mathcal{M}(T; \{e_1, \dots, e_n\}) = \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ & \ddots & \vdots \\ 0 & & A_{nn} \end{pmatrix}.$$

Claim. $\mathcal{M}(T)$ is a diagonal matrix.

We see that

$$\begin{aligned} \|Te_1\|^2 &= |A_{11}|^2 \\ \|T^*e_1\|^2 &= |A_{11}|^2 + |A_{12}|^2 + \cdots + |A_{1n}|^2. \end{aligned}$$

Since T is normal, $\|Te_1\| = \|T^*e_1\|$ (by 8.12). Thus the two equations above imply that all entries in the first row of $\mathcal{M}(T)$, except possibly the first entry A_{11}, equal 0.

Now since $A_{12} = 0$, we have

$$\begin{aligned} \|Te_2\|^2 &= |A_{22}|^2 \\ \|T^*e_2\|^2 &= |A_{22}|^2 + |A_{23}|^2 + \cdots + |A_{2n}|^2. \end{aligned}$$

Since T is normal, $\|Te_2\| = \|T^*e_2\|$. Thus the two equations above imply that all entries in the second row of $\mathcal{M}(T)$, except possibly the diagonal entry A_{22}, equal 0.

Continuing in this fashion, we see that all non-diagonal entries in $\mathcal{M}(T)$ equal 0. Thus $\mathcal{M}(T)$ is a diagonal matrix.

$\boxed{\text{(ii)} \implies \text{(i)}}$ Suppose T has a diagonal matrix with respect to some orthonormal basis of V. The matrix of T^* (with respect to the same basis) is obtained by taking the conjugate transpose of the matrix of T; hence T^* also has a diagonal matrix. Any two diagonal matrices commute; thus T commutes with T^*, which means that T is normal.

$\boxed{\text{(ii)} \iff \text{(iii)}}$ This follows from the definitions. ∎

8.3 Positive Operators

Definition 8.20 (Positive operator)**.** An operator $T \in \mathcal{L}(V)$ is called ***positive*** if

(i) T is self-adjoint, and

(ii) $\langle Tv, v\rangle \geq 0$ for all $v \in V$.

Definition 8.21 (Squared root)**.** An operator R is called a ***square root*** of an operator T if $R^2 = T$.

Lemma 8.22 (Characterisation of positive operators)**.** *Let $T \in \mathcal{L}(V)$. Then the following are equivalent:*

(i) T is a positive operator.

(ii) T is self-adjoint and all eigenvalues of T are non-negative.

(iii) With respect to some orthonormal basis of V, the matrix of T is a diagonal matrix with only non-negative numbers on the diagonal.

(iv) T has a positive square root.

(v) T has a self-adjoint square root.

(vi) $T = R^*R$ for some $R \in \mathcal{L}(V)$.

Proof.

$\boxed{\text{(i)} \implies \text{(ii)}}$ Suppose T is positive. Then T is self-adjoint.

Suppose λ is an eigenvalue of T, with corresponding eigenvector v. Then

$$0 \leq \langle Tv, v\rangle = \langle \lambda v, v\rangle = \lambda \langle v, v\rangle .$$

Thus λ is a non-negative number.

$\boxed{\text{(ii)} \implies \text{(iii)}}$ Suppose T is self-adjoint, and all eigenvalues of T are non-negative.

By the spectral theorem (8.18 and 8.19), there exists an orthonormal basis $\{e_1, \dots, e_n\}$ of V consisting of eigenvectors of T. Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of T corresponding to $e_1, \dots, e_n$; thus each λ_i is a non-negative number.

The matrix of T with respect to $\{e_1, \dots, e_n\}$ is the diagonal matrix with $\lambda_1, \dots, \lambda_n$ on the diagonal.

$\boxed{\text{(iii)} \implies \text{(iv)}}$ Suppose $\{e_1, \dots, e_n\}$ is an orthonormal basis of V such that the matrix of T with respect to this basis is a diagonal matrix, with non-negative $\lambda_1, \dots, \lambda_n$ on the diagonal.

By the linear map lemma, there exists $R \in \mathcal{L}(V)$ such that

$$Re_i = \sqrt{\lambda_i} e_i \quad (i = 1, \dots, n).$$

Claim. R is a positive square root of T.

As you should verify, R is a positive operator. Furthermore,

$$R^2 e_i = \lambda_i e_i = T e_i$$

for each i, which implies that $R^2 = T$. Thus R is a positive square root of T.

$\boxed{\text{(iv)} \implies \text{(v)}}$ Every positive operator is self-adjoint (by definition of positive operator).

$\boxed{\text{(v)} \implies \text{(vi)}}$ Suppose T has a self-adjoint square root. Then there exists a self-adjoint operator R on V such that $T = R^2$. Then $T = R^* R$ (since $R^* = R$).

$\boxed{\text{(vi)} \implies \text{(i)}}$ Let $R \in \mathcal{L}(V)$ be such that $T = R^* R$. Then

$$T^* = (R^* R)^* = R^* (R^*)^* = R^* R = T.$$

Hence T is self-adjoint. Now for every $v \in V$,

$$\langle Tv, v \rangle = \langle R^* Rv, v \rangle = \langle Rv, Rv \rangle \geq 0.$$

Thus T is positive. □

Every non-negative number has a unique non-negative square root. The next result shows that positive operators enjoy a similar property.

Lemma 8.23. *Every positive operator on V has a unique positive square root.*

Proof. Suppose $T \in \mathcal{L}(V)$ is positive. Suppose $v \in V$ is an eigenvector of T. Hence there exists a real number $\lambda \geq 0$ such that $Tv = \lambda v$.

Let R be a positive square root of T. We will prove that $Rv = \sqrt{\lambda} v$. This will imply that the behaviour of R on the eigenvectors of T is uniquely determined. Since there is a basis of V consisting of eigenvectors of T (by the spectral theorem), this will imply that R is uniquely determined.

To prove that $Rv = \sqrt{\lambda} v$, note that the spectral theorem asserts that there is an orthonormal basis $\{e_1, \dots, e_n\}$ of V consisting of eigenvectors of R. Since R is a positive operator, all its eigenvalues are non-negative. Thus there exist non-negative numbers $\lambda_1, \dots, \lambda_n$ such that $Re_i = \sqrt{\lambda_i} e_i$ for each $i = 1, \dots, n$.

Since $\{e_1, \dots, e_n\}$ is a basis of V, we can write

$$v = a_1 e_1 + \cdots + a_n e_n$$

for some $a_1, \dots, a_n \in \mathbf{F}$. Thus

$$Rv = a_1 \sqrt{\lambda_1} e_1 + \cdots + a_n \sqrt{\lambda_n} e_n.$$

Hence

$$\lambda v = Tv = R^2 v = a_1 \lambda_1 e_1 + \cdots + a_n \lambda_n e_n.$$

The equation above implies that

$$a_1\lambda e_1 + \cdots + a_n\lambda e_n = a_1\lambda_1 e_1 + \cdots + a_n\lambda_n e_n.$$

Thus $a_i(\lambda - \lambda_i) = 0$ for each $i = 1, \ldots, n$. Hence

$$v = \sum_{\{i|\lambda_i=\lambda\}} a_i e_i.$$

Thus

$$Rv = \sum_{\{i|\lambda_i=\lambda\}} a_i\sqrt{\lambda}e_i = \sqrt{\lambda}v$$

as desired. □

Notation. For a positive operator T, let $\sqrt{T}$ denotes the unique positive square root of T.

Corollary 8.24. *Suppose T is a positive operator on V, and $v \in V$ is such that $\langle Tv, v\rangle = 0$. Then $Tv = \mathbf{0}$.*

Proof. We have

$$0 = \langle Tv, v\rangle = \left\langle \sqrt{T}\sqrt{T}v, v\right\rangle = \left\langle \sqrt{T}v, \sqrt{T}v\right\rangle = \left\|\sqrt{T}v\right\|^2.$$

Hence $\sqrt{T}v = \mathbf{0}$. Thus $Tv = \sqrt{T}\left(\sqrt{T}v\right) = \mathbf{0}$, as desired. □

8.4 Isometries, Unitary Operators, and Matrix Factorisation

8.4.1 Isometries

Linear maps that preserve norms are sufficiently important to deserve a name.

Definition 8.25 (Isometry)**.** We call $S \in \mathcal{L}(V, W)$ an ***isometry*** if

$$\|Sv\| = \|v\| \quad (v \in V).$$

That is, an isometry preserves norms.

Lemma 8.26. *Every isometry is injective.*

Proof. Let $S \in \mathcal{L}(V, W)$ be an isometry. Then

$$v \in \ker S \iff Sv = \mathbf{0} \iff \|Sv\| = 0 \iff \|v\| = 0 \iff v = \mathbf{0}.$$

Thus $\ker S = \{\mathbf{0}\}$. □

Lemma 8.27 (Characterisation of isometries)**.** *Suppose $S \in \mathcal{L}(V, W)$. Suppose $\{e_1, \ldots, e_n\}$ is an orthonormal basis of V, and $\{f_1, \ldots, f_m\}$ is an orthonormal basis of W. Then the following are equivalent:*

(i) *S is an isometry.*

(ii) $S^*S = I$.

(iii) $\langle Su, Sv \rangle = \langle u, v \rangle$ *for all* $u, v \in V$.

(iv) $\{Se_1, \ldots, Se_n\}$ *is an orthonormal set in* W.

(v) *The columns of* $\mathcal{M}(S; \{e_1, \ldots, e_n\}, \{f_1, \ldots, f_m\})$ *form an orthonormal set in* $\mathbf{F}^m$ *with respect to the Euclidean inner product.*

Proof.

$\boxed{\text{(i)} \implies \text{(ii)}}$ Suppose S is an isometry. Let $v \in V$, then

$$\begin{aligned}\langle (I - S^*S)v, v \rangle &= \langle v, v \rangle - \langle S^*Sv, v \rangle \\ &= \|v\|^2 - \langle Sv, Sv \rangle \\ &= \|v\|^2 - \|Sv\|^2 = 0.\end{aligned}$$

Hence the self-adjoint operator $I - S^*S$ equals 0 (by 7.16). Thus $S^*S = I$.

$\boxed{\text{(ii)} \implies \text{(iii)}}$ Suppose $S^*S = I$. Let $u, v \in V$. Then

$$\langle Su, Sv \rangle = \langle S^*Su, u \rangle = \langle Iu, v \rangle = \langle u, v \rangle.$$

$\boxed{\text{(iii)} \implies \text{(iv)}}$ Suppose $\langle Su, Sv\rangle = \langle u, v\rangle$ for all $u, v \in V$.

If $i, j \in \{1, \dots, n\}$, then

$$\langle Se_i, Se_j\rangle = \langle e_i, e_j\rangle .$$

Hence $\{Se_1, \dots, Se_n\}$ is an orthonormal set in W.

$\boxed{\text{(iv)} \implies \text{(v)}}$ Suppose $\{Se_1, \dots, Se_n\}$ is an orthonormal set in W.

Let $A = \mathcal{M}(S; \{e_1, \dots, e_n\}, \{f_1, \dots, f_m\})$. If $j, k \in \{1, \dots, n\}$, the j-th and k-th columns of A are

$$\begin{pmatrix} A_{1j} \\ \vdots \\ A_{mj} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} A_{1k} \\ \vdots \\ A_{mk} \end{pmatrix}$$

so the Euclidean inner product in $\mathbf{F}^m$ of the two columns is

$$\begin{aligned} \sum_{i=1}^{m} A_{ij}\overline{A_{ik}} &= \left\langle \sum_{i=1}^{m} A_{ij} f_i, \sum_{i=1}^{m} A_{ik} f_i \right\rangle \\ &= \langle Se_j, Se_k\rangle \\ &= \begin{cases} 1 & (j = k) \\ 0 & (j \neq k) \end{cases} \end{aligned} \tag{I}$$

Thus the columns of A form an orthonormal set in $\mathbf{F}^m$.

$\boxed{\text{(v)} \implies \text{(i)}}$ Suppose the columns of the matrix A defined above form an orthonormal set in $\mathbf{F}^m$. Then (I) shows that $\{Se_1, \dots, Se_n\}$ is an orthonormal set in W.

Let $v \in V$. Then

$$v = \langle v, e_1\rangle e_1 + \cdots + \langle v, e_n\rangle e_n$$

and

$$\|v\|^2 = |\langle v, e_1\rangle|^2 + \cdots + |\langle v, e_n\rangle|^2 .$$

Applying S gives

$$Sv = \langle v, e_1\rangle Se_1 + \cdots + \langle v, e_n\rangle Se_n$$

so

$$\|Sv\|^2 = |\langle v, e_1\rangle|^2 + \cdots + |\langle v, e_n\rangle|^2 .$$

Thus $\|Sv\| = \|v\|$, so S is an isometry. $\square$

8.4.2 Unitary Operators

Definition 8.28 (Unitary operator)**.** An operator $S \in \mathcal{L}(V)$ is called ***unitary*** if S is an invertible isometry.

Lemma 8.29 (Characterisation of unitary operators)**.** *Suppose $S \in \mathcal{L}(V)$, $\{e_1, \dots, e_n\}$ is an orthonormal basis of V. Then the following are equivalent:*

(i) S is a unitary operator.

(ii) $S^*S = SS^* = I$.

(iii) S is invertible, and $S^{-1} = S^*$.

(iv) $\{Se_1, \dots, Se_n\}$ is an orthonormal basis of V.

(v) The rows of $\mathcal{M}(S; \{e_1, \dots, e_n\})$ form an orthonormal basis of $\mathbf{F}^n$ with respect to the Euclidean inner product.

(vi) S^* is a unitary operator.

Proof.

$\boxed{\text{(i)} \implies \text{(ii)}}$ Suppose S is a unitary operator. Then S is an isometry, so $S^*S = I$ (by 8.27). Multiplying both sides by S^{-1} on the right yields $S^* = S^{-1}$. Thus $SS^* = SS^{-1} = I$ as desired.

$\boxed{\text{(ii)} \implies \text{(iii)}}$ This follows from the definitions of invertibility and inverse.

$\boxed{\text{(iii)} \implies \text{(iv)}}$ Suppose S is invertible and $S^{-1} = S^*$. Thus $S^*S = I$. By 8.27, $\{Se_1, \dots, Se_n\}$ is an orthonormal set in V.

Since this set has length $\dim V$, we conclude that $\{Se_1, \dots, Se_n\}$ is an orthonormal basis of V.

$\boxed{\text{(iv)} \implies \text{(v)}}$ Suppose $\{Se_1, \dots, Se_n\}$ is an orthonormal basis of V. By 8.27, S is an isometry.

$\boxed{\text{(v)} \implies \text{(vi)}}$ Suppose the rows of $\mathcal{M}(S; \{e_1, \dots, e_n\})$ form an orthonormal basis of $\mathbf{F}^n$. Thus the columns of $\mathcal{M}(S^*; \{e_1, \dots, e_n\})$ form an orthonormal basis of $\mathbf{F}^n$. By 8.27, S^* is an isometry.

$\boxed{\text{(vi)} \implies \text{(i)}}$ Suppose S^* is a unitary operator. □

Lemma 8.30. *Suppose λ is an eigenvalue of a unitary operator. Then $|\lambda| = 1$.*

Proof. Suppose $S \in \mathcal{L}(V)$ is a unitary operator. Let λ be an eigenvalue of S, with corresponding eigenvector v. Then

$$|\lambda|\|v\| = \|\lambda v\| = \|Sv\| = \|v\|.$$

Thus $|\lambda| = 1$, as desired. □

The next result characterises unitary operators on finite-dimensional complex inner product spaces, using the complex spectral theorem as the main tool.

Lemma 8.31. *Suppose* $\mathbf{F} = \mathbb{C}$ *and* $S \in \mathcal{L}(V)$*. Then* S *is a unitary operator if and only if there exists an orthonormal basis of* V *consisting of eigenvectors of* S *whose corresponding eigenvalues all have absolute value* 1*.*

Proof.

$\boxed{\Longrightarrow}$ Suppose S is a unitary operator. By 8.29, $S^*S = SS^* = I$. Since S commutes with its adjoint, S is normal.

By the complex spectral theorem (8.19), there exists an orthonormal basis $\{e_1, \ldots, e_n\}$ of V consisting of eigenvectors of S. By 8.30, every eigenvalue has absolute value 1.

$\boxed{\Longleftarrow}$ Let $\{e_1, \ldots, e_n\}$ be an orthonormal basis of V consisting of eigenvectors of S, whose corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$ all have absolute value 1. Then $\{Se_1, \ldots, Se_n\}$ is an orthonormal basis of V, because

$$\begin{aligned}\langle Se_i, Se_j \rangle &= \langle \lambda_i e_i, \lambda_j e_j \rangle \\ &= \lambda_i \overline{\lambda_j} \langle e_i, e_j \rangle \\ &= \begin{cases} 0 & (i \neq j) \\ 1 & (i = j) \end{cases}\end{aligned}$$

for all $i, j \in \{1, \ldots, n\}$. By 8.29, S is a unitary operator. $\square$

8.4.3 QR Factorisation

We begin by making the following definition, transferring the notion of a unitary operator to a unitary matrix.

Definition 8.32 (Unitary matrix)**.** An $n \times n$ matrix is called ***unitary*** if its columns form an orthonormal set in $\mathbf{F}^n$.

Lemma 8.33 (Characterisation of unitary matrices)**.** *Suppose Q is an $n \times n$ matrix. Then the following are equivalent:*

(i) Q is a unitary matrix.

(ii) The rows of Q form an orthonormal set in $\mathbf{F}^n$.

(iii) $\|Qv\| = \|v\|$ for every $v \in \mathbf{F}^n$.

(iv) $Q^*Q = QQ^* = I_n$.

Proof. □

Theorem 8.34 (QR factorisation)**.** *Suppose A is a square matrix with linearly independent columns. Then there exist unique unitary matrix Q, upper-triangular matrix R with only positive numbers on its diagonal, such that*

$$A = QR. \tag{8.1}$$

Proof.

$\boxed{\text{Existence}}$ Let $v_1, \dots, v_n$ denote the columns of A (consider these as elements of $\mathbf{F}^n$). Apply the Gram–Schmidt procedure to the linearly independent set $\{v_1, \dots, v_n\}$ to obtain an orthonormal basis of $\{e_1, \dots, e_n\}$ of $\mathbf{F}^n$ such that

$$\operatorname{span}(v_1, \dots, v_i) = \operatorname{span}(e_1, \dots, e_i)$$

for each $i = 1, \dots, n$.

Claim. Let Q be the matrix whose columns are $e_1, \dots, e_n$.

Then Q is unitary.

Claim. Let R be the $n \times n$ matrix defined by $R_{ij} = \langle v_j, e_i \rangle$.

If $i > j$, then e_i is orthogonal to $\operatorname{span}(e_1, \dots, e_j)$ and hence e_i is orthogonal to v_j (by 7.59); thus $R_{ij} = \langle v_j, e_i \rangle = 0$, so R is upper-triangular.

$\boxed{\text{Uniqueness}}$ Suppose we also have $A = \hat{Q}\hat{R}$, where $\hat{Q}$ is unitary and $\hat{R}$ is upper-triangular with only positive numbers on its diagonal. □

8.4.4 Cholesky Factorisation

We begin this subsection with a characterisation of positive invertible operators in terms of inner products.

Lemma 8.35 (Positive invertible operator)**.** *A self-adjoint operator $T \in \mathcal{L}(V)$ is a positive invertible operator if and only if $\langle Tv, v\rangle > 0$ for every non-zero $v \in V$.*

Proof.

$\boxed{\implies}$ Suppose T is a positive invertible operator.

Let $v \in V \setminus \{\mathbf{0}\}$. Since T is invertible, T is injective so $Tv \neq \mathbf{0}$. This implies that $\langle Tv, v\rangle \neq 0$ (by 7.43). Hence $\langle Tv, v\rangle > 0$.

$\boxed{\impliedby}$ Suppose $\langle Tv, v\rangle > 0$ for every $v \in V \setminus \{\mathbf{0}\}$. Thus $Tv \neq \mathbf{0}$ for every $v \in V \setminus \{\mathbf{0}\}$.

Hence T is injective, and thus is invertible. □

Definition 8.36 (Positive definite)**.** A matrix $B \in \mathcal{M}_{n\times n}(\mathbf{F})$ is called ***positive definite*** if

(i) B is Hermitian ($B^* = B$), and

(ii) $\langle Bx, x\rangle > 0$ for every non-zero $x \in \mathbf{F}^n$.

A matrix is upper triangular if and only if its conjugate transpose is lower triangular (meaning that all entries above the diagonal are 0). The factorisation below writes a positive definite matrix as the product of a lower triangular matrix and its conjugate transpose.

Theorem 8.37 (Cholesky factorisation)**.** *Suppose B is a positive definite matrix. Then there exists a unique upper-triangular matrix R with only positive numbers on its diagonal such that*

$$B = R^*R. \tag{8.2}$$

Proof. □

8.5 Singular Value Decomposition

8.5.1 Singular Values

We will need the following result in this section.

Lemma 8.38 (Properties of T^*T)**.** *Suppose $T \in \mathcal{L}(V, W)$. Then*

(i) T^*T *is a positive operator on* V*;*

(ii) $\ker T^*T = \ker T$*;*

(iii) $\operatorname{im} T^*T = \operatorname{im} T$*;*

(iv) $\dim \operatorname{im} T = \dim \operatorname{im} T^* = \dim \operatorname{im} T^*T$.

Proof.

(i) Since
$$(T^*T)^* = T^*(T^*)^* = T^*T,$$
T^*T is self-adjoint.

Let $v \in V$. Then
$$\langle (T^*T)v, v\rangle = \langle T^*(Tv), v\rangle = \langle Tv, Tv\rangle = \|Tv\|^2 \geq 0.$$
Thus T^*T is a positive operator.

(ii) $\boxed{\subset}$ Let $v \in \ker T^*T$. Then
$$\|Tv\|^2 = \langle Tv, Tv\rangle = \langle T^*Tv, v\rangle = \langle \mathbf{0}, v\rangle = 0.$$
Thus $Tv = \mathbf{0}$. Hence $\ker T^*T \subset \ker T$.

$\boxed{\supset}$ Let $v \in \ker T$. Then $Tv = \mathbf{0}$, so $T^*Tv = \mathbf{0}$. Hence $\ker T \subset \ker T^*T$.

(iii) By (i), T^*T is self-adjoint. Thus
$$\operatorname{im} T^*T = (\ker T^*T)^\perp = (\ker T)^\perp = \ker T^*.$$

(iv) For the first equality,
$$\dim \operatorname{im} T = \dim(\ker T^*)^\perp = \dim W - \dim \ker T^* = \dim \operatorname{im} T^*.$$
The second equality $\dim \operatorname{im} T^* = \dim \operatorname{im} T^*T$ follows from (iii).

□

Definition 8.39 (Singular values)**.** Suppose $T \in \mathcal{L}(V, W)$. The ***singular values*** of T are the non-negative square roots of the eigenvalues of T^*T

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0,$$

each included as many times as the dimension of the corresponding eigenspace of T^*T.

Lemma 8.40 (Role of positive singular values)**.** *Suppose* $T \in \mathcal{L}(V, W)$. *Then*

(i) T *is injective* $\iff$ 0 *is not a singular value of* T*;*

(ii) the number of positive singular values of T *equals* $\dim \operatorname{im} T$*;*

(iii) T *is surjective* $\iff$ *number of positive singular values of* T *equals* $\dim W$.

Proof.

(i) We have

$$\begin{aligned} T \text{ is injective} &\iff \ker T = \{\mathbf{0}\} \\ &\iff \ker T^*T = \{\mathbf{0}\} \\ &\iff 0 \text{ is not an eigenvalue of } T^*T \\ &\iff 0 \text{ is not a singular value of } T. \end{aligned}$$

(ii) The spectral theorem applied to T^*T shows that $\dim \operatorname{im} T^*T$ equals the number of positive eigenvalues of T^*T (counting repetitions). Thus 7.64(d) implies that $\dim \operatorname{im} T$ equals the number of positive singular values of T.

(iii) This follows from (ii) and 2.39.

□

The next result characterises isometries in terms of singular values.

Lemma 8.41. *Suppose* $S \in \mathcal{L}(V, W)$. *Then*

$$S \text{ is an isometry} \iff \text{all singular values of } S \text{ equal } 1.$$

Proof. We have

$$\begin{aligned} S \text{ is an isometry} &\iff S^*S = I \\ &\iff \text{all eigenvalues of } S^*S \text{ equal } 1 \\ &\iff \text{all singular values of } S \text{ equal } 1 \end{aligned}$$

□

8.5.2 SVD for Linear Maps and for Matrices

The next result shows that every linear map from V to W has a remarkably clean description in terms of its singular values and orthonormal sets in V and W; this is known as *singular value decomposition* (SVD).

Theorem 8.42 (Singular value decomposition)**.** *Suppose $T \in \mathcal{L}(V, W)$ and the positive singular values of T are $\sigma_1, \ldots, \sigma_m$. Then there exist orthonormal sets $\{e_1, \ldots, e_m\}$ in V and $\{f_1, \ldots, f_m\}$ in W such that*

$$Tv = \sigma_1 \langle v, e_1 \rangle f_1 + \cdots + \sigma_m \langle v, e_m \rangle f_m \tag{8.3}$$

for every $v \in V$.

Proof. Let $\sigma_1, \ldots, \sigma_n$ denote the singular values of T (thus $n = \dim V$).

1. Since T^*T is a positive operator, by the spectral theorem, there exists an orthonormal basis $\{e_1, \ldots, e_n\}$ of V with
$$T^*Te_i = \sigma_i e_i^2 \quad (i = 1, \ldots, n).$$

2. For each $i = 1, \ldots, m$, let
$$f_i = \frac{Te_i}{\sigma_i}.$$
We check that $\{f_1, \ldots, f_m\}$ is an orthonormal set in W. If $i, j \in \{1, \ldots, m\}$, then
$$\begin{aligned}
\langle f_i, f_j \rangle &= \left\langle \frac{1}{\sigma_i} Te_i, \frac{1}{\sigma_j} Te_j \right\rangle \\
&= \frac{1}{\sigma_i \sigma_j} \langle Te_i, Te_j \rangle \\
&= \frac{1}{\sigma_i \sigma_j} \langle e_i, T^*Te_j \rangle \\
&= \frac{\sigma_j}{\sigma_i} \langle e_i, e_j \rangle \\
&= \begin{cases} 0 & (i \neq j) \\ 1 & (i = j) \end{cases}
\end{aligned}$$

3. If $i \in \{1, \ldots, n\}$ and $i > m$, then $\sigma_i = 0$ and hence $T^*Te_i = 0$, which implies that $Te_i = 0$.
Let $v \in V$. Then
$$\begin{aligned}
Tv &= T(\langle v, e_1 \rangle e_1 + \cdots + \langle v, e_n \rangle e_n) \\
&= \langle v, e_1 \rangle Te_1 + \cdots + \langle v, e_m \rangle Te_m \\
&= \sigma_1 \langle v, e_1 \rangle f_1 + \cdots + \sigma_m \langle v, e_m \rangle f_m.
\end{aligned}$$

□

Suppose $T \in \mathcal{T}(V, W)$, with singular values $\sigma_1, \ldots, \sigma_m$. Let $\{e_1, \ldots, e_m\}$ and $\{f_1, \ldots, f_m\}$ be such that (8.3) holds. Extend the orthonormal set $\{e_1, \ldots, e_m\}$ to an orthonormal basis $\{e_1, \ldots, e_{\dim V}\}$ of

V, and extend the orthonormal set $\{f_1, \dots, f_m\}$ to an orthonormal basis $\{f_1, \dots, f_{\dim W}\}$ of W. (8.3) shows that

$$Te_i = \begin{cases} \sigma_i f_i & (1 \le i \le m) \\ 0 & (m < i \le \dim V) \end{cases}$$

Thus the matrix of T with respect to the orthonormal bases $\{e_1, \dots, e_{\dim V}\}$ and $\{f_1, \dots, f_{\dim W}\}$ is

$$\mathcal{M}(T)_{ij} = \begin{cases} \sigma_i & (1 \le i = j \le m) \\ 0 & (\text{otherwise}) \end{cases}$$

If $\dim V = \dim W$ (when $V = W$), then the matrix described in the paragraph above is a diagonal matrix. Let us extend the definition of diagonal matrix to matrices that are not necessarily square:

> An $M \times N$ matrix A is called a *diagonal matrix* if all entries of the matrix are 0 except possibly A_{ii} for $i = 1, \dots, \min\{M, N\}$.

Then we have shown that *every linear map has a diagonal matrix with respect to some orthonormal bases*.

Theorem 8.43 (Singular value decomposition of adjoint and pseudoinverse). *Suppose $T \in \mathcal{L}(V, W)$ with singular values $\sigma_1, \dots, \sigma_m$. Suppose $\{e_1, \dots, e_m\}$ and $\{f_1, \dots, f_m\}$ are orthonormal sets in V and W such that*

$$Tv = \sigma_1 \langle v, e_1 \rangle f_1 + \cdots + \sigma_m \langle v, e_m \rangle f_m$$

for every $v \in V$. Then

$$T^* w = \sigma_1 \langle w, f_1 \rangle e_1 + \cdots + \sigma_m \langle w, f_m \rangle e_m \tag{8.4}$$

and

$$T^+ w = \frac{\langle w, f_1 \rangle}{\sigma_1} e_1 + \cdots + \frac{\langle w, f_m \rangle}{\sigma_m} e_m \tag{8.5}$$

for every $w \in W$.

Proof.

$\boxed{\text{Adjoint}}$ Let $v \in V$, $w \in W$. Then

$$\begin{aligned} \langle Tv, w \rangle &= \langle \sigma_1 \langle v, e_1 \rangle f_1 + \cdots + \sigma_m \langle v, e_m \rangle f_m, w \rangle \\ &= \sigma_1 \langle v, e_1 \rangle \langle f_1, w \rangle + \cdots + \sigma_m \langle v, e_m \rangle \langle f_m, w \rangle \\ &= \langle v, \sigma_1 \langle w, f_1 \rangle e_1 + \cdots + \sigma_m \langle w, f_m \rangle e_m \rangle . \end{aligned}$$

Thus (8.4) follows.

$\boxed{\text{Pseudoinverse}}$ Let $w \in W$. Let

$$v = \frac{\langle w, f_1 \rangle}{\sigma_1} e_1 + \cdots + \frac{\langle w, f_m \rangle}{\sigma_m} e_m .$$

Applying T to both sides gives

$$\begin{aligned} Tv &= \frac{\langle w, f_1 \rangle}{\sigma_1} Te_1 + \cdots + \frac{\langle w, f_m \rangle}{\sigma_m} Te_m \\ &= \langle w, f_1 \rangle f_1 + \cdots + \langle w, f_m \rangle f_m \\ &= P_{\operatorname{im} T} w. \end{aligned}$$

... Thus $v = T^+ w$, and (8.5) follows. □

Theorem 8.44 (Singular value decomposition, matrix)**.** *Suppose $A \in \mathcal{M}_{p\times n}(\mathbf{F})$ has rank $m \geq 1$. Then*

$$A = U\Sigma V^* \tag{8.6}$$

for some $U \in \mathcal{M}_{p\times m}(\mathbf{F})$ with orthonormal columns, $\Sigma \in \mathcal{M}_{m\times m}(\mathbf{F})$ with positive numbers on the diagonal, $V \in \mathcal{M}_{n\times m}(\mathbf{F})$ with orthonormal columns.

Proof. Let $T\colon \mathbf{F}^n \to \mathbf{F}^p$ be the linear map whose matrix with respect to the standard bases equals A. Then $\dim \operatorname{im} T = m$. Let

$$Tv = \sigma_1 \langle v, e_1 \rangle f_1 + \cdots + \sigma_m \langle v, e_m \rangle f_m$$

be a singular value decomposition of T.

Claim. Let

- U be the $p \times m$ matrix whose columns are $f_1, \dots, f_m$,
- Σ be the $m \times m$ diagonal matrix whose diagonal entries are $\sigma_1, \dots, \sigma_m$,
- V be the $n \times m$ matrix whose columns are $e_1, \dots, e_m$.

We now show (8.6) holds. Let $\{u_1, \dots, u_m\}$ denote the standard basis of $\mathbf{F}^m$. For each $i = 1, \dots, m$,

$$\begin{aligned} (AV - U\Sigma)u_i &= A(Vu_i) - U(\Sigma u_i) \\ &= Ae_i - U(\sigma_i u_i) \\ &= \sigma_i f_i - \sigma_i f_i = 0 \end{aligned}$$

implies $AV = U\Sigma$. Then multiply both sides by V^* on the right to get

$$AVV^* = U\Sigma V^*.$$

Claim. $AVV^* = A$.

Note that the rows of V^* are the complex conjugates of $e_1, \dots, e_m$. Thus if $i \in \{1, \dots, m\}$, then the definition of matrix multiplication shows that $V^* e_i = u_i$; hence $VV^* e_i = e_i$. Thus $AVV^* v = Av$ for all $v \in \operatorname{span}(e_1, \dots, e_m)$.

If $v \in (\operatorname{span}(e_1, \dots, e_m))^\perp$, then $Av = 0$ (as follows from 7.81) and $V^* v = 0$ (as follows from the definition of matrix multiplication). Hence $AVV^* v = Av$ for all $v \in (\operatorname{span}(e_1, \dots, e_m))^\perp$.

Since AVV^* and A agree on $\operatorname{span}(e_1, \ldots, e_m)$ and on $(\operatorname{span}(e_1, \ldots, e_m))^{\perp}$, we conclude that $AVV^* = A$. Thus (8.6) follows. $\square$

8.6 Consequences of Singular Value Decomposition

8.6.1 Norm of Linear Map

The singular value decomposition leads to an upper bound for $\|Tv\|$.

Lemma 8.45. *Suppose $T \in \mathcal{L}(V, W)$. Then $\|Tv\| \le \sigma_1 \|v\|$ for all $v \in V$.*

Proof. Suppose $T \in \mathcal{L}(V, W)$, with singular values $\sigma_1, \dots, \sigma_m$. Let $\{e_1, \dots, e_m\}$ be an orthonormal set in V, and $\{f_1, \dots, f_m\}$ be an orthonormal set in W that provide a singular value decomposition of T. Thus

$$Tv = \sigma_1 \langle v, e_1 \rangle f_1 + \cdots + \sigma_m \langle v, e_m \rangle f_m$$

for all $v \in V$. Then

$$\begin{aligned}
\|Tv\|^2 &= {\sigma_1}^2 |\langle v, e_1 \rangle|^2 + \cdots + {\sigma_m}^2 |\langle v, e_m \rangle|^2 \\
&\le {\sigma_1}^2 \left(|\langle v, e_1 \rangle|^2 + \cdots + |\langle v, e_m \rangle|^2 \right) \\
&\le {\sigma_1}^2 \|v\|^2,
\end{aligned}$$

where the last inequality follows from Bessel's inequality. Taking square roots on both sides yields the desired inequality. □

Definition 8.46 (Norm of linear map)**.** Suppose $T \in \mathcal{L}(V, W)$. Then the ***norm*** of T is

$$\|T\| := \sup_{\|v\| \le 1} \|Tv\|.$$

Lemma 8.47 (Basic properties of norm of linear map)**.** *Suppose $T \in \mathcal{L}(V, W)$. Then*

(i) $\|T\| \ge 0$, where equality holds if and only if $T = 0$; *(positive definiteness)*

(ii) $\|\lambda T\| = |\lambda| \|T\|$ for all $\lambda \in \mathbf{F}$; *(homogeneity)*

(iii) $\|S + T\| \le \|S\| + \|T\|$ for all $S \in \mathcal{L}(V, W)$. *(triangle inequality)*

Proof.

(i) Since $\|Tv\| \ge 0$ for every $v \in V$, the definition of $\|T\|$ implies $\|T\| \ge 0$.

(ii) $\boxed{\implies}$ Suppose $\|T\| = 0$. Thus $Tv = \mathbf{0}$ for all $v \in V$, $\|v\| \le 1$.

Let $u \in V$, $u \ne \mathbf{0}$. Then

$$Tu = \|u\| T \left(\frac{u}{\|u\|} \right) = \mathbf{0},$$

where the last equality holds since $u/\|u\|$ has norm 1. Since $Tu = \mathbf{0}$ for all $u \in V$, we have $T = 0$.

$\boxed{\impliedby}$ If $T = 0$, then $Tv = \mathbf{0}$ for all $v \in V$. Hence $\|T\| = 0$.

(iii) Let $\lambda \in \mathbf{F}$. Then

$$\|\lambda T\| = \sup_{\|v\|\leq 1} \|\lambda T v\| = |\lambda| \sup_{\|v\|\leq 1} \|Tv\| = |\lambda| \|T\|.$$

(iv) Let $S \in \mathcal{L}(V, W)$. The definition of $\|S+T\|$ implies there exists $v \in V$, $\|v\| \leq 1$ such that $\|S+T\| = \|(S+T)v\|$. Then

$$\|S+T\| = \|(S+T)v\| = \|Sv + Tv\| \leq \|Sv\| + \|Tv\| \leq \|S\| + \|T\|.$$

□

Hence $\mathcal{L}(V, W)$ is a metric space, with metric $d(S,T) = \|S - T\|$ for $S, T \in \mathcal{L}(V, W)$.

(i) $d(S,S) = \|S - S\| = 0$. If $S \neq T$, then $d(S,T) = \|S - T\| = \sigma_{\max}(S - T)$. Since $S - T \neq 0$, its largest singular value is nonzero and therefore $d(S,T) > 0$.

(ii) $d(S,T) = \|S - T\| = \|T - S\| = d(T,S)$.

(iii) $d(S,G) = \|S - G\| \leq \|S - T\| + \|T - G\| = d(S,T) + d(T,G)$.

Lemma 8.48 (Alternative formulae for $\|T\|$)**.** *Suppose $T \in \mathcal{L}(V, W)$. Then*

(i) $\|T\| =$ largest singular value of T;

(ii) $\|T\| = \sup_{\|v\|=1} \|Tv\|$;

(iii) $\|T\| =$ smallest number c such that $\|Tv\| \leq c\|v\|$ for all $v \in V$.

Proof.

(i) This follows from 8.45.

(ii) Let $v \in V$, $\|v\| \leq 1$. Let $u = v/\|v\|$. Then

$$\|u\| = \left\| \frac{v}{\|v\|} \right\|$$

and

$$\|Tu\| = \left\| T\left(\frac{v}{\|v\|}\right) \right\| = \frac{\|Tv\|}{\|v\|} \geq \|Tv\|.$$

Thus when finding the maximum of $\|Tv\|$ with $\|v\| \leq 1$, we can restrict our attention to vectors in V with norm 1, proving (ii).

(iii) Let $v \in V$, $v \neq \mathbf{0}$. Then the definition of $\|T\|$ implies that

$$\left\| T\left(\frac{v}{\|v\|}\right) \right\| \leq \|T\|,$$

which implies that

$$\|Tv\| \leq \|T\| \|v\|.$$

Now suppose $c \geq 0$ and $\|Tv\| \leq c\|v\|$ for all $v \in V$. This implies that

$$\|Tv\| \leq c$$

for all $v \in V$, $\|v\| \leq 1$. Taking sup on the LHS over all $v \in V$, $\|v\| \leq 1$ shows that $\|T\| \leq c$. Thus $\|T\|$ is the smallest number c such that $\|Tv\| \leq c\|v\|$ for all $v \in V$.

□

An important inequality during the proof is

$$\|Tv\| \leq \|T\|\|v\| \tag{8.7}$$

for all $v \in V$, $v \neq \mathbf{0}$.

A linear map and its adjoint have the same norm, as shown by the next result.

Lemma 8.49 (Norm of adjoint)**.** *Suppose $T \in \mathcal{L}(V, W)$. Then $\|T^*\| = \|T\|$.*

Proof. Suppose $w \in W$. Then

$$\|T^*w\|^2 = \langle T^*w, T^*w \rangle = \langle TT^*w, w \rangle \leq \|TT^*w\|\|w\| \leq \|T\|\|T^*w\|\|w\|.$$

The inequality above implies that

$$\|T^*w\| \leq \|T\|\|w\|.$$

But $\|T^*w\| \leq \|T^*\|\|w\|$, so we have $\|T^*\| \leq \|T\|$.

Replacing T with T^* shows that $\|T\| \leq \|T^*\|$. Thus $\|T^*\| = \|T\|$, as desired. □

8.6.2 Approximation by Linear Maps with Lower-Dimensional Range

Theorem 8.50 (Best approximation by linear map whose image has dimension $\leq k$)**.** *Suppose $T \in \mathcal{L}(V, W)$, and $\sigma_1, \ldots, \sigma_m$ are the singular values of T. Suppose $1 \leq k < m$. Then*

$$\min\{\|T - S\| \mid S \in \mathcal{L}(V, W) \textit{ and } \dim \operatorname{im} S \leq k\} = \sigma_{k+1}. \tag{8.8}$$

Furthermore, if

$$Tv = \sigma_1 \langle v, e_1 \rangle f_1 + \cdots + \sigma_m \langle v, e_m \rangle f_m$$

is a singular value decomposition of T and $T_k \in \mathcal{L}(V, W)$ is defined by

$$T_k v = \sigma_1 \langle v, e_1 \rangle f_1 + \cdots + \sigma_k \langle v, e_k \rangle f_k$$

for each $v \in V$, then $\dim \operatorname{im} T_k = k$ and $\|T - T_k\| = \sigma_{k+1}$.

8.6.3 Polar Decomposition

Every non-zero complex number $z \in \mathbb{C}$ can be written in the form

$$\begin{aligned} z &= \left(\frac{z}{|z|}\right)|z| \\ &= \left(\frac{z}{|z|}\right)\sqrt{\bar{z}z} \end{aligned}$$

where $z/|z|$ has absolute value 1, and $\sqrt{\bar{z}z}$ is positive.

Our analogy leads us to guess that every operator $T \in \mathcal{L}(V)$ can be written as a unitary operator times $\sqrt{T^*T}$. The corresponding result is called the *polar decomposition*, which gives a beautiful description of an arbitrary operator on V.

Theorem 8.51 (Polar decomposition)**.** *Suppose $T \in \mathcal{L}(V)$. Then there exists a unitary operator $S \in \mathcal{L}(V)$ such that*

$$T = S\sqrt{T^*T}. \tag{8.9}$$

Remark. This holds for both $\mathbb{C}$ and $\mathbb{R}$.

Proof. Let $\sigma_1, \ldots, \sigma_m$ be the positive singular values of T. Let $\{e_1, \ldots, e_m\}$ and $\{f_1, \ldots, f_m\}$ be orthonormal sets in V such that

$$Tv = \sigma_1 \langle v, e_1 \rangle f_1 + \cdots + \sigma_m \langle v, e_m \rangle f_m$$

for every $v \in V$. Extend $\{e_1, \ldots, e_m\}$ and $\{f_1, \ldots, f_m\}$ to orthonormal bases $\{e_1, \ldots, e_n\}$ and $\{f_1, \ldots, f_n\}$ of V.

Recall that the singular value decomposition of the adjoint is

$$T^*w = \sigma_1 \langle w, f_1 \rangle e_1 + \cdots + \sigma_m \langle w, f_m \rangle e_m$$

for all $w \in W$. Thus

$$T^*Tv = \sigma_1^2 \langle v, e_1 \rangle e_1 + \cdots + \sigma_m^2 \langle v, e_m \rangle e_m$$

for every $v \in V$. Then

$$\sqrt{T^*T}v = \sigma_1 \langle v, e_1 \rangle e_1 + \cdots + \sigma_m \langle v, e_m \rangle e_m$$

because the operator that sends v to the RHS of the equation above is a positive operator whose square equals T^*T.

Claim. Define $S \in \mathcal{L}(V)$ by

$$Sv = \langle v, e_1 \rangle f_1 + \cdots + \langle v, e_n \rangle f_n$$

for each $v \in V$.

Then

$$\begin{aligned}\|Sv\|^2 &= \|\langle v, e_1\rangle f_1 + \cdots + \langle v, e_n\rangle f_n\|^2 \\ &= |\langle v, e_1\rangle|^2 + \cdots + |\langle v, e_n\rangle|^2 \\ &= \|v\|^2.\end{aligned}$$

Thus S is a unitary operator. Now

$$\begin{aligned}S\sqrt{T^*T}v &= S\left(\sigma_1 \langle v, e_1\rangle e_1 + \cdots + \sigma_m \langle v, e_m\rangle e_m\right) \\ &= \sigma_1 \langle v, e_1\rangle f_1 + \cdots + \sigma_m \langle v, e_m\rangle f_m \\ &= Tv.\end{aligned}$$

□

8.6.4 Operators Applied to Ellipsoids and Parallelepipeds

Definition 8.52 (Ball)**.** The ***unit ball*** in V centred at $\mathbf{0}$ is

$$B := \{v \in V \mid \|v\| \leq 1\}.$$

You can think of the ellipsoid defined below as obtained by starting with the ball B, and then stretching by a factor of s_i along each f_i-axis.

Definition 8.53 (Ellipsoid)**.** Suppose $\{f_1, \ldots, f_n\}$ is an orthonormal basis of V, and $s_1, \ldots, s_n > 0$. The ***ellipsoid*** with principal axes $s_1 f_1, \ldots, s_n f_n$ is

$$E(s_1 f_1, \ldots, s_n f_n) := \left\{ v \in V \;\middle|\; \frac{|\langle v, f_1\rangle|^2}{s_1} + \cdots + \frac{|\langle v, f_n\rangle|^2}{s_n} < 1 \right\}.$$

Remark. If $\dim V = 2$, the word "disk" is sometimes used to denote ball and the word "ellipse" is sometimes used to denote ellipsoid.

The next result states that an invertible map takes a ball to an ellipsoid.

Proposition 8.54. *Suppose $T \in \mathcal{L}(V)$ is invertible. Then T maps the ball B in V to an ellipsoid in V.*

Proof. Suppose T has the singular value decomposition

$$Tv = \sigma_1 \langle v, e_1\rangle f_1 + \cdots + \sigma_n \langle v, e_n\rangle f_n$$

for all $v \in V$, where $\{e_1, \ldots, e_n\}$ and $\{f_1, \ldots, f_n\}$ are orthonormal bases of V. We will show that

$$T(B) = E(\sigma_1 f_1, \ldots, \sigma_n f_n).$$

$\boxed{\subset}$ Let $v \in B$. Since T is invertible, none of the singular values $\sigma_1, \ldots, \sigma_n$ equal 0. Thus

$\boxed{\supset}$ □

The next result states that an invertible map takes an ellipsoid to an ellipsoid.

Proposition 8.55. *Suppose $T \in \mathcal{L}(V)$ is invertible, and E is an ellipsoid in V. Then $T(E)$ is an ellipsoid in V.*

Definition 8.56 (Parallelepiped)**.** Suppose $\{v_1, \ldots, v_n\}$ is a basis of V. Let

$$P(v_1, \ldots, v_n) := \{a_1 v_1 + \cdots + a_n v_n \mid a_i \in (0, 1)\}.$$

A ***parallelepiped*** is a set of the form $u + P(v_1, \ldots, v_n)$ for some $u \in V$. The vectors $v_1, \ldots, v_n$ are called the *edges* of the parallelepiped.

The next result states an invertible operator takes a parallelepiped to a parallelepiped.

Proposition 8.57. *Suppose $u \in V$, and $\{v_1, \ldots, v_n\}$ is a basis of V. Suppose $T \in \mathcal{L}(V)$ is invertible. Then*

$$T(u + P(v_1, \ldots, v_n)) = Tu + P(Tv_1, \ldots, Tv_n).$$

Definition 8.58 (Box)**.** A ***box*** is of the form

$$u + P(r_1 e_1, \ldots, r_n e_n)$$

where $u \in V$, $r_1, \ldots, r_n > 0$, and $\{e_1, \ldots, e_n\}$ is an orthonormal basis of V.

8.6.5 Volume via Singular Values

8.6.6 Properties of an Operator as Determined by Its Eigenvalues

Exercises

7A 1-12 15 16 17 18

Exercise 8.1 ([Axl24] 7A Q12)**.** An operator $B \in \mathcal{L}(V)$ is called *skew* if

$$B^* = -B.$$

Suppose $T \in \mathcal{L}(V)$. Prove that T is normal if and only if there exist commuting operators A and B such that A is self-adjoint, B is a skew operator, and $T = A + B$.

Solution. □

Exercise 8.2 ([Axl24] 7A Q19)**.** Suppose $T \in \mathcal{L}(V)$ and $\|T^*v\| \leq \|Tv\|$ for every $v \in V$. Prove that T is normal.

Remark. This exercise fails on infinite-dimensional inner product spaces, leading to what are called *hyponormal operators*.

Solution. Let $\{e_1, \dots, e_n\}$ be an orthonormal basis of V. For $i = 1, \dots, n$, write

$$\begin{aligned} \|Te_i\|^2 &= |\langle Te_i, e_1 \rangle|^2 + \cdots + |\langle Te_i, e_n \rangle|^2 \\ &= |\langle e_i, T^*e_1 \rangle|^2 + \cdots + |\langle e_i, T^*e_n \rangle|^2 . \end{aligned}$$

Summing over i,

$$\begin{aligned} \sum_{i=1}^{n} \|Te_i\|^2 &= \sum_{i=1}^{n} |\langle e_i, T^*e_1 \rangle|^2 + \cdots + \sum_{i=1}^{n} |\langle e_i, T^*e_n \rangle|^2 \\ &= \|T^*e_1\|^2 + \cdots + \|T^*e_n\|^2. \end{aligned}$$

Since we are given $\|T^*v\| \leq \|Tv\|$ for every $v \in V$, and equality holds, we must have $\|Te_i\| = \|T^*e_i\|$ for each $i = 1, \dots, n$. Since the choice of orthonormal basis was arbitrary, we must have $\|Tu\| = \|T^*u\|$ for every unit vector $u \in V$.

For every $v \in V$, $\frac{1}{\|v\|}v \in V$ is a unit vector, so

$$\left\| T\left(\frac{1}{\|v\|}v\right) \right\| = \left\| T^*\left(\frac{1}{\|v\|}v\right) \right\|$$

which simplifies to $\|Tv\| = \|T^*v\|$. Hence by 8.12, T is normal. □

7A 20

Exercise 8.3 ([Axl24] 7A Q24)**.** Suppose $T \in \mathcal{L}(V)$ and

$$a_0 + a_1 z + a_2 z^2 + \cdots + a_{m-1} z^{m-1} + z^m$$

is the minimal polynomial of T. Prove that the minimal polynomial of T^* is

$$\overline{a_0} + \overline{a_1} z + \overline{a_2} z^2 + \cdots + \overline{a_{m-1}} z^{m-1} + z^m.$$

Remark. This exercise shows that the minimal polynomial of T^* equals the minimal polynomial of T if $\mathbf{F} = \mathbb{R}$.

Solution. Let p be the minimal polynomial of T.

Claim. If f is any polynomial, then $f(T^*) = \overline{f(T)}^*$.

Let $f(x) = c_n x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0$. Then

$$\begin{aligned} f(T^*) &= c_n(T^*)^n + c_{n-1}(T^*)^{n-1} + \cdots + c_1 T^* + c_0 \\ &= c_n(T^n)^* + \cdots + c_1 T^* + c_0 \\ &= (\overline{c_n}T^n + \cdots + \overline{c_1}T + \overline{c_0})^* \\ &= \overline{f(T)}^* \end{aligned}$$

as desired.

Since p is the minimal polynomial of T, we have $p(T) = 0$, so

$$\overline{p(T^*)} = p(T)^* = 0^* = 0$$

which implies $\overline{p}$ is a zero polynomial of T^*.

Let q be the minimal polynomial of T^*. Then $\overline{q}$ is the minimal polynomial of $(T^*)^* = T$. Since p is the minimal polynomial, $p \mid \overline{q}$ which implies $\overline{p} \mid q$. Hence $\overline{p} = q$ by minimality of q. □

Exercise 8.4 ([Axl24] 7A Q25)**.** Suppose $T \in \mathcal{L}(V)$. Prove that T is diagonalisable if and only if T^* is diagonalisable.

Solution. □

7A 27 28 29

Exercise 8.5 ([Axl24] 7B Q1)**.** Prove that a normal operator on a complex inner product space is self-adjoint if and only if all its eigenvalues are real.

Exercise 8.6 ([Axl24] 7B Q2)**.** Suppose $\mathbf{F} = \mathbb{C}$. Suppose $T \in \mathcal{L}(V)$ is normal and has only one eigenvalue. Prove that T is a scalar multiple of the identity operator.

Exercise 8.7 ([Axl24] 7B Q3)**.** Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathcal{L}(V)$ is normal. Prove that the set of eigenvalues of T is contained in $\{0, 1\}$ if and only if there is a subspace U of V such that $T = P_U$.

Exercise 8.8 ([Axl24] 7B Q4)**.** Prove that a normal operator on a complex inner product space is skew (meaning it equals the negative of its adjoint) if and only if all its eigenvalues are purely imaginary.

Exercise 8.9 ([Axl24] 7B Q6)**.** Suppose V is a complex inner product space and $T \in \mathcal{L}(V)$ is a normal operator such that $T^9 = T^8$. Prove that T is self-adjoint and $T^2 = T$.

Exercise 8.10 ([Axl24] 7B Q8)**.** Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathcal{L}(V)$. Prove that T is normal if and only if every eigenvector of T is also an eigenvector of T^*.

Exercise 8.11 ([Axl24] 7B Q9)**.** Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathcal{L}(V)$. Prove that T is normal if and only if there exists a polynomial $p \in \mathbb{C}[z]$ such that $T^* = p(T)$.

Solution.

$\boxed{\Longleftarrow}$ Suppose there exists a polynomial $p \in \mathbb{C}[z]$ such that $T^* = p(T)$. Since $Tp(T) = p(T)T$, this implies $TT^* = T^*T$ so T is normal.

$\boxed{\Longrightarrow}$ Let $\lambda_1, \dots, \lambda_r$ be eigenvalues of T. Then

$$V = \bigoplus_{i=1}^{r} E(\lambda_i, T).$$

Since T is normal, we have $E(\lambda_i, T) = E(\overline{\lambda_i}, T^*)$ for each i. Thus

$$V = \bigoplus_{i=1}^{r} E(\overline{\lambda_i}, T^*).$$

Let $V_i = E(\lambda_i, T) = E(\overline{\lambda_i}, T^*)$. Then

$$T|_{V_i} = \lambda_i I_{V_i}, \quad T^*|_{V_i} = \overline{\lambda_i} I_{V_i}.$$

We want to express T^* as a polynomial of T. Define

$$p(T) = \sum_{i=1}^{r} \overline{\lambda_i} \frac{(T-\lambda_1 I)\cdots(T-\lambda_{i-1}I)(T-\lambda_{i+1}I)\cdots(T-\lambda_r I)}{(\lambda_i-\lambda_1)\cdots(\lambda_i-\lambda_{i-1})(\lambda_i-\lambda_{i+1})\cdots(\lambda_i-\lambda_r)}.$$

For each $v_i \in E(\lambda_i, T)$,

$$\begin{aligned} p(T)v_i &= \overline{\lambda_i} \frac{(T-\lambda_1 I)\cdots(T-\lambda_{i-1}I)(T-\lambda_{i+1}I)\cdots(T-\lambda_r I)v_i}{(\lambda_i-\lambda_1)\cdots(\lambda_i-\lambda_{i-1})(\lambda_i-\lambda_{i+1})\cdots(\lambda_i-\lambda_r)} \\ &= \overline{\lambda_i} v_i = T^* v_i. \end{aligned}$$

T is normal implies T is diagonalisable. Pick a basis of eigenvectors $v_1, \dots, v_n$ of T. Then $p(T)v_i = T^*v_i$ for $i = 1, \dots, n$ implies $T^* = p(T)$. □

7B 10 11 12(use Q9 to prove)

Exercise 8.12 ([Axl24] 7B Q17)**.** Suppose $\mathbf{F} = \mathbb{R}$ and $\mathcal{E} \subset \mathcal{L}(V)$. Prove that there is an orthonormal basis of V with respect to which every element of $\mathcal{E}$ has a diagonal matrix if and only if S and T are commuting self-adjoint operators for all $S, T \in \mathcal{E}$.

This exercise extends the real spectral theorem to the context of a collection of commuting self-adjoint operators.

Exercise 8.13 ([Axl24] 7B Q19)**.** Suppose $T \in \mathcal{L}(V)$ is self-adjoint, and $U \leq V$ is invariant under T.

(i) Prove that $U^\perp$ is invariant under T.

(ii) Prove that $T|_U \in \mathcal{L}(V)$ is self-adjoint.

(iii) Prove that $T|_{U^\perp} \in \mathcal{L}(U^\perp)$ is self-adjoint.

Solution.

(i) Let $v \in U^\perp$. Then for all $w \in U$, $\langle v, w \rangle = 0$. Since U is invariant under V, $Tw \in U$, so

$$\langle Tv, w \rangle = \langle v, Tw \rangle = 0.$$

Thus $Tw \in U^\perp$. Hence $U^\perp$ is invariant under T.

(ii) For all $v, w \in U$, since $T \in \mathcal{L}(V)$ is self-adjoint,

$$\langle Tv, w \rangle = \langle v, Tw \rangle .$$

Restricting T to U gives

$$\langle T|_U v, w \rangle = \langle v, T|_U w \rangle .$$

Hence $T|_U$ is self-adjoint.

(iii) This follows from (i) and (ii).

□

Exercise 8.14 ([Axl24] 7B Q20)**.** Suppose $T \in \mathcal{L}(V)$ is normal, and $U \leq V$ is invariant under T.

(i) Prove that $U^\perp$ is invariant under T.

(ii) Prove that U is invariant under T^*.

(iii) Prove that $(T|_U)^* = (T^*)|_U$.

(iv) Prove that $T|_U \in \mathcal{L}(U)$ and $T|_{U^\perp} \in \mathcal{L}(U^\perp)$ are normal operators.

Solution.

(i) Let $v \in U^\perp$. Then $\langle v, w \rangle = 0$ for all $w \in U$. Since U is invariant under T, $Tw \in U$, so

$$\langle T^* v, w \rangle = \langle v, Tw \rangle = 0$$

which implies that $T^* v \in U^\perp$. Hence $U^\perp$ is invariant under T^*.

Using Exercise 9, T^* is a polynomial of T. Let $T = p(T^*)$, then $U^\perp$ is invariant under $p(T^*)$, which implies that $U^\perp$ is invariant under T.

Since T is normal, T is diagonalisable. Since U is invariant, the restriction $T|_U \in \mathcal{L}(U)$ is diagonalisable. Pick a basis of eigenvectors $\{u_1, \dots, u_m\}$ of U. Then

$$Tu_1 = \lambda_1 u_1, \quad \dots, \quad Tu_m = \lambda_m u_m.$$

T is normal implies

$$T^* u_1 = \overline{\lambda_1} u_1, \quad \dots, \quad T^* u_m = \overline{\lambda_m} u_m.$$

For each $v \in U^\perp$, $\langle v, u_1 \rangle = \cdots = \langle v, u_m \rangle = 0$, so

$$\langle Tv, u_i \rangle = \langle v, T^* u_i \rangle = \left\langle v, \overline{\lambda_i} u_i \right\rangle = \lambda_i \langle v, u_i \rangle = 0$$

for $i = 1, \dots, n$. Thus $Tv \in U^\perp$. Hence $U^\perp$ is invariant under T.

(ii) This follows from (i).

(iii) We know $T|_U, T^*|_U \in \mathcal{L}(U)$. For all $v, w \in U$,

$$\langle Tv, w \rangle = \langle v, T^*w \rangle$$

so

$$\langle T|_U v, w \rangle = \langle v, T^*|_U w \rangle .$$

Hence $(T|_U)^* w = T^*|_U w$.

(iv) For each $v \in U$, $T^*Tv = TT^*v$ implies

$$T^*|_U T|_U v = T|_U T^*|_U v$$

so

$$(T|_U)^* T|_U v = T|_U (T|_U)^* v.$$

□

Exercise 8.15 ([Axl24] 7B Q21)**.** Suppose that T is a self-adjoint operator on a finite-dimensional inner product space, and that 2 and 3 are the only eigenvalues of T. Prove that

$$T^2 - 5T + 6I = 0.$$

We say a matrix A is *symmetric* if $A^T = A$, and *Hermitian* if $A^* = A$.

Exercise 8.16 ([Axl24] 7B Q24)**.** Suppose U is a finite-dimensional vector space, and $T \in \mathcal{L}(U)$.

(i) Suppose $\mathbf{F} = \mathbb{R}$. Prove that T is diagonalisable if and only if there exists a basis of U such that the matrix of T with respect to this basis is symmetric.

(ii) Suppose $\mathbf{F} = \mathbb{C}$. Prove that T is diagonalisable if and only if there exists a basis of U such that the matrix of T with respect to this basis commutes with its conjugate transpose.

Solution.

(i)

(ii)

□

7C 1 3 5 6 7 11 13 14 15 16 17 18

Exercise 8.17 ([Axl24] 7C Q18)**.** Suppose S and T are positive operators on V. Prove that ST is a positive operator if and only if S and T commute.

Solution.

$\boxed{\implies}$ Suppose ST is positive. Then ST is self-adjoint, so

$$ST = (ST)^* = T^*S^* = TS.$$

Thus S and T commute.

$\boxed{\impliedby}$ Suppose $ST = TS$. Then

$$(ST)^* = T^*S^* = TS = ST$$

implies ST is self-adjoint. Let $v \in V$, then

$$\langle STv, v\rangle = \left\langle S\sqrt{T}v, \sqrt{T}v\right\rangle \geq 0,$$

since S is positive. □

Exercise 8.18 ([Axl24] 7C Q22)**.** Suppose $T \in \mathcal{L}(V)$ is a positive operator and $u \in V$ is such that $\|u\| = 1$, and $\|Tu\| \geq \|Tv\|$ for all $v \in V$ with $\|v\| = 1$. Show that u is an eigenvector corresponding to the largest eigenvalue of T.

Solution. Let $\{e_1, \dots, e_n\}$ be the orthogonal eigenbasis of V, with corresponding values $\lambda_1, \dots, \lambda_n$ sorted from smallest to largest.

Let $u = a_1e_1 + \cdots + a_ne_n$. Then

$$\begin{aligned} Tu &= a_1Te_1 + \cdots + a_nTe_n \\ &= a_1\lambda_1e_1 + \cdots + a_n\lambda_ne_n, \end{aligned}$$

so

$$\|Tu\|^2 = \left\|\sum_{i=1}^n a_i\lambda_ie_i\right\|^2 = \sum_{i=1}^n |a_i|^2{\lambda_i}^2.$$

We can take $v = e_n$, then we have that

$$\|Tu\|^2 = \sum_{i=1}^n |a_i|^2{\lambda_i}^2 \geq \sum_{i=1}^n |a_i|^2{\lambda_n}^2 = {\lambda_n}^2 = \|Tv\|^2$$

which shows the desired conclusion. □

Exercise 8.19 ([Axl24] 7C Q24)**.** Suppose $S, T \in \mathcal{L}(V)$ are positive operators. Prove that

$$\ker(S + T) = \ker S \cap \ker T.$$

Proof.

$\boxed{\supset}$

$$\begin{aligned} v \in \ker S \cap \ker T &\implies Sv = Tv = \mathbf{0} \\ &\implies (S + T)v = \mathbf{0} \\ &\implies v \in \ker(S + T) \end{aligned}$$

$\boxed{\subset}$

$$\begin{aligned}
&v \in \ker(S+T) \\
\implies &(S+T)v = \mathbf{0} \\
\implies &0 = \langle (S+T)v, v\rangle = \underbrace{\langle Sv, v\rangle}_{\geq 0} + \underbrace{\langle Tv, v\rangle}_{\geq 0} \\
\implies &\langle Sv, v\rangle = \langle Tv, v\rangle = 0 \\
\implies &\left\langle \sqrt{S}v, \sqrt{S}v\right\rangle = \left\langle \sqrt{T}v, \sqrt{T}v\right\rangle = 0 \\
\implies &\sqrt{S}v = \sqrt{T}v = \mathbf{0} \\
\implies &Sv = Tv = \mathbf{0}
\end{aligned}$$

□

7D 2 3 5 9 11

Exercise 8.20 ([Axl24] 7D Q18)**.** Prove that if A is a symmetric matrix with real entries, then there exists a unitary matrix Q with real entries such that Q^*AQ is a diagonal matrix.

Solution. □

7E 1 2 4 7 8 9 10 11 13

7F 1-7 19

Exercise 8.21 ([Axl24] 7F Q8)**.**

(i) Prove that if $T \in \mathcal{L}(V)$ and $\|T\| < 1$, then $I + T$ is invertible.

(ii) Suppose that $S \in \mathcal{L}(V)$ is invertible. Prove that if $T \in \mathcal{L}(V)$ and $\|S - T\| < \dfrac{1}{\|S^{-1}\|}$, then T is invertible.

Solution.

(i) Let $v \in \ker(I + T)$. Then $(I + T)v = \mathbf{0}$, so $Tv = -v$. Taking the norm gives

$$\|v\| = \|Tv\| \leq \|T\|\|v\|.$$

Thus

$$(1 - \|T\|)\|v\| \leq 0.$$

Since $1 - \|T\| > 0$, we must have $\|v\| \leq 0$. Hence $v = \mathbf{0}$.

(ii) Let $v \in \ker T$. Then $Tv = \mathbf{0}$, so

$$Sv = (S - T)v.$$

Since S is invertible,

$$v = S^{-1}(S + T)v.$$

Taking norm,

$$\begin{aligned}\|v\| &= \left\|S^{-1}(S+T)v\right\| \\ &= \left\|S^{-1}\right\|\|S-T\|\|v\|.\end{aligned}$$

Thus

$$\underbrace{(1-\left\|S^{-1}\right\|\|S-T\|)}_{>0}\|v\| \leq 0.$$

This implies $\|v\|=0$, so $v=\mathbf{0}$. Hence T is injective, so T is invertible.

□

Then $B_{\frac{1}{\|S^{-1}\|}}(S)$ is an open ball in $\mathcal{L}(V)$ consisting of invertible linear maps. Hence the set of invertible operators in $\mathcal{L}(V)$ is an open subset of $\mathcal{L}(V)$.

Exercise 8.22 ([Axl24] 7F Q9)**.** Suppose $T \in \mathcal{L}(V)$. Prove that

$$\forall \varepsilon>0, \quad \exists S \in \mathcal{L}(V) \text{ invertible}, \quad 0<\|T-S\|<\varepsilon.$$

Solution. Define $S=T+\delta I$ for some $0<\delta<\varepsilon$. Then we have

$$\|T-S\| = \|\delta I\| = \delta$$

which satisifes the desired condition. Note that if T is invertible, we can simply choose a sufficiently small $\delta < 1/\left\|T^{-1}\right\|$; if not, then any $\delta \in (0,1)$ can make S invertible. □

Exercise 8.23 ([Axl24] 7F Q10)**.** Suppose $\dim V>1$ and $T \in \mathcal{L}(V)$ is not invertible. Prove that

$$\forall \varepsilon>0, \quad \exists S \in \mathcal{L}(V) \text{ not invertible}, \quad 0<\|T-S\|<\varepsilon.$$

Solution. □

Exercise 8.24 ([Axl24] 7F Q11)**.** Suppose $\mathbf{F}=\mathbb{C}$ and $T \in \mathcal{L}(V)$. Prove that

$$\forall \varepsilon>0, \quad \exists S \in \mathcal{L}(V) \text{ diagonalisable}, \quad 0<\|T-S\|<\varepsilon.$$

Exercise 8.25 ([Axl24] 7F Q12)**.** Suppose $T \in \mathcal{L}(V)$ is a positive operator. Show that $\left\|\sqrt{T}\right\| = \sqrt{\|T\|}$.

Solution. Let $\|T\|=\sigma_1$, the largest singular value of T. Then $\left\|\sqrt{T}\right\| = \sqrt{\sigma_1} = \sqrt{\|T\|}$. □

Exercise 8.26 ([Axl24] 7F Q14)**.** Suppose $U, W \leq V$ are such that $\|P_U - P_W\| < 1$. Prove that $\dim U = \dim W$.

Solution. We prove the contrapositive. Suppose $\dim U < \dim W$. Then the orthogonal complement

$$\dim U^\perp = \dim V - \dim U > \dim V - \dim W,$$

or

$$\dim W + \dim U^\perp > \dim W.$$

We know

$$\begin{aligned}\dim V &\geq \dim(W + U^\perp) \\ &= \dim W + \dim U^\perp - \dim(W \cap U^\perp) \\ &> \dim V - \dim(W \cap U^\perp)\end{aligned}$$

which implies $\dim W \cap U^\perp > 0$. Pick $v \in W \cap U^\perp$, $\|v\| = 1$. Then $P_U v = \mathbf{0}$ and $P_W v = v$, so

$$\|(P_U - P_w)v\| = \|v\| = 1.$$

Hence $\|P_U - P_W\| \geq 1$, so $\dim U \geq \dim W$. Similarly, $\|P_W - P_U\| \leq 1$, so $\dim W \geq \dim U$. Therefore $\dim U = \dim W$. $\square$

Exercise 8.27 ([Axl24] 7F Q19)**.** Prove that if $T \in \mathcal{L}(V, W)$, then $\|T^*T\| = \|T\|^2$.

Solution. Let ${s_1}^2$ be the greatest eigenvalue of T^*T. Then $\|T\| = s_1$, so $\|T\|^2 = {s_1}^2 = \|T^*T\|$. $\square$

Chapter 9

Operators on Complex Vector Spaces

9.1 Generalised Eigenvectors and Nilpotent Operators

9.1.1 Kernels of Powers of an Operator

We begin this chapter with a study of kernels of powers of an operator. The following result provides a sequence of increasing kernels.

Lemma 9.1. *Suppose $T \in \mathcal{L}(V)$. Then*

$$\{\mathbf{0}\} = \ker T^0 \subset \ker T^1 \subset \cdots \subset \ker T^k \subset \ker T^{k+1} \subset \cdots .$$

Proof. Let $v \in \ker T^k$, for non-negative integer k. Then $T^k v = \mathbf{0}$, so $T^{k+1}v = T\left(T^k v\right) = T(\mathbf{0}) = \mathbf{0}$. Thus $v \in \ker T^{k+1}$.

Hence $\ker T^k \subset \ker T^{k+1}$ for non-negative integers k. □

The next result states that if two consecutive terms in the sequence are equal, then all later terms are equal.

Lemma 9.2. *Suppose $T \in \mathcal{L}(V)$, and $\ker T^m = \ker T^{m+1}$ for some non-negative integer m. Then*

$$\ker T^m = \ker T^{m+1} = \ker T^{m+1} = \ker T^{m+3} = \cdots .$$

Proof. Let k be a positive integer. We want to prove that

$$\ker T^{m+k} = \ker T^{m+k+1}.$$

$\boxed{\subset}$ This follows from 9.1.

$\boxed{\supset}$ Let $v \in \ker T^{m+k+1}$. Then

$$T^{m+1}(T^k v) = T^{m+k+1}v = \mathbf{0}.$$

Hence

$$T^k v \in \ker T^{m+1} = \ker T^m.$$

Thus $T^{m+k} v = T^m(T^k v) = \mathbf{0}$, which means that $v \in \ker T^{m+k}$. □

The result above raises the question of whether there exists a non-negative integer m such that $\ker T^m = \ker T^{m+1}$. The next result shows that this equality holds, at least when $m = \dim V$.

Lemma 9.3 (Kernels stop growing). *Suppose $T \in \mathcal{L}(V)$. Then*

$$\ker T^{\dim V} = \ker T^{\dim V+1} = \ker T^{\dim V+2} = \cdots.$$

Proof. By 9.2, it suffices to prove that $\ker T^{\dim V} = \ker T^{\dim V+1}$. Suppose, for a contradiction, that $\ker T^{\dim V} \neq \ker T^{\dim V+1}$. Then

$$\{\mathbf{0}\} = \ker T^0 \subsetneq \ker T^1 \subsetneq \cdots \subsetneq \ker T^{\dim V} \subsetneq \ker T^{\dim V+1},$$

where we have strict inclusions in the chain above. At each of the strict inclusions, the dimension increases by at least 1. Thus $\dim \ker T^{\dim V+1} \geq \dim V + 1$. Since $\ker T^{\dim V+1}$ is a subspace of V, it cannot have a larger dimension than the whole space. Hence we have reached a contradiction. □

It is not true that $V = \ker T \oplus \operatorname{im} T$ for every $T \in \mathcal{L}(V)$. However, the next result can be a useful substitute.

Proposition 9.4 (Direct sum decomposition). *Suppose $T \in \mathcal{L}(V)$. Then*

$$V = \ker T^{\dim V} \oplus \operatorname{im} T^{\dim V}.$$

Proof. Let $n = \dim V$.

- We first show that $\ker T^n + \operatorname{im} T^n$ is a direct sum. By 3.19, we will show that

  $$\ker T^n \cap \operatorname{im} T^n = \{\mathbf{0}\}.$$

  Let $v \in \ker T^n \cap \operatorname{im} T^n$. Then $T^n v = \mathbf{0}$, and there exists $u \in V$ such that $v = T^n u$. Applying T^n to both sides gives

  $$T^n v = T^{2n} u = \mathbf{0}$$

  which implies

  $$T^n u = \mathbf{0}.$$

  Thus $v = T^n u = \mathbf{0}$.

- Next we show that $V = \ker T^n \oplus \operatorname{im} T^n$. We shall use 3.42 to show that the subspace $\ker T^n \oplus \operatorname{im} T^n$ equals the whole space V:

  $$\begin{aligned} \dim(\ker T^n \oplus \operatorname{im} T^n) &= \dim \ker T^n + \dim \operatorname{im} T^n && \text{[by 4.58]} \\ &= \dim V && \text{[by fundamental theorem of linear maps]} \end{aligned}$$

□

9.1.2 Generalised Eigenvectors

Definition 9.5 (Generalised eigenvector)**.** Suppose $T \in \mathcal{L}(V)$, and $\lambda \in \mathbf{F}$ is an eigenvalue of T. We say $v \in V \setminus \{\mathbf{0}\}$ is a ***generalised eigenvector*** of T corresponding to λ, if

$$(T - \lambda I)^k v = \mathbf{0}$$

for some positive integer k.

That is, there exists $k \in \mathbb{Z}^+$ such that

$$v \in \ker(T - \lambda I)^k.$$

Remark. If $k = 1$, then this coincides with the usual eigenvector. Hence all eigenvectors are generalised eigenvectors.

Remark. We do not define generalised eigenvalues because they are no different from the usual eigenvalues. Reason: if $(T - \lambda I)^k$ is not injective for some positive integer k, then $T - \lambda I$ is not injective, and hence λ is an eigenvalue of T.

A non-zero vector $v \in V$ is a generalised eigenvector of T corresponding to λ if and only if

$$(T - \lambda I)^{\dim V} v = \mathbf{0},$$

as follows from applying 9.1 and 9.3 to the operator $T - \lambda I$.

As we know, an operator on a complex vector space may not have enough eigenvectors to form a basis of the domain. The next result shows that on a complex vector space, there are enough generalised eigenvectors to do this.

Proposition 9.6. *Suppose* $\mathbf{F} = \mathbb{C}$*, and* $T \in \mathcal{L}(V)$*. Then* T *has a basis of generalised eigenvectors in* V*.*

Proof. Let $n = \dim V$. We shall induct on n.

If $n = 1$, then every non-zero vector in V is an eigenvector of T. Thus the desired result holds.

Suppose $n > 1$, and the desired result holds for all smaller values of $\dim V$. Let λ be an eigenvalue of T. Applying 9.4 to $T - \lambda I$ shows that

$$V = \ker(T - \lambda I)^n \oplus \operatorname{im}(T - \lambda I)^n.$$

If $\ker(T - \lambda I)^n = V$, then every non-zero vector in V is a generalised eigenvector of T, and thus in this case there is a basis of V consisting of generalised eigenvectors of T. Hence we can assume that $\ker(T - \lambda I)^n \neq V$, which implies that $\operatorname{im}(T - \lambda I)^n \neq \{\mathbf{0}\}$. Since λ is an eigenvalue of T, we have $\ker(T - \lambda I)^n \neq \{\mathbf{0}\}$. Thus

$$0 < \dim \operatorname{im}(T - \lambda I)^n < n.$$

Note that $\operatorname{im}(T-\lambda I)^n$ is invariant under T. Let $S \in \mathcal{L}(\operatorname{im}(T-\lambda I)^n)$ be defined by

$$S = T|_{\operatorname{im}(T-\lambda I)^n}.$$

By induction hypothesis applied to S, there is a basis of $\operatorname{im}(T-\lambda I)^n$ consisting of generalised eigenvectors of S, which of course are generalised eigenvectors of T. Adjoining that basis of $\operatorname{im}(T-\lambda I)^n$ to a basis of $\ker(T-\lambda I)^n$ gives a basis of V consisting of generalised eigenvectors of T. □

Suppose $T \in \mathcal{L}(V)$. If v is an eigenvector of T, then the corresponding eigenvalue λ is uniquely determined by the equation $Tv = \lambda v$, which can be satisfied by only one $\lambda \in \mathbf{F}$ (since $v \neq \mathbf{0}$). The next result shows a similar result holds for generalised eigenvectors: if v is a generalised eigenvector of T, then the equation $(T-\lambda I)^{\dim V} v = \mathbf{0}$ can be satisfied by only one $\lambda \in \mathbf{F}$.

Lemma 9.7. *Suppose $T \in \mathcal{L}(V)$. Then each generalised eigenvector of T corresponds to only one eigenvalue of T.*

Proof. Suppose $T \in \mathcal{L}(V)$. Let $v \in V$ be a generalised eigenvector of T corresponding to eigenvalues λ and λ'.

Let m be the smallest positive integer such that $(T-\lambda' I)^m v = \mathbf{0}$. Let $n = \dim V$. Then

$$\begin{aligned} \mathbf{0} &= (T-\lambda I)^n v \\ &= \left((T-\lambda' I) + (\lambda' - \lambda)I\right)^n v \\ &= \sum_{i=1}^{n} \binom{n}{i} (\lambda' - \lambda)^{n-i} (T-\lambda' I)^i v. \end{aligned}$$

Applying the operator $(T-\lambda I)^{m-1}$ to both sides gives

$$\mathbf{0} = (\lambda' - \lambda)^n (T-\lambda' I)^{m-1} v.$$

Since $(T-\lambda' I)^{m-1} v \neq \mathbf{0}$, the equation above implies that $\lambda' = \lambda$, as desired. □

Recall that by 6.6, eigenvectors corresponding to distinct eigenvalues are linearly independent. We now prove a similar result for generalised eigenvectors.

Proposition 9.8. *Suppose $T \in \mathcal{L}(V)$. Then every set of generalised eigenvectors of T corresponding to distinct eigenvalues of T is linearly independent.*

Proof. Suppose, for a contradiction, that the desired result is false. Then there exists a smallest positive integer m such that $\{v_1, \dots, v_m\}$ are linearly dependent generalised eigenvectors of T corresponding to distinct eigenvalues $\lambda_1, \dots, \lambda_m$ of T.

Thus there exist $a_1, \dots, a_m \in \mathbf{F}$, none of which are 0 (because of the minimality of m), such that

$$a_1 v_1 + \cdots + a_m v_m = \mathbf{0}.$$

Let $n = \dim V$. Applying $(T - \lambda_m I)^n$ to both sides gives

$$a_1(T - \lambda_m I)^n v_1 + \cdots + a_{m-1}(T - \lambda_m I)^n v_{m-1} = \mathbf{0}. \tag{I}$$

Let $i \in \{1, \ldots, m-1\}$. Then

$$(T - \lambda_m I)^n v_i \neq \mathbf{0},$$

because otherwise v_i would be a generalised eigenvector of T corresponding to distinct eigenvalues λ_i and λ_m, contradicting 8.11. However

$$(T - \lambda_i I)^n \left((T - \lambda_m I)^n v_i\right) = (T - \lambda_m I)^n \left((T - \lambda_i I)^n v_i\right) = \mathbf{0}.$$

Thus the last two equations show that $(T - \lambda_m I)^n v_i$ is a generalised eigenvector of T corresponding to the eigenvalue λ_i. Hence

$$(T - \lambda_m I)^n v_1, \ldots, (T - \lambda_m I)^n v_{m-1}$$

is a linearly dependent set (by (I)) of $m-1$ generalised eigenvectors corresponding to distinct eigenvalues, contradicting the minimality of m. □

9.1.3 Nilpotent Operators

Definition 9.9 (Nilpotent operator)**.** An operator is ***nilpotent*** if some power of it equals 0.

Thus an operator $T \in \mathcal{L}(V)$ is nilpotent if and only if every non-zero vector in V is a generalised eigenvector of T corresponding to the eigenvalue 0.

Lemma 9.10. *Suppose $T \in \mathcal{L}(V)$ is nilpotent. Then $T^{\dim V} = 0$.*

Proof. Since T is nilpotent, there exists $k \in \mathbb{Z}^+$ such that $T^k = 0$.

Thus $\ker T^k = V$. Now 9.1 and 9.3 imply that $\ker T^{\dim V} = V$. Hence $T^{\dim V} = 0$. □

The next result concerns the eigenvalues of nilpotent operators.

Proposition 9.11. *Suppose $T \in \mathcal{L}(V)$.*

(i) If T is nilpotent, then 0 is the only eigenvalue of T.

(ii) If $\mathbf{F} = \mathbb{C}$ and 0 is the only eigenvalue of T, then T is nilpotent.

Proof.

(i) Suppose T is nilpotent. Then $T^m = 0$ for some $m \in \mathbb{Z}^+$. This implies that T is not injective. Thus 0 is an eigenvalue of T.

To show that T has no other eigenvalues, suppose λ is an eigenvalue of T. Then there exists $v \in V \setminus \{\mathbf{0}\}$ such that

$$\lambda v = Tv.$$

Repeatedly applying T to both sides of the equation gives

$$\lambda^m v = T^m v = \mathbf{0}.$$

Thus $\lambda = 0$, as desired.

(ii) Suppose $\mathbf{F} = \mathbb{C}$, and 0 is the only eigenvalue of T. By 6.16, the minimal polynomial of T equals z^m for some positive integer m. Thus $T^m = 0$. Hence T is nilpotent.

□

The next result provides a characterisation of a nilpotent operator, in terms of its minimal polynomial and matrix.

Proposition 9.12. *Suppose $T \in \mathcal{L}(V)$. Then the following are equivalent:*

(i) T is nilpotent.

(ii) The minimal polynomial of T is z^m, for some positive integer m.

(iii) T has a strictly upper-triangular matrix with respect to some basis of V.

Proof.

$\boxed{\text{(i)} \implies \text{(ii)}}$ Suppose T is nilpotent. Then $T^n = 0$ for some $n \in \mathbb{Z}^+$.

By 6.17, z^n is a polynomial multiple of the minimal polynomial of T. Thus the minimal polynomial of T is z^m for some $m \in \mathbb{Z}^+$.

$\boxed{\text{(ii)} \implies \text{(iii)}}$ Suppose the minimal polynomial of T must be z^m, for some $m \in \mathbb{Z}^+$.

Since 0 is the only zero of z^m, by 6.16, 0 is the only eigenvalue of T.

Thus by 6.27, T has an upper-triangular matrix with respect to some basis of V.

By 6.24, since the eigenvalues of T are the diagonal entries of the matrix, we conclude that all entries on the diagonal of this matrix are 0.

$\boxed{\text{(iii)} \implies \text{(i)}}$ Suppose T has a strictly upper-triangular matrix with respect to some basis of V.

Then the diagonal entries are all 0. By 6.24, $T^{\dim V} = 0$. Hence T is nilpotent. □

9.2 Generalised Eigenspace Decomposition

9.2.1 Generalised Eigenspaces

Definition 9.13 (Generalised eigenspace)**.** Suppose $T \in \mathcal{L}(V)$, and $\lambda \in \mathbf{F}$. The ***generalised eigenspace*** of T corresponding to λ is

$$G(\lambda, T) := \left\{ v \in V \mid \exists k \in \mathbb{Z}^+, (T - \lambda I)^k v = \mathbf{0} \right\}.$$

That is, the generalised eigenspace of T corresponding to λ is the set of generalised eigenvectors of T corresponding to λ, along with $\mathbf{0}$.

Remark. $E(\lambda, T) \subset G(\lambda, T)$.

A consequence of the next result is $G(\lambda, T)$ is a subspace of V.

Lemma 9.14. *Suppose $T \in \mathcal{L}(V)$, and $\lambda \in \mathbf{F}$. Then*

$$G(\lambda, T) = \ker(T - \lambda I)^{\dim V}.$$

Proof.

$\boxed{\supset}$ Let $v \in \ker(T - \lambda I)^{\dim V}$. Then $(T - \lambda I)^{\dim V} = \mathbf{0}$. By definition, $v \in G(\lambda, T)$.

$\boxed{\subset}$ Let $v \in G(\lambda, T)$. Then $(T - \lambda I)^k v = \mathbf{0}$ for some $k \in \mathbb{Z}^+$, so $v \in \ker(T - \lambda I)^k$.

From 9.1 and 9.3 (with $T - \lambda I$ replacing T), we get $v \in \ker(T - \lambda I)^{\dim V}$. □

Theorem 9.15 (Generalised eigenspace decomposition)**.** *Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathcal{L}(V)$. Let $\lambda_1, \dots, \lambda_m$ be distinct eigenvalues of T. Then*

$$V = G(\lambda_1, T) \oplus \cdots \oplus G(\lambda_m, T). \tag{9.1}$$

9.2.2 Multiplicity of an Eigenvalue

Definition 9.16 (Multiplicity)**.** Suppose $T \in \mathcal{L}(V)$. The ***multiplicity*** of an eigenvalue λ of T is defined as

$$\dim G(\lambda, T).$$

Equivalently, the multiplicity of an eigenvalue λ of T equals

$$\dim \ker(T - \lambda I)^{\dim V}.$$

Lemma 9.17. *Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathcal{L}(V)$. Then the sum of the multiplicities of all eigenvalues of T equals* $\dim V$.

Definition 9.18 (Characteristic polynomial)**.** Suppose $\mathbf{F} = \mathbb{C}$ and $T \in \mathcal{L}(V)$. Let $\lambda_1, \ldots, \lambda_m$ be the distinct eigenvalues of T, with multiplicities $d_1, \ldots, d_m$. The polynomial

$$(z - \lambda_1)^{d_1} \cdots (z - \lambda_m)^{d_m}$$

is called the ***characteristic polynomial*** of T.

The next result concerns the degree and zeros of the characteristic polynomial.

Lemma 9.19. *Suppose* $\mathbf{F} = \mathbb{C}$ *and* $T \in \mathcal{L}(V)$. *Then*

(i) the characteristic polynomial of T *has degree* $\dim V$*;*

(ii) the zeros of the characteristic polynomial of T *are the eigenvalues of* T.

Theorem 9.20 (Cayley–Hamilton theorem)**.** *Suppose* $\mathbf{F} = \mathbb{C}$ *and* $T \in \mathcal{L}(V)$. *Let* q *be the characteristic polynomial of* T. *Then*

$$q(T) = 0.$$

Proposition 9.21. *Suppose* $\mathbf{F} = \mathbb{C}$ *and* $T \in \mathcal{L}(V)$. *Then the characteristic polynomial of* T *is a polynomial multiple of the minimal polynomial of* T.

Theorem 9.22. *Suppose* $\mathbf{F} = \mathbb{C}$ *and* $T \in \mathcal{L}(V)$. *Suppose* $\mathcal{M}(T)$ *is upper-triangular with respect to a basis* $\{v_1, \ldots, v_n\}$ *of* V. *Then the number of times that each eigenvalue* λ *of* T *appears on the diagonal of* $\mathcal{M}(T)$ *equals the multiplicity of* λ *as an eigenvalue of* T.

9.2.3 Block Diagonal Matrices

Often we can understand a matrix better by thinking of it as composed of smaller matrices.

Definition 9.23 (Block diagonal matrix)**.** A ***block diagonal matrix*** is a square matrix of the form

$$\begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_m \end{pmatrix}$$

where $A_1, \ldots, A_m$ are square matrices lying along the diagonal, and all other entries of the matrix equal 0.

Proposition 9.24. *Suppose* $\mathbf{F} = \mathbb{C}$ *and* $T \in \mathcal{L}(V)$. *Let* $\lambda_1, \ldots, \lambda_m$ *be distinct eigenvalues of* T, *with multiplicities* $d_1, \ldots, d_m$. *Then there is a basis of* V *with respect to which* T *has a block* *diagonal matrix of the form*

$$\begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_m \end{pmatrix}$$

where each A_i *is a* $d_i \times d_i$ *upper-triangular matrix of the form*

$$A_i = \begin{pmatrix} \lambda_i & & * \\ & \ddots & \\ 0 & & \lambda_i \end{pmatrix}.$$

9.3 Consequences of Generalised Eigenspace Decomposition

9.3.1 Square Roots of Operators

9.3.2 Jordan Form

9.4 Trace: A Connection Between Matrices and Operators

Exercises

8A Q 2 4 5 6-9

Exercise 9.1 ([Axl24] 8A)**.**

III
Abstract Algebra

Algebra is the study of collections of objects (sets, groups, rings, fields, etc). In algebra, we are concerned about the structures of these collections and how these collections interact than about the objects themselves. In fact, with homomorphism and isomorphisms, the original objects become irrelevant.

Chapter 10

Groups

One of the simplest forms of abstract algebraic systems is a *group*, which is roughly a set of objects and a rule for multiplying them together. Groups arise all over mathematics, particularly where there is symmetry.

10.1 Groups

10.1.1 Definitions and Properties

A ***binary operation*** on a set G is a map $*\colon G \times G \to G$.

Notation. For any $a, b \in G$, if the operation is clear, we write ab for the image of (a, b) under $*$.

Definition 10.1 (Group)**.** A ***group*** $(G, *)$ consists of a set G and a binary operation $*$ on G satisfying the following properties:

(i) $a(bc) = (ab)c$ for all $a, b, c \in G$; (associativity)

(ii) there exists $e \in G$ such that $ae = ea = a$ for all $a \in G$; (identity)

(iii) for all $a \in G$, there exists $c \in G$ such that $ac = ca = e$. (invertibility)

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

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: $ab \in G$ for all $a, b \in G$.

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

We say G is ***abelian*** if the operation is commutative; otherwise, G is *non-abelian*.

Lemma 10.2. *A group has a unique identity.*

Proof. Suppose that e and e' are identities of G. Then

$$e = ee' = e'$$

where the first equality holds since e' is an identity, and the second equality holds since e is an identity. □

Notation. We denote *the* identity of G as 1_G, and omit the subscript if there is no ambiguity.

Lemma 10.3. *Each element of a group has a unique inverse.*

Proof. Suppose that b and c are both inverses of a. Then $ab = 1$, $ca = 1$, so

$$c = c1 = c(ab) = (ca)b = 1b = b.$$

□

Notation. We denote *the* inverse of $a \in G$ as a^{-1}.

Lemma 10.4. *Let G be a group.*

(i) $(a^{-1})^{-1} = a$ for all $a \in G$.

(ii) $(ab)^{-1} = b^{-1}a^{-1}$ for all $a, b \in G$.

(iii) For any $a_1, \dots, a_n \in G$, $a_1 \cdots a_n$ is independent of how we arrange the parantheses (generalised associative law).

Proof.

(i) 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).

(ii) Let $c = (ab)^{-1}$. Then $(ab)c = 1$, or $a(bc) = 1$ by associativity, which gives $bc = a^{-1}$. Applying b^{-1} on both sides gives $c = b^{-1}a^{-1}$.

(iii) Induct on n. 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

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

By inductive hypothesis, both terms are independent of parentheses since $k, n - k < n$. Hence by induction we are done.

□

Lemma 10.5 (Cancellation law)**.** *Let $a, b \in G$. Then the equations $ax = b$ and $ya = b$ have unique solutions for $x, y \in G$.*

This means that we can cancel on the left and right.

Proof. To solve $ax = b$, apply a^{-1} on both sides to get $x = a^{-1}b$. The uniqueness of x follows because a^{-1} is unique.

A similar case holds for $ya = b$. □

We now introduce notation for repeated application of the operation on an element.

Notation. For any $a \in G$, $n \in \mathbb{N}$, denote $a^n = \underbrace{a \cdot a \cdots a}_{n \text{ times}}$, $a^0 = 1$, and $a^{-n} = (a^{-1})^n$.

The usual rules of exponents hold true:

$$a^{m+n} = a^m a^n$$

$$(a^m)^n = a^{mn}$$

$$(a^n)^{-1} = (a^{-1})^n$$

Definition 10.6 (Order of a group)**.** Let G be a group. The ***order*** of G is its cardinality $|G|$. We say G is a *finite group* if $|G| < \infty$.

One way to represent a finite group is by means of a ***Cayley table***. Let $G = \{1, g_2, g_3, \ldots, g_n\}$. The Cayley 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.

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.

10.1.2 Examples

Example 10.7.

- $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ are abelian groups under addition.
- $\mathbb{Q}^\times = \mathbb{Q} \setminus \{0\}$, $\mathbb{R}^\times = \mathbb{R} \setminus \{0\}$, and $\mathbb{C}^\times = \mathbb{C} \setminus \{0\}$ are groups under multiplication.
- The complex numbers of absolute value 1 form a group under multiplication.
- $\{1, -1\}$ is a group under multiplication.
- $\{1, -1, i, -i\}$ is a group under multiplication.

Example 10.8 (Modular arithmetic)**.** For $n \in \mathbb{N}$, the set of (congruence classes of) integers modulo n, $\mathbb{Z}/n\mathbb{Z}$, is an abelian group under addition.

For $n \in \mathbb{N}$, $(\mathbb{Z}/n\mathbb{Z})^\times$ is an abelian group under multiplication.

Example 10.9 (Direct product)**.** Let G, H be groups. The cartesian product $G \times H$ is a group under the operation

$$(g_1, h_1) \cdot (g_2, h_2) = (g_1 g_2, h_1 h_2).$$

We call $G \times H$ the *direct product* of G and H.

One may also take a direct product of a finite number of groups. Thus if $G_1, \ldots, G_n$, we let

$$\prod_{i=1}^{n} G_i = G_1 \times \cdots \times G_n$$

be the set of all n-tuples $(x_1, \ldots, x_n)$ with $x_i \in G_i$. We define multiplication componentwise, and see at once that $G_1 \times \cdots \times G_n$ is a group. If 1_i is the identity of G_i, then $(1_1, \ldots, 1_n)$ is the identity of the product.

Example 10.10 (Dihedral groups)**.** An important family of groups is the ***dihedral groups***. For $n \in \mathbb{N}$, $n \geq 3$, let D_{2n} be the set of symmetries of a regular n-gon.

Let r be the rotation clockwise about the origin by $\frac{2\pi}{n}$ radians, s be the reflection about the line of symmetry through the first labelled vertex and the origin. (Read from right to left: for instance, sr means do r then s.)

Properties of D_{2n}:

- $1, r, r^2, \dots, r^{n-1}$ are all distinct and $r^n = 1$, so $|r| = n$.
- $s^2 = 1$ since we either reflect or do not reflect, so $|s| = 2$.
- $s \neq r^i$ for any i, since the effect of any reflection cannot be obtained from any form of rotation.
- $sr^i \neq sr^j$ for all $i \neq j$ ($0 \leq i, j \leq n-1$), so

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

  and thus $|D_{2n}| = 2n$.
- $rs = sr^{-1}$
- $r^i s = sr^{-i}$

  Proof: From above, 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.

Note that for each $n \in \mathbb{N}$, the generators of D_{2n} are r and s, and we have shown that they satisfy $r^n = 1$, $s^2 = 1$, and $rs = sr^{-1}$; these are called *relations*. Any other equation involving the generators can be derived from these relations.

Any such collection of generators S and relations $R_1, \dots, R_m$ for a group G is called a *presentation*, written

$$G = \langle S \mid R_1, \dots, R_m \rangle.$$

For example,

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

Example 10.11 (Matrix groups)**.** For $n \in \mathbb{N}$, 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. Since $\det AB = \det A \det B$, 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.

(i) Matrix multiplication is associative.

(ii) $\det(A) \neq 0$ 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.

(iii) Inverse

Example 10.12 (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.

Example 10.13 (Roots of unity)**.** Let $n \in \mathbb{Z}^+$. Consider the set of roots of unity

$$\mu_n = \{e^{\frac{2k\pi i}{n}} \mid k = 0, \dots, n-1\}.$$

This forms an abelian group under multiplication, of order n.

10.1.3 Subgroups

When given a set with certain properties, it is natural to consider its subsets that inherit the same properties.

Definition 10.14 (Subgroup)**.** Let G be a group. We say that a non-empty $H \subset G$ is a ***subgroup*** of G, denoted by $H \leq G$, if H is a group under the restricted operation from G.

Every group G has two obvious subgroups: the group G itself, and the *trivial subgroup* $\{1\}$. A subgroup is a *proper subgroup* if it is not one of those two.

Example 10.15.

- $(\mathbb{Q}, +)$ is a subgroup of $(\mathbb{R}, +)$.
- The group of complex numbers of absolute value 1 is a subgroup of $\mathbb{C}^\times$, under multiplication.
- $\{1, -1\}$ is a subgroup of $\{1, -1, i, -i\}$, under multiplication.

According to the definition, to prove that H is a subgroup of G, we need to make sure H satisfies all group axioms. However, this is often tedious. Instead, there are some simplified criteria to decide whether H is a subgroup.

Lemma 10.16. *Let G be a group. Then $H \leq G$ if and only if*

(i) $1 \in H$; *(identity)*

(ii) $ab \in H$ *for all* $a, b \in H$; *(closure)*

(iii) $a^{-1} \in H$ *for all* $a \in H$. *(inverses)*

Humans are lazy, and the test above is still too complicated. We thus come up with an even simpler test:

Lemma 10.17 (Subgroup criterion)**.** *Let* G *be a group. Then* $H \leq G$ *if and only if*

(i) $H \neq \emptyset$;

(ii) $ab^{-1} \in H$ *for all* $a, b \in H$.

Proof.

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

$\boxed{\impliedby}$ Check group axioms:

(i) Since $H \neq \emptyset$, there exists $a \in H$. Then $1 = aa^{-1} \in H$.

(ii) Since $1 \in H$ and $a \in H$, then $a^{-1} = 1a^{-1} \in H$.

(iii) For any $a, b \in H$, $a, b^{-1} \in H$, so by (ii), $a(b^{-1})^{-1} = ab \in H$.

□

The next result and its corollary show that the intersection of subgroups is a subgroup.

Proposition 10.18. *Let* G *be a group,* $H, K \leq G$*. Then* $H \cap K \leq G$.

Proof. Apply the subgroup criterion:

(i) Since $1 \in H$ and $1 \in K$, then $1 \in H \cap K$ 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 $H, K \leq G$, by the subgroup criterion, $ab^{-1} \in H$ and $ab^{-1} \in K$, so $ab^{-1} \in H \cap K$.

□

Corollary 10.19. *Let* G *be a group,* $\{H_i \mid i \in I\}$ *is a collection of subgroups of* G*. Then*

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

Proposition 10.20. *Let* $H, K \leq G$*. If* $H \cup K \leq G$*, then either* $H \subset K$ *or* $K \subset H$.

Proof. Suppose $H \cup K \leq G$. Suppose, for a contradiction, that $H \not\subset K$ and $K \not\subset H$.

Let $h \in H \setminus K$, $k \in K \setminus H$. Since $H \cup K \leq G$, we have $hk \in H \cup K$.

- Suppose $hk \in H$, and let $h' = hk$. Since $h \in H$ and $H \leq G$, we have $h^{-1} \in H$. Thus $h^{-1}h' = h^{-1}hk = k$. But $h^{-1}h' \in H$ and $k \notin H$, which is a contradiction.

- Suppose $hk \in K$. Then similarly we will arrive at a contradiction.

Therefore, either $H \setminus K = \emptyset$ or $K \setminus H = \emptyset$. Equivalently, $H \subset K$ or $K \subset H$. □

There is a general way of obtaining subgroups from a group.

Definition 10.21 (Subgroup generated by subset)**.** Let $S \subset G$ be non-empty. Let H be the set of elements of G consisting of all products $x_1 \cdots x_n$ such that $x_i \in S$ or $x_i^{-1} \in S$ for each i, and also containing the unit element.
We call H the ***subgroup generated*** by S. We also say that S is a set of *generators* of H, and denote

$$H = \langle S \rangle .$$

Lemma. *The subgroup generated by a subset is indeed a subgroup.*

Thus if elements $\{x_1, \dots, x_n\}$ form a set of generators for G, we write

$$G = \langle x_1, \dots, x_n \rangle .$$

Example 10.22. 1 is a generator for $\mathbb{Z}$, since every integer can be written in the form

$$1 + 1 + \cdots + 1$$

or

$$-1 - 1 - \cdots - 1,$$

or it is the 0 integer.

10.1.4 Cyclic Groups

We consider the subgroup generated by one element.

Definition 10.23 (Cyclic subgroup)**.** The ***cyclic subgroup*** H generated by $a \in G$ is the set of all powers of a:

$$H = \langle a \rangle = \{a^n \mid n \in \mathbb{Z}\}.$$

We say that a is a *generator* of H.
We say G is ***cyclic*** if there exists $a \in G$ such that $G = \langle a \rangle$.

We write C_n for the cyclic group of order n:

$$C_n = \langle a \mid a^n = 1 \rangle .$$

Lemma. $H = \langle a \rangle$ *is a subgroup of* G.

Proof.

(i) H contains the identity $1 = a^0$.

(ii) Let $a^n, a^m \in H$. Then $a^m a^n = a^{m+n} \in H$.

(iii) $(a^n)^{-1} = a^{-n} \in H$.

□

Example 10.24.

- $\mathbb{Z}$ is cyclic with generator 1 or -1. It is *the* infinite cyclic group.
- The multiplicative group $\{1, -1\}$ is cyclic with generator -1.
- $\mathbb{Z}/n\mathbb{Z}$ is cyclic, with all numbers coprime with n as generators.
- The multiplicative group $\{1, -1\}$ is cyclic of order 2.
- The complex numbers $\{1, i, -1, -i\}$ form a cyclic group of order 4. The number i is a generator.

Remark. A cyclic subgroup may have more than one generator. For example, if a is a generator, then a^{-1} is also a generator:

$$\{a^n \mid n \in \mathbb{Z}\} = \{(a^{-1})^n \mid n \in \mathbb{Z}\}.$$

Lemma 10.25. *Cyclic groups are abelian.*

Proof. Let G be a cyclic group. For $a^i, a^j \in G$, we have $a^i a^j = a^{i+j} = a^j a^i$. □

Proposition 10.26. *A subgroup of a cyclic group is cyclic.*

Proof. Let $a \in G$, $H \leq \langle a \rangle$. If $H = \{1\}$ then trivially H is cyclic.

Suppose that H contains some other element $b \neq 1$. Then $b = a^n$ for some integer n. Since H is a subgroup, $b^{-1} = a^{-n} \in H$. Since either n or $-n$ is positive, we can assume H contains positive powers of a and $n > 0$. Let m be the smallest positive integer such that $a^m \in H$ (such an m exist by the well-ordering principle).

Claim. $h = a^m$ is a generator for H.

We need to show that every $h' \in H$ can be written as a power of h. Since $h' \in H$ and $H \leq \langle a \rangle$, $h' = a^k$ for some integer k. By the division algorithm, there exist integers q, r such that $k = qm + r$ with $0 \leq r < m$. Hence

$$a^k = a^{qm+r} = (a^m)^q a^r = h^q a^r$$

so $a^r = a^k h^{-q}$. Since $a^k, h^{-q} \in H$, we must have $a^r \in H$. By the minimality of m, we must have $m = 0$ and so $k = qm$. Hence

$$h' = a^k = a^{qm} = h^q$$

and H is generated by h. □

A corollary concerns all the subgroups of $\mathbb{Z}$.

Corollary 10.27. *The subgroups of $\mathbb{Z}$ are exactly $n\mathbb{Z}$ for $n = 0, 1, 2, \ldots$.*

10.1.5 Order

Definition 10.28 (Order). Let G be a group, $a \in G$. If there is a positive integer k such that $a^k = 1$, then the ***order*** of g is defined as

$$o(a) := \min\{m > 0 \mid a^m = 1\}.$$

Otherwise we say that the order of a is infinite.

We have given two different meanings to the word "order". One is the order of a group and the other is the order of an element. Since mathematicians are usually (but not always) sensible, the name wouldn't be used twice if they weren't related. This is explained by the next result.

Lemma 10.29. *For $a \in G$, $o(a) = |\langle a \rangle|$.*

Proof. We consider the cases where $o(a)$ is finite or infinite.

Case 1: $o(a) = \infty$. Then $a^n \neq a^m$ for all $n \neq m$; otherwise $a^{m-n} = 1$. Thus $|\langle a \rangle| = \infty = o(a)$.

Case 2: $o(a) < \infty$. Suppose $o(a) = k$. Thus $a^k = 1$. We now claim that $\langle a \rangle = \{1, a^1, a^2, \ldots, a^{k-1}\}$.

Note that $\langle a \rangle$ does not contain higher powers of a, since $a^k = 1$ so higher powers will loop back to existing elements. There are also no repeating elements in the list provided since $a^m = a^n$ implies $a^{m-n} = 1$. Hence $|\langle a \rangle| = k = o(a)$.

□

Lemma 10.30. *If $a \in G$ and $o(a)$ is finite, then $a^n = 1$ if and only if $o(a) \mid n$.*

Proof.

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

$$a^n = \left(a^{o(a)}\right)^k = 1^k = 1.$$

$\boxed{\Longrightarrow}$ Suppose $a^n = 1$. By the division algorithm, there exists integers q, r such that $n = qo(a) + r$, where $0 \leq r < o(a)$. Then

$$a^r = a^{n-qo(a)} = a^n \left(a^{o(a)}\right)^{-q} = 1.$$

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

Corollary 10.31. *Let G be a cyclic group, $a \in G$. Then $a^k = a^m$ if and only if $m \equiv k \pmod{o(a)}$.*

10.2 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 *isomorphism* between two groups.

10.2.1 Definitions and Properties

When we talk about functions between groups it makes sense to limit our scope to functions that preserve the group operation (morphisms in the category of groups). More precisely:

Definition 10.32 (Homomorphism)**.** Let $(G, *)$ and $(H, \diamond)$ be groups. We say $\phi\colon G \to H$ is a ***homomorphism*** if

$$\phi(x * y) = \phi(x) \diamond \phi(y) \quad (x, y \in G).$$

When the group operations for G and H are understood, we omit them and simply write

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

Example 10.33.

- Let G be a commutative group. The map $x \mapsto x^{-1}$ from G into itself is a homomorphism.
- The map $z \mapsto |z|$ is a homomorphism from $\mathbb{C}^\times$ to $\mathbb{R}^+$.
- The map $x \mapsto e^x$ is a homomorphism from $(\mathbb{R}, +)$ to $(\mathbb{R}^+, \times)$. Its inverse map, the logarithm, is also a homomorphism.

Lemma 10.34 (Basic properties)**.** *Let $\phi\colon G \to H$ be a homomorphism. Let $g \in G$, $n \in \mathbb{Z}$. Then*

(i) $\phi(1_G) = 1_H$

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

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

Proof.

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

(ii) $\phi(g)\phi(g^{-1}) = \phi(gg^{-1}) = \phi(1_G) = 1_H$.

(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

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

□

Lemma 10.35. *Let $\phi\colon G \to H$ and $\psi\colon H \to K$ be homomorphisms. Then $\psi \circ \phi$ is a homomorphism.*

Proof. We have

$$(\psi \circ \phi)(xy) = \psi\left(\phi(xy)\right) = \psi\left(\phi(x)\phi(y)\right) = \psi(\phi(x))\psi(\phi(y)) = (\psi \circ \phi)(x)(\psi \circ \phi)(y).$$

□

Let $\mathrm{Hom}(G, H)$ denote the set of homomorphisms from G to H. Then $\mathrm{Hom}(G, H)$ is a group under addition.

(i) If $\phi, \psi \in \mathrm{Hom}(G, H)$, then for $x, y \in G$,

$$\begin{aligned}(\phi + \psi)(x + y) &= \phi(x + y) + \psi(x + y) \\ &= \phi(x) + \psi(x) + \phi(y) + \psi(y) \\ &= (\phi + \psi)(x) + (\phi + \psi)(y),\end{aligned}$$

so that $f + g$ is a homomorphism.

(ii) If $\phi, \psi, \gamma \in \mathrm{Hom}(G, H)$, then for all $x \in G$,

$$\left((\phi + \psi) + \gamma\right)(x) = (\phi + \psi)(x) + \gamma(x) = \phi(x) + \psi(x) + \gamma(x),$$

and

$$\left(\phi + (\psi + \gamma)\right)(x) = \phi(x) + (\psi + \gamma)(x) = \phi(x) + \phi(x) + \gamma(x).$$

Hence $(\phi + \psi) + \gamma = \phi + (\psi + \gamma)$.

(iii) The zero map is the identity element of $\mathrm{Hom}(G, H)$.

(iv) The inverse of $\phi \in \mathrm{Hom}(G, H)$ is $-\phi$ (which is a homomorphism).

Definition 10.36 (Isomorphism)**.** An ***isomorphism*** is a bijective homomorphism. If there exists an isomorphism $\phi\colon G \to H$, we say G and H are ***isomorphic***, denoted by $G \cong H$.

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.

Example 10.37. The exponential map $\exp\colon \mathbb{R} \to \mathbb{R}^+$ defined by $\exp(x) = e^x$ is an isomorphism from $(\mathbb{R}, +)$ to $(\mathbb{R}^+, \times)$.

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

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

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

10.2.2 Kernel and Image

We introduce two important groups related to every homomorphism.

Definition 10.38 (Kernel)**.** Let $\phi\colon G \to H$ be a homomorphism. The ***kernel*** of ϕ is

$$\ker\phi := \{g \in G \mid \phi(g) = 1_H\}.$$

Lemma. $\ker\phi \lhd G$.

Proof. Apply the subgroup criterion. Since $1_G \in \ker\phi$, $\ker\phi \neq \emptyset$. Let $x, y \in \ker\phi$; that is, $\phi(x) = \phi(y) = 1_H$. Then

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

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

Let $x \in \ker\phi$, $g \in G$. Then

$$\phi(gxg^{-1}) = \phi(g)\phi(x)\phi(g^{-1}) = 1,$$

so $gxg^{-1} \in \ker\phi$. Hence $\ker\phi \lhd G$. □

Definition 10.39 (Image)**.** Let $\phi\colon G \to H$ be a homomorphism. The ***image*** of G under ϕ is

$$\operatorname{im}\phi := \phi(G) = \{\phi(g) \mid g \in G\}.$$

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

Lemma. $\operatorname{im}\phi \leq H$.

Proof. Since $\phi(1_G) = 1_H$, $1_H \in \operatorname{im}\phi$ 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, $\operatorname{im}\phi \leq G$. □

The following result is a useful characterisation for injective homomorphisms.

Lemma 10.40. *Let $\phi\colon G \to H$ be a homomorphism. Then ϕ is injective if and only if* $\ker\phi = \{1_G\}$.

Proof.

$\boxed{\implies}$ Suppose ϕ is injective. Since $\phi(1_G) = 1_H$, $1_G \in \ker\phi$ so $\{1_G\} \subset \ker\phi$.

Conversely, let $x \in \ker\phi$, so $\phi(x) = 1_H$. Then $\phi(x) = 1_H = \phi(1_G)$, so by injectivity $x = 1_G$. Hence $\ker\phi \subset \{1_G\}$, so $\ker\phi = \{1_G\}$.

$\boxed{\impliedby}$ Suppose $\ker\phi = \{1_G\}$. Suppose $\phi(a) = \phi(b)$, then $\phi(ab^{-1}) = \phi(a)\phi(b^{-1}) = \phi(a)\phi(a)^{-1} = 1_H$. Hence $ab^{-1} \in \ker\phi = \{1_G\}$, so $ab^{-1} = 1_G$ and thus $a = b$. Therefore ϕ is injective. □

Lemma 10.41. *Let $\phi\colon G \to H$ be an isomorphism. Then its inverse $\phi^{-1}\colon H \to G$ is an isomorphism.*

Proof. The inverse of a bijective map is bijective. Hence it suffices to show that $\phi^{-1}(x)\phi^{-1}(y) = \phi^{-1}(xy)$ for all $x, y \in H$.

Let $a = \phi^{-1}(x)$, $b = \phi^{-1}(y)$, $c = \phi^{-1}(xy)$; we will show that $ab = c$. Since ϕ is bijective, it suffices to show that $\phi(ab) = \phi(c)$.

Since ϕ is a homomorphism,
$$\phi(ab) = \phi(a)\phi(b) = xy = \phi(c).$$
□

10.2.3 Cosets

Definition 10.42 (Coset)**.** Let $H \leq G$. For $a \in G$, a ***left coset*** and ***right coset*** of H in G are

$$\begin{aligned} aH &:= \{ah \mid h \in H\} \\ Ha &:= \{ha \mid h \in H\} \end{aligned}$$

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

Example 10.43. Consider the subgroup $2\mathbb{Z} \leq \mathbb{Z}$. Then $6 + 2\mathbb{Z} = \{\text{all even numbers}\} = 0 + 2\mathbb{Z}$, and $1 + 2\mathbb{Z} = \{\text{all odd numbers}\} = 17 + 2\mathbb{Z}$.

Notation. We denote the set of (left) cosets by G/H.

In what will follow, the analogous results hold similarly for right cosets.

Lemma 10.44. *Let $H \leq G$. Then $aH = H$ if and only if $a \in H$.*

Proof.

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

$\boxed{\impliedby}$ 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 10.45. *Let $H \leq G$, $a, b \in G$. Then $aH = bH$ if and only if $a^{-1}b \in H$.*

Proof.

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

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

Proposition 10.46. *Let $H \leq G$. Then G/H forms a partition of G.*

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 $H \leq G$, $1 \in H$. Thus for all $a \in G$, $a = a1 \in aH$ 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 $a^{-1}b = h_1h_2^{-1} \in H$ and thus $aH = bH$.

$\square$

The next result shows that the left cosets of H partition G into equal-sized parts.

Lemma 10.47. *The cosets of H in G are the same size as H; that is, for all $a \in G$, $|aH| = |H|$.*

Proof. Consider the mapping

$$\begin{aligned} f\colon H &\to aH \\ h &\mapsto ah \end{aligned}$$

We will show that f is bijective.

- Let $h_1, h_2 \in H$, then

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

  so f is injective.

- Note that f is surjective by the definition of aH.

Since f is bijective, $|H| = |aH|$. $\square$

10.2.4 Lagrange's Theorem

Definition 10.48 (Index)**.** Let $H \leq G$. The ***index*** of H in G is the number of left cosets of H in G, denoted by $|G : H|$.

Then $|G| = |G : 1|$; that is, the order of G is the index of the trivial subgroup in G.

Theorem 10.49 (Lagrange's theorem)**.** *Let G be a finite group, $H \leq G$. Then $|H|$ divides $|G|$; in particular,*

$$|G| = |H|\,|G : H|. \tag{10.1}$$

(10.1) is known as the *counting formula*.

Proof. Suppose that there are $|G : H|$ left cosets in total. Since the left cosets partition G, and each coset has size $|H|$, we have

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

$\square$

Corollary 10.50. *The order of an element of a finite group divides the order of the group.*

Proof. Consider the subgroup generated by a, which has order $o(a)$. Then by Lagrange's theorem, $o(a)$ divides $|G|$. $\square$

Corollary 10.51. *For any finite group G and $a \in G$, $a^{|G|} = 1$.*

Proof. We know that $|G| = k\,o(a)$ for some $k \in \mathbb{N}$. Then $a^{|G|} = \left(a^{o(a)}\right)^k = 1^k = 1$. $\square$

A special case of the above result is Fermat's little theorem, by taking $G = (\mathbb{Z}/p\mathbb{Z})^\times$.

> If p is prime, and a is any integer, then
>
> $$a^p \equiv a \pmod{p}.$$

Corollary 10.52. *A group of prime order is cyclic.*

Proof. Let $|G| = p$ be prime. Let $a \in G$, $a \neq 1$. We will show that $G = \langle a \rangle$.

Since $o(a) \mid |G| = p$ and $o(a) > 1$, we must have $o(a) = p$. Notice that this is also the order of $\langle a \rangle$. Since G has order p, thus $\langle a \rangle = G$. $\square$

This corollary classifies groups of prime order p. They form one isomorphism class: the class of the cyclic groups of order p.

The next result is of great interest in number theory. The *Euler ϕ-function* $\phi(n)$ is defined for all positive integers as follows:

$$\phi(n) = \begin{cases} 1 & (n = 1) \\ \text{number of positive integers less than } n\text{, relatively prime to } n & (n > 1) \end{cases}$$

Theorem 10.53 (Euler)**.** *If n is a positive integer, and a is coprime to n, then*

$$a^{\phi(n)} \equiv 1 \pmod{n}.$$

10.2.5 Counting Principle

We generalise the notion of cosets, as defined earlier.

Definition 10.54. Let $H, K \leq G$. Define

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

Lemma 10.55. *Let $H, K \leq G$. Then $HK \leq G$ if and only if $HK = KH$.*

Proof.

$\boxed{\Longleftarrow}$ Suppose $HK = KH$; that is, if $h \in H$ and $k \in K$, then $hk = k_1 h_1$ for some $k_1 \in K, h_1 \in H$. We now show that HK is a subgroup of G:

(i) $1 \in H$ and $1 \in K$, so $1 \in HK$.

(ii) Let $x = hk \in HK$, $y = h'k' \in HK$. then

$$xy = hkh'k'.$$

Note that $kh' \in KH = HK$, so $kh' = h_2k_2$ for some $h_2 \in H, k_2 \in K$. Then

$$xy = h(h_2k_2)k' = (hh_2)(k_2k') \in HK.$$

Thus HK is closed.

(iii) Let $x \in HK$, then $x = hk$ for some $h \in H, k \in K$. Thus

$$x^{-1} = (hk)^{-1} = k^{-1}h^{-1} \in KH = HK,$$

so $x^{-1} \in HK$.

$\boxed{\Longrightarrow}$ Suppose $HK \leq G$.

- Let $x \in KH$, so $x = kh$ for some $k \in K, h \in H$. Then

$$x = kh = (h^{-1}k^{-1})^{-1} \in HK.$$

Thus $KH \subset HK$.

- Let $x \in HK$. Since $HK \leq G$, HK is closed under inverses, so $x^{-1} = hk \in HK$. Then

$$x = (x^{-1})^{-1} = (hk)^{-1} = k^{-1}h^{-1} \in KH.$$

Thus $HK \subset KH$.

Hence $HK = KH$. □

An interesting special case is the situation when G is an abelian group, for in that case trivially $HK = KH$. Thus as a consequence we have the following result.

Corollary 10.56. *Let $H, K \leq G$, where G is abelian. Then $HK \leq G$.*

Proposition 10.57. *If $H, K \leq G$ are finite groups, then*

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

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

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

□

10.2.6 Normal Subgroups, Quotient Groups

Definition 10.58 (Normal subgroup)**.** Let G be a group. We say $H \leq G$ is a ***normal subgroup*** of G, denoted by $H \lhd G$, if

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

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 10.59. *The following are equivalent.*

(i) $H \lhd G$.

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

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

Proof.

$\boxed{\text{(i)} \iff \text{(ii)}}$ First suppose $aH = Ha$ 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 $gxg^{-1} = h'gg^{-1} = h' \in H$.

Conversely suppose $ghg^{-1} \in H$ 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 $gH = Hg$.

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

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

□

Remark. We frequently use (ii) to check if a subgroup is a normal subgroup.

Lemma 10.60. *A subgroup of an abelian group is normal.*

Proof. Let G be abelian, $H \leq G$. For all $g \in G$, $h \in H$, we have $ghg^{-1} = gg^{-1}h = h \in H$. Thus H is normal. □

Lemma 10.61. *Every subgroup of index 2 is normal.*

Proof. Suppose $H \leq G$ has index 2. Then there are only two possible cosets, namely H and $G \setminus H$.

Since $1H = H1$ and cosets partition G, the other left coset and right coset must be $G \setminus H$. Hence all left cosets and right cosets are the same. □

Proposition 10.62. *A group of order 6 is either cyclic or dihedral.*

Proof. Let $|G| = 6$. We will show that either $G \cong C_6$ or $G \cong D_6$.

By Lagrange's theorem, the possible element orders are 1, 2, 3 and 6. If there exists $a \in G$ of order 6, then $G = \langle a \rangle \cong C_6$.

Otherwise, we can only have elements of orders 2 and 3 other than the identity. If G only has elements of order 2, the order must be a power of 2 (why), which is not the case. So there must be an element r of order 3. So $\langle r \rangle \lhd G$ as it has index 2. Now G must also have an element s of order 2 (why).

Since $\langle r \rangle$ is normal, we know that $srs^{-1} \in \langle r \rangle$. If $srs^{-1} = 1$, then $r = 1$, which is not true. If $srs^{-1} = r$, then $sr = rs$ and sr has order 6 (lcm of the orders of s and r), which was ruled out above. Otherwise if $srs^{-1} = r^2 = r^{-1}$, then G is dihedral by definition of the dihedral group. □

The (left) cosets of a group form a group, known as the *quotient group*.

Definition 10.63 (Quotient group)**.** Let G be a group, $H \lhd G$. Then the ***quotient group*** of G by H is the set of left cosets of H in G:

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

Remark. Quotient groups are not subgroups of G; they contain different kinds of elements. For example, $\mathbb{Z}/n\mathbb{Z} \cong C_n$ are finite, but all subgroups of $\mathbb{Z}$ infinite.

Lemma. *G/H is a group under the operation $aH * bH = (ab)H$.*

Proof. First show that the operation is well-defined; that is, if $aH = a'H$ and $bH = b'H$, we want to show that $aH * bH = a'H * b'H$.

We know that $a' = ak_1$ and $b' = bk_2$ for some $k_1, k_2 \in H$. Then $a'b' = ak_1bk_2$. We know that $b^{-1}k_1b \in H$. Let $b^{-1}k_1b = k_3$. Then $k_1b = bk_3$. So $a'b' = abk_3k_2 \in (ab)H$. So picking a different representative of the coset gives the same product.

If aH and bH are cosets, then $(ab)H$ is also a coset, so the operation is closed.

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

$$(aH)(bHcH) = (aH)(bcH) = a(bc)H = (ab)cH = (aHbH)cH$$

so the operation is associative.

(ii) The identity is $1H = \{1h \mid h \in H\} = \{h \mid h \in H\} = H$.

(iii) The inverse of aH is $a^{-1}H$, since

$$(aH)(a^{-1}H) = aa^{-1}H = H \implies (aH)^{-1} = a^{-1}H.$$

□

Example 10.64 (Modular arithmetic)**.** Fix $n \in \mathbb{Z}^+$. Evidently $n\mathbb{Z}$ is a subgroup of $\mathbb{Z}$. Then the quotient group $\mathbb{Z}/n\mathbb{Z}$ consists of cosets of the form

$$n\mathbb{Z}, 1 + n\mathbb{Z}, 2 + n\mathbb{Z}, \dots, n - 1 + n\mathbb{Z}.$$

If we consider each coset as an equivalence class, we write

$$\mathbb{Z}/n\mathbb{Z} = \{[0], [1], \dots, [n-1]\}.$$

Addition on $\mathbb{Z}/n\mathbb{Z}$ is defined as

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

The next result concerns the order of the quotient group.

Lemma 10.65. *Let G be a finite group, $H \lhd G$. Then*

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

Proof. Since G/H has as its elements the left cosets of H in G, and there are precisely $|G : H|$ such cosets, the first equality holds.

The second equality holds by Lagrange's theorem. □

We now define a *canonical* homomorphism ("natural" map) from a group to its quotient group.

Definition 10.66 (Quotient map)**.** Let $H \lhd G$. The ***quotient map*** is the map

$$\begin{aligned}\pi\colon G &\to G/H\\ a &\mapsto aH\end{aligned}$$

Lemma 10.67. *Quotient maps are surjective homomorphisms.*

Proof. Let $\pi\colon G \to G/H$ which maps $a \mapsto aH$ be a quotient map.

- For all $a, b \in G$,
$$\pi(ab) = (ab)H = (aH)(bH) = \pi(a)\pi(b).$$
Thus π is a homomorphism.
- For all $aH \in G/H$, $\pi(a) = aK$. Thus π is surjective.

□

The next result provides a characterisation of normal subgroups.

Lemma 10.68. *$H \lhd G$ if and only if H is the kernel of some homomorphism.*

Proof.

$\boxed{\Longleftarrow}$ Suppose $H = \ker\phi$ for some homomorphism $\phi\colon G \to G'$.
Let $g \in G$, $h \in H$. Then
$$\phi(ghg^{-1}) = \phi(g)\phi(h)\phi(g)^{-1} = \phi(h) = 1.$$
Thus $ghg^{-1} \in \ker\phi = H$.

$\boxed{\Longrightarrow}$ The kernel of the quotient map is H itself:

$$\ker\pi = \{a \in G \mid aH = H\} = \{a \in G \mid a \in H\} = H.$$

□

10.2.7 Isomorphism Theorems

In this section, we will prove several isomorphism theorems.

Theorem 10.69 (First isomorphism theorem)**.** *Let $\phi\colon G \to H$ be a homomorphism. Then*

$$G/\ker\phi \cong \operatorname{im}\phi. \tag{10.2}$$

Proof. For ease of notation, denote $K = \ker\phi$. Consider the mapping

$$\begin{aligned}\theta\colon G/K &\to \operatorname{im}\phi\\ \forall x \in G, \quad xK &\mapsto \phi(x)\end{aligned}$$

We claim that θ is an isomorphism.

1. We check that θ is well-defined. Let $x, y \in G$, suppose $xK = yK$. Then

$$\begin{aligned} & xK = yK \\ \iff & x^{-1}y \in K \\ \iff & \phi(x^{-1}y) = 1_H \\ \iff & \phi(x)^{-1}\phi(y) = 1_H \\ \iff & \phi(x) = \phi(y) \end{aligned}$$

2. We show that θ is a homomorphism: for all $x, y \in G$,

$$\theta(xKyK) = \theta(xyK) = \phi(xy) = \phi(x)\phi(y) = \theta(xK)\theta(yK).$$

3. We show that θ is bijective:

   - θ is injective since

$$\theta(xK) = \theta(yK) \implies \phi(x) = \phi(y) \implies xK = yK.$$

   - θ is surjective, since

$$\operatorname{im}\theta = \{\theta(xK) \mid x \in G\} = \{\phi(x) \mid x \in G\} = \operatorname{im}\phi.$$

□

Corollary 10.70. *Any cyclic group is isomorphic to either $\mathbb{Z}$ or $\mathbb{Z}/n\mathbb{Z}$ for some $n \in \mathbb{N}$.*

Proof. Let $G = \langle g \rangle$ for some $g \in G$. Define the mapping

$$\begin{aligned} \phi\colon \mathbb{Z} &\to G \\ m &\mapsto g^m \end{aligned}$$

We claim that ϕ is a surjective homomorphism.

1. ϕ is a homomorphism, since $\phi(m_1 + m_2) = g^{m_1+m_2} = g^{m_1}g^{m_2} = \phi(m_1)\phi(m_2)$.
2. ϕ is surjective, since G is by definition all g^m for all m.

By surjectivity, $\operatorname{im}\phi = G$. We know that $\ker\phi \lhd \mathbb{Z}$. We have the following possibilities for the kernel:

Case 1: $\ker\phi = \{1\}$ This implies ϕ is injective, so ϕ is an isomorphism. Hence $G \cong \mathbb{Z}$.

Case 2: $\ker\phi = \mathbb{Z}$ By the first isomorphism theorem, $G \cong \mathbb{Z}/\mathbb{Z} = \{1\} = C_1$.

Case 3: $\ker\phi = n\mathbb{Z}$ (Since these are the only remaining proper subgroups of $\mathbb{Z}$.) By the first isomorphism theorem, $G \cong \mathbb{Z}/n\mathbb{Z}$.

□

Example 10.71 (Circle group)**.** Consider the subgroup $(\mathbb{Z},+)$ of $(\mathbb{R},+)$. The quotient group $\mathbb{R}/\mathbb{Z}$ is called the *circle group*.

Define a congruence relation on $\mathbb{R}$:

$$x \sim y \iff x - y \in \mathbb{Z}.$$

If $x \sim y$, we say $x, y \in \mathbb{R}$ are *congruent* mod $\mathbb{Z}$, and denote $x \equiv y \pmod{\mathbb{Z}}$. This congruence is an equivalence relation, and the congruence classes are precisely the cosets of $\mathbb{Z}$ in $\mathbb{R}$.

If $x \equiv y \pmod{\mathbb{Z}}$, then $e^{2\pi i x} = e^{2\pi i y}$, and conversely. Thus the map

$$\begin{aligned} f\colon \mathbb{R}/\mathbb{Z} &\to \mathbb{T} \\ x &\mapsto e^{2\pi i x} \end{aligned}$$

is an isomorphism, where $\mathbb{T} = \{z \in \mathbb{C} \mid |z| = 1\}$ is the multiplicative group of complex numbers having absolute value 1.

Remark. $2\pi x$ can be considered as the angle measured from the positive real axis of $\mathbb{C}$.

Example 10.72. Let $\mathbb{C}^\times$ be the multiplicative group of non-zero complex numbers, and $\mathbb{R}^+$ the multiplicative group of positive real numbers. Given a complex number $z \neq 0$, we can write

$$z = ru,$$

where $r \in \mathbb{R}^+$, and u has absolute value 1. (Let $u = z/|z|$.) Such an expression is uniquely determined, and the map

$$\begin{aligned} f\colon \mathbb{C}^\times &\to \mathbb{T} \\ z &\mapsto \frac{z}{|z|} \end{aligned}$$

is a homomorphism. Since $\ker f = \mathbb{R}^+$ and $\operatorname{im} f = \mathbb{T}$ (by surjectivity), by the first isomorphism theorem, $\mathbb{C}^\times/\mathbb{R}^+$ is isomorphic to $\mathbb{T}$.

Theorem 10.73 (Second isomorphism theorem)**.** *Let $H \leq G$, $K \lhd G$. Then*

$$HK/K \cong H/(H \cap K). \tag{10.3}$$

We first prove a few preliminary results.

Lemma. *Let $H \leq G$, $K \lhd G$. Then*

(i) $HK \leq G$;

(ii) $K \lhd HK$;

(iii) $H \cap K \lhd H$.

Proof.

(i) Since $1 \in H$ and $1 \in K$, we have $1 \in HK$, so $HK \neq \emptyset$. Let $hk, h'k' \in HK$. Then

$$h'k'(hk)^{-1} = h'k'k^{-1}h^{-1} = (\underbrace{h'h^{-1}}_{\in H})(\underbrace{hk'k^{-1}h^{-1}}_{\in K,\text{ by normality}}) \in HK.$$

By the subgroup criterion, $HK \leq G$.

(ii)

(iii) Since the intersection of subgroups is a subgroup, $H \cap K$ is a subgroup of N. It remains to be shown that $H \cap K$ is normal in H.

Let $h \in H$, $x \in H \cap K$. We will show that $hxh^{-1} \in H \cap K$.

$H \leq G$ and $h \in H$ imply $h \in G$. Since $K \lhd G$, $x \in H \cap K$ and $h \in H$ imply $hxh^{-1} \in H$.

□

We are now ready to prove the theorem.

Proof. Define the map

$$\begin{aligned} \phi\colon H &\to G/K \\ h &\mapsto hK \end{aligned}$$

This is easily seen to be a homomorphism. Let $x, y \in H$, then

$$\phi(xy) = (xy)K = (xK)(yK) = \phi(x)\phi(y).$$

The kernel and image of ϕ are

$$\begin{aligned} \ker\phi &= \{h \in H \mid hK = K\} = \{h \in H \mid h \in K\} = H \cap K, \\ \operatorname{im}\phi &= \{\phi(h) \mid h \in H\} = \{hK \mid h \in H\} = HK/K. \end{aligned}$$

Hence the desired result follows from the first isomorphism theorem. □

Theorem 10.74 (Third isomorphism theorem). *Let $H, K \lhd G$, $H \leq K$. Then*

$$(G/H)/(K/H) \cong G/K. \tag{10.4}$$

Lemma. $K/H \lhd G/H$.

Proof. We first show $K/H \leq G/H$:

(i) The identity of G/H is H, which is also the identity of K/H, since $1 \in K$.

(ii) Let $aH, bH \in K/H$. Since $ab \in K$ for all $a, b \in K$, we have $(aH)(bH) = (ab)H \in K/H$.

(iii) Let $aH \in K/H$. Since $a^{-1} \in K$, we have $a^{-1}H \in K/H$.

To show normality, let $gH \in G/H$, $kH \in K/H$. Then $(gH)(kH)(gH)^{-1} = (gkg^{-1})H$. Since $K \lhd G$, $gkg^{-1} \in K$. Thus $(gkg^{-1})H \in K/H$. □

We can now prove the theorem.

Proof. Define the *canonical* homomorphism

$$\begin{aligned} \phi\colon G/H &\to G/K \\ gH &\mapsto gK \end{aligned}$$

We claim that ϕ is a surjective homomorphism.

1. We check that ϕ is well-defined: If $gH = g'H$, then $g^{-1}g' \in H$. Since $H \subset K$, $g^{-1}g' \in K$. Thus $gK = g'K$.
2. ϕ is a homomorphism:
$$\phi(gH \cdot g'H) = \phi(gg'H) = gg'K = (gK)(g'K) = \phi(gH)\phi(g'H).$$
3. ϕ is clearly surjective, since any coset gK is the image $\phi(gH)$.

The kernel and image of ϕ are

$$\begin{aligned} \ker\phi &= \{gH \mid gK = K\} = \{gH \mid g \in K\} = K/H, \\ \operatorname{im}\phi &= G/K \quad \text{by surjectivity.} \end{aligned}$$

Hence the conclusion follows from the first isomorphism theorem. □

We now discuss an isomorphism theorem concerning pre-images of groups.

Theorem 10.75. *Let $\phi\colon G \to G'$ be a surjective homomorphism. Let $H' \lhd G'$ and $H = \phi^{-1}(H')$. Then*

$$G/H \cong G'/H'. \tag{10.5}$$

Lemma. $H = \phi^{-1}(H') \lhd G$.

Proof. We first show $H \le G$.

(i) Since $H' \le G'$, $1_{G'} \in H'$. Then $\phi(1_G) = 1_{G'} \in H'$, so $1_G \in \phi^{-1}(H')$.

(ii) Let $a, b \in H$. Then $\phi(a), \phi(b) \in H'$. By closure, $\phi(a)\phi(b) = \phi(ab) \in H'$, so $ab \in \phi^{-1}(H')$.

(iii) Let $a \in H$. Then $\phi(a) \in H'$. Since H' is closed under inverses, $\phi(a)^{-1} = \phi(a^{-1}) \in H'$, so $a^{-1} \in \phi^{-1}(H')$.

To show normality, let $g \in G$, $h \in H$, then $\phi(h) \in H'$. Since $\phi(ghg^{-1}) = \phi(g)\phi(h)\phi(g)^{-1}$, where $\phi(g) \in G'$ and $\phi(h) \in H'$, by normality $\phi(ghg^{-1}) \in H'$. Thus $ghg^{-1} \in \phi^{-1}(H')$. □

We now prove the theorem.

Proof. Consider the map

$$\begin{aligned}\psi\colon G &\to G'/H' \\ g &\mapsto \phi(g)H'\end{aligned}$$

Note that ψ can also be described as the composite map $\psi = \pi \circ \phi$:

$$G \xrightarrow{\phi} G' \xrightarrow{\pi} G'/H'$$

where $\pi\colon G' \to G'/H'$ is the quotient map.

We claim that ψ is a surjective homomorphism.

1. The composition of homomorphisms is a homomorphism, so ψ is a homomorphism.

2. The composition of surjective maps is surjective, so ψ is surjective.

The kernel and image of ψ are

$$\begin{aligned}\ker\psi &= \{g \in G \mid \psi(g) = H'\} = \{g \in G \mid \phi(g)H' = H'\} \\ &= \{g \in G \mid \phi(g) \in H'\} = \phi^{-1}(H') = H, \\ \operatorname{im}\psi &= G'/H' \quad \text{by surjectivity.}\end{aligned}$$

Hence the desired conclusion follows from the first isomorphism theorem. □

Theorem 10.76 (Fourth isomorphism theorem)**.** *Let $N \lhd G$. The canonical projection homomorphism $G \to G/N$ defines a bijective correspondence between the set of subgroups of G containing N and the set of (all) subgroups of G/N. Under this correspondence normal subgroups correspond to normal subgroups.*

10.2.8 Solvable Groups

Definition 10.77 (Solvable group)**.** A group G is ***solvable*** if there exists a sequence of normal subgroups (known as a *composition series*)

$$G = H_0 \rhd H_1 \rhd \cdots \rhd H_n = \{1\},$$

such that H_i/H_{i+1} (this quotient is called a *composition factor*) is abelian.

Proposition 10.78. *Let $K \lhd G$. If K and G/K are solvable, then G is solvable.*

Proof. By definition, and the assumption that K is solvable, it suffices to prove the existence of a sequence of normal subgroups

$$G = H_0 \rhd H_1 \rhd \cdots \rhd H_n = K$$

such that H_i/H_{i+1} is abelian. Let $\bar{G} = G/K$. By assumption, there exists a sequence of normal subgroups

$$\bar{G} = \bar{H}_0 \triangleright \bar{H}_1 \triangleright \cdots \triangleright \bar{H}_n = \{\bar{1}\}$$

such that $\bar{H}_i/\bar{H}_{i+1}$ is abelian.

Consider the quotient map

$$\begin{aligned} \pi\colon G &\to \bar{G} \\ g &\mapsto gK \end{aligned}$$

and let $H_i = \pi^{-1}(\bar{H}_i)$.

Claim. These H_i comprise a composition series for G.

Since the quotient map π is a surjective homomorphism, by 10.75, we have an isomorphism

$$H_i/H_{i+1} \cong \bar{H}_i/\bar{H}_{i+1}$$

and $K = \pi^{-1}(\bar{H}_n)$, so we have found the sequence of subgroups of G as we wanted, proving the result. □

Proposition 10.79. *Let $H \leq G$. If G is solvable, then H is solvable.*

Proof. Since G is solvable, there exists a sequence of normal subgroups

$$G = H_0 \triangleright H_1 \triangleright \cdots \triangleright H_n = \{1\}$$

such that H_i/H_{i+1} is abelian. Consider

$$H = H \cap H_0 \triangleright \cdots \triangleright H \cap H_n = \{1\}.$$

Now $(H \cap H_i) \cap H_{i+1} = H \cap H_{i+1}$. Since $H_i \triangleright H_{i+1}$, this implies $H \cap H_i \triangleright H \cap H_{i+1}$ by looking at the conjugacy relationship. By 10.75,

$$(H \cap H_i)/(H \cap H_{i+1}) \cong (H \cap H_i)H_{i+1}/H_{i+1} \leq H_i/H_{i+1}$$

which is abelian, by the following lemma:

Lemma. *Let G be abelian, and $\phi\colon G \to G'$ be a surjective homomorphism. Then G' is abelian.*

Proof. Let $x, y \in G'$. We can write $x = \phi(a)$ and $y = g(b)$, by surjectivity of ϕ. Thus $xy = \phi(a)\phi(b) = \phi(ab) = \phi(ba) = \phi(b)\phi(a) = yx$. □

□

Proposition 10.80. *If G is solvable and $\phi\colon G \to G'$ is a surjective homomorphism, then G' is solvable.*

Proof. Since G is solvable, there exists a sequence of normal subgroups

$$G = H_0 \triangleright H_1 \triangleright \cdots \triangleright H_n = \{1\}$$

such that H_i/H_{i+1} is abelian. Let $H_i' = \phi(H_i)$. Clearly $H_0' = \phi(G) = G'$ (by surjectivity), and $H_n' = \phi(\{1\}) = \{1'\}$.

Let $y \in H_{i+1}'$, and $y = \phi(a)$ for some $a \in H_{i+1}$. Since $H_{i+1} \subset H_i$, $a \in H_i$ and $y \in H_i'$. Thus $H_{i+1}' \leq H_i'$. To show normality, let $x = f(b) \in H_i'$ where $b \in H_i$. Then $xyx^{-1} = \phi(bab^{-1})$. Since $H_{i+1} \triangleleft H_i$, $bab^{-1} \in H_i$, so $\phi(bab^{-1}) \in H_{i+1}'$. Thus $H_{i+1}' \triangleleft H_i'$.

Consider the map

$$\begin{aligned} \psi\colon H_i/H_{i+1} &\to H_i'/H_{i+1}' \\ hH_{i+1} &\mapsto \phi(h)H_{i+1}' \end{aligned}$$

We claim that ψ is a surjective homomorphism.

1. We first check that ψ is well-defined. If $h_1H_{i+1} = h_2H_{i+1}$, then $h_1h_2^{-1} \in H_{i+1}$, so $\phi(h_1)\phi(h_2)^{-1} \in H_{i+1}'$. Thus $\phi(h_1)H_{i+1}' = \phi(h_2)H_{i+1}'$.
2. ψ is a homomorphism, since

$$\psi(h_1H_{i+1}h_2H_{i+1}) = \psi(h_1h_2H_{i+1}) = \phi(h_1h_2)H_{i+1}' = \phi(h_1)H_{i+1}'\phi(h_2)H_{i+1}' = \psi(h_1H_{i+1})\psi(h_2H_{i+1}).$$

3. If $h'H_{i+1}' \in H_i'/H_{i+1}'$, then since $h' \in H_i'$, $h' = \phi(h)$ for some $h \in H_i$ and $\psi(hH_{i+1}) = h'H_{i+1}'$. Thus ψ is surjective.

By the lemma in the proof of the previous result, H_i'/H_{i+1}' is abelian. □

In a sense, the objects having the "simplest" structure are the building blocks for the more complicated objects. For groups, these are the *simple groups*. All finite simple groups have been classified under the ATLAS of Finite Groups.

Definition 10.81 (Simple group)**.** A group is ***simple*** if it has no non-trivial proper normal subgroup

That is, G is simple if the only normal subgroups are $\{1\}$ and G.

Example 10.82. For prime p, the cyclic group C_p is simple, since it has no proper subgroups at all, let alone normal ones.

Let G be a finite group. Then one can find a sequence of normal subgroups

$$G = H_0 \triangleright H_1 \triangleright \cdots \triangleright H_n = \{1\},$$

such that H_i/H_{i+1} is simple. (This follows from a previous result and the third isomorphism theorem.)

A group may have more than one composition series. However, the Jordan–Hölder theorem states that any two composition series of a given group are equivalent. That is, they have the same composition length and the same composition factors, up to permutation and isomorphism.

Theorem 10.83 (Jordan–Hölder theorem)**.** *Let G be a finite group. Consider two composition series*

$$G = H_0 \triangleright H_1 \triangleright \cdots \triangleright H_n = \{1\}$$

$$G = K_0 \triangleright K_1 \triangleright \cdots \triangleright K_m = \{1\}$$

where each composition factor is simple. Then $n = m$, and the list of composition factors is unique up to permutation.

10.3 Symmetric Group

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

Lemma. $\mathrm{Sym}(S)$ *forms a group under function composition* $\circ$.

We call $\mathrm{Sym}(S)$ the ***symmetric group*** on S.

Proof. If $\sigma\colon S \to S$ and $\tau\colon S \to S$ are bijections, then the composition $\sigma \circ \tau$ is a bijection from S to S. Thus $\circ$ is a binary operation on $\mathrm{Sym}(S)$.

(i) Composition of functions is associative, so $\circ$ is associative.

(ii) The identity of $\mathrm{Sym}(S)$ is the identity map.

(iii) Every bijection has a bijective inverse.

$\square$

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

Lemma 10.84. $|S_n| = n!$

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

There are n choices for $\sigma(1)$, $n-1$ choices for $\sigma(2)$, ..., 1 choice for $\sigma(n)$. Hence $|S_n| = n(n-1)\cdots 1 = n!$. $\square$

There are two ways to denote a permutation (an element of the symmetric group). The first is the *two row notation*: if $\sigma \in S_n$, we write

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & \cdots & n \\ \sigma(1) & \sigma(2) & \sigma(3) & \cdots & \sigma(n) \end{pmatrix}.$$

Lemma 10.85. *If* $|S| \geq 3$, *then* $\mathrm{Sym}(S)$ *is non-abelian.*

Proof. S_3 consists of

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix},$$
$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}.$$

By considering the composition of any two of the above permutations, we see that they do not commute. Thus S_3 is not abelian.

For $n \geq 3$, since we can view S_3 as a subgroup of S_n by fixing $4, 5, 6, \ldots, n$, it follows that S_n is not abelian. $\square$

Theorem 10.86 (Cayley's theorem)**.** *Every finite group is isomorphic to some subgroup of some symmetric group.*

Proof. Let G be a finite group. For $g \in G$, $\sigma_g(h) = gh$ defines a permutation on G, and $\sigma_{g_1}\sigma_{g_2}(h) = \sigma_{g_1}(g_2 h) = g_1 g_2 h = \sigma_{g_1 g_2}(h)$. $\square$

A ***transposition*** τ is a permutation which interchanges two numbers and leaves the others fixed, i.e., there exist distinct $i, j \in J_n$ such that $\tau(i) = j$, $\tau(j) = i$, and $\tau(k) = k$ if $k \neq i$, $k \neq j$.

One sees at once that if τ is a transposition, then $\tau^{-1} = \tau$ and $\tau^2 = \mathrm{id}$. In particular, the inverse of a transposition is a transposition. We shall prove that the transpositions generate S_n.

Proposition 10.87. *Every permutation in S_n can be expressed as a product of transpositions.*

Proof. Induct on n. For $n = 1$, there is nothing to prove since there is only one element. Let $n > 1$ and assume the assertion proved for $n - 1$.

Let $\sigma \in S_n$. Let $\sigma(n) = k$. Let $\tau \in S_n$ be such that $\tau(k) = n$, $\tau(n) = k$. Then $\tau\sigma$ is a permutation such that

$$\tau\sigma(n) = \tau(k) = n.$$

In other words, $\tau\sigma$ leaves n fixed. We may therefore view $\tau\sigma$ as a permutation of J_{n-1}, and by induction, there exist transpositions $\tau_1, \ldots, \tau_s \in S_{n-1}$, leaving n fixed, such that

$$\tau\sigma = \tau_1 \cdots \tau_s.$$

We now write

$$\sigma = \tau^{-1}\tau_1 \cdots \tau_s,$$

thereby proving our proposition. $\square$

The two row notation is clumsy to write and wastes a lot of space. Hence we often use the *cycle notation*. Let $a_1, \ldots, a_k$ be distinct integers in J_n. By the symbol

$$(a_1 \cdots a_k)$$

we shall mean the permutation σ such that

$$\sigma(a_1) = a_2, \quad \sigma(a_2) = a_3, \quad \ldots, \quad \sigma(a_k) = a_1,$$

and σ leaves all other integers fixed. We call $(a_1, a_2, \ldots, a_k)$ a ***k-cycle***. (Thus a transposition is a 2-cycle.)

Example 10.88. (132) denotes the permutation σ such that $\sigma(1) = 3$, $\sigma(3) = 2$, $\sigma(2) = 1$, and σ leaves all other integers fixed.

If $\sigma = (a_1 \cdots a_k)$ is a cycle, then one verifies at once that the inverse σ^{-1} is also a cycle, and

$$\sigma^{-1} = (a_k \cdots a_1).$$

A product of cycles is easily determined, as illustrated by the following example.

Example 10.89. $(132)(34) = (2134)$. One sees this using the definition: If $\sigma = (132)$ and $\tau = (34)$. then

$$\begin{aligned}\sigma(\tau(3)) &= \sigma(4) = 4,\\ \sigma(\tau(4)) &= \sigma(3) = 2,\\ \sigma(\tau(2)) &= \sigma(2) = 1,\\ \sigma(\tau(1)) &= \sigma(1) = 3.\end{aligned}$$

Two cycles are said to be *disjoint* if no number appears in both cycles.

Lemma 10.90. *Disjoint cycles commute.*

Proof. Suppose $\sigma, \tau \in S_n$ are disjoint cycles. We will show that $\sigma(\tau(a)) = \tau(\sigma(a))$.

If a is in neither of σ and τ, then $\sigma(\tau(a)) = \tau(\sigma(a)) = a$.

Otherwise, WLOG assume that a is in τ but not in σ. Then $\tau(a) \in \tau$ and thus $\tau(a) \notin \sigma$. Thus $\sigma(a) = a$ and $\sigma(\tau(a)) = \tau(a)$. Hence we have $\sigma(\tau(a)) = \tau(\sigma(a)) = \tau(a)$.

Therefore τ and σ commute. □

We shall prove that for $n \geq 5$, the group S_n is not solvable. We need some prehminaries.

Lemma. *Let $H \lhd G$. Then G/H is abelian if and only if H contains all elements of the form $xyx^{-1}y^{-1}$, where $x, y \in G$.*

Proof.

$\boxed{\Longrightarrow}$ Consider the quotient map $\pi\colon G \to G/H$. Suppose G/H is abelian. For any $x, y \in G$, we have

$$\pi(xyx^{-1}y^{-1}) = \pi(x)\pi(y)\pi(x)^{-1}\pi(y)^{-1} = H,$$

since G/H is abelian. Hence $xyx^{-1}y^{-1} \in H$.

$\boxed{\Longleftarrow}$ Suppose H contains all elements of the form $xyx^{-1}y^{-1}$, where $x, y \in G$.

Let $\bar{x}, \bar{y} \in G/H$. Since π is surjective, there exist $x, y \in G$ such that $\bar{x} = \pi(x)$, $\bar{y} = \pi(y)$.

Let $\bar{1}$ denote the identity of G/H, and 1 denote the identity of G. Then

$$\bar{1} = \pi(1) = \pi(xyx^{-1}y^{-1}) = \pi(x)\pi(y)\pi(x)^{-1}\pi(y)^{-1} = \bar{x}\bar{y}\bar{x}^{-1}\bar{y}^{-1}.$$

Multiplying by $\bar{y}$ and then $\bar{x}$ on the right, we find

$$\bar{y}\bar{x} = \bar{x}\bar{y}.$$

Hence G/H is abelian. □

Theorem 10.91. *If $n \geq 5$, then S_n is not solvable.*

Proof. We need the following result.

Lemma. *Let $N \triangleleft H \leq S_n$. If H contains every 3-cycle and H/N is abelian, then N contains every 3-cycle.*

Proof. Let i, j, k, r, s be distinct integers in $\{1, \dots, n\}$, and let

$$\sigma = (ijk), \quad \tau = (krs).$$

Then

$$\begin{aligned} \sigma\tau\sigma^{-1}\tau^{-1} &= (ijk)(krs)(kji)(srk) \\ &= (rki). \end{aligned}$$

Since H contains every 3-cycle, $\sigma, \tau \in H$. Since H/N is abelian, by the above lemma, N contains all elements of the form $\sigma\tau\sigma^{-1}\tau^{-1}$. Thus $\sigma\tau\sigma^{-1}\tau^{-1} \in N$.

Since the choice of i, j, k, r, s was arbitrary, this implies $\sigma\tau\sigma^{-1}\tau^{-1}$ is an arbitrary 3-cycle. Hence N contains every 3-cycle. □

S_n contains all 3-cycles. Thus by induction on the previous lemma, a composition series

$$S_n = H_0 \triangleright H_1 \triangleright H_2 \triangleright \cdots \triangleright H_n$$

must have H_n containing all 3-cycles; thus $H_n \neq \{1\}$ (since the trivial subgroup does not contain any 3-cycles). □

10.4 Group Actions

10.4.1 Group Acting on Sets

We move now, from thinking of groups in their own right, to thinking of how groups can move sets around; for example, S_n permutes $\{1, 2, \ldots, n\}$, and matrix groups move vectors. This leads to the notion of a *group action*.

Definition 10.92 (Group action)**.** Let G be a group, S be a set. An ***action*** of G on S is a map $G \times S \to S$ satisfying the following properties:

(i) $g(hs) = (gh)s$ for all $g, h \in G$, $s \in S$; (associativity)

(ii) $1s = s$ for all $s \in S$. (identity)

Notation. If the group action $\cdot\colon G \times S \to S$ is not clear from context, we write $g \cdot s$ instead of gs.

Intuitively, an action of G on S means that every element $g \in G$ acts as a permutation on S in a manner consistent with the group operations in G.

There is another way of defining group actions, which is arguably a better way of thinking about group actions.

Lemma 10.93. *An action of G on S is equivalent to a homomorphism $\phi\colon G \to \mathrm{Sym}(S)$.*

Note that the statement by itself is useless, since it does not tell us how to translate between the homomorphism and a group action. The important part is the proof.

Proof. Let $(G, *)$ be a group. Let $\cdot\colon G \times S \to S$ be an action. Define

$$\begin{aligned} \phi\colon G &\to \mathrm{Sym}(S) \\ g &\mapsto (g \cdot *\colon S \to S) \end{aligned}$$

This is indeed a permutation □

Example 10.94.

- The *trivial action* is $gs = s$.
- S_n acts on $\{1, \ldots, n\}$ by permutation.
- D_{2n} acts on the vertices of a regular n-gon (or the set $\{1, \ldots, n\}$).

Definition 10.95 (Kernel of action)**.** Let G act on S. The ***kernel*** of the action is

$$\{g \in G \mid gs = s \; \forall s \in S\}.$$

We say an action is *faithful* if the kernel is just $\{1\}$.

10.4.2 Orbits and Stabilisers

Definition 10.96 (Orbit)**.** Let G act on S. The ***orbit*** of $s \in S$ is

$$Gs := \{gs \mid g \in G\}.$$

Intuitively, it is the elements that s can possibly get mapped to. An element $x \in G_s$ in an orbit is called a *representative* of the orbit, and we say that x *represents* the orbit.

In what follows, the formalism of orbits is similar to the formalism of cosets.

Lemma 10.97. *If $s \in G$, then $Gs = G$.*

Proof. Suppose $s \in G$. Then

$$\begin{aligned} x \in Gs &\iff x = gs,\ g \in G \\ &\iff x \in G \end{aligned}$$

□

The next result states when two orbits are equal.

Lemma 10.98. *If $t \in Gs$, then $Gt = Gs$.*

Proof. Suppose $t \in G_s$. Then $t = gs$ for some $g \in G$. Then

$$Gt = G(gs) = \{h(gs) \mid h \in G\} = \{(hg)s \mid h \in G\} = (Gg)s = Gs.$$

□

Lemma 10.99. *Suppose the group G acts on S. Then the orbits of the action partition S.*

We need to prove the following:

(i) The orbits are non-empty.

(ii) The union of orbits is S.

(iii) Two orbits are either equal or disjoint.

Proof.

(i) For every $s \in S$, $1s = s$ so $s \in Gs$. Hence every s is in some orbit.

(ii)

(iii) Let $x \in Gs$ and $x \in Gt$. Then $x = g_1 s = g_2 t$ for some $g_1, g_2 \in G$. Thus

$$Gs = Gg_1 s = Gg_2 t = Gt.$$

□

Hence S is a disjoint union of the distinct orbits, and we can write

$$S = \bigcup_{i \in I} G s_i$$

where I is some indexing set, and the s_i represent distinct orbits.

We say an action G on S is *transitive*, if for all $s \in S$, $Gs = S$. This means that we can reach any element from any element.

Example 10.100 (Left regular action)**.** Any group G acts on itself by left multiplication:

$$g \cdot s = gs.$$

This action is faithful and transitive.

Proof. If $g, s \in G$, then $g \cdot s = gs \in G$, so the operation is closed. We now show this is an action.

(i) For all $a \in G$, $1 \cdot a = 1a = a$ by definition of a group.

(ii) For all $g, h \in G$ and $a \in G$, $g(ha) = (gh)a$ by associativity.

To show that it is faithful, we want to show that for all $a \in A$, $g \cdot a = a$ implies $g = 1$; but this follows directly from the uniqueness of identity of the group G.

To show that it is transitive, for all $x, y \in G$, then $(yx^{-1}) \cdot x = y$. Thus any x can be sent to any y. □

Definition 10.101 (Stabiliser)**.** Let G act on S. The ***stabiliser*** of $s \in S$ is

$$G_s := \{g \in G \mid gs = s\}.$$

Intuitively, it is the elements in G that leave s unchanged.

Lemma. $G_s \leq G$.

Proof. Apply the subgroup criterion.

(i) By definition, $1s = s$, so $1 \in G_s$. Thus G_s is non-empty.

(ii) Let $g, h \in G_s$. Then $(gh^{-1})s = g(h^{-1}s) = gs = s$. Thus $gh^{-1} \in G_s$.

□

Since G_s is a subgroup of G, we can consider cosets of G_s in G.

Lemma 10.102. *Let G act on S. If $g, h \in G$ are in the same coset of G_s, then $gs = hs$.*

Proof. If $g, h \in G$ are in the same coset of G_s, then we can write $h = gk$ for some $k \in G_s$. Then

$$hs = (gk)s = g(ks) = gs.$$

□

Theorem 10.103 (Orbit–stabiliser theorem). *Let G act on S, and let $s \in S$. Then there exists a bijection between Gs and cosets of G_s in G. In particular, if G is finite, then*

$$|G| = |Gs|\,|G_s|. \tag{10.6}$$

Proof. We biject the cosets of G_s in G with elements in the Gs. Consider the mapping

$$\begin{aligned} \theta\colon G/G_s &\to Gs \\ gG_s &\mapsto gs \end{aligned}$$

We claim that θ is a bijection.

1. We check that θ is well-defined: Suppose $gG_s = hG_s$. Then $h = gk$ for some $k \in G_s$. Thus $\theta(hG_s) = hs = (gk)s = g(ks) = gs = \theta(gG_s)$.
2. θ is surjective: Let $x \in Gs$. Then there exists $g \in G$ such that $x = gs$. Thus $\theta(gG_s) = gs = x$.
3. θ is injective: Suppose $gs = hs$. Then $h^{-1}gs = s$, so $h^{-1}g \in G_s$. Thus $h^{-1}gG_s = G_s$, which implies $gG_s = hG_s$.

Since θ is a bijection, we have

$$|G/G_s| = |Gs|.$$

Then the result follows from Lagrange's theorem. □

An immediate corollary is a formula for the size of an orbit:

$$|Gs| = |G : G_s|.$$

Suppose S is a finite set. Then we get a decomposition of the order of S as a sum or orders of orbits, which we call the *orbit decomposition formula*:

$$|S| = \sum_{i=1}^{n} |G : G_{s_i}|. \tag{10.7}$$

An important application of the orbit–stabiliser theorem is determining group sizes. To find the order of the symmetry group of, say, a pyramid, we find something for it to act on, pick a favorite element, and find the orbit and stabiliser sizes.

Example 10.104. Suppose we want to know how big D_{2n} is. D_{2n} acts on the vertices $\{1, 2, \ldots, n\}$

transitively. Since

$$\begin{aligned}|\operatorname{orb}(1)| &= n\\ \operatorname{stab}(1) &= \{e, \text{reflection in the line through } 1\}\end{aligned}$$

we have that $|D_{2n}| = |\operatorname{orb}(1)|\,|\operatorname{stab}(1)| = 2n$.

10.4.3 More Actions

Given any group G, there are a few important actions we can define. In particular, we will define the *conjugation action*, which is a very important concept on its own.

Definition 10.105 (Conjugation of element)**.** The ***conjugation*** of $a \in G$ by $b \in G$ is

$$bab^{-1} \in G.$$

Two elements $a, b \in G$ are *conjugate* if there exists $g \in G$ such that $b = gag^{-1}$.

Lemma 10.106. *Conjugation is an equivalence relation.*

Proof. There are three properties to check:

(i) Since $a = 1a1^{-1}$, a is conjugate to a. (Reflexivity)

(ii) If a is conjugate to b, then $a = gbg^{-1}$ for some $g \in G$, so $b = g^{-1}ag$. (Symmetry)

(iii) Suppose a is conjugate to b, and b is conjugate to c. Then $a = gbg^{-1}$ for some $g \in G$, and $b = hch^{-1}$ for some $h \in G$. Thus $a = (gh)c(gh)^{-1}$. (Transitivity)

□

Lemma 10.107 (Conjugation action)**.** *Any group G acts on itself by conjugation:*

$$g \cdot h = ghg^{-1}$$

for all $g, h \in G$.

Proof. If $g, h \in G$ then $ghg^{-1} \in G$. We now show that this is an action:

(i) $1 \cdot s = 1s1^{-1} = s$.

(ii) $g \cdot (h \cdot k) = g \cdot (hkh^{-1}) = ghkh^{-1}g^{-1} = (gh)k(gh)^{-1} = (gh) \cdot k$.

□

We give special names for the orbits and stabilisers of the conjugation action.

Definition 10.108. The ***conjugacy classes*** are the orbits of the conjugacy action:

$$\operatorname{ccl}(a) := \{b \in G \mid \exists g \in G, gag^{-1} = b\}.$$

The ***centralisers*** are the stabilisers of the conjugation action:

$$C_G(a) := \{g \in G \mid gag^{-1} = a\} = \{g \in G \mid ga = ag\}.$$

By the orbit decomposition formula,

$$|G| = \sum_{i=1}^{n} |G : C_G(g_i)|,$$

where the g_i represent distinct centralisers.

The centraliser is defined as the elements that commute with a particular element h. For the whole group G, we can define the *center*.

Definition 10.109 (Center)**.** The ***center*** of G is the set of elements which commute with all the elements of G:

$$Z(G) := \{g \in G \mid gh = hg \, \forall h \in G\}.$$

Suppose $g_1, \ldots, g_m$ are representatives of the m conjugacy classes which contain more than one element. Note that an element $g \in G$ is in the center of G if and only if the orbit of g is $\{g\}$. In general, the order of the orbit of g is equal to the index of the centraliser of g. Then

$$|G| = |Z(G)| + \sum_{i=1}^{m} |G : C_G(g_i)|. \tag{10.8}$$

This is known as the *class equation*.

In many ways, conjugation is related to normal subgroups.

Lemma 10.110. *Let $H \triangleleft G$. Then G acts by conjugation on H.*

Proof. We only have to prove closure since the other properties follow from the conjugation action. However, by definition of a normal subgroup, for every $g \in G$, $h \in H$, we have $ghg^{-1} \in H$. So it is closed. □

Proposition 10.111. *Normal subgroups are exactly those subgroups which are unions of conjugacy classes.*

Proof. Let $H \triangleleft G$. If $h \in H$, then by definition for every $g \in G$, we get $ghg^{-1} \in H$. So $\operatorname{ccl}(h) \subset H$. So H is the union of the conjugacy classes of all its elements.

Conversely, if H is a union of conjugacy classes and a subgroup of G, then for all $h \in H$, $g \in G$, we have $ghg^{-1} \in H$. So H is normal. □

Lemma 10.112. *Let X be the set of subgroups of G. Then G acts by conjugation on X.*

Proof. We first show closure. If $H \leq G$, we need to show that gHg^{-1} is also a subgroup.

(i) We know that $1 \in H$ and thus $g1g^{-1} = 1 \in gHg^{-1}$, so gHg^{-1} is non-empty.

(ii) For any two elements gag^{-1} and $gbg^{-1} \in gHg^{-1}$, $(gag^{-1})(gbg^{-1})^{-1} = g(ab^{-1})g^{-1} \in gHg^{-1}$.

We now show that it is an action.

(i) $1H1^{-1} = H$.

(ii) $g_1(g_2Hg_2^{-1})g_1^{-1} = (g_1g_2)H(g_1g_2)^{-1}$.

□

Under this action, normal subgroups have singleton orbits.

Definition 10.113 (Normaliser of subgroup). The ***normaliser*** of a subgroup H is the stabiliser of the (group) conjugation action:

$$N_G(H) := \{g \in G \mid gHg^{-1} = H\}.$$

We clearly have $H \subset N_G(H)$. It is easy to show that $N_G(H)$ is the largest subgroup of G in which H is a normal subgroup, hence the name.

There is a connection between actions in general and conjugation of subgroups.

Lemma 10.114. *Stabilisers of the elements in the same orbit are conjugate, i.e., let G act on S and let $g \in G$, $s \in S$. Then $G_{gs} = g\, G_s\, g^{-1}$.*

10.4.4 Applications

Theorem 10.115 (Cauchy's theorem). *Let G be a finite group and prime p dividing $|G|$. Then G has an element of order p (in fact there must be at least $p-1$ elements of order p).*

Proof. Let G and p be fixed. Consider $G^p = G \times \cdots \times G$, the set of p-tuples of G. Let $X \subset G^p$ be

$$X = \{(a_1, \ldots, a_p) \in G^p \mid a_1 \cdots a_p = 1\}.$$

In particular, if an element b has order p, then $(b, b, \ldots, b) \in X$. In fact, if $(b, b, \ldots, b) \in X$ and $b \neq 1$, then b has order p, since p is prime.

Now let $H = \langle h \mid h^p = 1\rangle \cong C_p$ be a cyclic group of order p with generator h. Let H act on X by "rotation":

$$h(a_1, a_2 \ldots, a_p) = (a_2, a_3, \ldots, a_p, a_1).$$

For closure, if $a_1 \cdots a_p = 1$, then $a_1^{-1} = a_2 \cdots a_p$. So $a_2 \cdots_p a_1 = a^{-1}a_1 = 1$ thus $(a_2, a_3, \ldots, a_p, a_1) \in X$. This is an action:

(i) 1 acts as an identity by construction.

(ii) The associativity condition also works by construction.

As orbits partition X, the sum of all orbit sizes must be $|X|$. We know that $|X| = |G|^{p-1}$ since we can freely choose the first $p-1$ entries and the last one must be the inverse of their product.

Since p divides $|G|$, we see that p also divides $|X|$. We have $|\operatorname{orb}(a1, \dots, a_p)|\,|\operatorname{stab}_H(a_1, \dots, a_p)| = |H| = p$. So all orbits have size 1 or p, and they sum to $|X| = p \times$ something. We know that there is one orbit of size 1, namely $(1, 1, \dots, 1)$. So there must be at least $p-1$ other orbits of size 1 for the sum to be divisible by p.

In order to have an orbit of size 1, they must look like $(a, a, \dots, a)$ for some $a \in G$, which has order p. □

10.4.5 Sylow Subgroups

The Sylow theorems are a set of related theorems describing the subgroups of prime power order of a given finite group. They are very powerful, since they can apply to any finite group, and play an important role in the theory of finite groups.

Definition 10.116. Let p be a prime. By a *p-group*, we mean a finite group whose order is a power of p (i.e., p^α for some $\alpha \in \mathbb{N}$).
Let G be a finite group, $H \leq G$. We call H a ***p-subgroup*** of G if H is a p-group.

Proposition 10.117. *Let G be a non-trivial p-group. Then*

(i) $Z(G)$ is non-trivial;

(ii) G is solvable.

Proof.

(i) By the class equation,
$$|G| = |Z(G)| + \sum |G : G_{x_i}|$$
where the sum is taken over a finite number of elements x_i with $|G : G_{x_i}| \neq 1$.

Since G is a p-group, it follows that p divides $|G|$ and also $|G : G_{x_i}|$. Hence p divides $|Z(G)|$, so the center $Z(G)$ is not trivial.

(ii) $|G/Z(G)|$ divides $|G|$ so $G/Z(G)$ is a p-group, and by (i), we know that $|G/Z(G)| < |G|$. By induction $G/Z(G)$ is solvable. By 10.78, it follows that G is solvable.

□

Definition 10.118 (Sylow p-subgroup)**.** Let G be a finite group, $H \leq G$. We say H is a ***Sylow p-subgroup*** if $|H| = p^\alpha$ and p^α is the highest power of p dividing $|G|$.

We shall prove below that such subgroups always exist. For this we need a lemma.

Lemma 10.119. *Let G be a finite abelian group, $|G| = m$. Let p be a prime, $p \mid m$. Then there exists $H \leq G$ such that $|H| = p$.*

Proof. □

The frst Sylow theorem indicates existence of Sylow subgroups, the second Sylow theorem indicates that all Sylow subgroups are related by conjugation, and the third provides constraints on the number of such subgroups.

Theorem 10.120 (Sylow I)**.** *Let G be a finite group, $p \mid |G|$. Then there exists a p-Sylow subgroup of G.*

Proof. □

Theorem 10.121 (Sylow II)**.** *Let G be a finite group.*
If H is a p-subgroup of G, then H is contained in some p-Sylow subgroup.
All p-Sylow subgroups are conjugate.

Theorem 10.122 (Sylow III)**.** *The number of p-Sylow subgroups of G is $\equiv 1 \pmod{p}$.*

Exercises

Exercise 10.1. Show that any two cyclic groups of the same order are isomorphic.

Solution. 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 $\phi\colon \langle x\rangle \to \langle y\rangle$ which sends $x^k \mapsto y^k$ is an isomorphism.

Lemma. *Let G be a group, $g \in G$, let $m, n \in \mathbb{Z}$. Denote $d = \gcd(m, n)$. If $g^n = 1$ and $g^m = 1$, then $g^d = 1$.*

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

$$g^d = g^{qm+rn} = (g^m)^q\,(g^n)^r = 1.$$

□

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 $t \in \mathbb{Z}$, 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 $\phi\colon \mathbb{Z} \to \langle x\rangle$ 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 $a, b \in \mathbb{Z}$, 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. □

Exercise 10.2. The quotient group of a cyclic group is cyclic.

Proof. Let $G = C_n$ with $H \leq C_n$. We know that H is also cyclic; say $C_n = \langle c\rangle$ and $H = \left\langle c^k\right\rangle \cong C_\ell$, where $k\ell = n$. We have $C_n/H = \{H, cH, c^2H, \ldots, c^{k-1}H\} = \langle cH\rangle \cong C_k$. □

Chapter 11

Rings

Summary

- Prime ideals and maximal ideals, relation to fields and integral domains. Application of quotients to constructing fields by adjunction of elements. Degree of a field extension, the tower law.
- Euclidean domains. Principal ideal domains. EDs are PIDs. Unique factorisation for PIDs. Gauss's lemma and Eisenstein's criterion for irreducibility.

11.1 Rings

11.1.1 Definitions and Properties

Definition 11.1 (Ring)**.** A ***ring*** R is a set together with two binary operations $+$ and $\times$ (called addition and multiplication), satisfying the following axioms:

(i) $(R,+)$ is an abelian group, with identity 0;

(ii) $\times$ is associative: $(a\times b)\times c = a\times(b\times c)$ for all $a,b,c\in R$;

(iii) $\times$ distributes over $+$: for all $a,b,c\in R$,

$$\begin{aligned} a\times(b+c) &= (a\times b)+(a\times c),\\ (a+b)\times c &= (a\times c)+(b\times c). \end{aligned}$$

Notation. We simply write ab rather than $a\times b$, for $a,b\in R$.

Notation. Denote the additive identity of $a\in R$ by $-a$.

We say R is *commutative* if multiplication is commutative.

We say R has an *identity* if there exists $1\in R$ such that

$$1\times a = a\times 1 = a \quad (a\in R).$$

In general, a ring may not necessarily be commutative or have multiplicative inverses; when they do, we give such rings special names.

Definition 11.2. A ring R with identity 1, where $1 \neq 0$, is called a ***division ring*** if every $a \in R \setminus \{0\}$ has a multiplicative inverse, i.e., exists $b \in R$ such that $ab = ba = 1$.
A commutative division ring is called a ***field***.

$$\begin{tikzcd}
& \text{commutative ring} \arrow[dr, "\exists a^{-1}"] & \\
\text{ring} \arrow[ur, "ab=ba"] \arrow[dr, "\exists a^{-1}"'] & & \text{field} \\
& \text{division ring} \arrow[ur, "ab=ba"'] &
\end{tikzcd}$$

Example 11.3.

- $\mathbb{Z}$ is the prototypical ring; it is not a field.
- $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ are fields.
- $\mathbb{Z}/n\mathbb{Z}$ is a commutative ring with identity $\bar{1}$ under addition and multiplication of residue classes.
- $2\mathbb{Z}$ is a commutative ring without identity.
- The trivial ring $R = \{0\}$ is a commutative ring with identity $1 = 0$.
- The ring of (polynomial/continuous/differentiable) functions on $\mathbb{R}$.
- The endomorphism ring $\mathrm{End}_{\mathbb{R}}(V)$ of a vector space V over $\mathbb{R}$ is a non-commutative ring.
- The Hamilton Quaternions $\mathbb{H}$. Historically the first example of a non-commutative ring.

Lemma 11.4 (Basic properties)**.** *Let R be a ring.*

(i) $0a = a0 = 0$ *for all* $a \in R$.

(ii) $(-a)b = a(-b) = -(ab)$ *for all* $a, b \in R$.

(iii) $(-a)(-b) = ab$ *for all* $a, b \in R$.

(iv) If R has identity 1, then the identity is unique and $-a = (-1)a$.

Proof. These all follow from the distributive laws and cancellation in the additive group $(R, +)$.

(i) We have $0a = (0+0)a = 0a + 0a$. Then the cancellation law implies $0a = 0$.
Similarly, $a0 = a(0+0) = a0 + a0$. Thus $a0 = 0$.

(ii) We have $ab + (-a)b = (a + (-a))\,b = 0b = 0$. Thus $(-a)b = -(ab)$.
Similarly, $ab + a(-b) = a\,(b + (-b)) = a0 = 0$. Thus $a(-b) = -(ab)$.

(iii) Using (ii), $(-a)(-b) = -a(-b) = -\left(-(ab)\right) = ab$.

(iv) Suppose 1 and $1'$ are identities of R. Then $1 = 1 \times 1' = 1'$.
Since $(-1)a + a = (-1)a + 1a = \left((-1) + 1\right)a = 0a = 0$, it follows that $-a = (-1)a$.

$\square$

11.1.2 Subrings

Having defined the notion of a ring, there is a natural notion of a subring.

Definition 11.5 (Subring)**.** Let R be a ring. We say $S \subset R$ is a ***subring*** of R, if S is a subgroup of R that is closed under multiplication.

Example 11.6.

- $\mathbb{Z}$ is a subring of $\mathbb{Q}$, and $\mathbb{Q}$ is a subring of $\mathbb{R}$.
- $n\mathbb{Z} = \{nk \in \mathbb{Z} \mid k \in \mathbb{Z}\}$ is a subring of $\mathbb{Z}$.
- The real-valued differentiable functions on $\mathbb{R}$ form a subring of the ring of continuous functions.
- $\mathbb{Z}[i] = \{x + yi \mid x, y \in \mathbb{Z}\}$ is a subring of $\mathbb{C}$, called the ring of Guassian integers.
- $\mathbb{Q}[\sqrt{2}] = \{x + y\sqrt{2} \mid x, y \in \mathbb{Q}\}$ is a subring of $\mathbb{R}$.
- $S = \mathbb{Z} + \mathbb{Z}i + \mathbb{Z}j + \mathbb{Z}k$ form a subring of $\mathbb{H}$.

We will use the square brackets notation quite frequently. It should be clear what it should mean, and we will define it properly later.

If R contains 1, then S is a (unital) subring if $1_R \in S$. We assume subrings are unital unless otherwise specified.

Lemma 11.7 (Subring criterion)**.** *Let R be a ring, $S \subset R$. Then S is a subring of R if and only if*

(i) $1 \in S$; *(non-empty)*

(ii) $ab, a - b \in S$ *for all* $a, b \in S$. *(closed under multiplication and subtraction)*

Proof.

$\boxed{\Longrightarrow}$

$\boxed{\Longleftarrow}$ The condition that $a - b \in S$ for all $a, b \in S$ implies that S is an additive subgroup by the subgroup test (note that as $1 \in S$ we know that S is nonempty). The other conditions for a subring hold directly. $\square$

11.1.3 Units, Zero Divisors

Recall that in a ring we do not require that non-zero elements have a multiplicative inverse. Nevertheless, since the multiplication operation is associative and there is a multiplicative identity, the elements which happen to have multiplicative inverses form a group:

Definition 11.8. Let R be a ring, with identity $1 \neq 0$. We say $u \in R$ is a ***unit*** in R if there exists $v \in R$ such that $uv = vu = 1$.
We say $a \in R \setminus \{0\}$ is a ***zero divisor*** if there exists $b \in R \setminus \{0\}$ such that either $ab = 0$ or $ba = 0$.

Remark. A zero divisor can not be a unit.

Let $R^\times$ denote the set of units in R.

Lemma. *$R^\times$ forms a group under multiplication.*

We call $R^\times$ the ***group of units***.

Proof.

(i) Evidently $1 \in R^\times$.

(ii) Let $u_1, u_2 \in R^\times$. Then $u_1v_1 = 1$, $u_2v_2 = 1$ for some $v_1, v_2 \in R$. Thus $(u_1u_2)(v_2v_1) = 1$. Similarly $(v_2v_1)(u_1u_2) = 1$. Hence $u_1u_2 \in R^\times$.

(iii) Let $u \in R^\times$. Then $uv = vu = 1$ for some $v \in R$. Taking inverse gives $u^{-1}v^{-1} = v^{-1}u^{-1} = 1$. Hence $u^{-1} \in R^\times$.

□

Example 11.9.

- The ring $\mathbb{Z}$ has no zero divisors and its only units are ± 1.
- The group of units of $\mathbb{Z}/n\mathbb{Z}$ is $(\mathbb{Z}/n\mathbb{Z})^\times$. Recall that $(\mathbb{Z}/n\mathbb{Z})^\times = \{a \in \mathbb{Z}/n\mathbb{Z} \mid (a, n) = 1\}$. All elements not in $(\mathbb{Z}/n\mathbb{Z})^\times$ are zero divisors. In sum, every non-zero element of $\mathbb{Z}/n\mathbb{Z}$ is either a unit or a zero divisor.

Rings having some of the same characteristics as $\mathbb{Z}$ are given a name:

Definition 11.10 (Integral domain)**.** If a commutative ring with identity $1 \neq 0$ has no zero divisors, it is called an ***integral domain***.

Example 11.11.

- $\mathbb{Z}$ is an integral domain.
- All fields are integral domains.

The absence of zero divisors in integral domains give these rings a cancellation property:

Lemma 11.12.

(i) Let $a, b, c \in R$, a is not a zero divisor. If $ab = ac$, then either $a = 0$ or $b = c$.

(ii) In particular, for any a, b, c in an integral domain and $ab = ac$, then either $a = 0$ or $b = c$.

Proof.

(i) If $ab = ac$, then $a(b - c) = 0$. Since a is not a zero divisor, we have either $a = 0$ or $b - c = 0$.

(ii) This follows from (i) and the definition of an integral domain.

□

Corollary 11.13. *Any finite integral domain is a field.*

In this terminology, a field is a commutative ring F with identity $1 \neq 0$ in which every non-zero element is a unit, i.e., $F^\times = F \setminus \{0\}$.

Proof. Let R be a finite integral domain, let $a \in R \setminus \{0\}$.

By the cancellation law, the map $x \mapsto ax$ is an injective function. Since R is finite, this map is also surjective. In particular, there exists $b \in R$ such that $ab = 1$, i.e., a is a unit in R. Since a was an arbitrary non-zero element, R is a field. □

Corollary 11.14. *If p is a prime, $\mathbb{Z}/p\mathbb{Z}$ is a field, usually denoted by $\mathbf{F}_p$.*

11.1.4 Examples

Example 11.15 (Matrix rings)**.** Let R be a (often commutative) ring with 1. We define the matrix ring $M_{n\times n}(R)$ as the set consisting of

$$(a_{ij})_{n\times n}, \quad a_{ij} \in R.$$

Addition and multiplication on $M_{n\times n}(R)$ is defined following the matrix multiplication in linear algebra.

If we take $R = \mathbb{R}$, then $M_{n\times n}(\mathbb{R})$ the usual matrix algebra. We have the subring of diagonal matrices, and the subring of upper triangular matrices.

Example 11.16 (Group rings)**.** Let R be a commutative ring with 1. Let G be a finite group. We define the group ring $R[G]$ as the set consisting of

$$\sum_{g\in G} a_g g \quad (a_g \in R).$$

Addition on $R[G]$ is defined in the obvious/naive way. The multiplication is via the following example

$$(a_g g + a_h h)(a_{g'} g' + a_{h'} h') = a_g a_{g'} gg' + a_h a_{g'} hg' + a_g a_{h'} gh' + a_h a_{h'} hh',$$

where gg', hg', gh', hh' are the group multiplication in G.

Lemma. *Let R be a commutative ring with 1. Let G be a finite group.*

(i) Let $e \in G$ be the identity element. Then 1_e is the identity of the ring $R[G]$.

(ii) Let $e \neq g \in G$. Then $1 - g$ is a zero divisor.

(iii) Let H be a subgroup of G. Then $R[H]$ is a subring of $R[G]$.

(iv) The ring $R[G]$ is commutative if and only if G is commutative.

Example 11.17 (Product of rings)**.** Let R and S be two rings. We define the ring $R \times S$ as follows: as a set $R \times S$ is the same as the Cartesian product of sets; we define the addition and multiplication component wise:

$$(a, b) + (c, d) = (a + c, b + d),$$
$$(a, b) \times (c, d) = (ac, bd).$$

11.2 Homomorphisms and Isomorphisms

11.2.1 Ideals

Definition 11.18 (Ideal)**.** Let R be a ring. We say $I \subset R$ is a ***left ideal*** if

(i) $(I, +)$ is a subgroup of $(R, +)$; (additive subgroup)

(ii) $ax \in I$ for all $a \in R$, $x \in I$. (closed under left multiplication)

We define a ***right ideal*** similarly.
We say I is a (two-sided) ***ideal*** of R, if I is both a left ideal and a right ideal of R.

We say I is a *proper ideal* if $I \neq R$.

Remark. For commutative rings, left ideals, right ideals, and (two-sided) ideals coincide

Example 11.19.

- Trivial ideals: the zero ideal $\{0\}$ and the whole ring R are two-sided ideals.

  R is also called the *unit ideal*: if $x \in R^\times \cap I$, then $x^{-1}x = 1 \in I$, so $a \times 1 = a \in I$ for all $a \in R$. Thus $I = R$. (We have shown $I = R$ if and only if $1 \in I$.)

  This implies that in a field F, the only ideals are $\{0\}$ and F, since if $I \neq \{0\}$, let $x \in F \setminus \{0\}$, then x is a unit, so $I = F$.

- The even numbers $2\mathbb{Z} = (2)$ is an ideal of $\mathbb{Z}$.

The next definition provides a way to generate an ideal from an element of a ring.

Definition 11.20 (Principal ideal)**.** Let R be a ring, and let $a \in R$. The ***principal left ideal*** generated by a is

$$(a) := \{xa \mid x \in R\}.$$

More generally, let $a_1, \dots, a_n \in R$. Define

$$(a_1, \dots, a_n) := \{x_1 a_1 + \dots + x_n a_n \mid x_i \in R\}.$$

We call $a_1, \dots, a_n$ *generators* for this ideal.

Lemma. *$(a_1, \dots, a_n)$ is a left ideal.*

Proof. If $y_1, \dots, y_n, x_1, \dots, x_n \in R$ then

$$\begin{aligned}(x_1 a_1 + \dots + x_n a_n) + (y_1 a_1 + \dots + y_n a_n) &= x_1 a_1 + y_1 a_1 + \dots + x_n a_n + y_n a_n \\ &= (x_1 + y_1) a_1 + \dots + (x_n + y_n) a_n.\end{aligned}$$

If $z \in R$, then

$$z(x_1 a_1 + \dots + x_n a_n) = z x_1 a_1 + z x_n a_n.$$

Finally,

$$0 = 0a_1 + \cdots + 0a_n.$$

$\square$

Example 11.21. Let R be a ring. Let L, M be left ideals. Define the product

$$LM = \{x_1y_1 + \cdots + x_ny_n \mid x_i \in L, y_i \in M\}.$$

Then LM is also a left ideal.

If L, M, N are left ideals, then $(LM)N = L(MN)$.

Example 11.22. Let L, M be left ideals. Define the sum

$$L + M = \{x + y \mid x \in L, y \in M\}.$$

Then $L + M$ is a left ideal.

If L, M, N are left ideals, then $L(M + N) = LM + LN$.

Example 11.23. The ideals of $\mathbb{Z}$ are (n) for $n \in \mathbb{N}$, and $\{0\}$.

Proof. Let $I \neq \{0\}$ be an ideal of $\mathbb{Z}$. Let $a \in I$ be non-zero. Since $a, -a \in I$, I contains a natural number. By well-ordering, there is a minimal $n \in \mathbb{N} \cap I$.

Clearly $(n) \subset I$, since all multiples of n are contained in I. If $x \in I$ and $x \neq (n)$, by the division algorithm, we can write $x = qn + r$ for some $q, r \in \mathbb{Z}$, $0 < r < n$. But $qn \in I$, so $x - qn = r \in I$. Then r is a smaller natural number in I, which contradicts the minimality of n. Thus $I = (n)$. $\square$

This shows that $\mathbb{Z}$ is a principal ideal domain (all ideals of $\mathbb{Z}$ are principal ideals).

11.2.2 Homomorphisms

Definition 11.24. We say $\phi\colon R \to S$ is a ***homomorphism*** if it satisfies

(i) $\phi(a + b) = \phi(a) + \phi(b)$ for all $a, b \in R$;

(ii) $\phi(ab) = \phi(a)\phi(b)$ for all $a, b \in R$;

(iii) $\phi(1_R) = 1_S$.

An ***isomorphism*** is a bijective homomorphism. Two rings R and S are ***isomorphic***, denoted by $R \cong S$, if there exists an isomorphism between R and S.

Remark. For groups, condition (iii) is not required in the definition: $\phi(1)\phi(x) = \phi(1x) = \phi(x)$ then we can cancel on both sides due to the existence of (multiplicative) inverse.

An isomorphism between a ring with itself is called an *automorphism*.

An injective homomorphism $\phi\colon R \to S$ is called an ***embedding***[1]; we say R is *embedded* in S.

[1]If ϕ is injective, then $R \cong \operatorname{im}\phi$, where $\operatorname{im}\phi$ is a subring of S, so we can think of R as a "subring" of S; hence the term *embedding* to mean that R is "contained in" S.

Definition 11.25. Let $\phi\colon R \to S$ be a homomorphism. The ***kernel*** of ϕ is its kernel viewed as a homomorphism of additive groups:

$$\ker\phi := \{r \in R \mid \phi(r) = 0\}.$$

The ***image*** of ϕ is

$$\operatorname{im}\phi := \{s \in S \mid \exists r \in R, \phi(r) = s\}.$$

Example 11.26.

- Consider the quotient map $\pi\colon \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$; $\ker\pi = n\mathbb{Z}$.
- The embedding of the subring $n\mathbb{Z} \to \mathbb{Z}$. The kernel is trivial.
- The map

$$\begin{aligned}\phi\colon \mathbb{C}[x] &\to \mathbb{C}\\ f(x) &\mapsto f(a)\end{aligned}$$

  The kernel is

$$\ker\phi = \{f(x) \in \mathbb{C}[x] \mid f(a) = 0\} = \{(x-a)f(x) \mid f(x) \in \mathbb{C}[x]\}.$$

Lemma 11.27. *Let $\phi\colon R \to S$ be a homomorphism. Then*

(i) $\ker\phi$ *is a ideal of R;*

(ii) $\operatorname{im}\phi$ *is a subsring of S.*

Proof.

(i) Let $x, y \in \ker\phi$. Then

$$\phi(x-y) = \phi(x) - \phi(y) = 0 - 0 = 0$$

so $x - y \in \ker\phi$. Thus $\ker\phi$ is an additive subgroup of R.

Let $r, r' \in R$, $x \in \ker\phi$. Then

$$\phi(rxr') = \phi(r)\phi(x)\phi(r') = \phi(r)0\phi(r') = 0.$$

Thus $rxr' \in \ker\phi$, and so $\ker\phi$ is an ideal of R.

(ii)

□

Lemma 11.28. *Let $\phi\colon R \to S$ be a homomorphism. Then ϕ is injective if and only if* $\ker\phi = \{0\}$.

Proof. This follows from considering $(R, +)$ as an additive group. Then the result follows from group theory. □

11.2.3 Quotient Rings

Let $I \subset R$ be an ideal, and let $a \in R$. Define

$$a + I = \{a + x \mid x \in I\}.$$

This is usually not an ideal, but rather an *additive coset* of I (considering I as an additive subgroup of R). Any element of the coset is called a *representative* of the coset.

Definition 11.29 (Quotient ring). Let $I \subset R$ be an ideal. Then the ***quotient ring*** is

$$R/I := \{a + I \mid a \in R\}.$$

Lemma. *R/I is a ring, with addition and multiplication defined as*

$$(a + I) + (b + I) = (a + b) + I,$$
$$(a + I) \cdot (b + I) = ab + I.$$

Proof. Recall that by 10.60, a subgroup of an abelian group is normal. Since I is an additive subgroup of R, and $(R, +)$ is abelian, we have $I \lhd R$ under addition. Hence the quotient group $(R/I, +)$ is defined.

We now check that multiplication is well-defined. Suppose $a + I = a' + I$, $b + I = b' + I$. Then $a - a' = r \in I$, $b - b' = s \in I$. Thus

$$ab = (a' + r)(b' + s) = a'b' + a's + b'r + rs.$$

Note that $a's, b'r, rs \in I$. Hence $ab + I = a'b' + I$.

Check that R/I is a ring, with additive identity $0_R + I$ and multiplicative identity $1_R + I$. □

Example 11.30. Take $R = \mathbb{Z}$, $I = (n)$ for some $n \in \mathbb{N}$. We can write $(n) = n\mathbb{Z}$, so the quotient ring is $\mathbb{Z}/n\mathbb{Z}$.

As before, we give a name to the canonical homomorphism from R to R/I.

Definition 11.31 (Quotient map). Let $I \subset R$ be an ideal. The ***quotient map*** is

$$\begin{aligned} \pi \colon R &\to R/I \\ a &\mapsto a + I \end{aligned}$$

Lemma 11.32. *Quotient maps are surjective homomorphisms.*

Proof. Let $\pi \colon R \to R/I$ be a quotient map.

- Let $a, b \in R$. Then $\pi(a + b) = (a + b) + I = (a + I) + (b + I) = \pi(a) + \pi(b)$.
- Let $a, b \in R$. Then $\pi(ab) = ab + I = (a + I)(b + I) = \pi(a)\pi(b)$.
- $\pi(1_R) = 1_R + I$, which is the identity of R/I.

□

In addition,

$$\ker \pi = \{a \in R \mid a + I = 0_R + I\} = \{a \in R \mid a \in I\} = I.$$

11.2.4 Isomorphism Theorems

Theorem 11.33 (First isomorphism theorem)**.** *Let $\phi\colon R \to S$ be a homomorphism. Then*

$$R/\ker\phi \cong \operatorname{im}\phi. \tag{11.1}$$

Proof. Denote $K = \ker\phi$. Consider the map

$$\begin{aligned} \theta\colon R/K &\to \operatorname{im}\phi \\ a + K &\mapsto \phi(a) \end{aligned}$$

We claim that θ is an isomorphism.

1. We first check that θ is well-defined. If $a + K = a' + K$, then $a - a' \in K$, so $\phi(a - a') = 0$. Thus $\phi(a) = \phi(a')$.
2. θ is a homomorphism:

$$\begin{aligned} \theta((a+K)+(b+K)) &= \theta((a+b)+K) = \phi(a+b) = \phi(a)+\phi(b) = \theta(a+K)+\theta(b+K) \\ \theta((a+K)(b+K)) &= \theta(ab+K) = \phi(ab) = \phi(a)\phi(b) = \theta(a+K)\theta(b+K) \\ \theta(0_R+I) &= \phi(0_R) = 0_S \end{aligned}$$

3. θ is injective: $\theta(a+K) = \theta(b+K) \implies \phi(a) = \phi(b) \implies a+K = b+K$.
4. θ is surjective: Let $x \in \operatorname{im}\phi$. Then $x = \phi(a)$ for some $a \in R$. Thus $\theta(a+K) = \phi(a) = x$.

□

Theorem 11.34 (Second isomorphism theorem)**.** *Let A be a subring, and B be an ideal of R. Then*

$$(A+B)/B \cong A/(A\cap B). \tag{11.2}$$

Lemma.

(i) $A + B = \{a + b \mid a \in A, b \in B\}$ is a subring of R;

(ii) $A \cap B$ is an ideal of A.

Theorem 11.35 (Third isomorphism theorem)**.** *Let I and J be ideals of R, with $I \subset J$. Then*

$$(R/I)(J/I) \cong R/J. \tag{11.3}$$

Lemma. *J/I is an ideal of R/I.*

Theorem 11.36 (Fourth isomorphism theorem)**.** *Let I be an ideal of R. The correspondance $A \leftrightarrow A/I$ is an inclusion preserving bijection between the set of subrings of A of R that contain I and the set of subrings of R/I. Furthermore, A (a subring containing I) is an ideal of R if and only if A/I is an ideal of R/I.*

11.2.5 Chinese Remainder Theorem

Definition 11.37. Let R be a commutative ring. We say two ideals $I, J \subset R$ are *coprime* if

$$I + J = R.$$

In particular, there $i \in I$, $j \in J$ such that $i + j = 1$.

Theorem 11.38 (Chinese remainder theorem)**.** *Let R be a commutative ring. Suppose I and J are coprime ideals of R. Then for any $a, b \in R$, there exists $x \in R$ such that*

$$x \in (a + I) \cap (b + J).$$

Proof. Let $i \in I$ and $j \in J$ be such that $i + j = 1$.

Claim. $x = aj + bi$.

We can write

$$x = a(1 - i) + bi = a + (b - a)i \in a + I.$$

Similarly,

$$x = aj + b(1 - j) = b + (a - b)j \in b + J.$$

□

modular arithmetic

11.2.6 Prime and Maximal Ideals

Let R be a commutative ring.

Definition 11.39. An ideal $P \subsetneq R$ is ***prime*** if $ab \in P$ implies either $a \in P$ or $b \in P$.
An ideal $M \subsetneq R$ is ***maximal*** if there is no ideal between M and R, i.e., $M \subset I \subset R$ implies $I = M$ or $I = R$.

Example 11.40. In $\mathbb{Z}$, (p) is a prime ideal for prime p.

Further $p\mathbb{Z} \subset U = n\mathbb{Z} \subset \mathbb{Z}$, and $p \in U$, then $p = nq$ for some $q \in \mathbb{Z}$. But p is prime and $n \neq 1$ so $n = p$. Thus $U = p\mathbb{Z}$. Thus $p\mathbb{Z}$, for p prime, is a maximal ideal in $\mathbb{Z}$. Note that $0 \subset p\mathbb{Z} \subset \mathbb{Z}$, so 0 is not a maximal ideal in $\mathbb{Z}$.

Lemma 11.41. *Let R be a commutative ring.*

(i) A maximal ideal is prime.

(ii) An ideal P is prime if and only if R/P is integral.

(iii) An ideal M is maximal if and only if R/M is a field.

Proof.

(i) Suppose M is a maximal ideal. Let $ab \in M$, WLOG assume $a \notin M$. Then $M \subsetneq (a) + M = R$, since M is a maximal ideal.

Thus $xa + m = 1$ for some $x \in R$, $m \in M$. Then $b = xab + mb \in M$, since $ab, m \in M$. Hence M is prime.

□

11.2.7 Characteristic of Ring

In the following, let $R = \{0\}$ be a ring; let e denote the identity of R (to distinguish it from the identity of $\mathbb{Z}$). For any $a \in R$, $n \in \mathbb{Z}$, we can define an integer multiple of a ring element:

$$na = \begin{cases} \underbrace{a + \cdots + a}_{n \text{ times}} & (n > 0) \\ -(ka) & (n < 0, n = -k) \\ 0 & (n = 0) \end{cases}$$

Consider the map

$$\begin{aligned} f \colon \mathbb{Z} &\to R \\ n &\mapsto ne \end{aligned}$$

Then this is a homomorphism (this is a bit tedious, since we have to consider $n > 0$, $n < 0$ or $n = 0$). Now let $f \colon \mathbb{Z} \to R$ be any homomorphism. By definition, $f(1) = e$. Then if $n > 0$, $f(n) = f(1 + \cdots + 1) = f(1) + \cdots + f(1) = nf(1) = ne$. Hence there is one and only one homomorphism $\mathbb{Z} \to R$.

Assume $R \neq \{0\}$. Let $f \colon \mathbb{Z} \to R$ be *the* homomorphism. Since $\ker f$ is an ideal of $\mathbb{Z}$, $\ker f = n\mathbb{Z}$ for some integer $n \geq 0$. (Note that $n \neq 1$, otherwise $\ker f = \mathbb{Z}$ so $\operatorname{im} f = \{0\}$, but $f(1) = e \neq 0$.)

By the first isomorphism theorem, $\mathbb{Z}/n\mathbb{Z} \cong \operatorname{im} f$. In practice, we do not make any distinction between $\mathbb{Z}/n\mathbb{Z}$ and its image in R, and we agree to say that "R contains $\mathbb{Z}/n\mathbb{Z}$ as a subring".

Suppose $n \neq 0$. Then for all $a \in R$,

$$\underbrace{a + \cdots + a}_{n \text{ times}} = na = (ne)a = f(n)a = 0a = a.$$

We call n the ***characteristic*** of R, or say R has characteristic n, and denote $n = \operatorname{char}(R)$.

Remark. If $n = 0$, then $\mathbb{Z}/0\mathbb{Z} = \mathbb{Z}$, so rings of characteristic 0 are infinite (since it contains a subring isomorphic to $\mathbb{Z}$, which is infinite).

Note that $n \neq 0$ is the smallest positive integer m such that $me = 0$. This is because $m \in \ker f$, so $n \mid m$, which implies $n \leq m$.

Lemma 11.42. *Suppose R is an integral ring. Then $\operatorname{char}(R)$ is either 0 or prime.*

Proof. Suppose $n = \operatorname{char}(R) \neq 0$. Suppose, for a contradiction, that n is composite. Then $n = mk$, where $m, k > 1$. Then $m, k < n$.

By minimality of n, we have $me, ke \neq 0$. But $(me)(ke) = mke = ne = 0$. This implies that R has zero divisors, which contradicts the assumption that R is an integral ring. □

Lemma 11.43 (Freshman's dream). *Let R be commutative with prime characteristic p. Then $(x + y)^p = x^p + y^p$ for all $x, y \in R$.*

Proof. Since R is commutative, we have the binomial expansion:

$$(x + y)^p = \sum_{i=1}^{p} \binom{p}{i} x^i y^{p-i}.$$

(We require R to be commutative, so that we can freely move variables around in order to raise them by powers.) For $i \in \{1, \ldots, p - 1\}$, $\binom{p}{i}$ is divisible by p. Since $\operatorname{char}(R) = p$, multiples of p equal 0. Hence $(x + y)^p = x^p + 0 + \cdots + 0 + y^p = x^p + y^p$. □

Let K be a field, and let $f \colon \mathbb{Z} \to K$ be the homomorphism from the integers to K. If $\ker f = \{0\}$, then K contains $\mathbb{Z}$ as a subring, and we say that K has *characteristic* 0. If $\ker f = p\mathbb{Z}$ for some prime p, then we say K has *characteristic* p.

The field $\mathbb{Z}/p\mathbb{Z}$ is sometimes denoted by $\mathbf{F}_p$, and is called the *prime field*, of characteristic p. This prime field $\mathbf{F}_p$ is contained in every field of characteristic p.

11.2.8 Quotient Fields

Recall that we can construct $\mathbb{Q}$ from $\mathbb{Z}$, using equivalence classes of ordered pairs whose elements are in $\mathbb{Z}$. Instead of $\mathbb{Z}$, our discussion will apply to an arbitrary integral ring R.

Let $(a, b), (c, d) \in R \times R^*$, where $R^* = R \setminus \{0\}$; we call these ordered pairs *quotients*. Define a relation $R \times R^*$:

$$(a, b) \sim (c, d) \iff ad = bc.$$

Lemma. *$\sim$ is an equivalence relation on $R \times R^*$.*

Proof.

(i) Since $ab = ba$, we have $(a, b) \sim (a, b)$.

(ii) Suppose $(a, b) \sim (c, d)$. Then $ad = bc$, or $cb = da$. This implies $(c, d) \sim (a, b)$.

(iii) Suppose $(a, b) \sim (c, d)$ and $(c, d) \sim (e, f)$. Then

$$ad = bc, \quad cf = de.$$

Thus

$$adf = bcf = bde,$$

so $daf - dbe = 0$. Then $d(af - be) = 0$. Since R has no divisors of 0, and $d \neq 0$, it follows that $af - de = 0$, i.e., $af = be$. This means that $(a, b) \sim (e, f)$.

□

We denote the equivalence class of (a, b) by a/b; that is,

$$\frac{a}{b} = \{(c, d) \in R \times R^* \mid (a, b) \sim (c, d)\}.$$

Then the ***quotient field*** (or *field of fractions*) of R is the set of equivalence classes:

$$\mathrm{Frac}(R) := (R \times R^*)/\sim$$

with addition and multiplication defined by

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

Lemma 11.44. $\mathrm{Frac}(R)$ *is a field, with addition and multiplication being defined as above.*

Proof. We first check that addition and multiplication, as defined above, are well-defined.

Addition Suppose $a/b = a'/b'$ and $c/d = c'/d'$. We must show that

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

This is true if and only if

$$b'd'(ad + bc) = bd(a'd' + b'c'),$$

or in other words,

$$b'd'ad + b'd'bc = bda'd' + bdb'c'.$$

But $ab' = a'b$ and $cd' = c'd$ by assumption. Hence the above equation holds.

Multiplication Suppose $a/b = a'/b'$ and $c/d = c'/d'$.

The verification that $\mathrm{Frac}(R)$ is a commutative ring with identity is left as an exercise; note that the additive identity is $0/1$, and the multiplicative identity is $1/1$ (where 1 is the identity of R).

We now show that $\mathrm{Frac}(R)$ is a field. Note that if $a/b = 0/1$, then $(a, b) \sim (0, 1)$, so $a = 0$. Thus if $a/b \neq 0/1$ is a non-zero element, then $a \neq 0$. Then (b, a) and subsequently b/a is well-defined (since

$a \neq 0$). The multiplicative of a/b is then b/a:

$$\frac{a}{b}\frac{b}{a} = \frac{b}{a}\frac{a}{b} = \frac{ab}{ab} = \frac{1}{1}.$$

Hence every non-zero element in $\mathrm{Frac}(R)$ has a multiplicative inverse, so $\mathrm{Frac}(R)$ is a field. □

Example 11.45.

- $\mathbb{Q} = \mathrm{Frac}(\mathbb{Z})$.
- $\mathbb{Q}[i] = \mathrm{Frac}(\mathbb{Z}[i])$, the field of Gaussian rationals.
- The quotient field of a field is canonically isomorphic to the field itself.

Lemma 11.46. *R is embedded in* $\mathrm{Frac}(R)$.

Proof. Consider the map

$$\begin{aligned} \phi\colon R &\to \mathrm{Frac}(R) \\ a &\mapsto a/1 \end{aligned}$$

We claim that ϕ is an embedding (injective homomorphism).

1. ϕ is a homomorphism:

$$\begin{aligned} \phi(a+b) &= \frac{a+b}{1} = \frac{a}{1} + \frac{b}{1} = \phi(a) + \phi(b) \\ \phi(ab) &= \frac{ab}{1} = \frac{a}{1}\frac{b}{1} = \phi(a)\phi(b) \\ \phi(1) &= \frac{1}{1} \end{aligned}$$

2. ϕ is injective: $\phi(a) = \phi(b) \implies a/1 = b/1 \implies a = b$.

□

We often think of rationals as an integer dividing another non-zero integer, instead of considering them as equivalence classes. We now show this more generally.

Lemma 11.47. *Suppose R is a subring of a field F. (Thus R is an integral domain.) Then*

$$\mathrm{Frac}(R) \cong \{ab^{-1} \mid a, b \in R, b \neq 0\}.$$

Proof. We see that $\{ab^{-1} \mid a, b \in R, b \neq 0\}$ is a field, which is a subfield of F. Consider the map

$$a/b \mapsto ab^{-1}.$$

We claim this is an isomorphism. □

Hence we often call the field $\{ab^{-1} \mid a, b \in R, b \neq 0\}$ the *quotient field* of R in F; there can be no confusion with this terminology due to the above isomorphism. In view of this, the element ab^{-1} of F is also denoted by a/b.

Proposition 11.48. *Let $\phi\colon R \to F$ be an embedding of an integral domain R into a field F. Then there exists a unique extension $\phi^*\colon \mathrm{Frac}(R) \to F$ which is also an embedding.*
(ϕ^* being an extension of ϕ means $\phi^*|_R = \phi$.)

Proof.

$\boxed{\text{Existence}}$ Define the map

$$\begin{aligned} \phi^*\colon \mathrm{Frac}(R) &\to F \\ \frac{a}{b} &\mapsto \frac{\phi(a)}{\phi(b)} \end{aligned}$$

1. We first check that ϕ^* is well-defined. Suppose $a/b = c/d$. Then $ad = bc$, so $\phi(ad) = \phi(bc)$, or $\phi(a)\phi(d) = \phi(b)\phi(c)$, which implies $\frac{\phi(a)}{\phi(b)} = \frac{\phi(c)}{\phi(d)}$. (Note that $b, d \neq 0$, so $\phi(b), \phi(d) \neq 0$, since $\ker\phi = \{0\}$ due to injectivity.)

2. ϕ^* is a homomorphism:

$$\begin{aligned} \phi^*\left(\frac{a}{b} + \frac{c}{d}\right) &= \phi^*\left(\frac{ad+bc}{bd}\right) = \frac{\phi(ad+bc)}{\phi(bd)} = \frac{\phi(a)\phi(d) + \phi(b)\phi(c)}{\phi(b)\phi(d)} \\ &= \frac{\phi(a)}{\phi(b)} + \frac{\phi(c)}{\phi(d)} = \phi^*\left(\frac{a}{b}\right) + \phi^*\left(\frac{c}{d}\right) \\ \phi^*\left(\frac{a}{b}\frac{c}{d}\right) &= \phi^*\left(\frac{ac}{bd}\right) = \frac{\phi(ac)}{\phi(bd)} = \frac{\phi(a)\phi(c)}{\phi(b)\phi(d)} = \frac{\phi(a)}{\phi(b)}\frac{\phi(c)}{\phi(d)} = \phi^*\left(\frac{a}{b}\right)\phi^*\left(\frac{c}{d}\right) \\ \phi^*\left(\frac{1}{1}\right) &= \frac{\phi(1)}{\phi(1)} = \frac{1}{1} = 1 \end{aligned}$$

3. ϕ^* is injective: Let $a/b \in \ker\phi^*$. Then $\phi^*(a/b) = \phi(a)/\phi(b) = 0$, so $\phi(a) = 0$. By injectivity, $a = 0$, since $\ker\phi = \{0\}$. This implies $a/b = 0/1$, so $\ker\phi^* = \{0\}$.

4. ϕ^* is an extension of ϕ: Since $\phi(1) = 1$, we have $\phi^*(a/1) = \phi(a)/1 = \phi(a)$ for all $a \in R$.

$\boxed{\text{Uniqueness}}$ Suppose we have yet to define ϕ^* as above. Then

$$\phi^*\left(\frac{a}{b}\right) = \phi^*\left(\frac{a}{1}\frac{1}{b}\right) = \phi^*\left(\frac{a}{1}\right)\phi^*\left(\frac{1}{b}\right) = \phi^*\left(\frac{a}{1}\right)\phi^*\left(\frac{b}{1}\right)^{-1} = \phi(a)\phi(b)^{-1}.$$

Hence there is only one map ϕ^*, defined as above, which satisfies the above conditions. $\square$

11.3 Euclidean Domains, Principal Ideal Domains, and Unique Factorisation Domains

11.3.1 Euclidean Domains

Definition 11.49 (Euclidean domain)**.** An integral domain R is called a ***Euclidean domain*** if there exists $d : R \setminus \{0\} \to \mathbb{Z}_{\geq 0}$ that satisfies: for all $a, b \in R$, $b \neq 0$, there exists $q, r \in R$ such that $a = bq + r$ and $r = 0$ or $d(r) < d(b)$.

Example 11.50.

- Consider $\mathbb{Z}$. Let $a, b \in \mathbb{Z}$, $b \neq 0$. Then $a = bq + r$, $d : \mathbb{Z} \setminus \{0\} \to \mathbb{Z}_{\geq 0}$ and $d(x) = |x|$.
- Let F be a field, $F[x]$ be the ring of polynomials with elements of F as coefficients. Consider long division. $d : F[x] \setminus \{0\} \to \mathbb{Z}_{\geq 0}$ and $d(f(x)) := \deg f$.
- Consider $\mathbb{Z}[i] = \{a + bi \mid a, b \in \mathbb{Z}\}$ where $i^2 = -1$. Then $\mathbb{Z}[i]$ is an integral domain with unit $1 = 1 + 0i$ Then $d : \mathbb{Z}[i] \setminus \{0\} \to \mathbb{Z}_{\geq 0}$ under $d(a + bi) = a^2 + b^2$.

Theorem 11.51. *Let R be a Euclidean domain, I be an ideal in R. Then there exists $a_0 \in R$ such that $I = Ra_0$, i.e., I is a principal ideal.*

11.3.2 Principal Ideal Domains

Definition 11.52 (Principal ideal domain)**.** A commutative ring R such that all ideals are principal is called a ***principal ideal domain*** (PID).

Proposition 11.53. *Every Euclidean domain is a PID.*

Proposition 11.54. *Every field is a Euclidean domain.*

11.3.3 Unique Factorisation Domains

11.4 Polynomial Rings

11.4.1 Polynomials and Polynomial Functions

Let R be a commutative ring. Define the ***polynomial ring***

$$R[t] := \{a_0 + a_1 t + \cdots + a_n t^n \mid a_i \in R\}.$$

That is, $R[t]$ is the set of polynomials in t with coefficients in R.

Remark. A rigorous definition of the polynomial ring can be found in [Lan05].

Let R be a subring of a commutative ring S. If $f \in R[t]$ is a polynomial, then we may define the associated *polynomial function*

$$f_S \colon S \to S$$

by letting for $x \in S$

$$f_S(x) = f(x) = a_0 + a_1 x + \cdots + a_n x^n.$$

Hence f_S is a function (mapping) from S to itself, determined by the polynomial f.

Given $x \in S$, there is a homomorphism $R[t] \to S$ which maps $f \mapsto f_S(x)$. (show why)

11.4.2 Greatest Common Divisor

11.4.3 Unique Factorisation

11.4.4 Partial Fractions

11.4.5 Polynomials Over Rings and Over the Integers

11.4.6 Principal Rings and Factorial Rings

11.4.7 Polynomials in Several Variables

11.4.8 Symmetric Polynomials

11.4.9 The Mason–Stothers Theorem

11.4.10 The abc Conjecture

Chapter 12

Vector Spaces and Modules

12.1 Vector Spaces and Bases

Definition 12.1 (Vector space)**.** A vector space V over a field K is an additive (abelian) group, together with a multiplication of elements of V by elements of V by elements of K, i.e. an association

$$(x, v) \mapsto xv$$

from $K \times V$ to V, satisfying the following conditions:

12.1.1 Dimension of a Vector Space

12.1.2 Matrices and Linear Maps

12.2 Modules

We may consider a generalisation of the notion of vector space over a field, namely *module* over a ring.

Definition 12.2 (Module).

12.3 Factor Modules

12.4 Free Abelian Groups

12.5 Modules over Principal Rings

12.6 Eigenvectors and Eigenvalues

12.7 Polynomials of Matrices and Linear Maps

IV
Real Analysis

Real analysis deals with the real numbers and real-valued functions of a real variable.

A great part of analysis deals with inequalities and error terms. This is evident from the very beginning, in the theory of epsilons and deltas. Instead of obtaining precise values, it is sufficient to show that epsilon and delta are within a certain range. In order to show convergence, we just need to show that the error terms are small. Thus, there is often no perfect bound or best approximation, and there need not be; all that is needed is for the bound or the approximation to be good enough.

Chapter 13

Real and Complex Number Systems

Summary

- Supremum, infimum.
- Construction and properties of the real field $\mathbb{R}$.
- Construction and properties of the complex field $\mathbb{C}$.
- Construction and properties of the Euclidean space $\mathbb{R}^n$.

13.1 Ordered Sets and Boundedness

13.1.1 Definitions

Let S be a set.

Definition 13.1. An ***order*** on S is a binary relation $<$ such that

(i) for all $x, y \in S$, exactly one of $x < y$, $x = y$, or $y < x$ holds; (trichotomy)

(ii) if $x, y, z \in S$ are such that $x < y$ and $y < z$, then $x < z$. (transitivity)

S is an ***ordered set*** if it has an order; denote it by $(S, <)$.

Notation. We write $x \leq y$ if $x < y$ or $x = y$. We define $>$ and $\geq$ in the obvious way.

Definition 13.2 (Boundedness)**.** Let $E \subset S$, where S is an ordered set.

(i) E is ***bounded above*** if there exists $\beta \in S$ such that $x \leq \beta$ for all $x \in E$; we call β an *upper bound* of E.

(ii) E is ***bounded below*** if there exists $\beta \in S$ such that $x \geq \beta$ for all $x \in E$; we call β a *lower bound* of E.

E is ***bounded*** in S if it is bounded above and below.

Definition 13.3 (Supremum, infimum)**.** We say $\alpha \in S$ is the ***supremum*** of E if

(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).

Likewise, we say $\alpha \in S$ is the ***infimum*** of E if

(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).

Remark. It is not necessary for the supremum and infimum of E to be in E.

Lemma 13.4 (Uniqueness of suprenum)**.** *If E has a supremum, then it is unique.*

Proof. Suppose α are 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, so $\alpha \leq \beta$. Interchanging the roles of α and β gives $\beta \leq \alpha$. Hence $\alpha = \beta$. □

Since the supremum and infimum are unique, we can give them a notation.

Notation. Denote the supremum of E by $\sup E$, the infimum by $\inf E$.

Example 13.5. Let $E = \left\{\frac{1}{n} \,\middle|\, n \in \mathbb{N}\right\}$. Then $\sup E = 1$, $\inf E = 0$.

Proof. It is clear that 1 is an upper bound for E. Suppose $\beta < 1$. Since $1 \in E$, evidently β is not an upper bound for E. Hence $\sup E = 1$.

It is clear that 0 is a lower bound for E. Suppose $\beta > 0$. Pick $n = \left\lfloor \frac{1}{\beta} \right\rfloor + 1$, then $\beta > \frac{1}{n}$, so β is not a lower bound for E. Hence $\inf E = 0$. □

13.1.2 Least-upper-bound Property

Definition 13.6. An ordered set S has the ***least-upper-bound property*** (l.u.b.) if every non-empty subset of S that is bounded above has a supremum in S.
We define the *greatest-lower-bound property* similarly.

Proposition 13.7. *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,

$$L = \{y \in S \mid y \leq x \forall x \in B\}.$$

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 $\gamma \geq \sup L$. 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 13.8. *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.

13.1.3 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.

The next result shows that a set with a supremum contains numbers arbitrarily close to its supremum.

Lemma 13.9 (Approximation property). *Let $S \subset \mathbb{R}$ be non-empty, $b = \sup S$. Then for every $a < b$ there exists $x \in S$ such that*

$$a < x \le b.$$

Proof. We first show $x \le b$. Since $b = \sup S$ is an upper bound of S, $x \le b$ for all $x \in S$.

We now show there exist $x \in S$ such that $a < x$. Suppose otherwise, for a contradiction, that $x \le a$ 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.

Lemma 13.10 (Additive property). *Given non-empty $A, B \subset S$, 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

$$z = x + y \le a + b,$$

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) $a + b \ge c$, and (ii) $a + b \le c$.

(i) Since c is the *least* upper bound for C, and $a + b$ is an upper bound for C, we must have that $c \le a + b$.

(ii) Choose any $\varepsilon > 0$. By 13.9 there exist $x \in A$ and $y \in B$ such that

$$a - \varepsilon < x, \quad b - \varepsilon < y.$$

Adding these inequalities gives

$$a + b - 2\varepsilon < x + y \le c.$$

Thus $a + b < c + 2\varepsilon$ for every $\varepsilon > 0$. Hence $a + b \le c$.

□

Lemma 13.11 (Comparison property)**.** *Let non-empty $A, B \subset S$ such that $a \leq b$ for every $a \in A$, $b \in B$. If B has a supremum, then A has a supremum, and*

$$\sup A \leq \sup B.$$

Proof. Let $\beta = \sup B$. Since β is a supremum for B, then $b \leq \beta$ for all $b \in B$.

Let $a \in A$ and choose any $b \in B$. Since $a \leq b$ and $t \leq \beta$, $a \leq \beta$. 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 $\alpha \leq \beta$, as desired. ☐

Lemma 13.12. *Let $B \subset S$ be non-empty and bounded below. Let*

$$A = -B := \{-b \mid b \in B\}.$$

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 $b \geq \beta$ for all $b \in B$, which implies $-b \leq -\beta$ for all $b \in B$. Hence $a \leq -\beta$ 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 $a \leq \sup A$ for all $a \in A$, so $b \geq -\sup A$ 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 $-\beta \geq \sup A$ (since $\sup A$ is the least upper bound), so $\beta \leq -\sup A$. Therefore $-\sup A$ is the greatest lower bound of B. ☐

13.1.4 Ordered Fields

Definition 13.13 (Ordered field)**.** A field F is an ***ordered field*** if there exists an order $<$ on F such that for all $x, y, z \in F$,

(i) if $y < z$ then $x + y < x + z$;

(ii) if $x > 0$ and $y > 0$ then $xy > 0$.

If $x > 0$, we call x *positive*; if $x < 0$, x is *negative*.

All the familiar rules for working with inequalities apply in every ordered field: Multiplication by positive [negative] quantities preserves [reverses] inequalities, no square is negative, etc. The following result lists some of these.

Lemma 13.14 (Basic properties)**.** *Let F be an ordered field, $x, y, z \in F$.*

(i) If $x > 0$ then $-x < 0$, and vice versa.

(ii) If $x > 0$ and $y < z$ then $xy < xz$.

(iii) If $x < 0$ and $y < z$ then $xy > xz$.

(iv) If $x \neq 0$ then $x^2 > 0$. In particular, $1 > 0$.

(v) If $0 < x < y$ then $0 < \frac{1}{y} < \frac{1}{x}$.

Proof.

(i) If $x > 0$ then $0 = -x + x > -x + 0$, so that $-x < 0$.

If $x < 0$ then $0 = -x + x < -x + 0$, so that $-x > 0$.

(ii) Since $z > y$, we have $z - y > y - y = 0$, so $x(z - y) > 0$. Hence

$$xz = x(z - y) + xy > 0 + xy = xy.$$

(iii) By (i) and (ii),

$$-[x(z - y)] = (-x)(z - y) > 0,$$

so that $x(z - y) < 0$. Hence $xz < xy$.

(iv) If $x > 0$, part (ii) of the above definition gives $x^2 > 0$.

If $x < 0$, then $-x > 0$ so $(-x)^2 > 0$. But $x^2 = (-x)^2$.

Since $1 = 1^2$, $1 > 0$.

(v) If $y > 0$ and $v \leq 0$, then $yv \leq 0$. But $y\left(\frac{1}{y}\right) = 1 > 0$, so $\frac{1}{y} > 0$. Likewise, $\frac{1}{x} > 0$.

Multiplying both sides of the inequality $x < y$ by the positive quantity $\left(\frac{1}{x}\right)\left(\frac{1}{y}\right)$, we obtain $\frac{1}{y} < \frac{1}{x}$.

□

13.2 Real Numbers

13.2.1 Problems with $\mathbb{Q}$

$\mathbb{Q}$ has some problems, the first of which being *algebraic incompleteness*: there exists equations with coefficients in $\mathbb{Q}$ but do not have solutions in $\mathbb{Q}$ (in fact $\mathbb{R}$ has this problem too, but $\mathbb{C}$ is algebraically complete, by the fundamental theorem of algebra).

Lemma 13.15. $x^2 - 2 = 0$ *has no solution in* $\mathbb{Q}$*.*

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. $\square$

The second problem is *analytic incompleteness*: there exists a sequence of rational numbers that approach a point that is not in $\mathbb{Q}$; 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 13.16. *Let*

$$\begin{aligned} A &= \{p \in \mathbb{Q} \mid p > 0, p^2 < 2\}, \\ B &= \{p \in \mathbb{Q} \mid p > 0, p^2 > 2\}. \end{aligned}$$

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. $\square$

Proposition 13.17. $\mathbb{Q}$ *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 $A \subset \mathbb{Q}$ is bounded above but A has no least upper bound in $\mathbb{Q}$. □

13.2.2 Real Field

The sole objective of this subsection is to prove the following result.

Theorem 13.18 (Existence of real field)**.** *There exists an ordered field $\mathbb{R}$ that*

(i) contains $\mathbb{Q}$ as a subfield, and

(ii) has the least-upper-bound property (also known as the completeness axiom).

We want to construct $\mathbb{R}$ from $\mathbb{Q}$; one method to do so is using Dedekind cuts.

Definition 13.19 (Dedekind cut)**.** $\alpha \subset \mathbb{Q}$ is a ***Dedekind cut***, if

(i) $\alpha \neq \emptyset$, $\alpha \neq \mathbb{Q}$; (non-trivial)

(ii) if $p \in \alpha$, $q \in \mathbb{Q}$ 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 13.20. Let $r \in \mathbb{Q}$ and define

$$\alpha_r := \{p \in \mathbb{Q} \mid p < r\}.$$

We now check that this is indeed a Dedekind cut.

(i) $p = 1 + r \notin \alpha_r$ thus $\alpha_r \neq \mathbb{Q}$. $p = r - 1 \in \alpha_r$ 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 $\dfrac{q+r}{2} \in \mathbb{Q}$ 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.

Example 13.21. $\sqrt[3]{2}$ is not rational, but it is real. $\sqrt[3]{2}$ corresponds to the cut

$$\alpha = \{p \in \mathbb{Q} \mid p^3 < 2\}.$$

(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 13.22. The set of real numbers, denoted by $\mathbb{R}$, is the set of all Dedekind cuts:

$$\mathbb{R} := \{\alpha \subset \mathbb{Q} \mid \alpha \text{ is a Dedekind cut}\}.$$

Proposition 13.23. *$\mathbb{R}$ 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 13.24. *The ordered set $\mathbb{R}$ has the least-upper-bound property.*

Proof. Let non-empty $A \subset \mathbb{R}$ be bounded above. Let $\beta \in \mathbb{R}$ be an upper bound of A. We want to show that A has a supremum in $\mathbb{R}$.

Let

$$\gamma = \bigcup_{\alpha \in A} \alpha.$$

Then $p \in \gamma$ if and only if $p \in \alpha$ for some $\alpha \in A$.

Claim. $\gamma \in \mathbb{R}$ and $\gamma = \sup A$.

We first prove that $\gamma \in \mathbb{R}$ by checking that it is a Dedekind cut:

(i) Since $A \neq \emptyset$, there exists $\alpha_0 \in A$. Since $\alpha_0 \in \mathbb{R}$, 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 $\gamma \neq \mathbb{Q}$.

(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 $\alpha \leq \gamma$ 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 $\mathbb{R}$ is also known as the *completeness axiom* of $\mathbb{R}$.

We now define operations on $\mathbb{R}$.

Definition 13.25 (Addition). Given $\alpha, \beta \in \mathbb{R}$, define addition as

$$\alpha + \beta := \{r \in \mathbb{Q} \mid r = a + b, a \in \alpha, b \in \beta\}.$$

We first check if the above definition makes sense. We want to show that addition on $\mathbb{R}$ is closed: for all $\alpha, \beta \in \mathbb{R}$, $\alpha + \beta \in \mathbb{R}$.

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 $\alpha \neq \mathbb{Q}$ and $\beta \neq \mathbb{Q}$, 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 $\alpha + \beta \neq \mathbb{Q}$.

(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 $\beta \in \mathbb{R}$, $r' - a \in \beta$ 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 $r < r'$.

□

Lemma 13.26.

(i) Addition on $\mathbb{R}$ is commutative: $\alpha + \beta = \beta + \alpha$ for all $\alpha, \beta \in \mathbb{R}$.

(ii) Addition on $\mathbb{R}$ is associative: $\alpha + (\beta + \gamma) = (\alpha + \beta) + \gamma$ for all $\alpha, \beta, \gamma \in \mathbb{R}$.

(iii) Additive identity: Define $0^* := \{p \in \mathbb{Q} \mid p < 0\}$. Then $\alpha + 0^* = \alpha$ for all $\alpha \in \mathbb{R}$.

(iv) Additive inverse: Fix $\alpha \in \mathbb{R}$, define $\beta = \{p \in \mathbb{Q} \mid \exists r > 0,\ -p - r \notin \alpha\}$. 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 $\mathbb{Q}$. 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 $\mathbb{Q}$. 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 $\alpha \in \mathbb{R}$. 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 $\beta \neq \mathbb{Q}$.

(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 $\mathbb{Q}$, there exists $n \in \mathbb{N}$ 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 $-\alpha$.

Lemma 13.27. *If $\alpha, \beta, \gamma \in \mathbb{R}$ 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 $\mathbb{R}^+$ (the set of all $\alpha \in \mathbb{R}$ with $\alpha > 0^*$).

Definition 13.28. Given $\alpha, \beta \in \mathbb{R}^+$, define multiplication as

$$\alpha\beta := \{p \in \mathbb{Q} \mid p \leq rs,\ r \in \alpha, s \in \beta,\ r, s > 0\}.$$

We also define $1^* := \{q \in \mathbb{Q} \mid q < 1\}$.

As again, check if the above definition makes sense. We want to show that multiplication on $\mathbb{R}^+$ is closed: for all $\alpha, \beta \in \mathbb{R}$, $\alpha\beta \in \mathbb{R}$.

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 $rs \in \mathbb{Q}$ and $rs \leq rs$, so $rs \in \alpha\beta$. Hence $\alpha\beta \neq \emptyset$.

$\alpha \neq \mathbb{Q}$ means there exists $r' \notin \alpha$ such that $r' > r$ for all $r \in \alpha$. Similarly $\beta \neq \mathbb{Q}$ 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 $\alpha\beta \neq \mathbb{Q}$.

(ii) Let $p \in \alpha\beta$. Then $p \leq ab$ for some $a \in \alpha, b \in \beta$, $a, b > 0$.

If $q < p$, then $q < p \leq ab$ so $q \in \alpha\beta$.

(iii) Let $p \in \alpha\beta$. Then $p \leq ab$ 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 $a'b' > ab \geq p$, $a'b' \leq a'b'$ 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.

Lemma 13.29.

(i) Multiplication on $\mathbb{R}$ is commutative: $\alpha\beta = \beta\alpha$ for all $\alpha, \beta \in \mathbb{R}$.

(ii) Multiplication on $\mathbb{R}$ is associative: $(\alpha\beta)\gamma = \alpha(\beta\gamma)$ for all $\alpha, \beta, \gamma \in \mathbb{R}$.

(iii) Multiplicative identity: $1\alpha = \alpha$ for all $\alpha \in \mathbb{R}$.

(iv) Multiplicative inverse: If $\alpha \in \mathbb{R}$, $\alpha \neq 0^*$, then there exists $\beta \in \mathbb{R}$ such that $\alpha\beta = 1^*$.

We associate each $r \in \mathbb{Q}$ with the set

$$r^* = \{p \in \mathbb{Q} \mid p < r\}.$$

It is obvious that each r^* is a cut; that is, $r^* \in \mathbb{R}$.

Proposition 13.30. *The replacement of $r \in \mathbb{Q}$ by the corresponding "rational cuts" $r^* \in \mathbb{R}$ preserves sums, products, and order. That is, for all $r^*, s^* \in \mathbb{R}$,*

(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 $r \leq p < s$, so $r < s$.

$\square$

This shows that the ordered field $\mathbb{Q}$ is isomorphic to the ordered field $\mathbb{Q}^* = \{q^* \mid q \in \mathbb{Q}\}$ whose elements are rational cuts. It is this identification of $\mathbb{Q}$ with $\mathbb{Q}^*$ which allows us to regard $\mathbb{Q}$ as a subfield of $\mathbb{R}$.

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

Therefore we have proven 13.18.

13.2.3 Properties of $\mathbb{R}$

Proposition 13.31 (Archimedean property)**.** *For any $x \in \mathbb{R}^+$, $y \in \mathbb{R}$, there exists $n \in \mathbb{N}$ such that*

$$nx > y.$$

Proof. Suppose, for a contradiction, that $nx \le y$ for all $n \in \mathbb{N}$. Then y is an upper bound of the set

$$A = \{nx \mid n \in \mathbb{N}\}.$$

Since $A \subset R$ is non-empty and bounded above, by the l.u.b. property of $\mathbb{R}$, A has a supremum in $\mathbb{R}$, say $\alpha = \sup A$.

Consider $\alpha - x$. Since $\alpha - x < \alpha = \sup A$, $\alpha - x$ is not an upper bound of A. Then $\alpha - x \le n_0 x$ for some $n_0 \in \mathbb{N}$; rearranging gives $\alpha \le (n_0 + 1)x$. This implies that α is not an upper bound of A, which contradicts the fact that α is the supremum of A. □

Corollary 13.32. *Let $\varepsilon > 0$. Then there exists $n \in \mathbb{N}$ such that $0 < \frac{1}{n} < \varepsilon$.*

Proof. In 13.31, take $x = \varepsilon$ and $y = 1$. □

Proposition 13.33 ($\mathbb{Q}$ is dense in $\mathbb{R}$)**.** *For any $x, y \in \mathbb{R}$ with $x < y$, there exists $p \in \mathbb{Q}$ such that*

$$x < p < y.$$

Proof. We prove by construction (construct the required p from the given x and y).

Since $x < y$, we have $y - x > 0$. By 13.32, there exists $n \in \mathbb{N}$ such that

$$\frac{1}{n} < y - x.$$

Consider the set comprising multiples of $\frac{1}{n}$:

$$E = \left\{ \frac{k}{n} \;\middle|\; k \in \mathbb{N} \right\}.$$

Since E is unbounded, choose the first multiple $m \in \mathbb{N}$ such that $\frac{m}{n} > x$.

Claim. $x < \frac{m}{n} < y$.

It suffices to show 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 13.34** ($\mathbb{R}$ is closed under taking roots)**.** *For every $x \in \mathbb{R}^+$ and every $n \in \mathbb{N}$, there exists a unique $y \in \mathbb{R}^+$ so that $y^n = x$.*

We call the number y the positive n-th root of x, and denote it by $\sqrt[n]{x}$ or $x^{\frac{1}{n}}$.

Proof. Let $x \in \mathbb{R}^+$, fix $n \in \mathbb{N}$.

$\boxed{\text{Existence}}$ Let

$$E = \{t \in \mathbb{R}^+ \mid t^n < x\}.$$

Claim. $y = \sup E$ satisfies $y^n = x$.

We first show that E has a supremum.

(i) Let $t = \dfrac{x}{1+x}$. Then $0 \le t < 1$, so $t^n \le t < x$ implies $t^n < x$. Hence $t \in E$, so $E \neq \emptyset$.

(ii) We claim that $1 + x$ is an upper bound for E.

If $t > 1 + x$, then $t^n \ge t > x$ implies $t^n > x$, so $t \notin E$. [This is the contrapositive of $t \in E \implies t \le 1 + x$.] Hence $1 + x$ is an upper bound of E, so E is bounded above.

Hence E has a supremum; let $y = \sup E$.

To prove that $y^n = x$, we show that both the inequalities $y^n < x$ and $y^n > x$ lead to a contradiction. Consider the identity $b^n - a^n = (b-a)\left(b^{n-1} + b^{n-2}a + \cdots + a^{n-1}\right)$. If $0 < a < b$, then

$$b^n - a^n < (b-a)nb^{n-1}. \tag{1}$$

Case 1: $y^n < x$**.**

Idea. We can find a *small* $h > 0$ such that $(y+h)^n < x$.

Choose h so that $0 < h < 1$ and

$$h < \frac{x - y^n}{n(y+1)^{n-1}}.$$

In (1), take $b = y + h$, $a = y$. Then

$$\begin{aligned}
(y+h)^n - y^n &< hn(y+h)^{n-1} \\
&< hn(y+1)^{n-1} \\
&< \frac{x - y^n}{\cancel{n(y+1)^{n-1}}}\cancel{n(y+1)^{n-1}} \\
&= x - y^n.
\end{aligned}$$

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.

Case 2: $y^n > x$**.**

Idea. Similarly, we can find a *small* $k > 0$ such that $(y-k)^n > x$.

Let

$$k = \frac{y^n - x}{ny^{n-1}}.$$

Then $0 < k < y$, by (1). If $t \geq y - k$,

$$\begin{aligned} y^n - t^n &\leq y^n - (y-k)^n \\ &< kny^{n-1} \\ &= \frac{y^n - x}{\cancel{ny^{n-1}}} \cancel{ny^{n-1}} \\ &= y^n - x. \end{aligned}$$

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.

$\boxed{\text{Uniqueness}}$ Suppose, for a contradiction, that there exist distinct y_1, y_2 which are both n-th roots of x. WLOG assume that $0 < y_1 < y_2$. Then taking the n-th power gives ${y_1}^n < {y_2}^n$.

Since y_1 is a n-th root of x, then $x = {y_1}^n$, so $x < {y_2}^n$ implies $x \neq {y_2}^n$. Hence y_2 cannot be a n-th root of x, a contradiction. □

Corollary 13.35. *If $a, b \in \mathbb{R}^+$ and $n \in \mathbb{N}$, 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

$$(ab)^{\frac{1}{n}} = \alpha\beta = a^{\frac{1}{n}} b^{\frac{1}{n}}.$$

□

Lemma 13.36. *If $x \in \mathbb{R}^+$ and $m, n \in \mathbb{N}$, then*

$$(x^{\frac{1}{n}})^m = (x^m)^{\frac{1}{n}}.$$

Proof. Exercise. □

We can now define *rational exponents* x^r, where $x > 0$ and $r \in \mathbb{Q}$.

Definition 13.37 (Rational exponents). For $x > 0$ and $m, n \in \mathbb{N}$, define

$$x^{\frac{m}{n}} := \left(x^{\frac{1}{n}}\right)^m \quad \text{and} \quad x^{-\frac{m}{n}} := \frac{1}{x^{\frac{m}{n}}}.$$

(We also define $x^0 = 1$.)

We need to check that the above definition of x^r is well defined. That is, if $m, n, p, q \in \mathbb{N}$ are such that

$\frac{m}{n} = \frac{p}{q}$, then $(x^{\frac{1}{n}})^m = (x^{\frac{1}{q}})^p$. To see this, note that $mq = np$ and

$$\left((x^{\frac{1}{n}})^m\right)^q = (x^{\frac{1}{n}})^{mq} = (x^{\frac{1}{n}})^{np} = x^p.$$

Thus $(x^{\frac{1}{n}})^m$ is the q-th root of x^p, i.e.,

$$(x^{\frac{1}{n}})^m = (x^p)^{\frac{1}{q}}.$$

Lemma 13.38 (Properties of rational exponents)**.**

(i) If $a > 0$ and $r, s \in \mathbb{Q}$, then $a^{r+s} = a^r a^s$ and $(a^r)^s = a^{rs}$.

(ii) If $0 < a < b$ and $r \in \mathbb{Q}$ with $r > 0$, then $a^r < b^r$.

(iii) If $a > 1$, $r, s \in \mathbb{Q}$ with $r < s$, then $a^r < a^s$.

The next result shows that real numbers can be approximated to any desired degree of accuracy by rational numbers with finite decimal representations.

Proposition 13.39. *Let $x \geq 0$. Then for every integer $n \geq 1$ there exists a finite decimal $r_n = a_0.a_1a_2\cdots a_n$ such that*

$$r_n \leq x < r_n + \frac{1}{10^n}.$$

Proof. We prove by construction (construct the required finite decimal from x).

Let

$$S = \{k \in \mathbb{Z} \mid k \leq x\}.$$

S is non-empty (since $0 \in S$), and S is bounded above by x. Hence by the lub property of $\mathbb{R}$, S has a supremum in $\mathbb{R}$, 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

$$a_0 \leq x < a_0 + 1.$$

Now let $a_1 = \lfloor 10(x - a_0) \rfloor$. Since $0 \leq 10(x - a_0) < 10$, we have $0 \leq a_1 \leq 9$ and

$$a_1 \leq 10x - 10a_0 < a_1 + 1.$$

In other words, a_1 is the largest integer satisfying the inequalities

$$a_0 + \frac{a_1}{10} \leq x < a_0 + \frac{a_1 + 1}{10}.$$

More generally, having chosen $a_1, \dots, a_{n-1}$ with $0 \leq a_i \leq 9$, let a_n be the largest integer satisfying the inequalities

$$a_0 + \frac{a_1}{10} + \cdots + \frac{a_n}{10^n} \leq a_0 + \frac{a_1}{10} + \cdots + \frac{a_n + 1}{10^n}.$$

Then $0 \leq a_n \leq 9$ and we have

$$r_n \leq x < r_n + \frac{1}{10^n},$$

where $r_n = a_0.a_1a_2\cdots a_n$. □

Furthermore, it is easy to verify that $x = \sup_{n\in\mathbb{N}} r_n$.

13.2.4 Extended Real Number System

Definition 13.40 (Extended real number system)**.** The ***extended real number system*** is defined to be the union

$$\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\},$$

where we preserve the original order in $\mathbb{R}$, and define $-\infty < x < +\infty$ for all $x \in \mathbb{R}$.

Defining $\overline{\mathbb{R}}$ is convenient since the following result holds.

Proposition 13.41. *Any non-empty $E \subset \overline{\mathbb{R}}$ has a supremum and infimum in $\overline{\mathbb{R}}$.*

Proof. If E is bounded above in $\mathbb{R}$, then by the l.u.b. property of $\mathbb{R}$, it has a supremum in $\mathbb{R} \subset \overline{\mathbb{R}}$. If E is not bounded above in $\mathbb{R}$, then $\sup E = +\infty \in \overline{\mathbb{R}}$.

Exactly the same remarks apply to lower bounds. □

$\overline{\mathbb{R}}$ does not form a field, but it is customary to make the following conventions for arithmetic on $\overline{\mathbb{R}}$:

(i) If $x \in \mathbb{R}$ 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*.

13.3 Complex Field

Lemma 13.42. *Let $(a,b),(c,d) \in \mathbb{R}^2$. Define addition and multiplication on $\mathbb{R}^2$ as*

$$\begin{aligned}(a,b)+(c,d) &= (a+c,b+d),\\ (a,b)(c,d) &= (ac-bd,ad+bc).\end{aligned}$$

Then $\mathbb{R}^2$ is a field, with additive identity $(0,0)$ and multiplicative identity $(1,0)$.

We call this structure $\mathbb{C}$, the ***complex field***; its elements are called *complex numbers*.

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 $(a,0) \in \mathbb{C}$ with $a \in \mathbb{R}$. This identification implies that $\mathbb{R}$ is a subfield of $\mathbb{C}$.

Lemma 13.43. *For any $a,b \in \mathbb{R}$, 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 $a+bi$.

Define the imaginary number $i := (0,1)$. See that

$$i^2 = (0,1)(0,1) = (-1,0) = -1.$$

Lemma 13.44. *For $a,b \in \mathbb{R}$, $(a,b) = a+bi$.*

Proof.

$$\begin{aligned}a+bi &= (a,0)+(b,0)(0,1)\\ &= (a,0)+(0,b)\\ &= (a,b).\end{aligned}$$

□

For $a,b \in \mathbb{R}$, we write $z = a+bi$; 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.

Lemma 13.45 (Properties of conjugate)**.** *For* $z, w \in \mathbb{C}$*,*

(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) $z\overline{z} \in \mathbb{R}$ *and* $z\overline{z} \geq 0$

For $z \in \mathbb{C}$, tbe *absolute value* of z is defined as

$$|z| := (z\overline{z})^{\frac{1}{2}} .$$

Lemma 13.46 (Properties of absolute value)**.** *For* $z, w \in \mathbb{C}$*,*

(i) $|z| \geq 0$

(ii) $|\overline{z}| = |z|$

(iii) $|zw| = |z||w|$

(iv) $|\operatorname{Re}(z)| \leq |z|$

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 $a, b, c, d \in \mathbb{R}$. Then

$$\begin{aligned} |zw|^2 &= (ac - bd)^2 + (ad - bc)^2 \\ &= (a^2 + b^2)(c^2 + d^2) \\ &= |z|^2|w|^2 = (|z||w|)^2 \end{aligned}$$

and the desired result follows by taking square roots on both sides.

(iv) Let $z = a + bi$. Note that $a^2 \leq a^2 + b^2$, hence

$$|\operatorname{Re}(z)| = |a| = \sqrt{a^2} \leq \sqrt{a^2 + b^2} = |z|.$$

□

Proposition 13.47 (Triangle inequality)**.** *For* $z, w \in \mathbb{C}$*,*

$$|z + w| \leq |z| + |w|. \tag{13.1}$$

Proof. Let $z, w \in \mathbb{C}$. Note that the conjugate of $z\overline{w}$ is $\overline{z}w$, so $z\overline{w} + \overline{z}w = 2\operatorname{Re}(z\overline{w})$. Hence

$$\begin{aligned}|z+w|^2 &= (z+w)(\overline{z+w}) = (z+w)(\overline{z}+\overline{w}) \\ &= z\overline{z} + z\overline{w} + \overline{z}w + w\overline{w} \\ &= |z|^2 + 2\operatorname{Re}(z\overline{w}) + |w|^2 \\ &\leq |z|^2 + 2|z\overline{w}| + |w|^2 \\ &= |z|^2 + 2|z||w| + |w|^2 \\ &= (|z|+|w|)^2\end{aligned}$$

and taking square roots yields the desired result. □

Corollary 13.48 (Generalised triangle inequality)**.** *For $z_1, \dots, z_n \in \mathbb{C}$,*

$$|z_1 + \cdots + z_n| \leq |z_1| + \cdots + |z_n|.$$

Proof. We have proven the case $n = 2$. Assume the statement holds for $n-1$. Then

$$|z_1 + \cdots + z_{n-1} + z_n| \leq |z_1 + \cdots + z_{n-1}| + |z_n| \leq |z_1| + \cdots + |z_n|,$$

which establishes the claim by induction. □

Corollary 13.49. *For $x, y, z \in \mathbb{C}$,*

(i) $||x| - |y|| \leq |x-y|$;

(ii) $|x-y| \leq |x-z| + |z-y|$.

Proof.

(i) By the triangle inequality,

$$|x| = |(x-y)+y| \leq |x-y| + |y|$$

so that

$$|x| - |y| \leq |x-y|.$$

Interchanging the roles of x and y in the above gives

$$|y| - |x| \leq |x-y|.$$

Hence

$$||x| - |y|| \leq |x-y|.$$

(ii) In the triangle inequality, replace x by $x-y$ and y by $y-z$.

□

Proposition 13.50 (Cauchy–Schwarz inequality). *If $a_1, \dots, a_n, b_1, \dots, b_n \in \mathbb{C}$, then*

$$\left| \sum_{i=1}^{n} a_i \overline{b_i} \right|^2 \leq \sum_{i=1}^{n} |a_i|^2 \sum_{i=1}^{n} |b_i|^2. \tag{13.2}$$

Proof. For simplicity, we shall drop the upper and lower limits of the sums. Let

$$A = \sum |a_i|^2, \quad B = \sum |b_i|^2, \quad C = \sum a_i \overline{b_i}.$$

Then (13.2) becomes

$$|C|^2 \leq AB.$$

If $B = 0$, then $b_1 = \cdots = b_n = 0$, and the conclusion is trivial. Now assume that $B > 0$. Then consider the sum

$$\begin{aligned}
\sum |Ba_i - Cb_i|^2 &= \sum (Ba_i - Cb_i)(\overline{Ba_i - Cb_i}) \\
&= \sum (Ba_i - Cb_i)(B\overline{a_i} - \overline{Cb_i}) \\
&= B^2 \sum |a_i|^2 - B\overline{C} \sum a_i \overline{b_i} - BC \sum \overline{a_i} b_i + |C|^2 \sum |b_i|^2 \\
&= B^2 A - B|C|^2 \\
&= B(AB - |C|^2).
\end{aligned}$$

Each term in $\sum |Ba_i - Cb_i|^2$ is non-negative, so $\sum |Ba_i - Cb_i|^2 \geq 0$. Thus

$$B(AB - |C|^2) \geq 0.$$

Since $B > 0$, it follows that $AB - |C|^2 \geq 0$, or $|C|^2 \leq AB$. This is the desired inequality.

(when does equality hold?) □

Define

$$\mathbb{C}^n = \{(z_1, \dots, z_n) \mid z_i \in \mathbb{C}\}.$$

We can define an inner product on $\mathbb{C}^n$: for $\mathbf{a}, \mathbf{b} \in \mathbb{C}^n$,

$$\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i=1}^{n} a_i \overline{b_i}.$$

We can also define the norm of $\mathbf{a} \in \mathbb{C}^n$:

$$|\mathbf{a}| = \langle \mathbf{a}, \mathbf{a} \rangle^{\frac{1}{2}}.$$

13.4 Euclidean Space

For $n \in \mathbb{N}$, define

$$\mathbb{R}^n := \{(x_1, \dots, x_n) \mid x_i \in \mathbb{R}\}$$

where $\mathbf{x} = (x_1, \dots, x_n)$, x_i's are called the coordinates of $\mathbf{x}$. The elements of $\mathbb{R}^n$ are called *points*, or *vectors*.

Lemma 13.51. *Let* $\mathbf{x} = (x_1, \dots, x_n)$, $\mathbf{y} = (y_1, \dots, y_n)$. $\mathbb{R}^n$, *with addition and scalar multiplication defined as*

$$\begin{aligned} \mathbf{x} + \mathbf{y} &= (x_1 + y_1, \dots, x_n + y_n), \\ \alpha\mathbf{x} &= (\alpha x_1, \dots, \alpha x_n). \end{aligned}$$

for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$, *is a vector space over* $\mathbb{R}$. *Note that the zero element of* $\mathbb{R}^n$ *is* $\mathbf{0} = (0, \dots, 0)$.

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

$$\mathbf{x} \cdot \mathbf{y} := \sum_{i=1}^{n} x_i y_i,$$

and the *norm* of $\mathbf{x}$ by

$$\|\mathbf{x}\| := \sqrt{\mathbf{x} \cdot \mathbf{x}}.$$

The structure now defined (the vector space $\mathbb{R}^n$ with the above inner product and norm) is called the ***Euclidean n-space***.

Lemma 13.52 (Basic properties of norm)**.** *Suppose* $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$.

(i) $\|\mathbf{x}\| \geq 0$, *where equality holds if and only if* $\mathbf{x} = \mathbf{0}$ *(positive definiteness)*

(ii) $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ *(homogeneity)*

(iii) $\|\mathbf{x} \cdot \mathbf{y}\| \leq \|\mathbf{x}\|\|\mathbf{y}\|$ *(Cauchy–Schwarz inequality)*

(iv) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ *(triangle inequality)*

(v) $\|\mathbf{x} - \mathbf{z}\| \leq \|\mathbf{x} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{z}\|$ *(triangle inequality)*

Proof.

(i) Obvious from definition.

(ii) Obvious from definition.

(iii) We want to show

$$\sqrt{\sum_{i=1}^{n} x_i y_i} \leq \sqrt{\sum_{i=1}^{n} {x_i}^2} \sqrt{\sum_{i=1}^{n} {y_i}^2},$$

or, squaring both sides,

$$\sum_{i=1}^{n} x_i y_i \leq \left(\sum_{i=1}^{n} {x_i}^2 \right) \left(\sum_{i=1}^{n} {y_i}^2 \right).$$

But this is simply the Cauchy–Schwarz inequality (13.2).

(iv) By (iii) we have

$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\| &= (\mathbf{x} + \mathbf{y}) \cdot (\mathbf{x} + \mathbf{y}) \\ &= \mathbf{x} \cdot \mathbf{x} + 2\mathbf{x} \cdot \mathbf{y} + \mathbf{y} \cdot \mathbf{y} \\ &\leq \|\mathbf{x}\|^2 + 2\|\mathbf{x}\|\|\mathbf{y}\| + \|\mathbf{y}\|^2 \\ &= (\|\mathbf{x}\| + \|\mathbf{y}\|)^2 . \end{aligned}$$

(v) This follows directly from (iv) by replacing $\mathbf{x}$ by $\mathbf{x} - \mathbf{y}$, and $\mathbf{y}$ by $\mathbf{y} - \mathbf{z}$.

□

Exercises

Exercise 13.1 ([Rud76] 1.1)**.** If $r \in \mathbb{Q} \setminus \{0\}$ and $x \in \mathbb{R} \setminus \mathbb{Q}$, prove that $r + x \in \mathbb{R} \setminus \mathbb{Q}$ and $rx \in \mathbb{R} \setminus \mathbb{Q}$.

Solution. Prove by contradiction. If r and $r + x$ were both rational, then $x = (r + x) - r$ would also be rational. Similarly if rx were rational, then $x = \frac{rx}{r}$ would also be rational. □

Exercise 13.2 ([Rud76] 1.2)**.** Prove that there is no rational number whose square is 12.

Solution. Prove by contradiction. □

Exercise 13.3 ([Rud76] 1.4)**.** 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 $\alpha \le \beta$.

Solution. Since E is non-empty, there exists $x \in E$. By definition of lower and upper bounds, we have $\alpha \le x \le \beta$. □

Exercise 13.4 ([Rud76] 1.8)**.** Prove that no order can be defined in $\mathbb{C}$ that turns it into an ordered field. *Hint*: -1 is a square.

Solution. By 13.14, an order $<$ that makes $\mathbb{C}$ an ordered field would have to satisfy $-1 = i^2 > 0$, contradicting $1 > 0$. □

Exercise 13.5 ([Rud76] 1.9, lexicographic order)**.** Suppose $z = a + bi$, $w = c + di$. Define an order on $\mathbb{C}$ as follows:

$$z < w \iff \begin{cases} a < c, \text{ or} \\ a = c, b < d. \end{cases}$$

Prove that this turns $\mathbb{C}$ into an ordered set. Does this ordered set have the least upper bound property?

Solution. We show that this order turns $\mathbb{C}$ into an ordered set.

(i) Since the *real* numbers are ordered, we have $a < c$ or $a = c$ or $c < a$. In the first case $z < w$; in the third case $w < z$.

Now consider the second case where $a = c$. We must have $b < d$ or $b = d$ or $d < b$, which correspond to $z < w$, $z = w$, $w < z$ respectively.

Hence we have shown that either $z < w$ or $z = w$ or $w < z$.

(ii) We now show that if $z < w$ and $w < u$, then $z < u$. Let $u = e + fi$.

Since $z < w$, we have either $a < c$, or $a = c$ and $b < d$. Since $w < u$, we have either $c < f$, or $c = f$ and $d < g$. Hence there are four possible cases:

- $a < c$ and $c < f$. Then $a < f$ and so $z < u$, as required.
- $a < c$ and $c = f$, and $d < g$. Again $a < f$, so $z < u$.
- $a = c$, and $b < d$ and $c < f$. Once again $a < f$ so $z < u$.
- $a = c$ and $b < d$, and $c = f$ and $d < g$. Then $a = f$ and $b < g$, so $z < u$.

□

Exercise 13.6 ([Rud76] 1.10)**.** 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 $v \geq 0$ and that $\overline{z}^2 = w$ if $v \leq 0$. 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 if $v \geq 0$,

$$z^2 = \left(a^2 - b^2\right) + 2abi = u + |v|i = w;$$

if $v \leq 0$,

$$\overline{z}^2 = \left(a^2 - b^2\right) - 2abi = u - |v|i = w.$$

Hence every non-zero w has two square roots $\pm z$ or $\pm \overline{z}$. Of course, 0 has only one square root, itself. □

Exercise 13.7 ([Rud76] 1.11)**.** If $z \in \mathbb{C}$, prove that there exists $r \geq 0$ and $w \in \mathbb{C}$ with $|w| = 1$ such that $z = rw$. Are w and r always uniquely determined by z?

Solution. If $z = 0$, take $r = 0$ and $w = 1$; in this case w is not unique.

Otherwise take $r = |z|$ and $w = \frac{z}{|z|}$; these choices are unique, since if $z = rw$, we must have $r = r|w| = |rw| = |z|$ so $w = \frac{z}{r} = \frac{z}{|z|}$ are unique. □

Chapter 14

Basic Topology

Summary

- Metric space, subspace. Open ball, closed ball, boundedness. Open set, closed set. Interior, closure, boundary. Limit point.
- Compactness. Cantor intersection theorem, Heine–Borel theorem, Bolzano–Weierstrass theorem. Sequential compactness.
- Perfect sets. Cantor set.
- Connectedness.

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	$B'_r(p)$
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 14.1: Notation for topological structures in Chapter 14

14.1 Metric Spaces

14.1.1 Definitions and Examples

Definition 14.1 (Normed space). Let X be a vector space. A *norm* is a function $\|\cdot\| : X \to [0, \infty)$ if, for all $x, y \in X$ and constants α,

(i) $\|x\| \geq 0$, where equality holds if and only if $x = 0$; (positive definiteness)

(ii) $\|\alpha x\| = |\alpha|\|x\|$; (homogeneity)

(iii) $\|x + y\| \leq \|x\| + \|y\|$. (triangle inequality)

A ***normed space*** $(V, \|\cdot\|)$ is a vector space V together with a norm $\|\cdot\|$.

Definition 14.2 (Metric space). Let X be a set. A *metric* is a function $d : X \times X \to [0, \infty)$ if, for all $x, y, z \in X$,

(i) $d(x, y) \geq 0$, where equality holds if and only if $x = y$; (positive definiteness)

(ii) $d(x, y) = d(y, x)$; (symmetry)

(iii) $d(x, y) \leq d(x, z) + d(z, y)$. (triangle inequality)

A ***metric space*** (X, d) is a set X together with a metric d.

For the rest of the chapter, X is taken to be a metric space, unless specified otherwise.

Lemma 14.3 (Norm induces metric). *Let X be a normed space. Then X is a metric space, with the metric $d(x, y) = \|x - y\|$ for every $x, y \in X$.*

Proof. Trivial; check the conditions for a metric. □

Example 14.4 (Metrics on $\mathbb{R}^n$). Each of the following functions define metrics on $\mathbb{R}^n$.

$$d_1(x, y) = \sum_{i=1}^{n} |x_i - y_i|;$$
$$d_2(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)}$$
$$d_\infty(x, y) = \max_{i \in \{1,2,\dots,n\}} |x_i - y_i|.$$

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 $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is

$$\|x\|_2 := \left(\sum_{i=1}^{n} {x_i}^2\right)^{\frac{1}{2}} = \langle x, x\rangle^{\frac{1}{2}},$$

where the inner product is given by

$$\langle x, y\rangle := \sum_{i=1}^{n} x_i y_i.$$

Then $d_2(x, y) = \|x - y\|_2$, and so the triangle inequality is the statement that

$$\|w - y\|_2 \leq \|w - x\|_2 + \|x - y\|_2.$$

This follows immediately by taking $u = w - x$ and $v = x - y$ in the following lemma.

Lemma. *If* $u, v \in \mathbb{R}^n$ *then* $\|u + v\|_2 \leq \|u\|_2 + \|v\|_2$.

Proof. Since $\|u\|_2 \geq 0$ for all $u \in \mathbb{R}^n$, squaring both sides of the desired inequality gives

$$\|u + v\|_2{}^2 \leq \|u\|_2{}^2 + 2\|u\|_2\|v\|_2 + \|v\|_2{}^2.$$

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

$$|\langle u, v\rangle| \leq \|u\|_2\|v\|_2.$$

□

A metric space (X, d) naturally induces a metric on any of its subsets.

Definition 14.5 (Subspace)**.** Suppose (X, d) is a metric space, $Y \subset X$. Then the restriction of d to $Y \times Y$ gives Y a metric

$$d_Y = d|_{Y \times Y},$$

so that (Y, d_Y) is a metric space. We call Y equipped with this metric a ***subspace***.

14.1.2 Balls and Boundedness

Definition 14.6 (Balls)**.**

(i) The ***open ball*** centred at $x \in X$ with radius $r > 0$ is the set

$$B_r(x) := \{y \in X \mid d(y, x) < r\}.$$

(ii) The ***closed ball*** centred at x with radius r is

$$\overline{B}_r(x) := \{y \in X \mid d(y, x) \leq r\}.$$

(iii) The ***punctured ball*** is the open ball excluding its centre:

$$B'_r(x) := \{y \in X \mid 0 < d(y, x) < r\}.$$

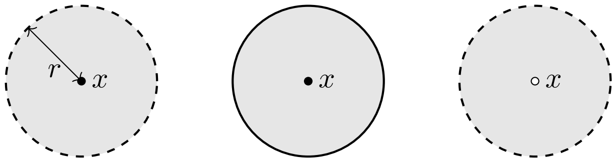


Figure 14.1: Open ball, closed ball, punctured ball

Example 14.7. Considering $\mathbb{R}^3$ 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 $B_1(x)$ contains only the point x, whereas the closed ball $\overline{B}_1(x)$ is the whole of X.

Definition 14.8 (Bounded)**.** We say $E \subset X$ is ***bounded*** if E is contained in some open ball; that is, there exists $M \in \mathbb{R}$ and $x \in X$ such that $E \subset B_M(x)$.

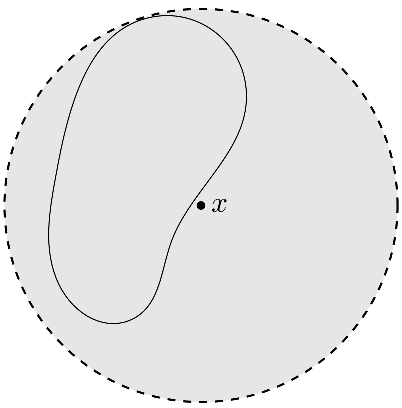


Figure 14.2: Bounded set

Proposition 14.9. *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 $\mathbb{R}$.*

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 $r \in \mathbb{R}$ such that $d(x,y) \leq r$ for all $x, y \in E$. If $E = \emptyset$, then E is certainly bounded. Otherwise, let $p \in E$ be an arbitrary point. Then $E \subset B_{r+1}(p)$. □

14.1.3 Open and Closed Sets

We say $N \subset X$ is a ***neighbourhood*** of $x \in X$ if there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subset N$.

Definition 14.10 (Open set)**.** We say $E \subset X$ is ***open*** (in X) if it is a neighbourhood of all its elements; that is, for all $x \in E$, there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subset E$.

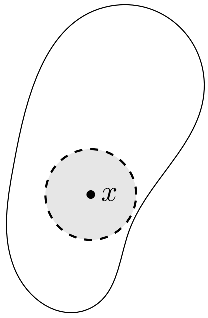
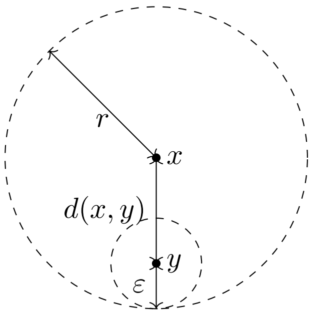


Figure 14.3: Open set

Lemma 14.11. *Any open ball is open.*



Proof. Let $B_r(x)$ be an open ball.

Let $y \in B_r(x)$. To show that $B_r(x)$ is open, we will show that $B_\varepsilon(y) \subset B_r(x)$ for some $\varepsilon > 0$.

Take $\varepsilon = r - d(x, y)$. Let $z \in B_\varepsilon(y)$. By the triangle inequality,

$$\begin{aligned} d(x,z) &\leq d(y,z) + d(x,y) \\ &< \varepsilon + d(x,y) = r \end{aligned}$$

so $z \in B_r(x)$, which implies $B_\varepsilon(y) \subset B_r(x)$. $\square$

Lemma 14.12.

(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 $x \in \bigcup_{i\in I} E_i$, then $x \in E_i$ for some $i \in I$. Since E_i is open, there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subset E_i$ and hence $B_\varepsilon(x) \in \bigcup_{i\in I} E_i$.

(iii) Suppose I is finite and $x \in \bigcap_{i\in I} E_i$. For each $i \in I$, we have $x \in E_i$ and so there exists ε_i such that $B_{\varepsilon_i}(x) \subset E_i$. Set $\varepsilon = \min_{i\in I} \varepsilon_i$, then $\varepsilon > 0$ (here it is, of course, crucial that I be finite), and $B_\varepsilon(x) \subset B_{\varepsilon_i}(x) \subset E_i$ for all i. Therefore $B_\varepsilon(x) \subset \bigcap_{i\in I} E_i$.

□

Remark. While the indexing set I in (ii) can be arbitrary, the indexing set in (iii) must be finite. For instance, $E_n = \left(-\frac{1}{n}, \frac{1}{n}\right)$ are open in $\mathbb{R}$, but their intersection $\bigcap_{n=1}^{\infty} E_n = \{0\}$ is not open.

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 $d : X \times X \to \mathbb{R}$ on X.)

Proposition 14.13. *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 $E = Y \cap G$.*

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

$$\bigcup_{p\in E} \left(B_{r_p}(p) \cap Y\right) \subset E.$$

Note that we can write

$$\bigcup_{p\in E} \left(B_{r_p}(p) \cap Y\right) = \left(\bigcup_{p\in E} B_{r_p}(p)\right) \cap Y \subset E.$$

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 14.12, 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.

[1] notice that the definition of an open ball depends on the metric space!

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. $\square$

The complement of an open set is a *closed* set.

Definition 14.14 (Closed set)**.** We say $E \subset X$ is ***closed*** if its complement $E^c = X \setminus E$ is open.

Lemma 14.15. *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,

$$\begin{aligned} d(p,q) &\geq d(p,s) - d(q,s) \\ &> r \end{aligned}$$

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. $\square$

Lemma 14.16.

(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 14.12, simply take complements and apply de Morgan's laws. $\square$

Remark. The indexing set in (iii) must be finite; for instane, the closed intervals $F_n = \left[-1 + \frac{1}{n}, 1 - \frac{1}{n}\right]$ are all closed in $\mathbb{R}$, but their union $\bigcup_{n=1}^{\infty} F_n = (-1,1)$ is open.

14.1.4 Interior, Closure, Boundary

Definition 14.17. Suppose $E \subset X$.

(i) The ***interior*** E° of the set E is the union of all open subsets of X contained in E; we call $p \in E^\circ$ an ***interior point*** of E.

(ii) The ***closure*** $\overline{E}$ of the set E is the intersection of all closed subsets of X containing E.

We say E is ***dense*** if $\overline{E} = X$.

(iii) The ***boundary*** of E is $\partial E = \overline{E} \setminus E^\circ$; we call $p \in \partial E$ 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.

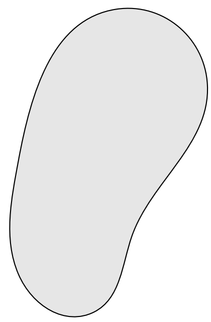


Figure 14.4: Interior, closure, boundary

Example 14.18.

- The interior of the closed interval $[a, b]$ is the open interval (a, b).
- $\mathbb{Q}$ is dense in $\mathbb{R}$.

Remark. E and E° do not necessarily have the same closures; for example, take $E = \mathbb{Q}$, then $\overline{E} = \mathbb{R}$ and $\overline{E^\circ} = \emptyset$.

Likewise, E and $\overline{E}$ do not necessarily have the same interiors; for example, take $E = (-1, 0) \cup (0, 1) \subset \mathbb{R}$. Then $E^\circ = (-1, 0) \cup (0, 1)$ and $(\overline{E})^\circ = [-1, 1]$.

Lemma 14.19. *Suppose $E \subset X$.*

(i) E is open if and only if $E = E^\circ$.

(That is, E is open if and only if every point of E is an interior point.)

(ii) E is closed if and only if $E = \overline{E}$.

Proof.

(i) $\boxed{\implies}$ Suppose E is open. By assumption, E is an open subset of X contained in E (since $E \subset E$), so $E \subset E^\circ$.

We now show the opposite containment. Let $x \in E^\circ$. Then x is in some open subset of X contained in E, so $x \in E$. Hence $E^\circ \subset E$.

Therefore $E = E^\circ$.

$\boxed{\Longleftarrow}$ Since an arbitrary union of open sets is open, E° is open. Since $E = E^\circ$, we have that E is open.

(ii) $\boxed{\Longrightarrow}$ Suppose E is closed. Then $E \subset \overline{E}$.

We now show the opposite containment. Let $x \in \overline{E}$. Then x is in every closed subset of X containing E, so $x \in E$. Hence $x \in E$.

Therefore $E = \overline{E}$.

$\boxed{\Longleftarrow}$ Since an arbitrary intersection of closed sets is closed, $\overline{E}$ is closed. Since $E = \overline{E}$, we have that E is closed.

□

Proposition 14.20. *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.

$\boxed{\Longrightarrow}$ Let $p \in \overline{E}$.

Suppose, for a contradiction, that there exists an 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 $p \notin \overline{E}$. 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.

Lemma 14.21 (Properties of closure and interior)**.** *Suppose $A, B \subset X$. Then*

(i) $\overline{A \cup B} = \overline{A} \cup \overline{B}$

(ii) $\overline{A \cap B} \subset \overline{A} \cap \overline{B}$

(iii) $(A \cup B)^\circ \supset A^\circ \cup B^\circ$

(iv) $(A \cap B)^\circ = A^\circ \cap B^\circ$

(v) $(A^\circ)^c = \overline{A^c}$

(vi) $(\overline{A})^c = (A^c)^\circ$

14.1.5 Limit Points

Definition 14.22.

(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*** 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 $B_\varepsilon(p) \cap E = \{p\}$).

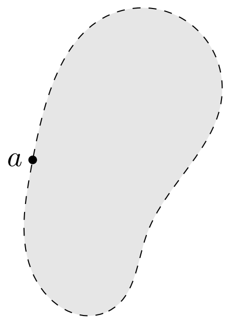


Figure 14.5: Adherent point, limit point, isolated point

Example 14.23 (Adherent point)**.**

- If $p \in E$, then p adheres to E because every ball contains p.
- If $E \subset \mathbb{R}$ is bounded above, then $\sup E$ is adherent to E.

Example 14.24 (Limit point)**.**

- The set $\left\{\frac{1}{n} \;\middle|\; n \in \mathbb{N}\right\}$ has 0 as a limit point.
- The set of rational numbers has every real number as a limit point.
- Every point of $[a, b]$ is a limit point of the set of numbers in (a, b).
- Consider $\mathbb{R}^2$. The set of limit points 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 $\mathbb{Q} \subset \mathbb{R}$. $\mathbb{Q}' = \overline{\mathbb{Q}} = \mathbb{R}$.

Proposition 14.25. *If p is a limit point of E, then every open ball of p contains infinitely many points of E.*

Proof. Suppose, for a contradiction, that there exists an open ball $B_r(p)$ which contains only a finite number of points of E distinct from p; let

$$B_r(p) = \{q_1, \dots, q_n\},$$

where $p \neq q_i$ for $i = 1, \dots, n$. Take

$$r = \min\{d(p, q_1), \dots, d(p, q_n)\},$$

then $B_r(p)$ contains no points of E distinct from p, which is a contradiction. □

Corollary 14.26. *A finite point set has no limit points.*

Remark. The converse is not true; for example, $\mathbb{N}$ 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 (14.49).

A closed set was defined to be the complement of an open set. The next result characterises closed sets in another way.

Lemma 14.27. *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 show $p \in E$.

Suppose, 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. □

Lemma 14.28. *Suppose $E \subset X$. Then E' is a closed subset of X.*

Proof. To prove that E' is closed, we will show its complement $(E')^c$ is open.

Let $p \in (E')^c$. Then $p \notin E'$, so p is not a limit point of E; thus, there exists a ball $B_\varepsilon(p)$ whose intersection with E is either empty or $\{p\}$ (depending on whether $p \in E$ or not).

We will show that $B_{\frac{\varepsilon}{2}}(p) \subset (E')^c$. Let $q \in B_{\frac{\varepsilon}{2}}(p)$.

Case 1: $q = p$. Then clearly $q \in (E')^c$.

Case 2: $q \neq p$. 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.

□

The next result provides a useful expression for the closure of a set; it states that every point of $\overline{E}$ is either a limit point of E, or in E.

Lemma 14.29. *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. to do

□

Lemma 14.30. *Suppose $E \subset X$. Then $\overline{E}$ is the smallest closed set containing E.*

Proof. Let $F \supset E$ be some closed set in X. We will show that $\overline{E} \subset F$.

Let p be a limit point of E. Then p is a limit point of F. But since F is closed, by 14.27, F contains all its limit points, so all the limit points of E are in F. Hence $\overline{E} \subset F$. □ to do

Lemma 14.31. *Suppose non-empty $E \subset \mathbb{R}$ 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

14.2 Compactness

14.2.1 Definitions and Properties

Definition 14.32 (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 14.33 (Compactness)**.** $K \subset X$ is ***compact*** if *every* open cover of K contains 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 $i_1, \dots, i_n \in I$ such that

$$K \subset \bigcup_{k=1}^{n} U_{i_k}.$$

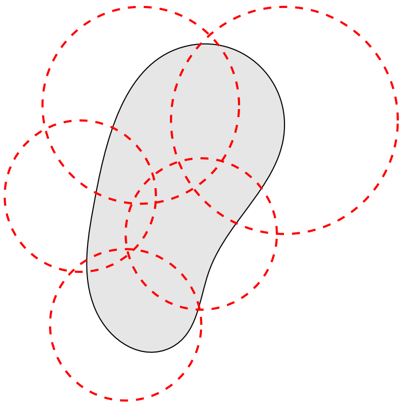


Figure 14.6: Compact set

Example 14.34.

- $\mathbb{R}$ is not compact; for instance, the open cover $\{(-n, n) \mid n \in \mathbb{N}\}$ has no finite subcover.
- $\mathbb{Z}$ is not compact in $\mathbb{R}$; for instance, the open cover $\left\{\left(n - \frac{1}{2}, n + \frac{1}{2}\right) \bigm| n \in \mathbb{Z}\right\}$ has no finite subcover.
- $[0, 1]$ is compact. (See 14.39 for the proof.)

Lemma 14.35. *Every finite set is compact.*

Proof. Let $E = \{p_1, \dots, p_n\}$. Let $\mathcal{U} = \{U_i \mid i \in I\}$ be an open cover of E. We will 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 $\{U_{i_k} \mid k = 1, \dots, n\}$ is a finite subcover of $\mathcal{U}$. $\square$

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 14.36. *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.

$\boxed{\implies}$ Suppose K is compact relative to X.

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 each U_i is open relative to Y, by 14.13, there exists V_i open relative to X such that $U_i = Y \cap V_i$. Consider the open cover $\{V_i \mid i \in I\}$ of K. Since K is compact relative to X, there exist finitely many indices $i_1, \dots, i_n$ 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 $\{U_{i_k} \mid k = 1, \dots, n\}$ forms a finite subcover of $\mathcal{U}$. Hence K is compact relative to Y.

$\boxed{\impliedby}$ 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 $i_1, \dots, i_n$ such that

$$K \subset \bigcup_{k=1}^{n} U_{i_k} \subset \bigcup_{k=1}^{n} V_{i_k}$$

since $U_i \subset V_i$. $\square$

Proposition 14.37. *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 $n \in \mathbb{N}$, let $U_n = B_n(p)$. Then $\{U_n \mid n \in \mathbb{N}\}$ is an open cover of K. By compactness of K, there exists a finite subcover

$$\{U_{n_i} \mid i = 1, \dots, m\}.$$

But note that $U_{n_1} \subset \dots \subset U_{n_m}$, so U_{n_m} contains K. Hence K is bounded. $\square$

Proposition 14.38. *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 $q_{i_1}, \dots, q_{i_n} \in K$ 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 $\min\{d(p, q_{i_k}) \mid k = 1, \dots, n\}$.

Claim. $\varepsilon = \min\{d(p, q_{i_k}) \mid k = 1, \dots, n\}$.

Note that $B_\varepsilon(p) \subset B_{r_{i_k}}(p)$ for all $k = 1, \dots, n$. 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,

$$B_\varepsilon(p) \cap B_{r_{i_k}}(q_{i_k}) = \emptyset \quad (k = 1, \dots, n)$$

Then

$$B_\varepsilon(p) \cap W = B_\varepsilon(p) \cap \left(\bigcup_{k=1}^{n} B_{r_{i_k}}(q_{i_k}) \right) = \bigcup_{k=1}^{n} \left(B_\varepsilon(p) \cap B_{r_{i_k}}(q_{i_k}) \right) = \emptyset$$

as desired. □

Proposition 14.39. *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 will construct 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

$$\Phi = \{U_{i_1}, \dots, U_{i_n}, F^c\}$$

which covers K, and hence F. Now remove F^c from Φ to obtain

$$\Phi' = \{U_{i_1}, \dots, U_{i_n}\},$$

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, $\mathbb{R}$ is closed in $\mathbb{R}$, 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 $\mathbb{R}^n$ (by 14.48).

Corollary 14.40. *If F is closed and K is compact, then $F \cap K$ is compact.*

Proof. Suppose F is closed, K is compact. By 14.38, K is closed. By 14.16, the intersection of two closed sets is closed, so $F \cap K$ is closed.

Since $F \cap K \subset K$ is a closed subset of a compact set K, by 14.39, $F \cap K$ is compact. $\square$

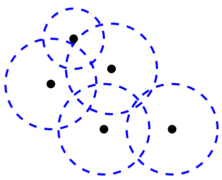
14.2.2 Heine–Borel Theorem

Proposition 14.41. *K is compact if and only if every infinite subset of K has a limit point in K.*

Proof.

$\boxed{\implies}$ Suppose K is compact. Let E be an infinite subset of K. Suppose, 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.

$\boxed{\impliedby}$ 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.

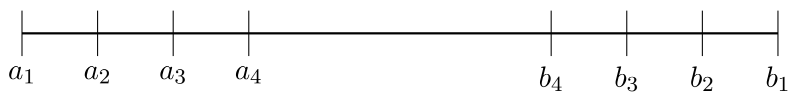
Before that, we will reindex $\mathcal{U}$ to make it more convenient, as follows. By the definition of a cover, every $p \in K$ is contained in some U_i. Pick *one* such U_i for each $p \in K$, and call it U_p. Then our open cover is now $\mathcal{U} = \{U_p \mid p \in K\}$, and for all $p \in K$ we have $p \in U_p$.

To complete proof

$\square$

Proposition 14.42 (Nested interval theorem)**.** *Suppose (I_n) is a decreasing sequence of closed and bounded intervals in $\mathbb{R}$; that is, $I_1 \supset I_2 \supset \cdots$. Then*

$$\bigcap_{n=1}^{\infty} I_n \neq \emptyset.$$



Proof. Let $I_n = [a_n, b_n]$, for $n = 1, 2, \dots$.

Let $E = \{a_n \mid n \in \mathbb{N}\}$. Since E is non-empty and bounded above (by b_1), it has a supremum in $\mathbb{R}$; let $x = \sup E$.

Claim. $x \in \displaystyle\bigcap_{n=1}^{\infty} I_n$.

Since x is the supremum, we have that $a_n \le x$ for all $n \in \mathbb{N}$. Note that for $m > n$, $I_n \supset I_m$ implies $a_n \le a_m \le b_m \le b_n$. This means b_n is an upper bound for all a_n; hence $x \le b_n$ for all $n \in \mathbb{N}$.

Therefore $x \in I_n$ for $n = 1, 2, \dots$ □

To generalise the notion of intervals, we define a *k-cell* as

$$\{(x_1, \dots, x_k) \in \mathbb{R}^k \mid a_i \le x_i \le b_i,\ 1 \le i \le k\}.$$

Example 14.43. 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 14.44. *Suppose (I_n) is a decreasing sequence of k-cells; that is, $I_1 \supset I_2 \supset \cdots$. Then $\bigcap_{n=1}^{\infty} I_n \ne \emptyset$.*

Proof. Let I_n consist of all points $\mathbf{x} = (x_1, \dots, x_k)$ such that

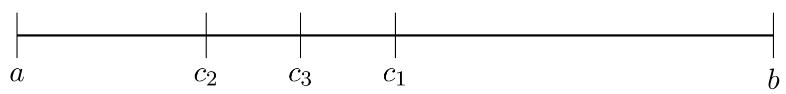
$$a_{n,i} \le x_i \le b_{n,i} \quad (1 \le i \le k;\ n = 1, 2, \dots),$$

and put $I_{n,i} = [a_{n,i}, b_{n,i}]$. For each i, the sequence $(I_{n,i})$ satisfies the hypotheses of 14.42. Hence there are real numbers x_i' $(1 \le i \le k)$ such that

$$a_{n,i} \le x_i' \le b_{n,i} \quad (1 \le i \le k;\ n = 1, 2, \dots).$$

Setting $\mathbf{x}' = (x_1', \dots, x_k')$, we see that $\mathbf{x}' \in I_n$ for $n = 1, 2, \dots$. Hence $\bigcap_{n=1}^{\infty} I_n \ne \emptyset$, as desired. □

Lemma 14.45. *Every closed interval is compact (in $\mathbb{R}$).*



Proof. Suppose, for a contradiction, that a closed interval $[a, b] \subset \mathbb{R}$ 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 (14.42), there exists $x' \in I_n$ for all $n \in \mathbb{N}$. 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 14.46. *Every k-cell is compact (in $\mathbb{R}^k$).*

Proof. We proceed in a similar manner to the proof the previous result.

Suppose I is a k-cell; that is,

$$I = \{(x_1, \dots, x_k) \mid a_i \le x_i \le b_i,\ 1 \le i \le k\}.$$

Write $\mathbf{x} = (x_1, \dots, x_k) \in \mathbb{R}^k$. Let

$$\delta = \left(\sum_{i=1}^{k} (b_i - a_i)^2 \right)^{1/2}$$

that is, the distance between the points $(a_1, \dots, a_k)$ and $(b_1, \dots, b_k)$, which is the maximum distance between two points in I: for all $\mathbf{x}, \mathbf{y} \in I$,

$$|\mathbf{x} - \mathbf{y}| \le \delta.$$

Suppose, 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 $1 \le i \le k$, 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) $|\mathbf{x} - \mathbf{y}| \le 2^{-n}\delta$ for all $\mathbf{x}, \mathbf{y} \in I_n$

By (i) and 14.44, 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 $2^n \le \frac{\delta}{r}$ for all positive integers n, which is absurd since $\mathbb{R}$ 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 14.47. *If $E \subset \mathbb{R}^k$ 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. By 14.46, I is compact. Since E is a closed subset of a compact set, by 14.39, E is compact.

$\boxed{\text{(ii)} \implies \text{(iii)}}$ This directly follows from 14.41.

$\boxed{\text{(iii)} \implies \text{(i)}}$ 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 $\mathbb{R}^k$, hence has none in E. Thus (iii) implies that E is bounded.

If E is not closed, then there is a point $\mathbf{x}_0 \in \mathbb{R}^k$ 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 $\mathbb{R}^k$. For if $\mathbf{y} \in \mathbb{R}^k$, $\mathbf{y} \neq \mathbf{x}_0$, then

$$\begin{aligned} |\mathbf{x}_n - \mathbf{y}| &\geq |\mathbf{x}_0 - \mathbf{y}| - |\mathbf{x}_n - \mathbf{x}_0| \\ &\geq |\mathbf{x}_0 - \mathbf{y}| - \frac{1}{n} \\ &\geq \frac{1}{2}|\mathbf{x}_0 - \mathbf{y}| \end{aligned}$$

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 14.48 (Heine–Borel theorem)**.** *$E \subset \mathbb{R}^n$ is compact if and only if E is closed and bounded.*

Proof. This is simply (i) $\iff$ (ii) in the previous result. □

14.2.3 Bolzano–Weierstrass Theorem

Theorem 14.49 (Bolzano–Weierstrass theorem)**.** *Every bounded infinite subset of $\mathbb{R}^n$ has a limit point in $\mathbb{R}^n$.*

Proof. Suppose E is a bounded infinite subset of $\mathbb{R}^n$.

Since E is bounded, there exists an n-cell $I \subset \mathbb{R}^n$ such that $E \subset I$. Since I is compact, by 14.41, E has a limit point in I and thus $\mathbb{R}^n$. $\square$

14.2.4 Cantor's Intersection Theorem

A collection $\mathcal{A}$ 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 14.50. *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, 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 $i_1, \dots, i_n$ 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 $K_1, K_{i_1}, \dots, K_{i_n}$ is a finite subcollection of $\mathcal{K}$ which has an empty intersection; this contradicts the finite intersection property of $\mathcal{K}$. □

Theorem 14.51 (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 is an immediate corollary of the previous result. □

The following result is a characterisation of compact sets.

Proposition 14.52. *K is compact if and only if every collection of closed subsets of K satisfies the finite intersection property.*

Proof.

$\boxed{\implies}$ 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

$\square$

14.2.5 Sequential Compactness

Definition 14.53 (Sequential compactness)**.** We say $K \subset X$ 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 14.54. *$K \subset X$ 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:

$$\left\{B_{r_{x_1}}(x_1), \dots, B_{r_{x_N}}(x_N)\right\}.$$

In particular, these open balls cover $\{y_n\}$. Hence there must be some x_i ($1 \le i \le N$) 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 $\Lambda = \mathbb{N}$. 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 $n \in \mathbb{N}$. 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 $N' \in \mathbb{N}$ such that if $n_k \ge N'$ 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. $\square$

14.3 Perfect Sets

14.3.1 Definition and Uncountability

Definition 14.55 (Perfect set)**.** E is ***perfect*** if

(i) E is closed, and

(ii) every point of E is a limit point of E.

Proposition 14.56. *Let non-empty $P \subset \mathbb{R}^k$ be perfect. Then P is uncountable.*

Proof. Since P has limit points, by 14.25, P is an infinite set.

Suppose, for a contradiction, that P is countable. This means we can list the points of P in a sequence:

$$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots$$

Consider a sequence (B_n) of open balls, where B_n is any open ball centred at $\mathbf{x}_n$:

$$B_n = \left\{ \mathbf{y} \in \mathbb{R}^k \mid |\mathbf{y} - \mathbf{x}_n| < r \right\}.$$

Then its closure $\overline{B_n}$ is the closed ball

$$\overline{B}_n = \left\{ \mathbf{y} \in \mathbb{R}^k \mid |\mathbf{y} - \mathbf{x}_n| \leq r \right\}.$$

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 B_{n+1} such that (i) $\overline{B}_{n+1} \subset B_n$, (ii) $\mathbf{x}_n \notin \overline{B}_{n+1}$, (iii) $B_{n+1} \cap P$ is not empty.

By (iii), B_{n+1} satisfies our induction hypothesis, and the construction can proceed.

Put $K_n = \overline{B}_n \cap P$. Since $\overline{B}_n$ is closed and bounded, $\overline{B}_n$ is compact. Since $\mathbf{x}_n \notin K_{n+1}$, no point of P lies in $\bigcap_{n=1}^{\infty} K_n$. Since $K_n \subset P$, this implies that $\bigcap_{n=1}^{\infty} K_n$ is empty. But each K_n is nonempty, by (iii), and $K_n \supset K_{n+1}$ by (i); this contradicts Cantor's intersection theorem (14.51). ☐

Corollary. *Every interval $[a, b]$ is uncountable. In particular, $\mathbb{R}$ is uncountable.*

14.3.2 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

$$C_2 = \left[0, \frac{1}{9}\right] \cup \left[\frac{2}{9}, \frac{3}{9}\right] \cup \left[\frac{6}{9}, \frac{7}{9}\right] \cup \left[\frac{8}{9}, 1\right].$$

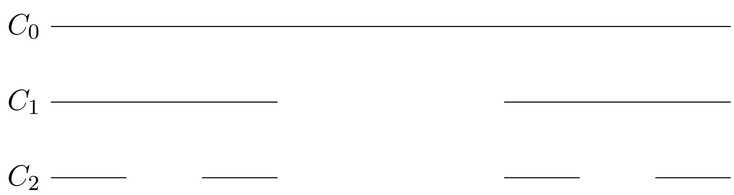


Figure 14.7: Cantor set

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

$$C := \bigcap_{n=1}^{\infty} C_n.$$

Lemma 14.57 (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 $n \in \mathbb{N}$ such that (a, b) is not contained in C_n and hence not contained in C.

□

Proposition 14.58. *C is a perfect set in $\mathbb{R}$ which contains no open interval.*

Proof. We will show that (i) C contains no open interval, and (ii) C is perfect.

(i) No open interval of the form

$$\left(\frac{3k+1}{3^m}, \frac{3k+2}{3^m}\right),$$

where $k, m \in \mathbb{Z}^+$, has a point in common with C. Since every open interval (α, β) contains a open interval of the above form, if

$$3^{-m} < \frac{\beta - \alpha}{6},$$

C contains no open interval.

(ii) Sincw we have shown that C is closed, it suffices to show that every point of C is a limit point.

Let $x \in C$, and let S be any open interval 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 14.59. *C is uncountable.*

One of the most interesting properties of the Cantor set is that it provides us with an example of an uncountable set of measure zero.

14.4 Connectedness

Definition 14.60 (Connectedness)**.** We say 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, $[0, 1]$ and $(1, 2)$ are not separated, since 1 is a limit point of $(1, 2)$. However $(0, 1)$ and $(1, 2)$ are separated.

Example 14.61. In $\mathbb{R}^2$, consider the set

$$E = \{(x, y) \mid x, y \in \mathbb{Q}\}.$$

Then E is not connected; if we let

$$\begin{aligned} A &= \left\{(x, y) \mid x, y \in \mathbb{Q}, x < \sqrt{2}\right\}, \\ B &= \left\{(x, y) \mid x, y \in \mathbb{Q}, x > \sqrt{2}\right\}, \end{aligned}$$

then note that $A \cup B = E$, as well as $A \cap \overline{B} = \emptyset$ and $\overline{A} \cap B = \emptyset$.

Lemma 14.62. *Closed intervals in $\mathbb{R}$ are connected.*

Proof. Suppose, for a contradiction, that a closed interval $[a, b]$ is not connected. Then 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 $s = \sup A$. By 14.31, $s \in \overline{A}$. 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. ∎

The connected subsets of the real line have a particularly simple structure:

Lemma 14.63. *$E \subset \mathbb{R}$ 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.

$\boxed{\implies}$ 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 14.31, $z \in \overline{A}$; hence $z \notin B$. In particular, $x \leq z < y$.

Case 1: $z \notin A$**.** It follows that $x < z < y$ and $z \notin E$.

Case 2: $z \in A$**.** Then $z \notin B$, hence there exists z_1 such that $z < z_1 < y$ and $z_1 \notin B$. Then $x < z_1 < y$ and $z_1 \notin E$.

□

14.4.1 Path Connectedness

14.5 Separable Spaces

Definition 14.64 (Separable space)**.** X is ***separable*** if it has a countable subset which is dense in X.

Example 14.65.

- $\mathbb{R}$ is separable.

  Proof. The set of rational numbers $\mathbb{Q}$ is countable and is dense in $\mathbb{R}$. □

- $\mathbb{C}$ is separable.

  Proof. A countable dense subset of $\mathbb{C}$ is the set of all complex numbers whose real and imaginary parts are both rational, i.e., the set $\{x + yi \mid x, y \in \mathbb{Q}\}$. □

- The *discrete metric space* X is separable if and only if X is countable.

  Proof. The kind of metric implies that no proper subset of X can be dense in X. Hence the only dense set in X is X itself, and the statement follows. □

- The *sequence space* ℓ^∞ is the set of all bounded complex sequences, with the metric defined by

  $$d(x, y) = \sup_{n \in \mathbb{N}} |x_n - y_n|.$$

  ℓ^∞ is not separable.

14.6 Baire Category Theorem

$E \subset X$ is called ***nowhere dense*** (in X) if the interior of the closure of A is empty, i.e., $(\overline{A})^\circ = \emptyset$.

Otherwise put, E is nowhere dense iff it is contained in a closed set with empty interior. Passing to complements, we can say equivalently that E is nowhere dense iff its complement contains a dense open set (why?).

Lemma 14.66. *Let X be a metric space.*

(i) Any subset of a nowhere dense set is nowhere dense.

(ii) The union of finitely many nowhere dense sets is nowhere dense.

(iii) The closure of a nowhere dense set is nowhere dense.

(iv) If X has no isolated points, then every finite set is nowhere dense.

Proof.

(i)

(ii)

(iii)

(iv)

□

Although the union of finitely many nowhere dense sets is nowhere dense, the union of countably many nowhere dense sets need not be nowhere dense: for instance, in $X = \mathbb{R}$, the rationals $\mathbb{Q}$ are the union of countably many nowhere dense sets (why?), but the rationals are certainly not nowhere dense (indeed, they are everywhere dense, i.e. $(\overline{\mathbb{Q}})^\circ = \overline{\mathbb{Q}} = \mathbb{R}$).

This observation motivates the introduction of a larger class of sets: $A \subset X$ is called ***meager*** (or of first category) in X if it can be written as a countable union of nowhere dense sets; otherwise, it is *non-meager* (or of second category). The complement of a meager set is called ***residual***.

We then have as an immediate consequence:

Lemma 14.67. *Let X be a metric space.*

(i) Any subset of a meager set is meager.

(ii) The union of countably many meager sets is meager.

(iii) If X has no isolated points, then every countable set is meager.

We are now ready to state the Baire category theorem.

Theorem 14.68 (Baire category theorem)**.** *Let X be a complete metric space.*

(i) A meager set has empty interior.

(ii) The complement of a meager set is dense. (That is, a residual set is dense.)

(iii) A countable intersection of dense open sets is dense.

You should carefully verify that (i), (ii) and (iii) are equivalent statements, obtained by taking complements.

In applications we frequently need only the weak form of the Baire category theorem that is obtained by weakening "is dense" in (b,c) to "is non-empty" (which is valid whenever X is itself non-empty):

Corollary 14.69 (Weak form of the Baire category theorem)**.** *Let X be a non-empty complete metric space.*

(i) X cannot be written as a countable union of nowhere dense sets. (In other words, X is nonmeager in itself.)

(ii) If X is written as a countable union of closed sets, then at least one of those closed sets has nonempty interior.

(iii) A countable intersection of dense open sets is nonempty.

Exercises

Exercise 14.1. 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 $\mathbb{Z}$, 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:

$$d(x,z) \leq \max\{d(x,y), d(y,z)\}.$$

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 $\mathbb{Q}$.)

(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 $P(\mathbb{R}^n)$ of one-dimensional subspaces of $\mathbb{R}^n$, 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*.)

Exercise 14.2 (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

$$\begin{aligned}&\sqrt{d_X(x_1,x_3)^2+d_Y(y_1,y_3)^2}+\sqrt{d_X(x_3,x_2)^2+d_Y(y_3,y_2)^2}\\&\geq\sqrt{d_X(x_1,x_2)^2+d_Y(y_1,y_2)^2}\end{aligned}\tag{1}$$

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

$$\sqrt{{a_2}^2+{b_2}^2}+\sqrt{{a_1}^2+{b_1}^2}\geq\sqrt{{a_3}^2+{b_3}^2}. \tag{2}$$

To prove this, note that from the triangle inequality we have $a_1 + a_2 \geq a_3$, $b_1 + b_2 \geq b_3$. Squaring and adding gives

$${a_1}^2+{b_1}^2+{a_2}^2+{b_2}^2+2(a_1a_2+b_1b_2)\geq{a_3}^2+{b_3}^2.$$

By Cauchy–Schwarz,

$$a_1a_2+b_1b_2\leq\sqrt{{a_1}^2+{b_1}^2}\sqrt{{a_2}^2+{b_2}^2}.$$

Substituting this into the previous line gives precisely the square of (2), and (1) follows. □

Chapter 15

Numerical Sequences and Series

Throughout, let (X, d) be a metric space.

15.1 Sequences

15.1.1 Convergence

A ***sequence*** (a_n) in X is a function $f\colon \mathbb{N} \to X$ which maps $n \mapsto a_n$.

The *range* of a sequence (a_n) is the set

$$\{x \in X \mid \exists n \in \mathbb{N}, x = a_n\}.$$

Note that the range of a sequence may be a finite set or it may be infinite. (a_n) is *bounded* if its range is bounded.

Definition 15.1. A sequence (a_n) ***converges*** to $a \in X$, denoted by $a_n \to a$, if

$$\forall \varepsilon > 0, \quad \exists N \in \mathbb{N}, \quad \forall n \geq N, \quad d(a_n, a) < \varepsilon.$$

We call a a *limit* of (a_n). If (a_n) does not converge, it is said to *diverge*.

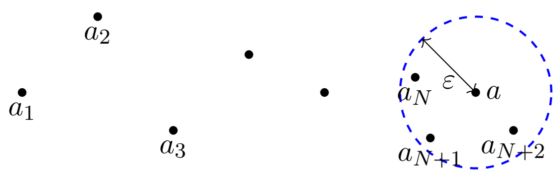


Figure 15.1: Convergence of sequence

Remark. This limit process conveys the intuitive idea that a_n can be made arbitrarily close to a, provided that n is sufficiently large. (Equivalently, if we remove more and more initial terms from the sequence, the *tail* of the sequence is increasingly closer to a.)

Remark. If $a_n \not\to a$, simply negate the definition for convergence:

$$\exists \varepsilon > 0, \quad \forall N \in \mathbb{N}, \quad \exists n \geq N, \quad d(a_n, a) \geq \varepsilon.$$

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 $\mathbb{R}$ (to 0), but fails to converge in $\mathbb{R}^+$. In cases of possible ambiguity, we shall specify "convergent in X" rather than "convergent".

Example 15.2. $\frac{1}{n} \to 0$.

Proof. Fix $\varepsilon > 0$. By the Archimedean property, there exists $N \in \mathbb{N}$ such that $\frac{1}{N} < \varepsilon$. Take $N = \lfloor \frac{1}{\varepsilon} \rfloor + 1$. Then for all $n \geq N$,

$$\left| \frac{1}{n} - 0 \right| = \frac{1}{n} \leq \frac{1}{N} = \frac{1}{\lfloor \frac{1}{\varepsilon} \rfloor + 1} < \frac{1}{\frac{1}{\varepsilon}} = \varepsilon$$

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 15.3. Let $a_n = 1 + (-1)^n \frac{1}{\sqrt{n}}$. Then $a_n \to 1$.

Before our proof, we aim to find some $N \in \mathbb{N}$ such that if $n \geq N$ then

$$\begin{aligned} & |a_n - 1| < \varepsilon \\ \iff & \frac{1}{\sqrt{n}} = \left| (-1)^n \frac{1}{\sqrt{n}} \right| < \varepsilon \\ \iff & \frac{1}{n} < \varepsilon^2 \\ \iff & n > \frac{1}{\varepsilon^2} \end{aligned}$$

Hence take $N = \lfloor \frac{1}{\varepsilon^2} \rfloor + 1$.

Proof. Let $\varepsilon > 0$ be given. Take $N = \lfloor \frac{1}{\varepsilon^2} \rfloor + 1$. If $n \geq N$, then

$$\begin{aligned} |a_n - 1| &= \left| (-1)^n \frac{1}{\sqrt{n}} \right| = \frac{1}{\sqrt{n}} \\ &\leq \frac{1}{\sqrt{N}} = \frac{1}{\sqrt{\lfloor \frac{1}{\varepsilon^2} \rfloor + 1}} \\ &< \frac{1}{\sqrt{\frac{1}{\varepsilon^2}}} = \varepsilon \end{aligned}$$

as desired. Therefore $a_n \to 1$. □

Lemma 15.4 (Uniqueness of limit)**.** *If a sequence converges, then its limit is unique.*

Proof. Let (a_n) be a sequence in X. Suppose that $a_n \to a$ and $a_n \to a'$ for $a, a' \in X$. We will show that $a' = a$.

Let $\varepsilon > 0$ be given. Then there exists $N, N' \in \mathbb{N}$ such that

$$\begin{aligned} n \geq N &\implies d(a_n, a) < \frac{\varepsilon}{2} \\ n \geq N' &\implies d(a_n, a') < \frac{\varepsilon}{2} \end{aligned}$$

Take $N_1 := \max\{N, N'\}$. If $n \geq N_1$, then both hold. By the triangle inequality,

$$d(a, a') \leq d(a, a_n) + d(a_n, a') < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Since this holds for all $\varepsilon > 0$, we must have $d(a, a') = 0$. Hence $a = a'$. □

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

Notation. If (a_n) converges to a, denote $\lim\limits_{n\to\infty} a_n = a$.

We now outline some important properties of convergent sequences in metric spaces.

Lemma 15.5. *Let (a_n) be a sequence in X.*

(i) $a_n \to a$ if and only if every open ball of a contains a_n for all but finitely many n.

(ii) Every convergent sequence is bounded.

(iii) Suppose $E \subset X$. Then a is a limit point of E if and only if there exists a sequence (a_n) in $E \setminus \{a\}$ such that $a_n \to a$.

Proof.

(i) $\boxed{\implies}$ Suppose $a_n \to a$. Let $\varepsilon > 0$ be given, there exists $N \in \mathbb{N}$ such that

$$n \geq N \implies d(a_n, a) < \varepsilon \implies B_\varepsilon(a).$$

Hence $n \geq N$ implies $a_n \in B_\varepsilon(a)$.

$\boxed{\impliedby}$ Suppose every open ball of a contains all but finitely many of the a_n.

Let $\varepsilon > 0$ be given. Consider the open ball $B_\varepsilon(a)$. Since $B_\varepsilon(a)$ is a open ball of a, it will also eventually contain all a_n; that is, there exists $N \in \mathbb{N}$ such that if $n \geq N$, then $a_n \in B_\varepsilon(a)$, i.e. $d(a_n, a) < \varepsilon$. Hence $a_n \to a$.

(ii) Suppose $a_n \to a$. Take $\varepsilon = 1$, there exists $N \in \mathbb{N}$ such that

$$n \geq N \implies d(a_n, a) < 1.$$

Let

$$r = \max\{1, d(a_1, a), \dots, d(a_N, a)\},$$

then $d(a_n, a) \leq r$, so the range of a_n is bounded by $B_r(a)$. Hence (a_n) is bounded.

(iii) $\boxed{\implies}$ Suppose a is a limit point of E.

Consider a sequence of open balls $\left(B_{\frac{1}{n}}(a)\right)$, for $n \in \mathbb{N}$. Since a is a limit point, each open ball intersects with E at some point which is not a. We pick one such point a_n from each $B_{\frac{1}{n}}(a) \cap E$. Then

$$d(a_n, a) < \frac{1}{n}.$$

Let $\varepsilon > 0$ be given. By the Archimedean property, there exists $N \in \mathbb{N}$ such that $\frac{1}{N} < \varepsilon$. If $n \geq N$,

$$d(a_n, a) \leq \frac{1}{n} \leq \frac{1}{N} < \varepsilon,$$

which shows that $a_n \to a$.

$\boxed{\Longleftarrow}$ Suppose that there exists a sequence (a_n) in $E \setminus \{a\}$ such that $a_n \to a$. Then for each open ball $B_\varepsilon(a)$, we can find some $N \in \mathbb{N}$ such that if $n \in \mathbb{N}$ then

$$a_n \in B_\varepsilon(a).$$

Since $a_n \in E \setminus \{a\}$, this shows that a is a limit point of E.

□

Remark. A consequence of (ii) is its contrapositive: any unbounded sequence is divergent. Note that the converse is not true; a counterexample is $(-1)^n$.

Lemma 15.6 (Ordering)**.** *Suppose (a_n) and (b_n) are convergent sequences, and $a_n \leq b_n$. Then*

$$\lim_{n\to\infty} a_n \leq \lim_{n\to\infty} b_n.$$

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 $N_1, N_2 \in \mathbb{N}$ such that

$$n \geq N_1 \implies |a_n - a| < \frac{\varepsilon}{2},$$
$$n \geq N_2 \implies |b_n - b| < \frac{\varepsilon}{2}.$$

Let $N = \max\{N_1, N_2\}$, then $n \geq N$ 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.

Lemma 15.7 (Arithmetic properties)**.** *Suppose (a_n) and (b_n) are convergent seqeunces in $\mathbb{C}$; 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 $N \in \mathbb{N}$ such that

$$n \geq N \implies |a_n - a| < \frac{\varepsilon}{|c|}.$$

Then if $n \geq N$,

$$|ca_n - ca| = |c|\,|a_n - a| < \varepsilon.$$

(ii) Let $\varepsilon > 0$ be given. Since $a_n \to a$ and $b_n \to b$, there exists $N_1, N_2 \in \mathbb{N}$ such that

$$n \geq N_1 \implies |a_n - a| < \frac{\varepsilon}{2},$$
$$n \geq N_2 \implies |b_n - b| < \frac{\varepsilon}{2}.$$

Let $N = \max\{N_1, N_2\}$, then $n \geq N$ implies

$$\begin{aligned} |(a_n + b_n) - (a + b)| &\leq |a_n - a| + |b_n - b| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

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 $N_1, N_2 \in \mathbb{N}$ such that

$$n \geq N_1 \implies |a_n - a| < \sqrt{\varepsilon},$$
$$n \geq N_2 \implies |b_n - b| < \sqrt{\varepsilon}.$$

Let $N = \max\{N_1, N_2\}$. Then $n \geq N$ 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 $m \in \mathbb{N}$ such that

$$n \geq m \implies |b_n - b| < \frac{1}{2}|b|.$$

Let $\varepsilon > 0$ be given. There exists $N \in \mathbb{N}$, $N > m$ such that

$$n \geq N \implies |b_n - b| < \frac{1}{2}|b|^2\varepsilon.$$

Hence for $n \geq N$,

$$\left|\frac{1}{b_n} - \frac{1}{b}\right| = \left|\frac{b - b_n}{b_n b}\right| < \frac{2}{|b|^2}|b_n - b| < \varepsilon.$$

$\square$

We now prove the analogue for Euclidean spaces.

Lemma 15.8.

(i) Suppose $\mathbf{x}_n \in \mathbb{R}^k$ $(n = 1, 2, \dots)$ *and*

$$\mathbf{x}_n = (\alpha_{1,n}, \dots, \alpha_{k,n}).$$

Then $(\mathbf{x}_n)$ *converges to* $\mathbf{x} = (\alpha_1, \dots, \alpha_k)$ *if and only if*

$$\lim_{n\to\infty} \alpha_{i,n} = \alpha_i \quad (1 \leq i \leq k).$$

(ii) Suppose $(\mathbf{x}_n)$ *and* $(\mathbf{y}_n)$ *are sequences in* $\mathbb{R}^k$, (β_n) *is a sequence of real numbers, and* $\mathbf{x}_n \to \mathbf{x}$, $\mathbf{y}_n \to \mathbf{y}$, $\beta_n \to \beta$. *Then*

$$\lim_{n\to\infty} (\mathbf{x}_n + \mathbf{y}_n) = \mathbf{x} + \mathbf{y}, \quad \lim_{n\to\infty} \mathbf{x}_n \cdot \mathbf{y}_n = \mathbf{x} \cdot \mathbf{y}, \quad \lim_{n\to\infty} \beta_n \mathbf{x}_n = \beta \mathbf{x}.$$

Proof.

(i) $\boxed{\implies}$ Suppose $\mathbf{x}_n \to \mathbf{x}$. From the definition of the norm in $\mathbb{R}^k$, the inequalities

$$|\alpha_{i,n} - \alpha_i| \leq \|\mathbf{x}_n - \mathbf{x}\|$$

follow immediately, which show that

$$\lim_{n\to\infty} \alpha_{i,n} = \alpha_i \quad (1 \leq i \leq k).$$

$\boxed{\impliedby}$ Suppose $\lim_{n\to\infty} \alpha_{i,n} = \alpha_i$ for $i = 1, \dots, k$. Then to each $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $n \geq N$ implies

$$|\alpha_{i,n} - \alpha_i| < \frac{\varepsilon}{\sqrt{k}} \quad (i = 1, \dots, k).$$

Hence $n \geq N$ implies

$$\|\mathbf{x}_n - \mathbf{x}\| = \left(\sum_{i=1}^{k} |\alpha_{i,n} - \alpha_i|^2 \right)^{1/2} < \varepsilon,$$

so that $\mathbf{x}_n \to \mathbf{x}$.

(ii) This follows from (i) and 15.7.

□

The next result provides a useful method to evaluate limits of sequences.

Lemma 15.9 (Squeeze theorem)**.** *Let $a_n \leq c_n \leq b_n$ 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 $N_1, N_2 \in \mathbb{N}$ such that

$$n \geq N_1 \implies |a_n - L| < \varepsilon,$$
$$n \geq N_2 \implies |b_n - L| < \varepsilon.$$

In particular, we have

$$a_n > L - \varepsilon, \quad b_n < L + \varepsilon.$$

Let $N = \max\{N_1, N_2\}$. Then $n \geq N$ implies

$$L - \varepsilon < a_n \leq c_n \leq b_n < L + \varepsilon$$

or

$$|c_n - L| < \varepsilon.$$

Hence (c_n) is convergent, and $c_n \to L$. □

The following example is a classic application of the squeeze theorem.

Example 15.10. Show that $\lim_{n\to\infty} \frac{\sin n}{n} = 0$.

Proof. We have $-1 \leq \sin n \leq 1$, so

$$-\frac{1}{n} \leq \frac{\sin n}{n} \leq \frac{1}{n}.$$

Now

$$\lim_{n\to\infty} \frac{1}{n} = \lim_{n\to\infty} \left(-\frac{1}{n}\right) = 0,$$

so the squeeze theorem yields the desired result. □

15.1.2 Subsequences

Definition 15.11 (Subsequence)**.** Given a sequence (a_n), consider a sequence (n_k) of positive integers such that $n_1 < n_2 < \cdots$. Then (a_{n_k}) is called a ***subsequence*** of (a_n).
If (a_{n_k}) converges, its limit is called a *subsequential limit* of (a_n).

Proposition 15.12. *(a_n) converges to a if and only if every subsequence of (a_n) converges to a.*

Proof.

$\boxed{\implies}$ Suppose $a_n \to a$. Let $\varepsilon > 0$ be give. Then there exists $N \in \mathbb{N}$ such that

$$n \geq N \implies d(a_n, a) < \varepsilon.$$

Every subsequence of (a_n) can be written in the form (a_{n_k}) where $n_1 < n_2 < \cdots$ is a strictly increasing sequence of positive integers. Pick M such that $n_M \geq N$. Then

$$k > M \implies n_k > n_M \geq N \implies d(a_{n_k}, a) < \varepsilon.$$

Hence every subsequence of (a_n) converges to a.

$\boxed{\impliedby}$ Suppose every subsequence of (a_n) converges to a. Since (a_n) is a subsequence of itself, we must have $a_n \to a$. ☐

Proposition 15.13. *In a compact metric space, any sequence has a convergent subsequence.*

Proof. Suppose (a_n) is a sequence in a compact metric space X.

Let E be the range of (a_n). We consider two cases:

Case 1: E is finite. Notice that there are infinitely many terms in the sequence (a_n), but only finitely many distinct terms in E. By the pigeonhole principle, at least one term of E appears infinitely many times in the sequence.

That is, there exists $a \in E$ and a sequence (n_k) with $n_1 < n_2 < \cdots$ such that

$$a_{n_1} = a_{n_2} = \cdots = a.$$

This subsequence (a_{n_k}) evidently converges to a.

Case 2: E is infinite. If E is infinite, then E is an infinite subset of a compact set. By 14.41, E has a limit point $a \in X$.

We now construct a subsequence (a_{n_k}) of (a_n) such that $a_{n_k} \to a$.

- Choose n_1 so that $d(a, a_{n_1}) < 1$.
- Having chosen $n_1, \dots, n_{k-1}$, choose n_k where $n_k > n_{k-1}$ such that $d(a, a_{n_k}) < \frac{1}{k}$ (such n_k exists due to 14.25).

Then $a_{n_k} \to a$.

□

Corollary 15.14 (Bolzano–Weierstrass). *Every bounded sequence in $\mathbb{R}^k$ has a convergent subsequence.*

Proof. By 14.47, every bounded sequence in $\mathbb{R}^k$ lives in a compact subset of $\mathbb{R}^k$, and therefore it lives in a compact metric space. Hence by the previous result, it contains a convergent subsequence converging to a point in $\mathbb{R}^k$. □

Lemma 15.15. *Suppose (a_n) is a sequence in X. Then the subsequential limits of (a_n) form a closed subset of X.*

Proof. Let E be the set of all subsequential limits of (a_n), let q be a limit point of E. We want to show that $q \in E$.

Choose n_1 so that $a_{n_1} \neq q$. (If no such n_1 exists, then E has only one point, and there is nothing to prove.) Put $\delta = d(q, a_{n_1})$. Suppose $n_1, \ldots, n_{i-1}$ are chosen. Since q is a limit point of E, there is an $a \in E$ with $d(a, q) < 2^{-1}\delta$. Since $a \in E$, there is an $n_i > n_{i-1}$ such that $d(a, a_{n_k}) < 2^{-i}\delta$. Thus

$$d(q, a_{n_k}) < 2^{1-i}\delta$$

for $i = 1, 2, 3, \ldots$. This says that (a_{n_k}) converges to q. Hence $q \in E$. □

15.1.3 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. Subsequently we 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 15.16 (Cauchy sequence)**.** A sequence (a_n) in X is a ***Cauchy sequence*** if

$$\forall \varepsilon > 0, \quad \exists N \in \mathbb{N}, \quad \forall n, m \geq N, \quad d(a_n, a_m) < \varepsilon.$$

Remark. Intuitively, the distances between any two terms becomes sufficiently small after a certain point.

A natural question is regarding the relationship between convergent sequences and Cauchy sequences. We now address this.

Proposition 15.17.

(i) In any metric space, every convergent sequence is a Cauchy sequence.

(ii) If X is a compact metric space and if (a_n) is a Cauchy sequence in X, then (a_n) converges to some point of X.

(iii) In $\mathbb{R}^k$, every Cauchy sequence converges.

Remark. The converse of (i) is not true. For instance, the sequence $\{3, 3.1, 3.14, 3.141, 3.1415, \dots\}$ is a Cauchy sequence but does not converge in $\mathbb{Q}$.

Proof.

(i) Suppose $a_n \to a$. Let $\varepsilon > 0$. There exists $N \in \mathbb{N}$ such that for all $n \geq N$,

$$d(a_n, a) < \frac{\varepsilon}{2}.$$

Then for all $n, m \geq N$,

$$d(a_n, a_m) \leq d(a_n, a) + d(a_m, a) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

as desired. Hence (a_n) is a Cauchy sequence.

(ii) Let (a_n) be a Cauchy sequence in X. Since X is compact, it is sequentially compact. Then there exists a subsequence (a_{n_k}) such that $a_{n_k} \to a$.

Claim. $a_n \to a$.

Let $\varepsilon > 0$. Since (a_n) is a Cauchy sequence, there exists $N_1 \in \mathbb{N}$ such that

$$n, m \geq N_1 \implies d(a_n - a_m) < \frac{\varepsilon}{2}.$$

$a_{n_k} \to a$ implies there exists $N_2 \in \mathbb{N}$ such that

$$n_k \geq N_2 \implies d(a_{n_k}, a) < \frac{\varepsilon}{2}.$$

Let $N = \max\{N_1, N_2\}$, fix some $n_k \geq N$. Then $n \geq N$ implies

$$d(a_n, a) \leq d(a_n, a_{n_k}) + d(a_{n_k}, a) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

(iii) Suppose (a_n) is a Cauchy sequence.

We perform three steps:

- We first show that (a_n) is bounded:
  Pick $N \in \mathbb{N}$ such that $|a_n - a_N| \leq 1$ for all $n \geq N$. Then

  $$|a_n| \leq \max\{1 + |a_N|, |a_1|, \dots, |a_{N-1}|\}.$$

- Since (a_n) is bounded, by Bolzano–Weierstrass, (a_n) contains a subsequence (a_{n_k}) which converges to a.
- We now show that $a_n \to a$.
  Let $\varepsilon > 0$ be given. Since (a_n) is a Cauchy sequence, there exists $N_1 \in \mathbb{N}$ such that

  $$n, m \geq N_1 \implies |a_n - a_m| < \frac{\varepsilon}{2}.$$

  Since $a_{n_k} \to a$, there exists $M \in \mathbb{N}$ such that for all $k > M$,

  $$n_k > n_M \implies |a_{n_k} - a| < \frac{\varepsilon}{2}.$$

  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 $n \geq N_1$, by setting $m = n_k$ we obtain

  $$|a_n - a_{n_k}| < \frac{\varepsilon}{2}, \quad |a_{n_k} - a| < \frac{\varepsilon}{2}.$$

  Hence

  $$|a_n - a| \leq |a_n - a_{n_k}| + |a_{n_k} - a| < \varepsilon.$$

  Therefore (a_n) is convergent, and $a_n \to a$.

□

Definition 15.18 (Complete metric space)**.** 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 $a \in X$, and actually $a \in E$ since E is closed.)

Example 15.19. The sequence (a_n) is defined as follows:

$$a_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}.$$

(a_n) does not converge in $\mathbb{R}$.

Proof. We claim that (a_n) is not a Cauchy sequence. WLOG assume $n > m$. Consider

$$|a_n - a_m| = \frac{1}{m+1} + \frac{1}{m+2} + \cdots + \frac{1}{n} \geq \frac{n-m}{n} = 1 - \frac{m}{n}.$$

Let $n = 2m$, then

$$|a_n - a_m| = |a_{2m} - a_m| > \frac{1}{2}.$$

Hence (a_n) is not a Cauchy sequence, so it does not converge. $\square$

15.1.4 Monotonic Sequences

Definition 15.20 (Monotonic sequence)**.** A sequence (a_n) in $\mathbb{R}$ is

(i) *monotonically increasing* if $a_n \leq a_{n+1}$ for $n \in \mathbb{N}$;

(ii) *monotonically decreasing* if $a_n \geq a_{n+1}$ for $n \in \mathbb{N}$;

(iii) ***monotonic*** if it is either monotonically increasing or monotonically decreasing.

Lemma 15.21 (Monotone convergence theorem)**.** *A monotonic sequence in* $\mathbb{R}$ *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.

$\boxed{\Longrightarrow}$ We already proved that a convergent sequence is bounded.

$\boxed{\Longleftarrow}$ Suppose (a_n) is a monotonically increasing sequence bounded above.

Let E be the range of a_n. Since E is bounded above, let $a = \sup E$.

Claim. $a_n \to a$.

By definition of supremum, $a_n \leq a$ for all $n \in \mathbb{N}$. For every $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that

$$a - \varepsilon < a_N \leq a,$$

otherwise $a - \varepsilon$ would be an upper bound of E. Since (a_n) is monotically increasing, $n \geq N$ implies $a_N \leq a_n \leq a$, so

$$a - \varepsilon < a_n \leq a,$$

which implies $|a_n - a| < \varepsilon$. Hence $a_n \to a$. ☐

15.1.5 Limit Superior and Inferior

For properly divergent sequences, we make the following definition.

Definition 15.22. Suppose (a_n) is a sequence in $\mathbb{R}$. We write $a_n \to \infty$ if

$$\forall M \in \mathbb{R}, \quad \exists N \in \mathbb{N}, \quad \forall n \geq N, \quad a_n \geq M.$$

Similarly, we write $a_n \to -\infty$ if

$$\forall M \in \mathbb{R}, \quad \exists N \in \mathbb{N}, \quad \forall n \geq N, \quad a_n \leq M.$$

Definition 15.23. Suppose (a_n) is a sequence in $[-\infty, \infty]$. Define respectively the ***limit superior*** and ***limit infimum*** of (a_n) as

$$\limsup_{n\to\infty} a_n := \lim_{n\to\infty} \sup_{m\geq n} (a_m),$$
$$\liminf_{n\to\infty} a_n := \lim_{n\to\infty} \inf_{m\geq n} (a_m).$$

Remark. The limit superior and limit infimum exist due to the existence of supremum and infimum in $\overline{\mathbb{R}}$.

Remark. [Rud76] defines the limit superior and infimum in another manner, using subsequential limits; both definitons are equivalent.

Example 15.24. Let $a_n = \dfrac{(-1)^n}{1+\frac{1}{n}}$. Then

$$\limsup_{n\to\infty} a_n = 1, \quad \liminf_{n\to\infty} a_n = -1.$$

Lemma 15.25.

$$\liminf_{n\to\infty} a_n = -\limsup_{n\to\infty}(-a_n).$$

Proof. Exercise; use the definitions and 13.12. □

Lemma 15.26. *A sequence (a_n) in $[-\infty, \infty]$ converges if and only if*

$$\limsup_{n\to\infty} a_n = \liminf_{n\to\infty} a_n = \lim_{n\to\infty} a_n.$$

Proof.

$\boxed{\Longrightarrow}$ Suppose $a_n \to a$. Let $\varepsilon > 0$ be given, there exists $n \in \mathbb{N}$ such that for all $m \geq n$,

$$|a_m - a| < \varepsilon,$$

or,

$$a - \varepsilon < a_m < a + \varepsilon.$$

It follows that

$$a - \varepsilon < \inf_{m \geq n} a_m \leq \sup_{m \geq n} a_m < a + \varepsilon.$$

Since ε was arbitrary, we must have

$$\liminf_{n \to \infty} a_n = \limsup_{n \to \infty} a_n = a.$$

$\boxed{\Longleftarrow}$ Suppose $\liminf_{n \to \infty} a_n = \limsup_{n \to \infty} a_n = a$. Let $\varepsilon > 0$ be give, there exists $n \in \mathbb{N}$ such that

$$\inf_{m \geq n} a_m > a - \varepsilon, \quad \sup_{m \geq n} a_m < a + \varepsilon.$$

It follows that for all $m \geq n$, we have $|a_m - a| < \varepsilon$. ☐

Proposition 15.27. *Suppose (a_n) is a sequence in $\mathbb{R}$. Then*

(i) $\limsup_{n \to \infty} a_n \in E$;

(ii) if $a > \limsup_{n \to \infty} a_n$, there exists $N \in \mathbb{N}$ such that $a_n < a$ for all $n \geq N$.

Moreover, $\limsup_{n \to \infty} a_n$ is the only number that satisfies (i) and (ii).

Proof.

(i) We consider three cases for the value of $\limsup_{n \to \infty} a_n$:

- If $\limsup_{n \to \infty} a_n = +\infty$, then $\sup E = +\infty$, so E is not bounded above. Hence (a_n) is not bounded above, so (a_n) has a subsequence (a_{n_k}) such that $a_{n_k} \to \infty$
- If $\limsup_{n \to \infty} a_n \in \mathbb{R}$, then $\sup E \in \mathbb{R}$, so E is bounded above. Hence at least one subsequential limit exists, so that (i) follows from Theorems 3.7 and 2.28.
- If $\limsup_{n \to \infty} a_n = -\infty$, then $\sup E = -\infty$, so E contains only one element, namely $-\infty$. Hence (a_n) has no subsequential limit. Thus for any $M \in \mathbb{R}$, $a_n > M$ for at most a finite number of values of n, so that $a_n \to -\infty$.

(ii) We prove by contradiction.

Suppose there is a number $a > \limsup_{n \to \infty} a_n$ such that $a_n \geq a$ for infinitely many values of n. In that case, there is a number $y \in E$ such that $y \geq a > \limsup_{n \to \infty} a_n$, contradicting the definition of $\limsup_{n \to \infty} a_n$.

We now show uniqueness. Suppose, for a contradiction, that two numbers p and q satisfy (i) and (ii). WLOG assume $p < q$. Then choose a such that $p < a < q$. Since p satisfies (i), we have $a_n < a$ for all $n \geq N$. But then q cannot satisfy (i). ☐

Of course, an analogous result is true for $\liminf_{n \to \infty} a_n$.

Lemma 15.28 (Comparison)**.** *If $a_n \leq b_n$ for $n \geq N$ (where N is fixed), then*

$$\liminf_{n\to\infty} a_n \leq \liminf_{n\to\infty} b_n,$$
$$\limsup_{n\to\infty} a_n \leq \limsup_{n\to\infty} b_n.$$

Lemma 15.29 (Arithmetic properties)**.**

(i) *If* $k > 0$, $\limsup\limits_{n\to\infty} ka_n = k \limsup\limits_{n\to\infty} a_n$.

If $k < 0$, $\limsup\limits_{n\to\infty} ka_n = k \liminf\limits_{n\to\infty} a_n$.

(ii) $\limsup(a_n + b_n) \leq \limsup a_n + \limsup b_n$

Moreover, $\limsup\limits_{n\to\infty}(a_n + b_n)$ *may be bounded from below as follows:*

$$\limsup_{n\to\infty}(a_n + b_n) \geq \limsup_{n\to\infty} a_n + \liminf_{n\to\infty} b_n.$$

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

15.1.6 o-notation and Big o-notation

The o-notation and O-notation are used to compare the size of some given sequence relative to some well known sequence.

The o-notation is, roughly speaking, used to compare two sequences when one is much smaller than the other. x_n is said to be much smaller than y_n, denoted $x_n \ll y_n$, if $\lim_{n\to\infty} \frac{x_n}{y_n} = 0$, denoted by $x_n = o(y_n)$. We usually use this notation when both sequences approach 0, for example $\frac{1}{n^2} = o\left(\frac{1}{n}\right)$.

The O-notation tries to measure how fast a sequence grows or shrinks We denote $x_n = O(y_n)$ if there exists a constant $M > 0$ and a natural number $N \in \mathbb{N}$ such that $|x_n| \leq My_n$ for all $n > N$. For example $2n^2 \sin n = O(n^2)$.

Example 15.30. Let (r_k) be a sequence of all rational numbers strictly between 0 and 1 where each rational number appears exactly once. Show that $\limsup r_k = 1$ and $\liminf r_k = 0$.

15.2 Series

Definition 15.31 (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 15.32 (Convergence of series)**.** We say that the series *converges* if $s_n \to s$ (the sequence of partial sums converges), and write $\sum\limits_{n=1}^{\infty} a_n = s$; that is,

$$\forall \varepsilon > 0, \quad \exists N \in \mathbb{N}, \quad \forall n \geq N, \quad \left| \sum_{k=1}^{n} a_k - s \right| < \varepsilon.$$

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 $\sum\limits_{n=1}^{\infty} a_n$ simply as $\sum a_n$.

The Cauchy criterion can be restated in the following form:

Lemma 15.33 (Cauchy criterion)**.** *$\sum a_n$ converges if and only if*

$$\forall \varepsilon > 0, \quad \exists N \in \mathbb{N}, \quad \forall n \geq m \geq N, \quad \left| \sum_{k=m}^{n} a_k \right| \leq \varepsilon.$$

15.2.1 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 (15.34)
- Boundedness of partial sums (15.35, for series of non-negative terms)
- Comparison test (15.36)
- Root test (15.40)
- Ratio test (15.41)
- Absolute convergence (15.42)

Lemma 15.34 (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 $|a_n| \leq \varepsilon$ for all $n \geq N$. □

Remark. The converse is not true; a counterexample of the harmonic series.

Lemma 15.35. *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 15.36 (Comparison test). *Consider two sequences (a_n) and (b_n).*

(i) Suppose $|a_n| \leq b_n$ for all $n \geq N_0$ (where N_0 is some fixed integer). If $\sum b_n$ converges, then $\sum a_n$ converges.

(ii) Suppose $a_n \geq b_n \geq 0$ for all $n \geq N_0$. 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 $N \in \mathbb{N}$, $N \geq N_0$ such that for $n \geq m \geq N$,

$$\sum_{k=m}^{n} b_k \leq \varepsilon.$$

By the triangle inequality,

$$\left|\sum_{k=m}^{n} a_k\right| \leq \sum_{k=m}^{n} |a_k| \leq \sum_{k=m}^{n} b_k \leq \varepsilon,$$

so $\sum a_n$ converges, by the Cauchy criterion.

(ii) We prove the contrapositive. If $\sum a_n$ converges, and since $|b_n| \leq a_n$ for all $n \geq N_0$, 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 15.37 (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 $|x| \geq 1$*, 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 $|x| \geq 1$, $x^n \not\to 0$. By the divergence test, $\sum x^n$ diverges.

□

Example 15.38 (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 $a_1 \ge a_2 \ge \cdots \ge 0$. 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 $(2^k a_{2^k})$ respectively; 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

$$\begin{aligned} s_n &= a_1 + a_2 + \cdots + a_n \\ &\le a_1 + (a_2 + a_3) + \cdots + (a_{2^k} + \cdots + a_{2^{k+1}-1}) \\ &\le a_1 + 2a_2 + \cdots + 2^k a_{2^k} \\ &= t_k. \end{aligned}$$

  By comparison test, if (t_k) converges, then (s_n) converges.

- For $n > 2^k$,

$$\begin{aligned} s_n &\ge a_1 + a_2 + (a_3 + a_4) + \cdots + (a_{2^{k-1}+1} + \cdots + a_{2^k}) \\ &\ge \frac{1}{2}a_1 + a_2 + 2a_4 + \cdots + 2^{k-1} a_{2^k} \\ &= \frac{1}{2} t_k. \end{aligned}$$

  By comparison test, if (s_n) converges, then (t_k) converges.

$\square$

Proposition (p-test)**.**

(i) If $p > 1$, $\sum \frac{1}{n^p}$ converges.

(ii) If $p \le 1$, $\sum \frac{1}{n^p}$ diverges.

Proof. Note that if $p \le 0$, 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 $\frac{1}{n^p}$. $\square$

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 15.39 (The number e)**.** Consider the series

$$\sum_{n=0}^{\infty} \frac{1}{n!}.$$

We first show that the above series converges. Consider the n-th partial sum:

$$\begin{aligned}\sum_{k=0}^{n} \frac{1}{k!} &= \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{n!} \\ &\leq 1 + 1 + \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^{n-1}} \\ &< 1 + 1 + \frac{1}{2} + \frac{1}{2^2} + \cdots = 3.\end{aligned}$$

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:

$$e := \sum_{n=0}^{\infty} \frac{1}{n!}$$

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 $q \geq 1$, 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 $t_n \leq s_n$. By 15.28, we have that

$$\limsup_{n\to\infty} t_n \leq \limsup_{n\to\infty} s_n = e.$$

Next, if $n \geq m$,

$$t_n \geq 1 + 1 + \frac{1}{2!}\left(1 - \frac{1}{n}\right) + \cdots + \frac{1}{m!}\left(1 - \frac{1}{n}\right)\cdots\left(1 - \frac{m-1}{n}\right).$$

Let $n \to \infty$, keeping m fixed. We get

$$\liminf_{n\to\infty} t_n \geq 1 + 1 + \frac{1}{2!} + \cdots + \frac{1}{m!},$$

so that

$$s_m \leq \liminf_{n\to\infty} t_n.$$

Letting $m \to \infty$, we get

$$e \leq \liminf_{n\to\infty} t_n.$$

Thus it follows that

$$\limsup_{n\to\infty} t_n = \liminf_{n\to\infty} t_n = e,$$

so the desired result follows. ☐

Lemma 15.40 (Root test)**.** *Given $\sum a_n$, let $\alpha = \limsup\limits_{n\to\infty} \sqrt[n]{|a_n|}$.*

(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.

Remark. We use limsup since the limsup of a sequence always exists (in $\overline{\mathbb{R}}$), while the limit may not necessarily exist.

Proof.

(i) If $\alpha < 1$, choose β such that $\alpha < \beta < 1$. Since $\beta > \limsup_{n\to\infty} \sqrt[n]{|a_n|}$, there exists $n \in \mathbb{N}$ such that for all $n \geq N$,
$$\sqrt[n]{|a_n|} < \beta,$$
or
$$|a_n| < \beta^n.$$
Note that $\sum \beta^n$ converges since $0 < \beta < 1$. By the comparison test, $\sum a_n$ converges.

(ii) If $\alpha > 1$, $\limsup_{n\to\infty} \sqrt[n]{|a_n|} > 1$, so there exists a subsequence (a_{n_k}) such that
$$\sqrt[n_k]{|a_{n_k}|} \to \alpha.$$
Thus $|a_n| > 1$ for infinitely many values of n. Hence $a_n \not\to 0$, so by the divergence test, $\sum a_n$ diverges.

(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.

$\square$

Lemma 15.41 (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 $\left|\frac{a_{n+1}}{a_n}\right| \geq 1$ for all $n \geq N_0$ (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 $N \in \mathbb{N}$ such tht for all $n \geq N$,
$$\left|\frac{a_{n+1}}{a_n}\right| < \beta.$$

In particular, from $n = N$ to $n = N + k$,

$$\begin{aligned} |a_{N+1}| &< \beta |a_N| \\ |a_{N+2}| &< \beta |a_{N+1}| < \beta^2 |a_N| \\ &\vdots \\ |a_{N+k}| &< \beta^k |a_N| \end{aligned}$$

Hence for all $n \ge N$,

$$\begin{aligned} |a_n| &< |a_N| \beta^{n-N} \\ &= \left(|a_N| \beta^{-N} \right) \beta^n \end{aligned}$$

and taking the sum gives

$$\sum |a_n| < |a_N| \beta^{-N} \sum \beta^n.$$

Since $\beta < 1$, $\sum \beta^n$ converges. By the comparison test, $\sum a_n$ converges.

(ii) Suppose $\left| \frac{a_{n+1}}{a_n} \right| \ge 1$ for all $n \ge N_0$. Then $|a_{n+1}| \ge |a_n|$ for $n \ge N_0$, so $a_n \not\to 0$. By the divergence test, $\sum a_n$ diverges.

□

Remark. The ratio test is easier to apply than the root test (since it is usually easier to compute ratios than n-th roots), but the root test is more powerful, as shown by Theorem 3.37 in [Rud76].

The series $\sum a_n$ is said to *converge absolutely* if the series $\sum |a_n|$ converges.

Lemma 15.42 (Absolute convergence)**.** *If $\sum a_n$ converges absolutely, then $\sum a_n$ converges.*

Proof. Suppose $\sum a_n$ converges absolutely; that is, $\sum |a_n|$ converges. Using the Cauchy criterion, fix $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $n \ge m \ge N$,

$$\left| \sum_{k=m}^{n} |a_k| \right| < \varepsilon.$$

Since all the terms are non-negative, we can simply write

$$\sum_{k=m}^{n} |a_k| < \varepsilon.$$

By the triangle inequality,

$$\left| \sum_{k=m}^{n} a_k \right| \le \sum_{k=m}^{n} |a_k| < \varepsilon.$$

Hence by the Cauchy criterion, $\sum a_n$ converges. □

Note that the converse may not necessarily be true. We say that $\sum a_n$ is *conditionally convergent* if it converges, but does not converge absolutely.

Example 15.43. The alternating harmonic series given by

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}$$

converges to $\ln 2$, but it is not absolutely convergent (since the harmonic series diverges).

15.2.2 Summation by Parts

Proposition 15.44 (Partial summation formula). *Given two sequences (a_n) and (b_n), let the n-partial sum of (a_n) be denoted by*

$$A_n = \sum_{k=0}^{n} a_k$$

for $n \geq 0$; let $A_{-1} = 0$. Then, if $0 \leq p \leq q$, 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. □

Suppose that we have a series $\sum a_n b_n$ and we wish to show that it converges, then there are these two strategies:

Lemma 15.45 (Dirichlet's test). *Suppose (a_n) and (b_n) are sequences such that*

- *the partial sums A_n of $\sum a_n$ form a bounded sequence,*
- $b_0 \geq b_1 \geq b_2 \geq \cdots$,
- $b_n \to 0$.

Then $\sum a_n b_n = 0$.

Proof. Since the partial sums A_n form a bounded sequence, there exists M such that

$$|A_n| \leq M \quad (\forall n \in \mathbb{N})$$

Since $b_n \to 0$, fix $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that

$$b_N \leq \frac{\varepsilon}{2M}.$$

For $q \geq p \geq N$, by the partial summation formula, we have

$$\begin{aligned}\left|\sum_{n=p}^{q} a_n b_n\right| &= \left|\sum_{n=p}^{q-1} A_n(b_n - b_{n+1}) + A_q b_q - A_{p-1} b_p\right| \\ &\leq M\left|\sum_{n=p}^{q-1}(b_n - b_{n+1}) + b_q + b_p\right| \qquad [\because |A_n| \leq M] \\ &= M\left|(b_p - b_q) + b_q + b_p\right| = 2Mb_p \leq 2Mb_n \leq \varepsilon.\end{aligned}$$

By the Cauchy criterion, $\sum a_n b_n$ converges to 0. □

The following is a convenient application of Dirichlet's test.

Corollary 15.46 (Alternating series test)**.** *Suppose* (c_n) *is a sequence such that*

- $|c_1| \geq |c_2| \geq |c_3| \geq \cdots$,
- $c_{2m-1} \geq 0, c_{2m} \leq 0$ *for* $m = 1, 2, 3, \ldots$,
- $c_n \to 0$.

Then $\sum c_n = 0$.

Proof. Let

$$a_n = (-1)^{n+1}, \quad b_n = |c_n|.$$

Note that

- the partial sums of (a_n) are 0s and 1s, so they are bounded;
- $b_0 \geq b_1 \geq b_2 \geq \cdots$ holds by assumption;
- $c_n \to 0$ implies $|c_n| \to 0$, so $b_n \to 0$.

Then by 15.45, we have that $\sum a_n b_n = 0$, so $\sum c_n = 0$. □

Lemma 15.47 (Abel's test)**.** *If* $\sum a_n$ *converges, and* (b_n) *is monotonic and bounded, then* $\sum a_n b_n$ *converges.*

Proof. Suppose that $\sum a_n$ is convergent, and (b_n) is monotonic and bounded. Now what we do is that we try to transform b_n into the one that we see in Dirichlet's test.

Since (b_n) is monotonic, it's either monotonically increasing or decreasing. If (b_n) were to be monotonically increasing, we multiply the sequences (a_n) and (b_n) by -1; $\sum a_n$ is still convergent here. Thus assume (b_n) is monotonically decreasing.

Since (b_n) is monotonically decreasing, the sequence (b_n) has an infimum which is also its limit, so we may let $b = \lim b_n = \inf b_n$. Then $(b_n - b)$ is a monotonically decreasing sequence which converges to 0

Since $\sum a_n$ is convergent, its partial sums A_n is bounded

By Dirichlet's test, this shows that $\sum a_n(b_n - b)$ is convergent

However, we have the additional requirement that $\sum a_n$ is convergent, so we can easily deal with the extra '$-b$' term just by noting that $\sum a_n b$ is convergent and hence

$$\sum a_n b_n = \sum a_n(b_n - b) + \sum a_n b$$

is convergent. $\square$

15.2.3 Addition and Multiplication of Series

Lemma 15.48. *If* $\sum a_n = A$ *and* $\sum b_n = B$*, then*

(i) $\sum (a_n + b_n) = A + B$, *(addition)*

(ii) $\sum ca_n = cA$ *for some constant* c. *(scalar multiplication)*

Proof.

(i) Let the n-th partial sums be denoted by

$$A_n = \sum_{k=0}^{n} a_k, \quad B_n = \sum_{k=0}^{n} b_k.$$

Then

$$\lim_{n\to\infty} \sum_{k=0}^{n} (a_k + b_k) = \lim_{n\to\infty} (A_n + B_n) = \lim_{n\to\infty} A_n + \lim_{n\to\infty} B_n = A + B.$$

(ii) Simply factor out the constant c:

$$\lim_{n\to\infty} \sum_{k=0}^{n} ca_k = c \lim_{n\to\infty} \sum_{k=0}^{n} a_k = cA.$$

□

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 15.49 (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

$$\begin{aligned}\left(\sum_{n=0}^{\infty} a_n z^n\right)\left(\sum_{n=0}^{\infty} b_n z^n\right) &= \left(a_0 + a_1 z + a_2 z^2 + \cdots\right)\left(b_0 + b_1 z + b_2 z^2 + \cdots\right) \\ &= a_0 b_0 + (a_0 b_1 + a_1 b_0) z + (a_0 b_2 + a_1 b_1 + a_2 b_0) z^2 + \cdots \\ &= c_0 + c_1 z + c_2 z^2 + \cdots\end{aligned}$$

Setting $z = 1$, we arrive at the above definition.

Note that $\sum c_n$ may not converge, even if $\sum a_n$ and $\sum b_n$ do. However $\sum c_n$ converges if an additional condition is imposed: at least one of the two series converges absolutely.

Proposition 15.50 (Mertens' theorem)**.** *Suppose $\sum a_n = A$, $\sum b_n = B$, and $\sum a_n$ converges absolutely. Then their Cauchy product converges to AB.*

Proof. Let $\sum c_n$ be the Cauchy product of $\sum a_n$ and $\sum b_n$. Let the n-th partial sums be denoted by

$$A_n = \sum_{k=0}^{n} a_k, \quad B_n = \sum_{k=0}^{n} b_k, \quad C_n = \sum_{k=0}^{n} c_k.$$

Also let $\beta_n = B_n - B$. Then

$$\begin{aligned} C_n &= a_0 b_0 + (a_0 b_1 + a_1 b_0) + \cdots + (a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0) \\ &= a_0 B_n + a_1 B_{n-1} + \cdots + a_n B_0 \\ &= a_0(B + \beta_n) + a_1(B + \beta_{n-1}) + \cdots + a_n(B + \beta_0) \\ &= A_n B + (a_0 \beta_n + a_1 \beta_{n-1} + \cdots + a_n \beta_0) \end{aligned}$$

Our goal is to show that $C_n \to AB$. Since $A_n B \to AB$, it suffices to show that

$$\gamma_n = a_0 \beta_n + a_1 \beta_{n-1} + \cdots + a_n \beta_0 \to 0.$$

We now use the absolute convergence of (a_n); let $\alpha = \sum |a_n|$. Fix $\varepsilon > 0$, there exists $N_1 \in \mathbb{N}$ such that

$$n \geq N_1 \implies \sum_{k=0}^{n} |a_k| - \alpha < \varepsilon$$

since the terms are non-negative. Since $B_n \to B$, $\beta_n \to 0$. Then there exists $N_2 \in \mathbb{N}$ such that

$$n \geq N_2 \implies |\beta_n| \leq \varepsilon.$$

Let $N = \max\{N_1, N_2\}$. Then for $n \geq N$, by triangle inequality,

$$\begin{aligned} |\gamma_n| &= |\beta_0 a_n + \cdots + \beta_n a_0| \\ &\leq |\beta_0 a_n + \cdots + \beta_N a_{n-N}| + |\beta_{N+1} a_{n-N-1} + \cdots + \beta_n a_0| \\ &\leq |\beta_0 a_n + \cdots + \beta_N a_{n-N}| + \varepsilon(|a_{n-N-1}| + \cdots + |a_0|) \\ &\leq |\beta_0 a_n + \cdots + \beta_N a_{n-N}| + \varepsilon\alpha. \end{aligned}$$

Keeping N fixed, and letting $n \to \infty$, we get

$$\limsup_{n\to\infty} |\gamma_n| \leq \varepsilon\alpha,$$

sine $a_n \to 0$. Since ε is arbitrary, we have $\gamma_n \to 0$, as desired. □

Proposition 15.51 (Abel's theorem)**.** *Let $\sum a_n = A$, $\sum b_n = B$, $\sum c_n = C$, where $\sum c_n$ is the Cauchy product of $\sum a_n$ and $\sum b_n$. Then $C = AB$.*

15.2.4 Rearrangements

Definition 15.52 (Rearrangement)**.** Let (k_n) be a sequence in which every positive integer appears once and only once. Let

$$a'_n = a_{k_n} \quad (\forall n \in \mathbb{N})$$

We say that $\sum a'_n$ is a *rearrangement* of $\sum a_n$.

If (s_n) and (s'_n) are the sequences of partial sums of (a_n) and (a'_n) respectively, it is easily seen that, in general, these two sequences consist of entirely different numbers. We are thus led to the problem of determining under what conditions all rearrangements of a convergent series will converge and whether the sums are necessarily the same.

Theorem 15.53 (Riemann series theorem)**.** *Let $\sum a_n$ be a series of real numbers which converges, but not absolutely. Suppose $-\infty \leq \alpha \leq \beta \leq \infty$. Then there exists a rearrangement $\sum a'_n$ with partial sums s'_n such that*

$$\liminf_{n\to\infty} s'_n = \alpha, \quad \limsup_{n\to\infty} s'_n = \beta.$$

Proof. Let

$$p_n = \frac{|a_n| + a_n}{2}, \quad q_n = \frac{|a_n| - a_n}{2} \quad (n = 1, 2, \dots).$$

Then $p_n - q_n = a_n$, $p_n + q_n = |a_n|$, $p_n \geq 0$, $q_n \geq 0$.

Claim. The series $\sum p_n$ and $\sum q_n$ must both diverge.

If both were convergent, then

$$\sum (p_n + q_n) = \sum |a_n|$$

would converge, contrary to hypothesis. Since

$$\sum_{n=1}^{N} a_n = \sum_{n=1}^{N} (p_n - q_n) = \sum_{n=1}^{N} p_n - \sum_{n=1}^{N} q_n,$$

divergence of $\sum p_n$ and convergence of $\sum q_n$ (or vice versa) implies divergence of $\sum a_n$, again contrary to hypothesis.

Now let $P_1, P_2, \dots$ denote the non-negative terms of $\sum a_n$, in the order which they occur, and let $Q_1, Q_2, \dots$ be the absolute values of the negative terms of $\sum a_n$, also in their original order.

The series $\sum P_n$ and $\sum Q_n$ differ from $\sum p_n$ and $\sum q_n$ only by zero terms, and are therefore divergent.

We shall construct sequences (m_n) and (k_n), such that the series

$$\begin{aligned} &(P_1 + \cdots + P_{m_1}) - (Q_1 + \cdots + Q_{k_1}) + \\ &(P_{m_1+1} + \cdots + P_{m_2}) - (Q_{k_1+1} + \cdots + Q_{k_2}) + \cdots \end{aligned} \tag{1}$$

which clearly is a rearrangement of $\sum a_n$, satisfies $\liminf\limits_{n\to\infty} s'_n = \alpha$, $\limsup\limits_{n\to\infty} s'_n = \beta$.

Choose real-valued sequences (α_n) and (β_n) such that $\alpha_n \to \alpha$, $\beta_n \to \beta$, $\alpha_n < \beta_n$, $\beta_1 > 0$.

Let m_1, k_1 be the smallest integers such that

$$P_1 + \cdots + P_{m_1} > \beta_1,$$
$$P_1 + \cdots + P_{m_1} - (Q_1 + \cdots + Q_{k_1}) < \alpha_1;$$

let m_2, k_2 be the smallest integers such that

$$(P_1 + \cdots + P_{m_1}) - (Q_1 + \cdots + Q_{k_1}) + (P_{m_1+1} + \cdots + P_{m_2}) > \beta_2$$
$$(P_1 + \cdots + P_{m_1}) - (Q_1 + \cdots + Q_{k_1}) + (P_{m_1+1} + \cdots + P_{m_2}) - (Q_{k_1+1} + \cdots + Q_{k_2}) < \alpha_2;$$

and continue in this way. This is possible since $\sum P_n$ and $\sum Q_n$ diverge.

If x_n, y_n denote the partial sums of (1) whose last terms are $P_{m_n}, -Q_{k_n}$, then

$$|x_n - \beta_n| \leq P_{m_n}, \quad |y_n - \alpha_n| \leq Q_{k_n}.$$

Since $P_n \to 0$ and $Q_n \to 0$ as $n \to \infty$, wee see that $x_n \to \beta$, $y_n \to \alpha$.

Finally, it is clear that no number less than α or greater than β can be a subsequential limit of the partial sums of (1). to review

□

Theorem 15.54. *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.*

Proof. Let $\sum a'_n$ be a rearrangement, with partial sums s'_n. Since $\sum a_n$ converges absolutely, given $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $m \geq n \geq N$ implies

$$\sum_{i=n}^{m} |a_i| < \varepsilon. \tag{1}$$

Now choose p such that the integers $1, 2, \ldots, N$ are all contained in the set $k_1, \ldots, k_p$ (we use the notation of Definition 15.52). Then if $n > p$, the numbers $a_1, \ldots, a_N$ will cancel in the difference $s_n - s'_n$, so that $|s_n - s'_n| < \varepsilon$, by (1) Hence (s'_n) converges to the same sum as (s_n). □ to review

Exercises

Exercise 15.1. Show the following:

(i) $\lim\limits_{n\to\infty} \dfrac{1}{n^p} = 0 \ (p > 0)$

(ii) $\lim\limits_{n\to\infty} \sqrt[n]{p} = 1 \ (p > 0)$

(iii) $\lim\limits_{n\to\infty} \sqrt[n]{n} = 1$

(iv) $\lim\limits_{n\to\infty} \dfrac{n^\alpha}{(1+p)^n} = 0 \ (p > 0, \alpha \in \mathbb{R})$

(v) $\lim\limits_{n\to\infty} x^n = 0 \ (|x| < 1)$

Solution.

(i) Let $\varepsilon > 0$ be given. Take $N = \left\lfloor \left(\frac{1}{\varepsilon}\right)^{\frac{1}{p}} \right\rfloor + 1$. Then $n \geq N$ implies

$$\left|\frac{1}{n^p} - 0\right| = \frac{1}{n^p} \leq \frac{1}{N^p} < \frac{1}{\left(\left(\frac{1}{\varepsilon}\right)^{\frac{1}{p}}\right)^p} = \varepsilon.$$

(ii) We need to consider cases corresponding to different values of p.

Case 1: $p > 1$**.** Put $x_n = \sqrt[n]{p} - 1$. Then $x_n > 0$, and, by the binomial theorem,

$$1 + nx_n \leq (1 + x_n)^n = p,$$

so that

$$0 < x_n \leq \frac{p-1}{n}.$$

Hence $x_n \to 0$.

Case 2: $p = 1$**.** Trivial.

Case 3: $0 < p < 1$**.** The result is obtained by taking reciprocals.

(iii) Put $x_n = \sqrt[n]{n} - 1$. Then $x_n \geq 0$, and, by the binomial theorem,

$$n = (1 + x_n)^n \geq \frac{n(n-1)}{2} x_n^2.$$

Hence

$$0 \leq x_n \leq \sqrt{\frac{2}{n-1}} \quad (n \geq 2.)$$

(iv) Let k be an integer such that $k > \alpha$, $k > 0$. For $n > 2k$,

$$(1+p)^n > \binom{n}{k} p^k = \frac{n(n-1)\cdots(n-k+1)}{k!} p^k > \frac{n^k p^k}{2^k k!}.$$

Hence

$$0 < \frac{n^\alpha}{(1+p)^n} < \frac{2^k k!}{p^k} n^{\alpha - k} \quad (n > 2k).$$

Since $\alpha - k < 0$, by (i), $n^{\alpha-k} \to 0$.

(v) Take $\alpha = 0$ in (iv).

□

Exercise 15.2. Let (x_n) be a real sequence, let $\alpha \geq 2$ 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.

Exercise 15.3 ([Rud76] 3.1)**.** Prove that the convergence of (a_n) implies the convergence of $(|a_n|)$. Is the converse true?

Solution. Let $\varepsilon > 0$ be given. Since (a_n) is a Cauchy sequence, there exists $N \in \mathbb{N}$ such that for all $n, m \geq N$,

$$|a_n - a_m| < \varepsilon.$$

See that

$$||a_n| - |a_m|| \leq |a_n - a_m| < \varepsilon,$$

so $(|a_n|)$ is a Cauchy sequence, and therefore must converge.

The converse is not true, as shown by the sequence (a_n) with $a_n = (-1)^n$. □

Exercise 15.4 ([Rud76] 3.2)**.** Calculate $\lim\limits_{n\to\infty} \left(\sqrt{n^2+n} - n\right)$.

Solution. □

Exercise 15.5 ([Rud76] 3.3)**.** The sequence (a_n) is recursively defined by

$$\begin{cases} a_0 = \sqrt{2}, \\ a_{n+1} = \sqrt{2 + a_n} \quad n \geq 0. \end{cases}$$

Show that (a_n) converges.

Solution. We first prove by induction that $a_n \leq a_{n+1} \leq 2$ for all $n \in \mathbb{N}$. For $n = 0$,

$$a_0 = \sqrt{2} \leq \sqrt{2 + \sqrt{2}} = a_1 \leq \sqrt{2 + \sqrt{4}} = 2.$$

If $a_{n-1} \leq a_n \leq 2$, then

$$a_n = \sqrt{2 + a_{n-1}} \leq \sqrt{2 + a_n} = a_{n+1} \leq \sqrt{2 + 2} = 2.$$

Hence (a_n) is monotonically increasing and bounded above by 2. By the monotone convergence theorem, (a_n) converges; let $a_n \to a$. Applying the limit on both sides of $a_{n+1} = \sqrt{2 + a_n}$,

$$\begin{aligned} \lim_{n\to\infty} a_{n+1} &= \lim_{n\to\infty} \sqrt{2 + a_n} \\ a &= \sqrt{2 + a} \\ a &= 2 \text{ or } 1 \end{aligned}$$

Since all $a_n \geq 0$, we must have $a = 2$. □

Exercise 15.6 (Contractive sequence)**.** A complex sequence (x_n) is *contractive* if there exists $k \in [0, 1)$ such that

$$|a_{n+2} - a_{n+1}| \leq k|a_{n+1} - a_n| \quad (\forall n \in \mathbb{N})$$

Show that every contractive sequence is convergent.

Solution. By induction on n, we have

$$|a_{n+1} - a_n| \leq k^{n-1}|a_2 - a_1| \quad (\forall n \in \mathbb{N})$$

Thus

$$\begin{aligned} |a_{n+p} - a_n| &\leq |a_{n+1} - a_n| + |a_{n+2} - a_{n+1}| + \cdots + |a_{n+p} - a_{n+p-1}| \\ &\leq \left(k^{n-1} + k^n + \cdots + k^{n+p-2}\right)|a_2 - a_1| \\ &\leq k^{n-1}\left(1 + k + k^2 + \cdots + k^{p-1}\right)|a_2 - a_1| \\ &\leq \frac{k^{n-1}}{1-k}|a_2 - a_1| \end{aligned}$$

for all $n, p \in \mathbb{N}$. Since $k^{n-1} \to 0$ as $n \to \infty$ (independently of p), this implies (a_n) is a Cauchy sequence, so it is convergent. □

Exercise 15.7 ([Rud76] 3.4)**.** Find the limit superior and limit inferior of the sequence (a_n) defined by

$$a_1 = 0, \quad a_{2m} = \frac{a_{2m-1}}{2}, \quad a_{2m+1} = a_{2m} + \frac{1}{2}.$$

Solution. We shall prove by induction that

$$a_{2m} = \frac{1}{2} - \frac{1}{2^m}, \quad a_{2m+1} = 1 - \frac{1}{2^m}$$

for $m = 1, 2, \ldots$. The second of these equalities is a direct consequence of the first, and so we need only prove the first. Immediate computation shows that $a_2 = 0$ and $a_3 = \frac{1}{2}$. Hence assume that both formulae holds for $m \leq r$. Then

$$a_{2r+2} = \frac{1}{2}a_{2r+1} = \frac{1}{2}\left(1 - \frac{1}{2^r}\right) = \frac{1}{2} - \frac{1}{2^{r+1}}.$$

This completes the induction. We thus have $\limsup_{n\to\infty} a_n = 1$ and $\liminf_{n\to\infty} a_n = \frac{1}{2}$. □

Exercise 15.8 ([Rud76] 3.7)**.** Prove that the convergence of $\sum a_n$ implies the convergence of

$$\sum \frac{\sqrt{a_n}}{n}$$

if $a_n \geq 0$.

Exercise 15.9 ([Rud76] 3.13)**.** Prove that the Cauchy product of two absolutely convergent series converges absolutely.

Exercise 15.10 ([Rud76] 3.23)**.** Suppose (a_n) and (b_n) are Cauchy sequences in a metric space X. Show that the sequence $(d(a_n, b_n))$ converges.

Exercise 15.11. Prove that $\sum_{n=1}^{\infty} \frac{\sin n\alpha}{n}$ is convergent for all real α, and that $\sum_{n=1}^{\infty} \frac{\cos n\alpha}{n}$ is convergent if and only if $\alpha \neq 2k\pi$ for integer k.

Hint:

$$\begin{aligned}
&\sin x + \sin 2x + \sin 3x \\
&= \frac{\left(\sin \frac{x}{2}\right)\left(\sin x + \sin 2x + \sin 3x\right)}{\sin \frac{x}{2}} \\
&= \frac{-\frac{1}{2}\left(\cos \frac{3x}{2} - \cos \frac{x}{2} - \cdots\right)}{\sin \frac{x}{2}}
\end{aligned}$$

Exercise 15.12. Determine whether the series

$$\sum_{n=1}^{\infty} \ln\left(1 + \frac{(-1)^n}{n^p}\right)$$

is absolutely convergent, conditionally convergent or divergent for each $p > 0$.

Hint: $\ln(1+x) \to x$ as $x \to 0$.

Exercise 15.13. Suppose that $\sum_{n=1}^{\infty} a_n$ is absolutely convergent; for $n = 1$ to ∞, define the subsequences

$$b_n = a_{2n-1}, \quad c_n = a_{2n}.$$

(i) Show that $\sum b_n$ and $\sum c_n$ are absolutely convergent.

(ii) Show that $\sum a_n = \sum b_n + \sum c_n$.

(iii) Using the fact that $\sum \frac{1}{n^2} = \frac{\pi^2}{6}$, find $\sum \frac{(-1)^{n-1}}{n^2}$.

Exercise 15.14. Given a function $f\colon \mathbb{R} \to \mathbb{R}$ such that for any $x \in \mathbb{R}$, there exists $\varepsilon > 0$ such that f is monotically increasing on $(x - \varepsilon, x + \varepsilon)$. Prove that f is monotonically increasing on $\mathbb{R}$.

Chapter 16

Continuity

Let (X, d_X) and (Y, d_Y) be metric spaces. Let $E \subset X$, then the metric d_X induces a metric on E. Consider a function $f : E \to Y$. In particular, if $Y = \mathbb{R}$, f is called a *real-valued function*; if $Y = \mathbb{C}$, f is called a *complex-valued function*.

16.1 Limit of Functions

Recall that we have previously defined limits for sequences. Now, we will define limits for functions.

Definition 16.1 (Limit of function)**.** Let p be a limit point of E. We say $\lim_{x \to p} f(x) = q$ if there exists $q \in Y$ such that

$$\forall \varepsilon > 0, \quad \exists \delta > 0, \quad \forall x \in E, \quad 0 < d_X(x, p) < \delta \implies d_Y(f(x), q) < \varepsilon.$$

That is, we can get $f(x)$ close to q as desired, by choosing x sufficiently close to, but not equal to, p.

Remark. Note that $p \in X$, but it is not necessary that $p \in E$ in the above definition. Moreover, even if $p \in E$, we may very well have $f(p) \neq \lim_{x \to p} f(x)$.

We can recast the above definition in terms of limits of sequences:

Lemma 16.2. *Let p be a limit point of E. Then*

$$\lim_{x \to p} f(x) = q \tag{I}$$

if and only if

$$\lim_{n \to \infty} f(p_n) = q \tag{II}$$

for every sequence (p_n) in $E \setminus \{p\}$ where $p_n \to p$.

Proof.

$\boxed{\implies}$ Suppose (I) holds. Then fix $\varepsilon > 0$, there exists $\delta > 0$ such that for all $x \in E$,

$$0 < d_X(x, p) < \delta \implies d_Y(f(x), q) < \varepsilon.$$

Let (p_n) be a sequence in $E \setminus \{p\}$. Since $p_n \to p$, for the same $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $n \geq N$,

$$0 < d_X(p_n, p) < \delta.$$

This implies that for $n \geq N$, $d_Y\left(f(p_n), q\right) < \varepsilon$. Hence by definition $\lim_{n\to\infty} f(p_n) = q$.

$\boxed{\Longleftarrow}$ Suppose, for a contradiction, (II) holds and (I) does not hold. Then $\lim_{x\to p} f(x) \neq q$, so

$$\exists \varepsilon > 0, \quad \forall \delta > 0, \quad \exists x \in E, \quad 0 < d_X(x,p) < \delta \quad \text{and} \quad d_Y\left(f(x), q\right) \geq \varepsilon.$$

Since (II) holds, taking $\delta_n = \frac{1}{n}$ $(n = 1, 2, \dots)$, we thus find a sequence (p_n) in $E \setminus \{p\}$ such that

$$0 < d_X(p_n, p) < \frac{1}{n} \quad \text{and} \quad d_Y\left(f(p_n), q\right) \geq \varepsilon.$$

Clearly $p_n \to p$ but $f(p_n) \not\to q$, contradicting (II). □

Corollary 16.3. *If f has a limit at p, this limit is unique.*

Proof. Suppose $\lim_{x\to p} f(x) = q$ and $\lim_{x\to p} f(x) = q'$. We will show that $q = q'$.

By 16.2, for every sequence (p_n) in $E \setminus \{p\}$ where $p_n \to p$, we have

$$f(p_n) \to q \quad \text{and} \quad f(p_n) \to q'.$$

But the limit of a sequence is unique, so we must have $q = q'$. □

Suppose $f, g \colon E \to \mathbb{C}$. Define

$$(f+g)(x) = f(x) + g(x) \quad (x \in E).$$

We define the difference $f - g$, the product fg, and the quotient f/g similarly, with the understanding that the quotient is defined only at $x \in E$ at which $g(x) \neq 0$.

Similarly, if $\mathbf{f}, \mathbf{g} \colon E \to \mathbb{R}^k$, we define

$$(\mathbf{f} + \mathbf{g})(x) = \mathbf{f}(x) + \mathbf{g}(x), \quad (\mathbf{f} \cdot \mathbf{g})(x) = \mathbf{f}(x) \cdot \mathbf{g}(x);$$

and if λ is a real number, $(\lambda \mathbf{f})(x) = \lambda \mathbf{f}(x)$.

Lemma 16.4 (Arithmetic properties)**.** *Suppose $E \subset X$, p is a limit point of E. Let $f, g \colon E \to \mathbb{C}$, $\lim_{x\to p} f(x) = A$, $\lim_{x\to p} g(x) = B$. Then*

(i) $\lim_{x\to p} (f+g)(x) = A + B$ *(sum)*

(ii) $\lim_{x\to p} (fg)(x) = AB$ *(product)*

(iii) $\lim_{x\to p} \left(\frac{f}{g}\right)(x) = \frac{A}{B}$ $(B \neq 0)$ *(quotient)*

Proof. These follow from 16.2 and analagous limit properties of sequences in $\mathbb{C}$. $\square$

If $\mathbf{f}, \mathbf{g}\colon E \to \mathbb{R}^k$, then (i) remains true, and (ii) becomes $\lim_{x\to p}(\mathbf{f} \cdot \mathbf{g})(x) = \mathbf{A} \cdot \mathbf{B}$.

16.2 Continuous Functions

Definition 16.5 (Continuity)**.** Suppose $E \subset X$. We say $f\colon E \to Y$ is ***continuous*** at $p \in E$ if

$$\forall \varepsilon > 0, \quad \exists \delta > 0, \quad \forall x \in E, \quad d_X(x,p) < \delta \implies d_Y(f(x), f(p)) < \varepsilon.$$

If f is continuous at every point of E, we say that f is *continuous on* E.

This definition reflects the intuitive idea that for any arbitrary target distance around $f(p)$, we can always find points $x \in E$ that are sufficiently close to p, such that their images under f are within the target distance around $f(p)$.

Remark. For f to be continuous at p, we require f to be defined at p. (Compare this with the remark following Definition 16.1.)

Notation. Let X and Y be metric spaces. We denote the space of continuous bounded functions from X to Y as $\mathcal{C}(X,Y)$.

We will often use the next result to show that a function is continuous at a point.

Lemma 16.6. *Let p be a limit point of E. Then f is continuous at p if and only if*

$$\lim_{x \to p} f(x) = f(p).$$

Proof. Compare Definitions 16.1 and 16.5. □

Corollary 16.7 (Sequential criterion for continuity)**.** *$f\colon E \subset X \to Y$ is continuous on E if and only if for every convergent sequence (p_n) in E,*

$$\lim_{n \to \infty} f(p_n) = f\left(\lim_{n \to \infty} p_n\right).$$

Remark. This means that for continuous functions, the limit symbol can be interchanged with the function symbol. Some care is needed in interchanging these symbols because sometimes $(f(p_n))$ converges when (p_n) diverges.

Lemma 16.8. *Let $f, g\colon X \to \mathbb{C}$ be continuous on X. Then the following are continuous on X:*

(i) $f + g$ *(sum)*

(ii) fg *(product)*

(iii) f/g *(provided $g(x) \neq 0$ for all $x \in X$)* *(quotient)*

Proof. At isolated points of X, there is nothing to prove.

At limit points, the statement follows from 16.4 and 16.6. □

Example 16.9. It is a trivial exercise to show that the following complex-valued functions are continuous on $\mathbb{C}$:

- constant functions, defined by $f(z) = c$ for all $z \in \mathbb{C}$;
- the identity function, defined by $f(z) = z$ for all $z \in \mathbb{C}$.

Repeated application of the previous result establishes the continuity of every polynomial

$$f(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_n z^n$$

where $a_i \in \mathbb{C}$.

We now prove the analogue for Euclidean spaces.

Lemma 16.10.

(i) Let $f_1, \dots, f_k : X \to \mathbb{R}$, and let $\mathbf{f} : X \to \mathbb{R}^k$ be defined by

$$\mathbf{f}(x) = (f_1(x), \dots, f_k(x)) \quad (x \in X).$$

Then $\mathbf{f}$ is continuous if and only if each of its components $f_1, \dots, f_k$ is continuous.

(ii) Let $\mathbf{f}, \mathbf{g} : X \to \mathbb{R}^k$ be continuous on X. Then $\mathbf{f} + \mathbf{g}$ and $\mathbf{f} \cdot \mathbf{g}$ are continuous on X.

Proof. (i) follows from the inequalities

$$|f_j(x) - f_j(y)| \le |\mathbf{f}(x) - \mathbf{f}(y)| = \left(\sum_{i=1}^{k} |f_i(x) - f_i(y)|^2 \right)^{1/2}$$

for $j = 1, \dots, k$.

(ii) follows from (i) and 16.8. □

We now consider the composition of functions. The following result shows that a continuous function of a continuous function is continuous.

Proposition 16.11. *Suppose X, Y, Z are metric spaces, $E \subset X$. Let*

- *$f \colon E \to Y$,*
- *$g \colon f(E) \subset Y \to Z$,*
- *$h \colon E \to Z$ is defined by $h = g \circ f$.*

If f is continuous at $p \in E$, and g is continuous at $f(p)$, then h is continuous at p.

Proof. Let $\varepsilon > 0$ be given. Since g is continuous at $f(p)$, there exists $\eta > 0$ such that for all $y \in f(E)$,

$$d_Y(y, f(p)) < \eta \implies d_Z(g(y), g(f(p))) < \varepsilon. \tag{1}$$

Since f is continuous at p, there exists $\delta > 0$ such that for all $x \in E$,

$$d_X(x, p) < \delta \implies d_Y(f(x), f(p)) < \eta. \tag{2}$$

Combining (1) and (2), it follows that for all $x \in E$,

$$d_X(x, p) < \delta \implies d_Z\left(h(x), h(p)\right) = d_Z\left(g\left(f(x)\right), g\left(f(p)\right)\right) < \varepsilon.$$

Therefore h is continuous at p. $\square$

Notation. While functions are technically defined on a subset E of a metric space, the complement of E plays no role in the definition of continuity, so we can safely ignore the complement, and think of continuous functions as mappings from one metric space to another.

16.2.1 Continuity and Pre-images of Open or Closed Sets

The next result is a characterisation of continuity; it states that continuous functions are maps whose pre-image of open sets are open.

> **Lemma 16.12.** *$f\colon X \to Y$ is continuous on X if and only if $f^{-1}(U)$ is open in X for every open set $U \subset Y$.*

Proof.

$\boxed{\Longrightarrow}$ Suppose f is continuous on X. Let $U \subset Y$ be open. Let $p \in f^{-1}(U)$.

Since $p \in f^{-1}(U)$, there exists $y \in U$ such that $f(p) = y$. By openness of U, there exists $\varepsilon > 0$ such that $B_\varepsilon(y) \subset U$.

Since f is continuous at p, for the same ε, there exists $\delta > 0$ such that for all $x \in X$,

$$d_X(x, p) < \delta \implies d_Y\left(f(x), y\right) < \varepsilon,$$

or

$$f\left(B_\delta(p)\right) \subset B_\varepsilon(y).$$

Hence

$$B_\delta(p) \subset f^{-1}\left(f(B_\delta(p))\right) \subset f^{-1}\left(B_\varepsilon(y)\right) \subset f^{-1}(U),$$

so $f^{-1}(U)$ is open in X.

$\boxed{\Longleftarrow}$ Suppose $f^{-1}(U)$ is open in X for every open set $U \subset Y$. Fix $p \in X$, let $y = f(p)$. We will show that f is continuous at p.

For every $\varepsilon > 0$, the ball $B_\varepsilon(y)$ is open in Y, so $f^{-1}\left(B_\varepsilon(y)\right)$ is open in X (by assumption). Now $p \in f^{-1}\left(B_\varepsilon(y)\right)$, so by openness of $f^{-1}\left(B_\varepsilon(y)\right)$, there exists $\delta > 0$ such that $B_\delta(p) \subset f^{-1}\left(B_\varepsilon(y)\right)$. Hence $f\left(B_\delta(p)\right) \subset B_\varepsilon(y)$; that is,

$$d_X(x, p) < \delta \implies d_Y\left(f(x), y\right) < \varepsilon.$$

Therefore f is continuous at p. □

> **Corollary 16.13.** *$f\colon X \to Y$ is continuous on X if and only if $f^{-1}(C)$ is closed in X for every closed set $C \subset Y$.*

Proof. This follows from the above result, since a set is closed if and only if its complement is open, and since $f^{-1}(E^c) = [f^{-1}(E)]^c$ for every $E \subset Y$. □

16.2.2 Continuity and Compactness

We say $\mathbf{f}: E \to \mathbb{R}^k$ is *bounded* if there exists $M \in \mathbb{R}$ such that $\|\mathbf{f}(x)\| \leq M$ for all $x \in E$.

The next result shows that continuous functions preserve compactness.

Proposition 16.14. *Suppose $f: X \to Y$ is continuous on X, where X is compact. Then $f(X)$ is compact.*

Proof. Let $\{U_i \mid i \in I\}$ be an open cover of $f(X)$. Since f is continuous on X, by 16.12, each of the sets $f^{-1}(U_i)$ is open.

Consider the open cover $\{f^{-1}(U_i) \mid i \in I\}$. Since X is compact, there exist finitely many indices $i_1, \dots, i_n$ such that

$$X \subset \bigcup_{k=1}^{n} f^{-1}(U_{i_k}).$$

Since $f(f^{-1}(E)) \subset E$ for every $E \subset Y$, we have that

$$f(X) \subset \bigcup_{k=1}^{n} U_{i_k}.$$

Hence $f(X)$ is compact. □

Corollary 16.15. *If $\mathbf{f}: X \to \mathbb{R}^k$ is continuous on X, where X is compact, then $\mathbf{f}(X)$ is closed and bounded. Thus, $\mathbf{f}$ is bounded.*

Proof. By 16.14, $\mathbf{f}(X)$ is compact. Since $\mathbf{f}(X) \subset \mathbb{R}^k$, by the Heine–Borel theorem, $\mathbf{f}(X)$ is closed and bounded. □

The result is particularly important when f is a real-valued function; the next result states that a continuous real-valued function on a compact set must attain its minimum and maximum.

Theorem 16.16 (Extreme value theorem)**.** *Suppose $f: X \to \mathbb{R}$ is continuous, X is compact. Let*

$$M = \sup_{p \in X} f(p), \quad m = \inf_{p \in X} f(p).$$

Then there exist $p, q \in X$ such that $f(p) = M$ and $f(q) = m$.

Proof. From the previous corollary, $f(X)$ is a closed and bounded set in $\mathbb{R}$. Hence $f(X)$ contains its supremum and infimum, by 14.31. □

Proposition 16.17. *Suppose $f: X \to Y$ is continuous on X and bijective, X is compact. Then its inverse $f^{-1}: Y \to X$ is continuous on Y.*

Proof. By 16.12, it suffices to prove that $f(U)$ is open in Y for every open set U in X.

Let U be an open set in X. Then its complement U^c is closed in X. Since U^c is a closed subset of a compact set X, U^c is compact. Thus by 16.14, $f(U^c)$ is a compact subset of Y, so $f(U^c)$ is closed in Y. Since f is bijective and thus surjective, $f(U)$ is the complement of $f(U^c)$. Hence $f(U)$ is open. $\square$

16.2.3 Oscillation

If $f\colon X \to \mathbb{R}$ is bounded, the extent to which f fails to be continuous at $p \in X$ can be measured in a precise way. For $\delta > 0$ let

$$M(p, f, \delta) = \sup\{f(x) \mid x \in X, |x - p| < \delta\},$$
$$m(p, f, \delta) = \inf\{f(x) \mid x \in X, |x - p| < \delta\}.$$

Definition 16.18 (Oscillation)**.** The ***oscillation*** $o(f, p)$ of f at p is defined by

$$o(f, p) := \lim_{\delta \to 0} [M(p, f, \delta) - m(p, f, \delta)].$$

This limit always exists, since $M(p, f, \delta) - m(p, f, \delta)$ decreases as δ decreases.

There are two important facts about $o(f, p)$.

Lemma 16.19. *The bounded function f is continuous at p if and only if $o(f, p) = 0$.*

Proof.

$\boxed{\implies}$ Suppose f is continuous at p.

Let $\varepsilon > 0$ be given, there exists $\delta > 0$ such that for all $x \in X$,

$$|x - p| < \delta \implies |f(x) - f(p)| < \varepsilon.$$

Thus

$$M(p, f, \delta) - m(p, f, \delta) \le 2\varepsilon.$$

Since this is true for every ε, we have $o(f, p) = 0$.

$\boxed{\impliedby}$ ☐

Lemma 16.20. *Let $E \subset \mathbb{R}^n$ be closed. If $f\colon E \to \mathbb{R}$ is bounded, then*

$$\{x \in E \mid o(f, x) \ge \varepsilon\}$$

is closed for any $\varepsilon > 0$.

Proof. Let $F = \{x \in E \mid o(f, x) \ge \varepsilon\}$. To show that E is closed, we want to show that its complement F^c is open.

Let $x \in F^c$. Then either $x \notin E$, or $x \in E$ and $o(f, x) < \varepsilon$.

Case 1: $x \notin E$ Since E is closed, E^c is open, so there exists an open ball B containing x such that $B \subset E^c \subset F^c$.

Case 2: $x \in E$ **and** $o(f, x) < \varepsilon$ Then there exists $\delta > 0$ such that $M(x, f, \delta) - m(x, f, \delta) < \varepsilon$.

Consider the open ball $B_\delta(x)$. Then if $y \in B_\delta(x)$, $|x - y| < \delta$. There exists δ_1 such that $|x - z| < \delta$ for all z satisfying $|z - y| < \delta_1$.

Thus $M(y, f, \delta_1) - m(y, f, \delta_1) < \varepsilon$, and consequently $o(y, f) < \varepsilon$. Hence $B_\delta(x) \subset F^c$.

$\square$

16.2.4 Bolzano's Theorem

Lemma 16.21 (Sign-preserving property). *Let $f\colon [a,b] \to \mathbb{R}$ be continuous at $c \in [a,b]$, $f(c) \neq 0$. Then there exists $\delta > 0$ such that $f(x)$ has the same sign as $f(c)$ for $c-\delta < x < c+\delta$.*

Proof. Assume $f(c) > 0$. Let $\varepsilon > 0$ be given. By continuity of f, there exists $\delta > 0$ such that

$$c - \delta < x < c + \delta \implies f(c) - \varepsilon < f(x) < f(c) + \varepsilon.$$

Take the δ corresponding to $\varepsilon = \frac{f(c)}{2}$. Then

$$\frac{1}{2}f(c) < f(x) < \frac{3}{2}f(c) \quad (c - \delta < x < c + \delta)$$

so $f(x)$ has the same sign as $f(c)$ for $c - \delta < x < c + \delta$.

The proof is similar if $f(c) < 0$, except that we take $\varepsilon = -\frac{1}{2}f(c)$. □

The next result states that if the graph of $f\colon [a,b] \to \mathbb{R}$ lies above the x-axis at a and below the x-axis at b, then the graph must cross the axis somewhere in between. (This should be intuitively obvious.)

Theorem 16.22 (Bolzano). *Suppose $f\colon [a,b] \to \mathbb{R}$ is continuous, and $f(a)f(b) < 0$ (that is, $f(a)$ and $f(b)$ have opposite signs). Then there exists $c \in (a,b)$ such that $f(c) = 0$.*

Proof. For definiteness, assume $f(a) > 0$ and $f(b) < 0$. Let

$$A = \{x \in [a,b] \mid f(x) \geq 0\}.$$

Then A is non-empty since $a \in A$, and A is bounded above by b, so A has a supremum in $\mathbb{R}$; let $c = \sup A$. Then $a < c < b$.

Claim. $f(c) = 0$.

If $f(c) \neq 0$, by the previous result, there exists $\delta > 0$ such that $f(x)$ has the same sign as $f(c)$ for $c - \delta < x < c + \delta$.

- If $f(c) > 0$, there are points $x > c$ at which $f(x) > 0$, contradicting the definition of c.
- If $f(c) < 0$, then $c - \frac{\delta}{2}$ is an upper bound for A, again contradicting the definition of c.

Therefore we must have $f(c) = 0$. □

16.2.5 Continuity and Connectedness

Proposition 16.23. *Suppose $f\colon X \to Y$ is continuous. If $E \subset X$ is connected, then $f(E)$ is connected.*

Proof. We prove the contrapositive. Suppose $f(E)$ is not connected, then $f(E) = A \cup B$ for some $A, B \subset Y$ where $\overline{A} \cap B = \overline{B} \cap A = \emptyset$.

Consider $\overline{A}$ and $\overline{B}$, which are closed in Y. Since f is continuous, by 16.13, $f^{-1}(\overline{A})$ and $f^{-1}(\overline{B})$ are closed in X; let $K_A = f^{-1}(\overline{A})$, $K_B = f^{-1}(\overline{B})$. We now want to construct a separation of E.

Let $E_1 = f^{-1}(A) \cap E$, $E_2 = f^{-1}(B) \cap E$. Since $A \cap B = \emptyset$, we have that $E_1 \cap E_2 = \emptyset$. Since $A, B \neq \emptyset$, we have that $E_1, E_2 \neq \emptyset$.

Claim. E_1 and E_2 is a separation of E.

Notice $E_1 \subset K_A$ (which is closed) and $E_2 \subset K_B$ (which is closed). Then $\overline{E_1} \subset K_A$ and $\overline{E_2} \subset K_B$. Note that

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

so $K_A \cap E_2 = \emptyset$. Similarly $K_B \cap E_1 = \emptyset$.

Therefore E is separated. □

The next result says that a continuous real-valued function assumes all intermediate values on an interval.

Theorem 16.24 (Intermediate value theorem). *Suppose $f\colon [a, b] \to \mathbb{R}$ is continuous. If $f(a) < f(b)$ and $f(a) < c < f(b)$, then there exists $x \in (a, b)$ such that $f(x) = c$.*

Proof. By 14.62, $[a, b]$ is connected. By the previous result, $f([a, b])$ is a connected subset of $\mathbb{R}$. Then apply 14.63 and we are done. □

Remark. The converse is not necessarily true. For instance, the *topologist's sine curve*

$$f(x) = \begin{cases} 0 & (x = 0) \\ \sin\left(\frac{1}{x}\right) & (x \neq 0) \end{cases}$$

satisfies the intermediate value property, but f is not continuous.

Example 16.25. $x^5 - 2x^3 + 3x^2 - 1 = 0$ has a solution in the interval $[0, 1]$.

Proof. Let $f(x) = x^5 - 2x^3 + 3x^2 - 1$. f is continuous on $[0, 1]$. In addition,

$$f(0) = -1 < 0, \quad 0 < 1 = f(1).$$

By the intermediate value theorem, there exists x, with $0 < x < 1$, for which $x^5 - 2x^3 + 3x^2 - 1 = 0$. □

16.3 Uniform Continuity

Definition 16.26 (Uniform continuity)**.** We say $f\colon X \to Y$ is ***uniformly continuous*** on X if

$$\forall \varepsilon > 0, \quad \exists \delta > 0, \quad \forall p, q \in X, \quad d_X(p,q) < \delta \implies d_Y\left(f(p), f(q)\right) < \varepsilon.$$

Remark. The difference between continuity and uniform continuity is that of one between a local and global property.

- Continuity can be defined at a single point, as δ depends on ε as well as the point p.
- Uniform continuity is a property of a function on a set, as the same δ has to work for *all* $p \in X$ (which ensures a *uniform* rate of closeness across the entire domain.).

Hence uniform continuity is a stronger continuity condition than continuity; a function that is uniformly continuous is continuous but a function that is continuous is not necessarily uniformly continuous.

Example 16.27.

- Let $f(x) = \frac{1}{x}$. Then f is continuous on $(0,1]$ but not uniformly continuous on $(0,1]$. To prove this, let $\varepsilon = 10$, and suppose we could find a δ ($0 \le \delta < 1$) that satisfies the condition of the definition. Taking $p = \delta$, $q = \frac{\delta}{11}$, we obtain $|p-q| < \delta$ and

  $$|f(p) - f(q)| = \frac{11}{\delta} - \frac{1}{\delta} = \frac{10}{\delta} > 10.$$

  Hence, for these two points we would always have $|f(p) - f(q)| > 10$, contradicting the definition of uniform continuity.

- Let $f(x) = x^2$. Then f is uniformly continuous on $(0,1]$. To prove this, observe that

  $$|f(p) - f(q)| = |p^2 - q^2| = |(p+q)(p-q)| < 2|p-q|.$$

  If $|p-q| < \delta$, then $|f(p) - f(q)| < 2\delta$. Hence, for any given ε, we need only take $\delta = \frac{\varepsilon}{2}$ to guarantee that $|f(p) - f(q)| < \varepsilon$ for every $p, q \in (0,1]$ with $|p-q| < \delta$.

The next result concerns the relationship between continuity and uniform continuity.

Lemma 16.28.

(i) If $f\colon X \to Y$ is uniformly continuous on X, then f is continuous on X.

(ii) (Heine–Cantor theorem) If $f\colon X \to Y$ is continuous on X, and X is compact, then f is uniformly continuous on X.

Proof.

(i)

(ii) Let $\varepsilon > 0$ be given. Since f is continuous on X, for each $p \in X$, we can associate some $\phi(p) > 0$ such that for all $q \in X$,

$$d_X(p,q) < \phi(p) \implies d_Y(f(p), f(q)) < \frac{\varepsilon}{2}.$$

Consider the collection of open balls centred at each $p \in X$:

$$\left\{ B_{\frac{1}{2}\phi(p)}(p) \mid p \in X \right\}.$$

Since $p \in B_{\frac{1}{2}\phi(p)}(p)$, the above collection of open balls forms an open cover of X. Since X is compact, there exists finitely many points $p_1, \dots, p_n \in X$ such that

$$X \subset \bigcup_{k=1}^{n} B_{\frac{1}{2}\phi(p_k)}(p_k).$$

Let

$$\delta = \min \left\{ \frac{1}{2}\phi(p_1), \dots, \frac{1}{2}\phi(p_n) \right\}.$$

We claim that this value of δ works in the definition of uniform continuity. Note that $\delta > 0$. (This is one point where the finiteness of the covering, inherent in the definition of compactness, is essential. The minimum of a finite set of positive numbers is positive, whereas the inf of an infinite set of positive numbers may very well be 0.)

Let $p, q \in X$ such that $d_X(p,q) < \delta$. SInce X is covered by finitely many open balls, $p \in B_{\frac{1}{2}\phi(p_m)}(p_m)$ for some m ($1 \le m \le n$); thus

$$d_X(p, p_m) < \frac{1}{2}\phi(p_m).$$

We also have

$$\begin{aligned} d_X(q, p_m) &\le d_X(p,q) + d_X(p, p_m) \\ &< \delta + \frac{1}{2}\phi(p_m) \\ &\le \phi(p_m). \end{aligned}$$

Finally, invoking the continuity of f,

$$\begin{aligned} d_Y(f(p), f(q)) &\le d_Y(f(p), f(p_m)) + d_Y(f(q), f(p_m)) \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

□

Lemma 16.29 (Lebesgue covering lemma)**.** *Suppose $\{U_i \mid i \in I\}$ is an open cover of a compact metric space X. Then there exists $\delta > 0$ such that for all $x \in X$,*

$$B_\delta(x) \subset U_i$$

for some $i \in I$; δ *is called a* Lebesgue number *of the cover.*

Proof. Since X is compact, there exist finitely many indices $i_1, \dots, i_n$ such that

$$X \subset \bigcup_{k=1}^{n} U_{i_k}.$$

For any closed set A, define the distance

$$d(x, A) = \inf_{a \in A} d(x, a).$$

Claim. $d(x, A)$ is a continuous function of x.

Then let the average distance from each x to the complements of U_{i_k} be the function

$$f(x) = \frac{1}{n} \sum_{k=1}^{n} d(x, {U_{i_k}}^c).$$

Since f is a sum of continuous functions, f is continuous. Since f is continuous on a compact set, f attains its minimum value; call it δ. See that $\delta > 0$ since $\{U_{i_1}, \dots, U_{i_n}\}$ is an open cover (so $x \in U_{i_k}$ implies $d(x, {U_{i_k}}^c) > 0$).

For each x, $f(x) \geq \delta$ implies that at least one of the distances $d(x, {U_{i_k}}^c) \geq \delta$. Hence $B_\delta(x) \subset U_{i_k}$, as desired. □

16.4 Discontinuities

We now focus our attention on real-valued functions defined on intervals of the real line.

Definition 16.30 (One-sided limits)**.** Let $f\colon (a,b) \to \mathbb{R}$. Let $x \in [a,b)$. The ***right-hand limit***, denoted by $f(x+)$ or $\lim_{t\to x^+} f(t)$, exists if

$$\forall \varepsilon > 0, \quad \exists \delta > 0, \quad x < t < x+\delta < b \implies |f(t) - f(x+)| < \varepsilon.$$

If f is defined at x and if $f(x+) = f(x)$, we say that f is *continuous from the right* at x.

Similarly, let $x \in (a,b]$. The ***left-hand limit***, denoted by $f(x-)$ or $\lim_{t\to x^-} f(t)$, exists if

$$\forall \varepsilon > 0, \quad \exists \delta > 0, \quad a < x-\delta < t < x \implies |f(t) - f(x-)| < \varepsilon.$$

If f is defined at x and if $f(x+) = f(x)$, we say that f is *continuous from the left* at x.

Remark. Compare the above definition with Definition 16.1; for one-sided limits, we are only concerned with half open balls around t (since we only require x to approach t from either the right or left side).

Remark. An equivalent formulation using limits of sequences is presented in [Rud76].

Lemma 16.31. *If $a < x < b$, then f is continuous at c if and only if*

$$f(x) = f(x+) = f(x-).$$

If f is not continuous at x, we say that f is *discontinuous* at x, or that f has a *discontinuity* at x.

Example 16.32 (Dirichlet function)**.** The *Dirichlet function*, defined by

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

is discontinuous everywhere; that is, f is not continuous at any point in $\mathbb{R}$.

Proof. We consider two cases.

- If $x \in \mathbb{Q}$, then $f(x) = 1$. Take $\varepsilon = \frac{1}{2}$. Since the irrational numbers are dense in the reals, for any $\delta > 0$, we can always find an irrational $y \in \mathbb{R} \setminus \mathbb{Q}$ such that

$$|x - y| < \delta \quad \text{and} \quad |f(x) - f(y)| = 1 \geq \frac{1}{2}.$$

- If $x \in \mathbb{R} \setminus \mathbb{Q}$, then $f(x) = 0$. Again take $\varepsilon = \frac{1}{2}$. Since $\mathbb{Q}$ is dense in $\mathbb{R}$, for any $\delta > 0$, we can always find $y \in \mathbb{Q}$ such that

$$|x - y| < \delta \quad \text{and} \quad |f(x) - f(y)| = 1 \geq \frac{1}{2}.$$

$\square$

If f is defined on an interval, it is customary to divide discontinuities into two types.

Definition 16.33 (Discontinuities)**.** Let $f\colon (a,b) \to \mathbb{R}$. Suppose f is discontinuous at $x \in (a,b)$.

(i) We say f has a *discontinuity of the first kind* (or a *simple discontinuity*) at x, if $f(x+)$ and $f(x-)$ exist;

(ii) we say f has a *discontinuity of the second kind* if otherwise.

There are two ways in which a function can have a simple discontinuity: either $f(x+) \neq f(x)$ [in which case the value $f(x)$ is immaterial], or $f(x+) = f(x-) \neq f(x)$.

Example 16.34.

- The function
$$f(x) = \begin{cases} x+2 & (-3 < x < -2) \\ -x-2 & (-2 \leq x < 0) \\ x+2 & (0 \leq x < 1) \end{cases}$$
has a simple discontinuity at $x = 0$, and is continuous at every other point of $(-3,1)$.

- The Dirichlet function has a discontinuity of the second kind at every $x \in \mathbb{R}$, since both $f(x+)$ and $f(x-)$ do not exist.

- The topologist's sine curve has a discontinuity of the second kind at $x = 0$, since $f(x+)$ does not exist.

16.5 Monotonic Functions

We now study those functions which never decrease (or never increase) on a given interval.

Definition 16.35 (Monotonicity). We say $f\colon (a,b) \to \mathbb{R}$ is

(i) *monotonically increasing*, if $f(x_1) \leq f(x_2)$ for any $a < x_1 \leq x_2 < b$;

(ii) *monotonically decreasing*, if $f(x_1) \geq f(x_2)$ for any any $a < x_1 \leq x_2 < b$;

(iii) ***monotonic*** if it is either monotonically increasing or monotonically decreasing.

Proposition 16.36. *Let $f\colon (a,b) \to \mathbb{R}$ be monotonically increasing. Then $f(x+)$ and $f(x-)$ exist for all $x \in (a,b)$; more precisely,*

$$\sup_{t\in(a,x)} f(t) = f(x-) \leq f(x) \leq f(x+) = \inf_{t\in(x,b)} f(t).$$

Furthermore, if $a < x < y < b$, then

$$f(x+) \leq f(y-).$$

Analogous results evidently hold for monotonically decreasing functions.

Proof. We will prove the first half of the given statement; the second half can be proven in precisely the same way.

Let $x \in (a,b)$. Since f is monotonically increasing, the set

$$A = \{f(t) \mid a < t < x\}$$

is bounded above by the number $f(x)$. Hence A has a supremum in $\mathbb{R}$; let $\alpha = \sup A$. Evidently $\alpha \leq f(x)$.

Claim. $f(x-) = \alpha$.

To prove this, we need to show that for all $\varepsilon > 0$, there exists $\delta > 0$ such that

$$x - \delta < t < x \implies |f(t) - \alpha| < \varepsilon.$$

Let $\varepsilon > 0$ be given. Since $\alpha = \sup A$, there exists $\delta > 0$ such that $a < x - \delta < x$ and

$$\alpha - \varepsilon < f(x - \delta) \leq \alpha. \tag{1}$$

Since f is monotonic, we have

$$f(x - \delta) \leq f(t) \leq \alpha \quad (x - \delta < t < x) \tag{2}$$

Combining (1) and (2) gives

$$|f(t) - \alpha| < \varepsilon \quad (x - \delta < t < x)$$

as desired. Hence $f(x-) = \alpha$.

Next, if $a < x < y < b$, we see from the given statement that

$$f(x+) = \inf_{t \in (x,b)} f(t) = \inf_{t \in (x,y)} f(t)$$

where the last equality is obtained by applying the given statement to (a, y) in place of (a, b). Similarly,

$$f(y-) = \sup_{t \in (a,y)} f(t) = \sup_{t \in (x,y)} f(t).$$

Comparing these two equations, we conclude that $f(x+) \leq f(y-)$. □

Corollary 16.37. *Monotonic functions have no discontinuities of the second kind.*

Proposition 16.38. *Let $f\colon (a, b) \to \mathbb{R}$ be monotonic. Then the set of points of (a, b) at which f is discontinuous is at most countable.*

Proof. Suppose, for the sake of definiteness, that f is monotonically increasing. Let D be the set of points at which f is discontinuous.

For every $x \in D$, we associate a rational number $r(x)$, where

$$f(x-) < r(x) < f(x+).$$

We now check that the rationals picked for two distinct points of discontinuities are different: since $x_1 < x_2$ implies $f(x_1+) \leq f(x_2-)$ (from the previous result), we see that $r(x_1) \neq r(x_2)$ if $x_1 \neq x_2$.

We have thus established a 1-1 correspondence between D and a subset of $\mathbb{Q}$ (which we know is at most countable). Hence D is at most countable. □

16.6 Lipschitz Continuity

Definition 16.39 (Lipschitz continuity)**.** We say $f\colon X \to Y$ is ***Lipschitz continuous*** if there exists $K \geq 0$ such that

$$\forall x, y \in X, \quad d_Y\left(f(x), f(y)\right) \leq K d_X(x, y).$$

K is called a *Lipschitz constant* for f; we also refer to f as *K-Lipschitz*.

Lemma 16.40. *Lipschitz continuity implies uniform continuity.*

Proof. Let $f\colon X \to Y$ be K-Lipschitz continuous.

Let $\varepsilon > 0$ be given, let $x, y \in X$. We consider two cases.

Case 1: $K \leq 0$**.** Then

$$d_X(x, y) \leq 0 d_Y\left(f(x), f(y)\right)$$

so

$$d_X(x, y) \leq 0 \implies d_X(x, y) = 0 \implies x = y$$

for all $x, y \in X$. Hence f is a constant function, which is uniformly continuous.

Case 2: $K > 0$**.** Take $\delta = \frac{\varepsilon}{K}$. If $d_X(x, y) < \delta$, then

$$K d_X(x, y) < \varepsilon.$$

By Lipschitz continuity of f,

$$d_Y\left(f(x), f(y)\right) \leq K d_X(x, y).$$

These last two statements together imply $d_Y\left(f(x), f(y)\right) < \varepsilon$. Hence f is uniformly continuous on X.

$\square$

We say $f\colon X \to Y$ is a *contraction* if it is a K-Lipschitz map for some $K < 1$.

Let $f\colon X \to X$, we say $x \in X$ is a *fixed point* if $f(x) = x$.

Theorem 16.41 (Contraction mapping theorem)**.** *Let X be a complete metric space, and $f\colon X \to X$ be a contraction. Then f has a unique fixed point.*

Remark. The hypotheses "complete" and "contraction" are necessary. For example, $f\colon (0,1) \to (0,1)$ defined by $f(x) = Kx$ for any $0 < K < 1$ is a contraction with no fixed point. Also, $f\colon \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x + 1$ is not a contraction ($K = 1$) and has no fixed point.

Proof. Pick any $x_0 \in X$. Define a sequence (x_n) by $x_{n+1} = f(x_n)$. Since f is a contraction, we have

$$\begin{aligned} d(x_{n+1}, x_n) &= d\left(f(x_n), f(x_{n-1})\right) \\ &\leq K d(x_n, x_{n-1}) \\ &\leq \cdots \\ &\leq K^n d(x_1, x_0) \end{aligned}$$

by induction. Suppose $m \geq n$, then

$$\begin{aligned} d(x_m, x_n) &\leq \sum_{i=n}^{m-1} d(x_{i+1}, x_i) \\ &\leq \sum_{i=n}^{m-1} K^i d(x_1, x_0) \\ &= K^n d(x_1, x_0) \sum_{i=0}^{m-n-1} k^i \\ &\leq K^n d(x_1, x_0) \sum_{i=0}^{\infty} K^i = \frac{K^n}{1-K} d(x_1, x_0). \end{aligned}$$

Thus (x_n) is a Cauchy sequence. Since X is complete, (x_n) converges; let $\lim_{n\to\infty} x_n = x$ for some $x \in X$.

Claim. x is our unique fixed point.

Note that f is continuous because it is a contraction. Hence

$$f(x) = \lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_{n+1} = x,$$

so x is a fixed point.

Let x' also be a fixed point. Then

$$d(x, x') = d\left(f(x), f(x')\right) = K d(x, x').$$

As $K < 1$ this means that $d(x, x') = 0$ and hence $x = x'$. The theorem is proved. □

Note that the proof is constructive. Not only do we know that a unique fixed point exists. We also know how to find it.

16.7 Infinite Limits and Limits at Infinity

To enable us to operate in the extended real number system, we shall now enlarge the scope of Definition 16.1, reformulating it in terms of open balls.

For any real number x, we have already defined an open ball of x to be any open interval $(x-\delta, x+\delta)$.

Definition 16.42. Let $c \in \mathbb{R}$. A *neighbourhood of* $+\infty$ is

$$(c, +\infty) := \{x \in \mathbb{R} \mid x > c\}.$$

Similarly, the set $(-\infty, c)$ is a neighbourhood of $-\infty$.

Definition 16.43. Let $f\colon E \subset \mathbb{R} \to \mathbb{R}$. We say that $\lim_{t\to x} f(t) = A$ where A and x are in the extended real number system, if for every neighbourhood of U of A there is a neighbourhood V of x such that $V \cap E$ is not empty, and such that $f(t) \in U$ for all $t \in V \cap E$, $t \neq x$.

Remark. When A and x are real, Definition 16.43 coincides with Definition 16.1.

to do

Lemma 16.44 (Uniqueness of limit)**.** *Let $f\colon E \subset \mathbb{R} \to \mathbb{R}$. The limit of f at a point x is unique.*

Proof. Suppose

$$\lim_{t\to x} f(t) = A, \quad \lim_{t\to x} f(t) = A'.$$

We will show that $A' = A$. □

The analogue of Theorem 4.4 is still true, and the proof offers nothing new. We state it, for the sake of completeness.

Lemma 16.45. *Let $f, g\colon E \subset \mathbb{R} \to \mathbb{R}$. Suppose $\lim_{t\to x} f(t) = A$, $\lim_{t\to x} g(t) = B$. Then*

(i) $\lim_{t\to x}(f+g)(t) = A + B$

(ii) $\lim_{t\to x}(fg)(t) = AB$

(iii) $\lim_{t\to x}(f/g)(t) = A/B$

provided the RHS are defined.

Note that $\infty - \infty$, $0 \cdot \infty$, ∞/∞, $A/0$ are not defined (see Definition 1.23).

Exercises

Exercise 16.1 ([Rud76] 4.1)**.** Suppose $f\colon \mathbb{R} \to \mathbb{R}$ satisfies

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

for every $x \in \mathbb{R}$. Does this imply that f is continuous?

Exercise 16.2 ([Rud76] 4.2)**.** If $f\colon X \to Y$ is continuous, prove that

$$f(\overline{E}) \subset \overline{f(E)}$$

for every $E \subset X$.

Exercise 16.3 ([Rud76] 4.3)**.** Let $f\colon X \to \mathbb{R}$ be continuous. Let the *zero set* of f be

$$Z(f) = \{x \in X \mid f(x) = 0\}.$$

Prove that $Z(f)$ is closed.

Exercise 16.4 ([Rud76] 4.8)**.** Let f be a real uniformly continuous function on the bounded set $E \subset \mathbb{R}$. Prove that f is bounded on E.

Show that the conclusion is false if boundedness of E is omitted from the hypothesis.

Exercise 16.5 ([Rud76] 4.11)**.** Suppose $f\colon X \to Y$ is uniformly continuous on X. Prove that $(f(x_n))$ is a Cauchy sequence in Y for every Cauchy sequence (x_n) in X.

Exercise 16.6 ([Rud76] 4.12)**.** A uniformly continuous function of a uniformly continuous function is uniformly continuous.

Exercise 16.7 ([Rud76] 4.14)**.** Let $I = [0,1]$ be the closed unit interval. Suppose f is a continuous mapping of I into I. Prove that $f(x) = x$ for at least one $x \in I$.

Exercise 16.8 ([Rud76] 4.15)**.** $f\colon X \to Y$ is said to be *open* if $f(V)$ is an open set in Y whenever V is an open set in X.

Prove that every continuous open mapping of $\mathbb{R}$ into $\mathbb{R}$ is monotonic.

Exercise 16.9 ([Rud76] 4.16)**.** Let $[x]$ denote the largest integer contained in x, and let $\{x\} = x - [x]$ denote the fractional part of x. What discontinuities do the functions $[x]$ and $\{x\}$ have?

Exercise 16.10 ([Rud76] 4.18)**.** Every rational x can be written in the form $x = \frac{m}{n}$, where $m \in \mathbb{Z}$, $n \in \mathbb{N}$, $\gcd(m,n) = 1$. When $x = 0$, we take $n = 1$. Consider the function f defined on $\mathbb{R}$ by

$$f(x) = \begin{cases} 0 & (x \in \mathbb{R} \setminus \mathbb{Q}) \\ \frac{1}{n} & (x = \frac{m}{n}) \end{cases}$$

Prove that f is continuous at every irrational point, and that f has a simple discontinuity at every rational point.

Exercise 16.11 ([Rud76] 4.26)**.** Suppose X, Y, Z are metric spaces, and Y is compact. Let $f\colon X \to Y$, $g\colon Y \to Z$ be continuous and injective, and $h = g \circ f$.

Prove that f is uniformly continuous if h is uniformly continuous. *Hint*: g^{-1} has compact domain $g(Y)$, and $f(x) = g^{-1}(h(x))$.

Prove also that f is continuous if h is continuous.

Exercise 16.12. Show that $f\colon [0, +\infty) \to [0, +\infty)$, $f(x) = \sqrt{x}$ is uniformly continuous.

Chapter 17

Differentiation

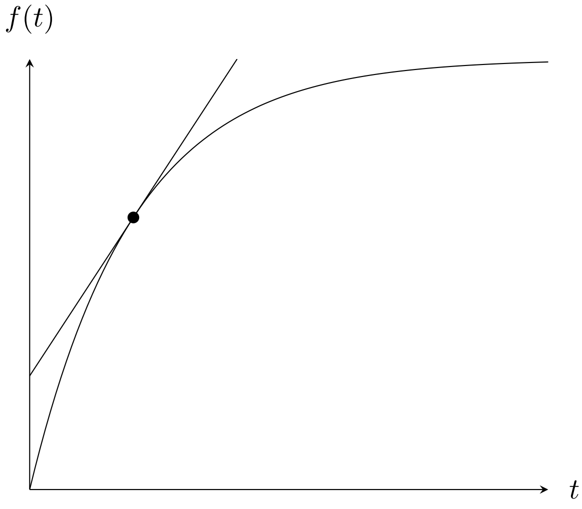
17.1 The Derivative of A Real Function

17.1.1 Definitions and Properties

The derivative of a function $f : [a, b] \to \mathbb{R}$ at a point x can be intuitively thought of as the gradient of the tangent line at x. Consider the gradient of the secant line:

$$\phi(t) = \frac{f(t) - f(x)}{t - x}.$$

Taking the limit $t \to x$, the secant line approaches the tangent line at x, and so the value of $\phi(t)$ approaches the gradient of the tangent line.



Definition 17.1 (Derivative)**.** Suppose $f\colon [a, b] \to \mathbb{R}$. For any $x \in [a, b]$, if the limit

$$\lim_{t\to x} \frac{f(t) - f(x)}{t - x} \quad (a < t < b, t \neq x) \tag{17.1}$$

exists, we call it the ***derivative*** of f, and denote it by f'.
If f' is defined at x, we say that f is ***differentiable*** at x. If f' is defined at every point of $E \subset$

$[a, b]$, we say that f is *differentiable on* E.
We say that f is *continuously differentiable* on E if f' exists at every point of E, and f' is continuous on E.

Equivalently, (17.1) can be written as

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

or,

$$\frac{f(x+h)-f(x)}{h} = f'(x) + \varepsilon(h),$$

where $\varepsilon(h) \to 0$ as $h \to 0$. Rearranging gives

$$f(x+h) = f(x) + hf'(x) + h\varepsilon(h).$$

Using the small-o notation, we write $o(h)$ for a function that satisfies $o(h)/h \to 0$ as $h \to 0$. Hence we have

$$f(x+h) = f(x) + hf'(x) + o(h). \tag{17.2}$$

We can interpret (17.2) as an *approximation* of $f(x+h)$:

$$f(x+h) = \underbrace{f(x) + hf'(x)}_{\text{linear approximation}} + \underbrace{o(h)}_{\text{error term}} .$$

Lemma 17.2 (Differentiability implies continuity). *If $f\colon [a, b] \to \mathbb{R}$ is differentiable at $x \in [a, b]$, then f is continuous at x.*

Proof. Suppose $f\colon [a, b] \to \mathbb{R}$ is differentiable at $x \in [a, b]$. Then the limit $\lim_{t\to x} \frac{f(t)-f(x)}{t-x}$ exists. Thus by arithmetic properties of limits,

$$\begin{aligned}
\lim_{t\to x}[f(t)-f(x)] &= \lim_{t\to x}\left[\frac{f(t)-f(x)}{t-x}\cdot(t-x)\right] \\
&= \lim_{t\to x}\frac{f(t)-f(x)}{t-x}\cdot\lim_{t\to x}(t-x) \\
&= f'(x)\cdot 0 = 0.
\end{aligned}$$

Since $\lim_{t\to x} f(t) = f(x)$, by 16.6, f is continuous at x. $\square$

Remark. The converse is not true; it is easy to construct continuous functions which fail to be differentiable at isolated points.

Example 17.3 (Weierstrass function). Let $0 < a < 1$, let $b > 1$ be an odd integer, and $ab > 1 + \frac{3}{2}\pi$. Then the function

$$W(x) = \sum_{n=0}^{\infty} a^n \cos(b^n \pi x)$$

is continuous and nowhere differentiable on $\mathbb{R}$.

Example 17.4. One family of pathological examples in calculus is functions of the form

$$f(x) = x^p \sin \frac{1}{x}.$$

For $p = 1$, the function is continuous and differentiable everywhere other than $x = 0$; for $p = 2$, the function is differentiable everywhere, but the derivative is discontinuous.

Lemma 17.5 (Differentiation rules)**.** *Suppose $f, g : [a, b] \to \mathbb{R}$ are differentiable at $x \in [a, b]$. Then*

(i) For a constant α, αf is differentiable at x, and *(scalar multiplication)*

$$(\alpha f)'(x) = \alpha f'(x).$$

(ii) $f + g$ is differentiable at x, and *(addition)*

$$(f + g)'(x) = f'(x) + g'(x).$$

(iii) fg is differentiable at x, and *(product rule)*

$$(fg)'(x) = f'(x)g(x) + f(x)g'(x).$$

(iv) f/g (when $g(x) \neq 0$) is differentiable at x, and *(quotient rule)*

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}.$$

Proof.

(i)
$$(\alpha f)'(x) = \lim_{t \to x} \frac{(\alpha f)(t) - (\alpha f)(x)}{t - x} = \alpha \lim_{t \to x} \frac{f(t) - f(x)}{t - x} = \alpha f'(x).$$

(ii)
$$\begin{aligned}
(f \pm g)'(x) &= \lim_{t \to x} \frac{(f + g)(t) - (f + g)(x)}{t - x} \\
&= \lim_{t \to x} \frac{f(t) + g(t) - f(x) - g(x)}{t - x} \\
&= \lim_{t \to x} \frac{f(t) - f(x)}{t - x} + \lim_{t \to x} \frac{g(t) - g(x)}{t - x} \\
&= f'(x) + g'(x)
\end{aligned}$$

(iii)

$$\begin{aligned}(fg)'(x) &= \lim_{t\to x} \frac{(fg)(t)-(fg)(x)}{t-x} \\ &= \lim_{t\to x} \frac{f(t)g(t)-f(x)g(x)}{t-x} \\ &= \lim_{t\to x} \frac{[f(t)-f(x)]\,g(t)+f(x)\,[g(t)-g(x)]}{t-x} \\ &= \lim_{t\to x} \frac{f(t)-f(x)}{t-x}\cdot g(t) + \lim_{t\to x} f(x)\cdot \frac{g(t)-g(x)}{t-x} \\ &= f'(x)g(x)+f(x)g'(x)\end{aligned}$$

(iv)

$$\begin{aligned}\left(\frac{f}{g}\right)'(x) &= \lim_{t\to x} \frac{\left(\frac{f}{g}\right)(t)-\left(\frac{f}{g}\right)(x)}{t-x} \\ &= \lim_{t\to x} \frac{\frac{f(t)}{g(t)}-\frac{f(x)}{g(x)}}{t-x} \\ &= \lim_{t\to x} \frac{1}{g(t)g(x)}\left[g(x)\cdot\frac{f(t)-f(x)}{t-x} - f(x)\cdot\frac{g(t)-g(x)}{t-x}\right] \\ &= \frac{f'(x)g(x)-f(x)g'(x)}{g(x)^2}\end{aligned}$$

□

By induction, we can obtain the following extensions of the differentiation rules.

Corollary. *Suppose $f_1, f_2, \dots, f_n : [a,b] \to \mathbb{R}$ are differentiable at $x \in [a,b]$. Then*

(i) $f_1+f_2+\cdots+f_n$ is differentiable at x, and

$$(f_1+f_2+\cdots+f_n)'(x) = {f_1}'(x)+{f_2}'(x)+\cdots+{f_n}'(x).$$

(ii) $f_1f_2\cdots f_n$ is differentiable at x, and

$$\begin{aligned}(f_1f_2\cdots f_n)'(x) =& {f_1}'(x)f_2(x)\cdots f_n(x)+f_1(x){f_2}'(x)\cdots f_n(x) \\ &+\cdots+f_1(x)f_2(x)\cdots {f_n}'(x).\end{aligned}$$

The next result concerns the derivative of composition of functions.

Lemma 17.6 (Chain rule). *Suppose f is continuous on $[a,b]$, $f'(x)$ exists at $x \in [a,b]$, g is defined on I that contains $f([a,b])$, and g is differentiable at $f(x)$. Then $h = g\circ f$ is differentiable at x, and*

$$h'(x) = g'\left(f(x)\right)f'(x). \tag{17.3}$$

Proof. By the definition of the derivative, we have

$$f(t)-f(x) = (t-x)[f'(x)+u(t)] \tag{1}$$

$$g(s) - g(f(x)) = (s - f(x))[g'(f(x)) + v(s)] \tag{2}$$

where $t \in [a, b]$, $s \in I$, $\lim_{t \to x} u(t) = 0$, $\lim_{s \to f(x)} v(s) = 0$. ($u(t)$ and $v(s)$ can be viewed as some small error terms which eventually go to 0.) Using first (2) and then (1), we obtain

$$\begin{aligned} h(t) - h(x) &= g\left(f(t)\right) - g\left(f(x)\right) \\ &= [f(t) - f(x)] \cdot [g'(f(x)) + v(s)] \\ &= (t - x)[f'(x) + u(t)][g'(f(x)) + v(s)], \end{aligned}$$

or, if $t \neq x$,

$$\frac{h(t) - h(x)}{t - x} = [g'(f(x)) + v(s)][f'(x) + u(t)].$$

Taking limits $t \to x$, we see that $u(t)$ and $v(s)$ eventually go to 0, so

$$h'(x) = \lim_{t \to x} \frac{h(t) - h(x)}{t - x} = g'\left(f(x)\right) f'(x)$$

as desired. □

Later on when we talk about properties of differentiation such as the intermediate value theorems, we usually have the following requirement on the function:

> f is continuous on $[a, b]$, differentiable on (a, b).

17.1.2 Derivatives of Higher Order

If f has a derivative f' on an interval, and if f' is itself differentiable, we denote the derivative of f' by f'', and call f'' the *second derivative* of f. Continuing in this manner, we obtain functions

$$f, f', f'', f^{(3)}, f^{(4)}, \ldots, f^{(n)},$$

each of which is the derivative of the preceding one. $f^{(n)}$ is called the n-th derivative (or the derivative or order n) of f.

Notation. $\mathcal{C}_1[a, b]$ denotes the set of differentiable functions over $[a, b]$ whose derivative is continuous. More generally, $\mathcal{C}_n[a, b]$ denotes the set of functions whose n-th derivative is continuous. In particular, $\mathcal{C}_0[a, b]$ is the set of continuous functions over $[a, b]$.

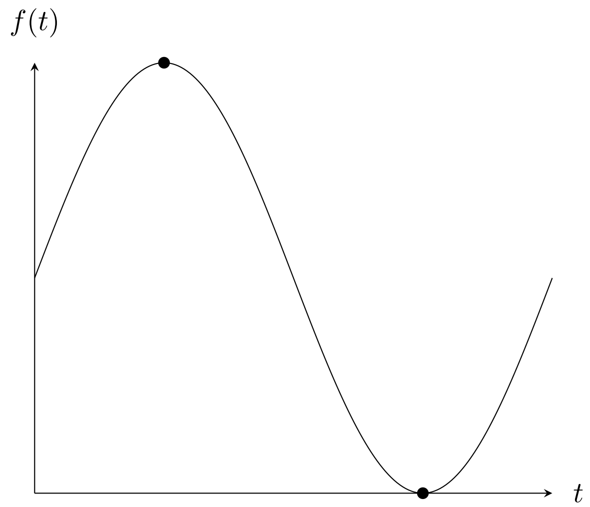
17.2 Mean Value Theorems

Let (X, d) be a metric space.

Definition 17.7 (Local maximum and minimum)**.** We say $f\colon X \to \mathbb{R}$ has

(i) a ***local maximum*** at $x \in X$ if there exists $\delta > 0$ such that $f(x) \geq f(t)$ for all $t \in B_\delta(x)$;

(ii) a ***local minimum*** at $x \in X$ if there exists $\delta > 0$ such that $f(x) \leq f(t)$ for all $t \in B_\delta(x)$.



Our next result is the basis of many applications of differentiation.

Lemma 17.8 (Fermat's theorem)**.** *Suppose $f\colon [a, b] \to \mathbb{R}$. If f has a local maximum or minimum at $x \in (a, b)$, and if $f'(x)$ exists, then*

$$f'(x) = 0.$$

Proof. We prove the case for local maxima; the proof for the case for local minima is similar.

Since x is a local maximum, choose $\delta > 0$ such that

$$a < x - \delta < x < x + \delta < b,$$

and $f(x) \geq f(t)$ for all $t \in (x - \delta, x + \delta)$. Then

$$\frac{f(t) - f(x)}{t - x} \begin{cases} \geq 0 & (x - \delta < t < x) \\ \leq 0 & (x < t < x + \delta) \end{cases}$$

Letting $t \to x$, we obtain

$$f'(x) \begin{cases} \geq 0 & (x - \delta < t < x) \\ \leq 0 & (x < t < x + \delta) \end{cases}$$

Hence we must have $f'(x) = 0$. $\square$

Theorem 17.9 (Rolle's theorem)**.** *Suppose f is continuous on $[a,b]$ and differentiable in (a,b). If $f(a)=f(b)$, then there exists $c\in(a,b)$ such that*

$$f'(c)=0.$$

The idea is to show that f has a local maximum/minimum, then by Fermat's theorem this will then be the stationary point that we're trying to find.

Proof. Since f is continuous on $[a,b]$, by the extreme value theorem (16.16), f attains its maximum M and minimum m.

- If M and m both equal $f(a)=f(b)$, then f is simply a constant function; hence $f'(x)=0$ for all $x\in[a,b]$.
- Otherwise, f has a maximum/minimum that does not equal $f(a)=f(b)$. Then there exists $c\in(a,b)$ such that $f(c)$ is a local maximum/minimum. Since f is differentiable on (a,b), $f'(c)$ exists, so by Fermat's theorem, $f'(c)=0$.

□

Theorem 17.10 (Generalised mean value theorem)**.** *Suppose f and g are continuous on $[a,b]$ and differentiable in (a,b). Then there exists $c\in(a,b)$ such that*

$$[f(b)-f(a)]g'(c)=[g(b)-g(a)]f'(c). \tag{17.4}$$

Proof. For $t\in[a,b]$, consider the *auxilliary function*

$$h(t)=[f(b)-f(a)]g(t)-[g(b)-g(a)]f(t).$$

Then h is continuous on $[a,b]$, and h is differentiable on (a,b). Moreover,

$$h(a)=f(b)g(a)-f(a)g(b)=h(b).$$

By Rolle's theorem, there exists $c\in(a,b)$ such that $h'(c)=0$; that is,

$$[f(b)-f(a)]g'(c)=[g(b)-g(a)]f'(c)$$

as desired. □

Theorem 17.11 (Mean value theorem)**.** *Suppose f is continuous on $[a,b]$ and differentiable in (a,b). Then there exists $c\in(a,b)$ such that*

$$f(b)-f(a)=f'(c)(b-a). \tag{17.5}$$

Proof. Take $g(x)=x$ in 17.10. □

Lemma 17.12. *Suppose f is differentiable in (a, b).*

(i) If $f'(x) \geq 0$ for all $x \in (a, b)$, then f is monotonically increasing.

(ii) If $f'(x) = 0$ for all $x \in (a, b)$, then f is constant.

(iii) If $f'(x) \leq 0$ for all $x \in (a, b)$, then f is monotonically decreasing.

Proof. All conclusions can be read off from the equation

$$f'(x) = \frac{f(x_2) - f(x_1)}{x_2 - x_1},$$

which is valid, for each pair of numbers x_1, x_2 in (a, b), for some x between x_1 and x_2. $\square$

Theorem 17.13 (Generalised Rolle's theorem)**.** *Suppose $f \in \mathcal{C}[a, b]$ is n times differentiable on (a, b). If $f(x) = 0$ at the $n + 1$ distinct numbers $a \leq x_0 < x_1 < \cdots < x_n \leq b$, then there exists $c \in (x_0, x_n)$, and hence in (a, b), such that*

$$f^{(n)}(c) = 0.$$

17.3 Continuity of Derivatives

The following result implies some sort of a "intermediate value" property of derivatives that is similar to continuous functions.

> **Theorem 17.14** (Darboux's theorem)**.** *Suppose f is differentiable on $[a,b]$, and suppose $f'(a) < c < f'(b)$. Then there exists $x \in (a,b)$ such that $f'(x) = c$.*

Proof. For $t \in (a,b)$, consider the auxilliary function

$$g(t) = f(t) - ct.$$

Then

$$g'(a) = f'(a) - c < 0,$$

so there exists $t_1 \in (a,b)$ such that $g(t_1) < g(a)$. Similarly,

$$g'(b) = f'(b) - c > 0,$$

so there exists $t_2 \in (a,b)$ such that $g(t_2) < g(b)$.

By the extreme value theorem, g attains its minimum on $[a,b]$. From above, $g(a)$ and $g(b)$ cannot be minimums, so g attains its minimum at $x \in (a,b)$. By Fermat's theorem, $g'(x) = 0$. Hence $f'(x) = c$, as desired. $\square$

> **Corollary 17.15.** *If f is differentiable on $[a,b]$, then f' cannot have any simple discontinuities on $[a,b]$.*

17.4 L'Hopital's Rule

The following result is frequently used in the evaluation of limits.

Lemma 17.16 (L'Hopital's rule)**.** *Suppose f and g are differentiable over (a,b), with $g'(x) \neq 0$ for all $x \in (a,b)$, where $-\infty \leq a < b \leq +\infty$. If either*

(i) $\lim_{x\to a} f(x) = 0$ *and* $\lim_{x\to a} g(x) = 0$*; or*

(ii) $\lim_{x\to a} g(x) = +\infty$,

and

$$\lim_{x\to a} \frac{f'(x)}{g'(x)} = A,$$

then

$$\lim_{x\to a} \frac{f(x)}{g(x)} = A.$$

The analogous statement is of course also true if $x > b$, or if $g(x) \to -\infty$ in (ii).

Note that we now use the limit concept in the extended sense of Definition 16.43.

Proof. We first consider the case in which $-\infty \leq A < +\infty$. Choose $q \in \mathbb{R}$ such that $A < q$, and choose $r \in \mathbb{R}$ such that $A < r < q$. By (13) there exists $c \in (a,b)$ such that $a < x < c$ implies

$$\frac{f'(x)}{g'(x)} < r.$$

If $a < x < y < c$, then by the generalised mean value theorem (17.10), there exists $t \in (x,y)$ such that

$$\frac{f(x) - f(y)}{g(x) - g(y)} = \frac{f'(t)}{g'(t)} < r.$$

(i) Suppose $\lim_{x\to a} f(x) = 0$ and $\lim_{x\to a} g(x) = 0$. Let $x \to a$ in (18), we see that

$$\frac{f(y)}{g(y)} \leq r < q \quad (a < y < c).$$

(ii) Next, suppose $\lim_{x\to a} g(x) = +\infty$. Keeping y fixed in (18), we can choose a point $c_1 \in (a,y)$ such that $g(x) > g(y)$ and $g(x) > 0$ if $a < x < c_1$. Multiplying (18) by $[g(x) - g(y)]/g(x)$, we obtain

$$\frac{f(x)}{g(x)} < r - r\frac{g(y)}{g(x)} + \frac{f(y)}{g(x)} \quad (a < x < c_1).$$

If we let $x \to a$ in (20), (15) shows that there exists $c_2 \in (a, c_1)$ such that

$$\frac{f(x)}{g(x)} < q \quad (a < x < c_2).$$

Summing up, (19) and (21) show that for any q, subject only to the condition $A < q$, there is a point c_2 such that $f(x)/g(x) < q$ if $a < x < c_2$.

In the same manner, if $-\infty < A \leq +\infty$, and p is chosen so that $p < A$, we can find a point c_3 such that

$$p < \frac{f(x)}{g(x)} \quad (a < x < c_3),$$

and (16) follows from these two statements. $\square$

to review proof

17.5 Taylor's Theorem

Theorem 17.17 (Taylor's theorem)**.** *Suppose $f\colon [a,b] \to \mathbb{R}$, $f^{(n-1)}$ is continuous on $[a,b]$, $f^{(n)}$ exists on (a,b). Assume that $c \in [a,b]$. Let the* Taylor polynomial *of degree $n-1$ of f at $x = c$ be*

$$\begin{aligned} P_{n-1}(x) &= \sum_{k=0}^{n-1} \frac{f^{(k)}(c)}{k!}(x-c)^k \\ &= f(c) + f'(c)(x-c) + \frac{f''(c)}{2!}(x-c)^2 + \cdots + \frac{f^{(n-1)}(c)}{(n-1)!}(x-c)^{n-1}. \end{aligned}$$

Then for every $x \in [a,b]$, $x \neq c$, there exists z_x between x and c such that

$$f(x) = P_{n-1}(x) + \frac{f^{(n)}(z_x)}{n!}(x-c)^n. \tag{17.6}$$

For $n = 1$, this is just the mean value theorem. In general, the theorem shows that f can be approximated by a polynomial of degree $n-1$, and that (17.6) allows us to accurately estimate the error.

Proof. Let M be the number defined by

$$f(x) = P_{n-1}(x) + M(x-c)^n.$$

We claim that $n!M = f^{(n)}(z_x)$ for some z_x between x and c.

For all $x \in [a,b]$, let

$$g(x) = f(x) - P_{n-1}(x) - M(x-c)^n.$$

Then for all $x \in (a,b)$,

$$g^{(n)}(x) = f^{(n)}(x) - n!M.$$

Hence our proof will be complete if we can show that $g^{(n)}(z_x) = 0$ for some z_x between c and x.

Since $P_{n-1}^{(k)}(c) = f^{(k)}(c)$ for $k = 0, \dots, n-1$, we have

$$g(c) = g'(c) = \cdots = g^{(n-1)}(c) = 0.$$

By our choice of M, we have that $g(x) = 0$. By the mean value theorem, there exists x_1 between x and c such that $g'(x_1) = 0$. Since $g'(c) = 0$, we conclude similarly that $g''(x_2) = 0$ for some x_2 between x_1 and c. After n steps we arrive at the conclusion that $g^{(n)}(x_n) = 0$ for some x_n between x_{n-1} and c, that is, between x and c. ∎

17.6 Differentiation of Vector-valued Functions

Definition 5.1 applies without any change to complex functions f defined on $[a,b]$, and Theorems 5.2 and 5.3, as well as their proofs, remain valid. If f_1 and f_2 are the real and imaginary parts of f, that is, if

$$f(t) = f_1(t) + i f_2(t)$$

for $a \leq t \leq b$, where $f_1(t)$ and $f_2(t)$ are real, then we clearly have

$$f'(x) = f_1'(x) + i f_2'(x);$$

also, f is differentiable at x if and only if both f_1 and f_2 are differentiable at x.

Passing to vector-valued functions $\mathbf{f} : [a,b] \to \mathbb{R}^k$, we may still apply Definition 5.1 to define $\mathbf{f}'(x)$. The term $\phi(t)$ in (1) is now, for each t, a point in $\mathbb{R}^k$, and the limit in (2) is taken with respect to the norm of $\mathbb{R}^k$. In other words, $\mathbf{f}'(x)$ is that point of $\mathbb{R}^k$ (if there is one) for which

$$\lim_{t\to x} \left| \frac{\mathbf{f}(t) - \mathbf{f}(x)}{t-x} - \mathbf{f}'(x) \right| = 0,$$

and $\mathbf{f}'$ is again a function with values in $\mathbb{R}^k$.

If $f_1, \dots, f_k$ are the components of $\mathbf{f}$, as defined in Theorem 4.10, then

$$\mathbf{f}' = (f_1, \dots, f_k),$$

and $\mathbf{f}$ is differentiable at a point x if and only if each of the functions $f_1, \dots, f_k$ is differentiable at x.

Theorem 5.2 is true in this context as well, and so is Theorem 5.3(a) and (b), if fg is replaced by the inner product $\mathbf{f} \cdot \mathbf{g}$ (see Definition 4.3).

When we turn to the mean value theorem, however, and to one of its consequences, namely, L'Hospital's rule, the situation changes. The next two examples will show that each of these results fails to be true for complex-valued functions.

Example 17.18. Define, for real x,

$$f(x) := e^{ix}.$$

Then $f(x) = \cos x + i \sin x$, so

$$f(2\pi) - f(0) = 1 - 1 = 0,$$

but $f'(x) = ie^{ix}$, so $|f'(x)| = 1$ for all real x.

Hence the mean value theorem fails to hold in this case.

Example 17.19. On $(0,1)$ define

$$f(x) := x, \quad g(x) := x + x^2 e^{i/x^2}.$$

Since $|e^{it} = 1|$, we see that

$$\lim_{x\to 0} \frac{f(x)}{g(x)} = 1.$$

Next,

$$g'(x) = 1 + \left(2x - \frac{2i}{x}\right) e^{i/x^2} \quad (0 < x < 1),$$

so that

$$|g'(x)| \geq \left|2x - \frac{2i}{x}\right| - 1 \geq \frac{2}{x} - 1.$$

Hence

$$\left|\frac{f'(x)}{g'(x)}\right| = \frac{1}{|g'(x)|} \leq \frac{x}{2-x}$$

and so

$$\lim_{x\to 0} \frac{f'(x)}{g'(x)} = 0.$$

By (36) and (40), L'Hospital's rule fails in this case. Note also that $g'(x) \neq 0$ on $(0,1)$, by (38).

However, there is a consequence of the mean value theorem which, for purposes of applications, is almost as useful as Theorem 5.10, and which remains true for vector-valued functions: From Theorem 5.10 it follows that

$$|f(b) - f(a)| \leq (b-a) \sup_{x\in[a,b]} |f'(x)|.$$

Theorem 17.20. *Suppose* $\mathbf{f} : [a,b] \to \mathbb{R}^k$ *is continuous on* $[a,b]$ *and differentiable in* (a,b)*. Then there exists* $x \in (a,b)$ *such that*

$$\|\mathbf{f}(b) - \mathbf{f}(a)\| \leq (b-a)\left\|\mathbf{f}'(x)\right\|. \tag{17.7}$$

Proof. Put $\mathbf{z} = \mathbf{f}(b) - \mathbf{f}(a)$, and define

$$\phi(t) = \mathbf{z} \cdot \mathbf{f}(t) \quad (a \leq t \leq b).$$

Then ϕ is a real-valued continuous function on $[a,b]$ which is differentiable in (a,b). By the mean value theorem (17.11), there exists $x \in (a,b)$ such that

$$\phi(b) - \phi(a) = (b-a)\phi'(x) = (b-a)\mathbf{z} \cdot \mathbf{f}'(x).$$

On the other hand,

$$\begin{aligned}\phi(b) - \phi(a) &= \mathbf{z} \cdot \mathbf{f}(b) - \mathbf{z} \cdot \mathbf{f}(a) \\ &= \mathbf{z} \cdot (\mathbf{f}(b) - \mathbf{f}(a)) \\ &= \mathbf{z} \cdot \mathbf{z} = \|\mathbf{z}\|^2.\end{aligned}$$

By the Cauchy–Schwarz inequality, we obtain

$$\|\mathbf{z}\|^2 = (b-a)\left\|\mathbf{z} \cdot \mathbf{f}'(x)\right\| \leq (b-a)\|\mathbf{z}\| \left\|\mathbf{f}'(x)\right\|.$$

Hence $\|\mathbf{z}\| \leq (b-a)\|\mathbf{f}'(x)\|$, which is the desired conclusion. $\square$

Exercises

Exercise 17.1. Let f and g be continuous on $[a, b]$ and differentiable on (a, b). If $f'(x) = g'(x)$, then $f(x) = g(x) + C$.

Exercise 17.2. Given that $f(x) = x^{\alpha}$ where $0 < \alpha < 1$. Prove that f is uniformly continuous on $[0, +\infty)$.

Exercise 17.3. Let f be continuous on $[0, 1]$ and differentiable on $(0, 1)$ where $f(0) = f(1) = 0$. Prove that there exists $c \in (0, 1)$ such that

$$f(x) + f'(x) = 0.$$

Chapter 18

Riemann–Stieltjes Integral

The present chapter is based on a definition of the Riemann integral which depends very explicitly on the order structure of the real line. Accordingly, we begin by discussing integration of real-valued functions on intervals. Extensions to complex- and vector-valued functions on intervals follow in later sections.

18.1 Definition of Riemann–Stieltjes Integral

To approximate the area under the curve of a function, we partition the interval into finitely many sub-intervals, then multiply the width of each sub-interval by its height.

- For the height, we can choose to either use the supremum of the function over the interval or the infimum. Obviously, using the supremum will provide an upper bound, and using the infimum will provide a lower bound.
- For the width, we use the difference between the two endpoints in their output values when input into a monotonically increasing function α.

The upper Riemann integral is the infimum of upper bounds over all possible partitions. The lower Riemann integral is similarly defined. If they are equal, then the function is said to be Riemann–Stieltjes integrable.

18.1.1 Notation and Preliminaries

A ***partition*** P of a closed interval $[a, b]$ is a finite set of points $\{x_0, x_1, \dots, x_n\}$, where

$$a = x_0 \le x_1 \le \cdots \le x_{n-1} \le x_n = b.$$

Notation. Denote the set of all partitions of $[a, b]$ by $\mathcal{P}[a, b]$.

Let $f\colon [a, b] \to \mathbb{R}$ be bounded. Denote

$$M_i = \sup_{x\in[x_{i-1},x_i]} f(x), \quad m_i = \inf_{x\in[x_{i-1},x_i]} f(x) \quad (i = 1, \dots, n).$$

Let α be a monotonically increasing function on $[a,b]$. Denote

$$\Delta\alpha_i = \alpha(x_i) - \alpha(x_{i-1}) \quad (i = 1, \dots, n).$$

(These suprema and infima are well-defined, finite real numbers due to the boundedness of f.)

The *upper sum* and *lower sum* of f with respect to the partition P and α are respectively

$$U(f,\alpha;P) = \sum_{i=1}^{n} M_i \Delta\alpha_i,$$
$$L(f,\alpha;P) = \sum_{i=1}^{n} m_i \Delta\alpha_i.$$

insert diagram

The partition P' is a ***refinement*** of P if $P' \supset P$. Given two partitions P_1 and P_2, we say that P' is their *common refinement* if $P' = P_1 \cup P_2$.

Intuitively, a refinement will give a better estimation than the original partition, so the upper and lower sums of a refinement should be more restrictive.

Lemma 18.1. *If P' is a refinement of P, then*

(i) $L(f,\alpha;P) \le L(f,\alpha;P')$

(ii) $U(f,\alpha;P') \le U(f,\alpha;P)$

Proof.

(i) Suppose first that P' contains just one point more than P. Let this extra point be x', and suppose $x_{i-1} < x' < x_i$ for some i ($1 \le i \le n$), where $x_{i-1}, x_i \in P$. Let

$$w_1 = \inf_{x \in [x_{i-1}, x']} f(x), \quad w_2 = \inf_{x \in [x', x_i]} f(x).$$

Let, as before,

$$m_i = \inf_{x \in [x_{i-1}, x_i]} f(x).$$

Clearly $w_1 \ge m_i$ and $w_2 \ge m_i$. Then

$$\begin{aligned}
&L(f,\alpha;P') - L(f,\alpha;P) \\
&= w_1\big(\alpha(x') - \alpha(x_{i-1})\big) + w_2\big(\alpha(x_i) - \alpha(x')\big) - m_i\big(\alpha(x_i) - \alpha(x_{i-1})\big) \\
&= \underbrace{(w_1 - m_i)}_{\ge 0}\underbrace{\big(\alpha(x') - \alpha(x_{i-1})\big)}_{>0} + \underbrace{(w_2 - m_i)}_{\ge 0}\underbrace{\big(\alpha(x_i) - \alpha(x')\big)}_{>0} \\
&\ge 0
\end{aligned}$$

and hence $L(f,\alpha;P) \le L(f,\alpha;P')$.

If P' contains k more points than P, we repeat this reasoning k times.

(ii) Analogous to the proof of (i).

$\square$

Since f is bounded, there exist m and M such that $m \leq f(x) \leq M$ for all $x \in [a,b]$. Hence for every partition P,

$$m\,(\alpha(b) - \alpha(a)) \leq L(f,\alpha;P) \leq U(f,\alpha;P) \leq M\,(\alpha(b) - \alpha(a))$$

so that the numbers $L(f,\alpha;P)$ and $U(f,\alpha;P)$ form a bounded set. This shows that the upper and lower integrals are defined for every bounded function f. We now define the *upper and lower Riemann–Stieltjes integrals* respectively as

$$\begin{aligned} \overline{\int_a^b} f \,\mathrm{d}\alpha &:= \inf_{P \in \mathcal{P}[a,b]} U(f,\alpha;P) \\ \underline{\int_a^b} f \,\mathrm{d}\alpha &:= \sup_{P \in \mathcal{P}[a,b]} L(f,\alpha;P) \end{aligned}$$

where we take inf and sup over all partitions.

One would expect the lower RS integral to be less than or equal to the upper RS integral. We now show this.

Lemma 18.2.

$$\underline{\int_a^b} f \,\mathrm{d}\alpha \leq \overline{\int_a^b} f \,\mathrm{d}\alpha \,.$$

Proof. Let P' be the common refinement of partitions P_1 and P_2; that is, $P' = P_1 \cup P_2$. Clearly $P' \supset P_1$; by 18.1,

$$L(f,\alpha;P_1) \leq L(f,\alpha;P').$$

Similarly, $P' \supset P_2$, so

$$U(f,\alpha;P') \leq U(f,\alpha;P_2).$$

Clearly $L(f,\alpha;P') \leq U(f,\alpha;P')$. Thus combining the above two equations gives

$$L(f,\alpha;P_1) \leq U(f,\alpha;P_2).$$

Fix P_2 and take sup over all P_1 gives

$$\underline{\int_a^b} f \,\mathrm{d}\alpha \leq U(f,\alpha;P_2).$$

Then taking inf over all P_2 gives

$$\underline{\int_a^b} f \,\mathrm{d}\alpha \leq \overline{\int_a^b} f \,\mathrm{d}\alpha \,.$$

□

18.1.2 Defining the Integral

Definition 18.3 (Riemann–Stieltjes integral)**.** We say $f : [a,b] \to \mathbb{R}$ is ***Riemann–Stieltjes integrable*** with respect to α over $[a,b]$, if

$$\underline{\int_a^b} f \, d\alpha = \overline{\int_a^b} f \, d\alpha \,.$$

We call the common value the ***Riemann–Stieltjes integral*** of f with respect to α over $[a,b]$, and denote it as

$$\int_a^b f \, d\alpha \,.$$

The functions f and α are referred to as the *integrand* and the *integrator*, respectively.

Notation. $\mathcal{R}(\alpha)$ denotes the set of Riemann–Stieltjes integrable functions with respect to α.

In particular, when $\alpha(x) = x$, we call the corresponding Riemann–Stieltjes integral the *Riemann integral*, and use $\mathcal{R}$ to denote the set of Riemann integrable functions.

Notation. Since x is a "dummy variable" and may be replaced by any other variable, we shall omit it.

Example 18.4 (Dirichlet function)**.** The *Dirichlet function* is defined over $[0,1]$ by

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

For each subinterval $[x_{i-1}, x_i]$, due to the density of rationals and irrationals, $[x_{i-1}, x_i]$ contains both rationals and irrationals, so $M_i = 1$ and $m_i = 0$. Thus for any partition P,

$$U(f;P) = 1, \quad L(f;P) = 0.$$

Therefore,

$$1 = \overline{\int} f \, d\alpha \neq \underline{\int} f \, d\alpha = 0$$

so the Dirichlet function is not Riemann–Stieltjes integrable.

The next result is particularly useful in determining the Riemann–Stieltjes integrability of a function. We will use it many times later.

Lemma 18.5 (Integrability criterion)**.** $f \in \mathcal{R}(\alpha)$ *if and only if*

$$\forall \varepsilon > 0, \quad \exists P, \quad U(f,\alpha;P) - L(f,\alpha;P) < \varepsilon.$$

Proof.

$\boxed{\Longrightarrow}$ Suppose $f \in \mathcal{R}(\alpha)$. Let $\varepsilon > 0$ be given. Then there exists partitions P_1 and P_2 such that

$$U(f,\alpha;P_2) - \int_a^b f \, d\alpha < \frac{\varepsilon}{2}$$

and
$$\int_a^b f\,\mathrm{d}\alpha - L(f,\alpha;P_1) < \frac{\varepsilon}{2}.$$
Choose P to be the common refinement of P_1 and P_2. Then

$$\begin{aligned}
U(f,\alpha;P) &\le U(f,\alpha;P_2) \\
&< \int_a^b f\,\mathrm{d}\alpha + \frac{\varepsilon}{2} \\
&< L(f,\alpha;P_1) + \varepsilon \\
&\le L(f,\alpha;P) + \varepsilon.
\end{aligned}$$

Hence for this partition P, we have

$$U(f,\alpha;P) - L(f,\alpha;P) < \varepsilon.$$

$\boxed{\Longleftarrow}$ By 18.2, for every partition P,

$$L(f,\alpha;P) \le \underline{\int_a^b} f\,\mathrm{d}\alpha \le \overline{\int_a^b} f\,\mathrm{d}\alpha \le U(f,\alpha;P).$$

Since $U(f,\alpha;P) - L(f,\alpha;P) < \varepsilon$, we have

$$0 \le \overline{\int_a^b} \mathrm{d}\alpha - \underline{\int_a^b} f\,\mathrm{d}\alpha < \varepsilon.$$

Since this holds for all $\varepsilon > 0$, we have

$$\overline{\int_a^b} f\,\mathrm{d}\alpha = \underline{\int_a^b} f\,\mathrm{d}\alpha\,.$$

Hence $f \in \mathcal{R}(\alpha)$. □

18.1.3 Useful Identities

Proposition 18.6 (Cauchy criterion)**.**

(i) If $U(f,\alpha;P) - L(f,\alpha;P) < \varepsilon$ holds for some P and some $\varepsilon > 0$, then $U(f,\alpha;P') - L(f,\alpha;P') < \varepsilon$ holds (with the same ε) for every refinement of P, P'.

(ii) If $U(f,\alpha;P) - L(f,\alpha;P) < \varepsilon$ holds for $P = \{x_0,\dots,x_n\}$, and

$$s_i, t_i \in [x_{i-1}, x_i] \quad (i = 1,\dots,n)$$

then

$$\sum_{i=1}^{n} |f(s_i) - f(t_i)|\, \Delta\alpha_i < \varepsilon.$$

(iii) If $f \in \mathcal{R}(\alpha)$ and the hypotheses of (ii) hold, then

$$\left| \sum_{i=1}^{n} f(t_i)\Delta\alpha_i - \int_a^b f \,\mathrm{d}\alpha \right| < \varepsilon.$$

Proof.

(i) Suppose $U(f,\alpha;P) - L(f,\alpha;P) < \varepsilon$ holds for some partition P and some $\varepsilon > 0$. By 18.1, for any refinement P',

$$U(f,\alpha;P') \leq U(f,\alpha;P), \quad L(f,\alpha;P) \leq L(f,\alpha;P').$$

Hence

$$U(f,\alpha;P') - L(f,\alpha;P') \leq U(f,\alpha;P) - L(f,\alpha;P) < \varepsilon.$$

(ii) Since

$$f(s_i), f(t_i) \in [m_i, M_i] \quad (i = 1,\dots,n)$$

it follows that

$$|f(s_i) - f(t_i)| \leq M_i - m_i.$$

Hence

$$\sum_{i=1}^{n} |f(s_i) - f(t_i)|\, \Delta\alpha_i \leq U(f,\alpha;P) - L(f,\alpha;P) < \varepsilon.$$

(iii) The desired result follows from the two inequalities

$$L(f,\alpha;P) \leq \sum_{i=1}^{n} f(t_i)\Delta\alpha_i \leq U(f,\alpha;P)$$

$$L(f,\alpha;P) \leq \int_a^b f \,\mathrm{d}\alpha \leq U(f,\alpha;P)$$

□

The next result states that all continuous functions are integrable.

Proposition 18.7 (Continuity implies integrability). *If f is continuous on $[a, b]$, then $f \in \mathcal{R}(\alpha)$.*

Proof. Let $\varepsilon > 0$ be given. Choose $\eta > 0$ such that

$$(\alpha(b) - \alpha(a))\,\eta < \varepsilon.$$

Since f is continuous on $[a, b]$ which is compact, by 16.28, f is uniformly continuous on $[a, b]$. Thus there exists $\delta > 0$ such that for all $x, y \in [a, b]$,

$$|x - y| < \delta \implies |f(x) - f(y)| < \eta.$$

If P is any partition of $[a, b]$ such that $\Delta x_i < \delta$ for $i = 1, \ldots, n$, then

$$M_i - m_i \leq \eta \quad (i = 1, \ldots, n).$$

Hence

$$\begin{aligned} U(f, \alpha; P) - L(f, \alpha; P) &= \sum_{i-1}^{n} (M_i - m_i)\Delta\alpha_i \\ &\leq \eta \sum_{i=1}^{n} \Delta\alpha_i = \eta\,(\alpha(b) - \alpha(a)) < \varepsilon. \end{aligned}$$

Therefore $f \in \mathcal{R}(\alpha)$, by the integrability criterion (18.5). $\square$

Proposition 18.8. *If f is monotonic on $[a, b]$, and if α is continuous on $[a, b]$, then $f \in \mathcal{R}(\alpha)$.*

Proof. Let $\varepsilon > 0$ be given. For any positive integer n, choose a partition P such that

$$\Delta\alpha_i = \frac{\alpha(b) - \alpha(a)}{n} \quad (i = 1, \ldots, n).$$

This is possible by the intermediate value theorem, due to the continuity of α.

Suppose that f is monotonically increasing (the proof is analogous in the other case). Then

$$M_i = f(x_i), \quad m_i = f(x_{i-1}) \quad (i = 1, \ldots, n).$$

Hence

$$\begin{aligned} U(f, \alpha; P) - L(f, \alpha; P) &= \sum_{i=1}^{n} (M_i - m_i)\Delta\alpha_i \\ &= \frac{\alpha(b) - \alpha(a)}{n} \sum_{i=1}^{n} (f(x_i) - f(x_{i-1})) \\ &= \frac{\alpha(b) - \alpha(a)}{n} (f(b) - f(a)) < \varepsilon \end{aligned}$$

if n is taken large enough. Hence $f \in \mathcal{R}(\alpha)$, by the integrability criterion. □

Proposition 18.9. *Suppose f is bounded on $[a, b]$, f has only finitely many points of discontinuity on $[a, b]$, and α is continuous at every point at which f is discontinuous. Then $f \in \mathcal{R}(\alpha)$.*

Proof. Let $\varepsilon > 0$ be given. Since f is bounded, let $M = \sup |f(x)|$. Let E be the set of points at which f is discontinuous.

Since E is finite, and α is continuous at every point of E, we can cover E by finitely many disjoint intervals $[u_j, v_j] \subset [a, b]$ such that the sum of the corresponding differences $\sum_j (\alpha(v_j) - \alpha(u_j)) < \varepsilon$. Furthermore, we can place these intervals in such a way that every point of $E \cap (a, b)$ lies in the interior of some $[u_j, v_j]$.

Remove the segments (u_j, v_j) from $[a, b]$. The remaining set K is compact. Hence f is uniformly continuous on K, so there exists $\delta > 0$ such that for all $s, t \in K$,

$$|s - t| < \delta \implies |f(x) - f(t)| < \varepsilon.$$

Now form a partition $P = \{x_0, x_1, \ldots, x_n\}$ of $[a, b]$ as follows: Each u_j occurs in P. Each v_j occurs in P. No point of any segment (u_j, v_j) occurs in P. If x_{i-i} is not one of the u_j, then $\Delta x_i < \delta$.

Note that $M_i - m_i \leq 2M$ for every i, and that $M_i - m_i < \varepsilon$ unless x_{i-i} is one of the u_j. Hence

$$\begin{aligned} U(f, \alpha; P) - L(f, \alpha; P) &= \sum_{i=1}^{n} (M_i - m_i) \Delta\alpha_i \\ &\leq (\alpha(b) - \alpha(a))\, \varepsilon + 2M\varepsilon. \end{aligned}$$

Since ε is arbitrary, we have $f \in \mathcal{R}(\alpha)$, by the integrability criterion. □

The next result states that a uniformly continuous function of an integrable function is also integrable.

Proposition 18.10. *Suppose $f \in \mathcal{R}(\alpha)$, $m \leq f \leq M$, and ϕ is continuous on $[m, M]$. Then $\phi \circ f \in \mathcal{R}(\alpha)$.*

Proof. Let $h = \phi \circ f$. Let $\varepsilon > 0$ be given. Since ϕ is uniformly continuous on $[m, M]$, there exists $\delta > 0$ such that $\delta < \varepsilon$, and for all $s, t \in [m, M]$,

$$|s - t| \leq \delta \implies |\phi(s) - \phi(t)| < \varepsilon.$$

Since $f \in \mathcal{R}(\alpha)$, by 18.5, there exists a partition $P = \{x_0, \ldots, x_n\}$ of $[a, b]$ such that

$$U(f, \alpha; P) - L(f, \alpha; P) < \delta^2. \tag{1}$$

Let

$$\begin{aligned} M_i &= \sup_{x \in [x_{i-1}, x_i]} f(x), & M_i^* &= \sup_{x \in [x_{i-1}, x_i]} h(x), \\ m_i &= \inf_{x \in [x_{i-1}, x_i]} f(x), & m_i^* &= \inf_{x \in [x_{i-1}, x_i]} h(x). \end{aligned}$$

Divide the numbers $1, \dots, n$ into two classes:

$$A = \{i \mid M_i - m_i < \delta\},$$
$$B = \{i \mid M_i - m_i \geq \delta\}.$$

- For $i \in A$, our choice of δ shows that $M_i^* - m_i^* \leq \varepsilon$.
- For $i \in B$, $M_i^* - m_i^* \leq 2K$, where $K = \sup_{m \leq t \leq M} |\phi(t)|$.

By (1), we have

$$\delta \sum_{i \in B} \Delta\alpha_i \leq \sum_{i \in B} (M_i - m_i)\Delta\alpha_i < \delta^2$$

so that $\sum_{i \in B} \Delta\alpha_i < \delta$. It follows that

$$\begin{aligned} U(h, \alpha; P) - L(h, \alpha; P) &= \sum_{i \in A} (M_i^* - m_i^*)\Delta\alpha_i + \sum_{i \in B} (M_i^* - m_i^*)\Delta\alpha_i \\ &\leq \varepsilon\left(\alpha(b) - \alpha(a)\right) + 2K\delta \\ &< \varepsilon\left(\alpha(b) - \alpha(a) + 2K\right). \end{aligned}$$

Since ε was arbitrary, by the integrability criterion, $h \in \mathcal{R}(\alpha)$. $\square$

18.2 Properties of the Integral

Lemma 18.11.

(i) If $f_1, f_2 \in \mathcal{R}(\alpha)$, then $f_1 + f_2 \in \mathcal{R}(\alpha)$, and

$$\int_a^b (f_1 + f_2)\,\mathrm{d}\alpha = \int_a^b f_1\,\mathrm{d}\alpha + \int_a^b f_2\,\mathrm{d}\alpha\,.$$

(ii) If $f \in \mathcal{R}(\alpha)$, then $cf \in \mathcal{R}(\alpha)$ for every $c \in \mathbb{R}$, and

$$\int_a^b (cf)\,\mathrm{d}\alpha = c\int_a^b f\,\mathrm{d}\alpha\,.$$

(iii) If $f_1, f_2 \in \mathcal{R}(\alpha)$ and $f_1 \le f_2$, then

$$\int_a^b f_1\,\mathrm{d}\alpha \le \int_a^b f_2\,\mathrm{d}\alpha\,.$$

(iv) If $f \in \mathcal{R}(\alpha)$ and $c \in [a,b]$, then $f \in \mathcal{R}_\alpha[a,c]$ and $f \in \mathcal{R}_\alpha[c,b]$, and

$$\int_a^b f\,\mathrm{d}\alpha = \int_a^c f\,\mathrm{d}\alpha + \int_c^b f\,\mathrm{d}\alpha\,.$$

(v) If $f \in \mathcal{R}(\alpha)$ and $|f| \le M$, then

$$\left|\int_a^b f\,\mathrm{d}\alpha\right| \le M\left(\alpha(b) - \alpha(a)\right).$$

(vi) If $f \in R_{\alpha_1}[a,b]$ and $f \in R_{\alpha_2}[a,b]$, then $f \in R_{\alpha_1+\alpha_2}[a,b]$, and

$$\int_a^b f\,\mathrm{d}(\alpha_1 + \alpha_2) = \int_a^b f\,\mathrm{d}\alpha_1 + \int_a^b f\,\mathrm{d}\alpha_2\,;$$

if $f \in \mathcal{R}(\alpha)$ and c is a positive constant, then $f \in R_{c\alpha}[a,b]$, and

$$\int_a^b f\,\mathrm{d}(c\alpha) = c\int_a^b f\,\mathrm{d}\alpha\,.$$

(vii) If $f \in \mathcal{R}(\alpha)$ and $g \in \mathcal{R}(\alpha)$, then $fg \in \mathcal{R}(\alpha)$.

Proof.

(i) If $f = f_1 + f_2$ and P is any partition of $[a,b]$, we have

$$L(f_1,\alpha;P) + L(f_2,\alpha;P) \le L(f,\alpha;P) \le U(f,\alpha;P) \le U(f_1,\alpha;P) + U(f_2,\alpha;P). \quad (1)$$

If $f_1 \in \mathcal{R}(\alpha)$ and $f_2 \in \mathcal{R}(\alpha)$, let $\varepsilon > 0$ be given. There are partitions P_1 and P_2 such that

$$U(f_1, \alpha; P_1) - L(f_1, \alpha; P_1) < \frac{\varepsilon}{2}$$
$$U(f_2, \alpha; P_2) - L(f_2, \alpha; P_2) < \frac{\varepsilon}{2}$$

Let P be the common refinement of P_1 and P_2. Then (1) implies

$$U(f, \alpha; P) - L(f, \alpha; P) < \varepsilon$$

which proves that $f \in \mathcal{R}(\alpha)$.

With this same P we have

$$U(f_1, \alpha; P) < \int_a^b f_1 \,\mathrm{d}\alpha + \frac{\varepsilon}{2}$$
$$U(f_2, \alpha; P) < \int_a^b f_2 \,\mathrm{d}\alpha + \frac{\varepsilon}{2}$$

Hence (1) implies

$$\int_a^b f \,\mathrm{d}\alpha \le U(f, \alpha; P) < \int_a^b f_1 \,\mathrm{d}\alpha + \int_a^b f_2 \,\mathrm{d}\alpha + \varepsilon.$$

Since ε was arbitrary, we conclude that

$$\int_a^b f \,\mathrm{d}\alpha \le \int_a^b f_1 \,\mathrm{d}\alpha + \int_a^b f_2 \,\mathrm{d}\alpha \,.$$

If we replace f_1 and f_2 in the above equation by $-f_1$ and $-f_2$, the inequality is reversed, and the equality is proved.

(ii) The case where $c = 0$ is trivial. Given $\varepsilon > 0$, there exists P such that $U(f, \alpha; P) - L(f, \alpha; P) < \varepsilon$. If $c > 0$ write

$$U(cf, \alpha; P) = \sum_{i=1}^{n} cM_i\alpha_i = c\sum_{i=1}^{n} M_i\alpha_i = cU(f, \alpha; P).$$

Similarly,

$$L(cf, \alpha; P) = cL(f, \alpha; P).$$

Then

$$U(cf, \alpha; P) - L(cf, \alpha; P) = c\,(U(f, \alpha; P) - L(f, \alpha; P)) < c\varepsilon$$

and since ε is arbitrary, we are done. The case where $c < 0$ is similar. Therefore $cf \in \mathcal{R}(\alpha)$.

With this same P we have

$$U(f, \alpha; P) - \int_a^b f \,\mathrm{d}\alpha < \varepsilon.$$

Then if $c > 0$,

$$\int_a^b cf \,\mathrm{d}\alpha \le U(cf, \alpha; P) = cU(f, \alpha; P) < c\int_a^b f \,\mathrm{d}\alpha + c\varepsilon$$

so

$$\int_a^b cf \, \mathrm{d}\alpha \leq c \int_a^b f \, \mathrm{d}\alpha \, .$$

If we replace f in the above equation by $-f$, the inequality is reversed, and the equality is proved.

(iii) For every partition P, we have

$$U(f_1, \alpha; P) = \sum_{i=1}^{n} M_i(f_1)\Delta\alpha_i \leq \sum_{i=1}^{n} M_i(f_2)\Delta\alpha_i = U(f_2, \alpha; P)$$

since α is monotonically increasing on $[a, b]$.

(iv)

(v)

(vi)

(vii) Take $\phi(t) = t^2$. By 18.10, $f^2 \in R_\alpha[a, b]$ if $f \in R_\alpha[a, b]$. Write

$$fg = \frac{1}{4}\left((f+g)^2 - (f-g)^2\right) .$$

Then the desired result follows.

□

Lemma 18.12 (Triangle inequality)**.** *Suppose $f \in \mathcal{R}(\alpha)$. Then $|f| \in \mathcal{R}(\alpha)$, and*

$$\left| \int_a^b f \, \mathrm{d}\alpha \right| \leq \int_a^b |f| \, \mathrm{d}\alpha \, .$$

Proof. Take $\phi(t) = |t|$, which is a continuous function. By 18.10, we have that $|f| = \phi \circ f \in \mathcal{R}(\alpha)$. Choose $c = \pm 1$, so that

$$c \int_a^b f \, \mathrm{d}\alpha \geq 0.$$

Then

$$\left| \int_a^b f \, \mathrm{d}\alpha \right| = c \int_a^b f \, \mathrm{d}\alpha = \int_a^b cf \, \mathrm{d}\alpha \leq \int_a^b |f| \, \mathrm{d}\alpha \, ,$$

since $cf \leq |f|$. □

Example 18.13 (Heaviside step function)**.** The *Heaviside step function* is defined by

$$H(x) = \begin{cases} 0 & (x \leq 0) \\ 1 & (x > 0) \end{cases}$$

Proposition. *Suppose f is bounded on $[a, b]$, continuous at $s \in (a, b)$. Let $\alpha(x) = H(x - s)$, then*

$$\int_a^b f \, \mathrm{d}\alpha = f(s).$$

Proof. Consider partitions $P = \{x_0, x_1, x_2, x_3\}$, where $x_0 = a$, and $x_1 = s < x_2 < x_3 = b$. Then

$$U(f, \alpha; P) = M_2, \quad L(f, \alpha; P) = m_2.$$

Since f is continuous at s, we see that M_2 and m_2 converge to $f(s)$ as $x_2 \to s$. □

Proposition. *Suppose $c_n \geq 0$ for $n = 1, 2, \ldots$, $\sum c_n$ converges, (s_n) is a sequence of distinct points in (a, b), and*

$$\alpha(x) = \sum_{n=1}^{\infty} c_n H(x - s_n).$$

Let f be continuous on $[a, b]$. Then

$$\int_a^b f \, \mathrm{d}\alpha = \sum_{n=1}^{\infty} c_n f(s_n).$$

Proof. Since $0 \leq c_n H(x - s_n) \leq c_n$ for $n = 1, 2, \ldots$ and $\sum c_n$ converges, by the comparison test, $\alpha(x) = \sum c_n H(x - s_n)$ converges for every x. Its sum $\alpha(x)$ is evidently monotonic (since each term in the sum is non-negative), and $\alpha(a) = 0$, $\alpha(b) = \sum c_n$.

Let $\varepsilon > 0$ be given. Since $\sum c_n$ converges, choose $N \in \mathbb{N}$ so that

$$\sum_{n=N+1}^{\infty} c_n < \varepsilon.$$

Let

$$\alpha_1(x) = \sum_{n=1}^{N} c_n H(x - s_n), \quad \alpha_2(x) = \sum_{n=N+1}^{\infty} c_n H(x - s_n).$$

By the previous result,

$$\int_a^b f \, \mathrm{d}\alpha_1 = \sum_{n=1}^{N} c_n f(s_n).$$

Since $\alpha_2(b) - \alpha_2(a) < \varepsilon$,

$$\left| \int_a^b f \, \mathrm{d}\alpha_2 \right| \leq M\varepsilon,$$

where $M = \sup |f(x)|$. Since $\alpha = \alpha_1 + \alpha_2$,

$$\int_a^b f \, \mathrm{d}\alpha = \int_a^b f \, \mathrm{d}\alpha_1 + \int_a^b f \, \mathrm{d}\alpha_2$$

so it follows that

$$\left| \int_a^b f \, \mathrm{d}\alpha - \sum_{n=1}^{N} c_n f(s_n) \right| \leq M\varepsilon.$$

Since ε was arbitrary, and taking $N \to \infty$, we obtain

$$\int_a^b f \, \mathrm{d}\alpha = \sum_{n=1}^{\infty} c_n f(s_n).$$

□

In this case, we call $\alpha(x)$ a *step function*; then the integral reduces to a finite or infinite series.

The next result states that if α has an integrable derivative, then the integral reduces to an ordinary Riemann integral.

Proposition 18.14. *Assume α increases monotonically, $\alpha' \in \mathcal{R}$. Let $f\colon [a,b] \to \mathbb{R}$ be bounded, then $f \in \mathcal{R}(\alpha)$ if and only if $f\alpha' \in \mathcal{R}$. In that case*

$$\int_a^b f \,\mathrm{d}\alpha = \int_a^b f(x)\alpha'(x)\,\mathrm{d}x\,. \tag{18.1}$$

Proof. Let $\varepsilon > 0$ be given and apply 18.5 to α': There exists a partition $P = \{x_0, \ldots, x_n\}$ of $[a,b]$ such that

$$U(\alpha'; P) - L(\alpha'; P) < \varepsilon. \tag{1}$$

By the mean value theorem, there exist points $t_i \in [x_{i-1}, x_i]$ such that

$$\Delta\alpha_i = \alpha'(t_i)\Delta x_i \quad (i = 1, \ldots, n).$$

If $s_i \in [x_{i-1}, x_i]$, then by 18.6,

$$\sum_{i=1}^n \left|\alpha'(s_i) - \alpha'(t_i)\right| \Delta x_i < \varepsilon. \tag{2}$$

Let $M = \sup|f(x)|$. Since

$$\sum_{i=1}^n f(s_i)\Delta\alpha_i = \sum_{i=1}^n f(s_i)\alpha'(t_i)\Delta x_i$$

it follows from (2) that

$$\begin{aligned}
\left|\sum_{i=1}^n f(s_i)\Delta\alpha_i - \sum_{i=1}^n f(s_i)\alpha'(s_i)\Delta x_i\right| &= \left|\sum_{i=1}^n f(s_i)\left(\alpha'(t_i) - \alpha'(s_i)\right)\Delta x_i\right| \\
&\le \sum_{i=1}^n \left|f(s_i)\left(\alpha'(t_i) - \alpha'(s_i)\right)\Delta x_i\right| \\
&= \sum_{i=1}^n |f(s_i)|\left|\alpha'(t_i) - \alpha'(s_i)\right|\Delta x_i \\
&\le M\sum_{i=1}^n \left|\alpha'(t_i) - \alpha'(s_i)\right|\Delta x_i \\
&\le M\varepsilon. \qquad (3)
\end{aligned}$$

In particular, for all choices of $s_i \in [x_{i-1}, x_i]$,

$$\sum_{i=1}^n f(s_i)\Delta\alpha_i \le U(f\alpha'; P) + M\varepsilon$$

so taking sup for $f(s_i)$ gives

$$U(f, \alpha; P) \le U(f\alpha'; P) + M\varepsilon.$$

The same argument leads from (3) to

$$U(f\alpha';P) \le U(f,\alpha;P) + M\varepsilon.$$

Hence

$$\left|U(f,\alpha;P) - U(f\alpha';P)\right| \le M\varepsilon. \tag{4}$$

Since (1) holds true for any refinement of P, hence (4) also remains true. We conclude that

$$\left| \overline{\int_a^b} f\,\mathrm{d}\alpha - \overline{\int_a^b} f(x)\alpha'(x)\,\mathrm{d}x \right| \le M\varepsilon.$$

But ε is arbitrary. Hence

$$\overline{\int_a^b} f\,\mathrm{d}\alpha = \overline{\int_a^b} f(x)\alpha'(x)\,\mathrm{d}x$$

for any bounded f. The equality of the lower integrals follows from

$$\begin{aligned} \overline{\int_a^b} -f\,\mathrm{d}\alpha &= \overline{\int_a^b} -f\alpha'\,\mathrm{d}x \\ -\underline{\int_a^b} f\,\mathrm{d}\alpha &= -\underline{\int_a^b} f\alpha'\,\mathrm{d}x \\ \underline{\int_a^b} f\,\mathrm{d}\alpha &= \underline{\int_a^b} f(x)\alpha'(x)\,\mathrm{d}x \end{aligned}$$

Therefore the theorem follows. □

Proposition 18.15 (Change of variables)**.** *Suppose $\phi\colon [A,B] \to [a,b]$ is strictly increasing and continuous. Suppose α is monotonically increasing on $[a,b]$, $f \in \mathcal{R}(\alpha)$. Define β and g on $[A,B]$ by*

$$\beta(y) = \alpha(\phi(y)), \quad g(y) = f(\phi(y)).$$

Then $g \in \mathcal{R}(\beta)$, and

$$\int_A^B g\,\mathrm{d}\beta = \int_a^b f\,\mathrm{d}\alpha\,. \tag{18.2}$$

Proof. To each partition $P = \{x_0,\dots,x_n\}$ of $[a,b]$ corresponds a partition $Q = \{y_0,\dots,y_n\}$ of $[A,B]$, where

$$x_i = \phi(y_i) \quad (i = 1,\dots,n).$$

All partitions of $[A,B]$ are obtained in this way. Since the values taken by f on $[x_{i-i}, x_i]$ are exactly the same as those taken by g on $[y_{i-i}, y_i]$, we see that

$$\begin{aligned} U(g,\beta;Q) &= U(f,\alpha;P), \\ L(g,\beta;Q) &= L(f,\alpha;P). \end{aligned} \tag{1}$$

Since $f \in \mathcal{R}(\alpha)$, P can be chosen so that both $U(f,\alpha;P)$ and $L(f,\alpha;P)$ are close to $\int f\,\mathrm{d}\alpha$. Hence

(1), combined with 18.5, shows that $g \in \mathcal{R}_\beta[A, B]$ and

$$\int_A^B g \,\mathrm{d}\beta = \int_a^b f \,\mathrm{d}\alpha \,.$$

□

Note the following special case: Take $\alpha(x) = x$. Then $\beta = \phi$. Assume $\phi' \in \mathcal{R}$. Applying 18.14 to the LHS of

$$\int_A^B g \,\mathrm{d}\beta = \int_a^b f \,\mathrm{d}\alpha \,,$$

we obtain

$$\int_a^b f(x) \,\mathrm{d}x = \int_A^B f\left(\phi(y)\right) \phi'(y) \,\mathrm{d}y \,.$$

18.3 Integration and Differentiation

We shall show that integration and differentiation are, in a certain sense, inverse operations.

Theorem 18.16. *Suppose $f \in \mathcal{R}(\alpha)$. For $a \le x \le b$, let the* cumulative function *be*

$$F(x) = \int_a^x f(t)\,\mathrm{d}t\,.$$

Then F is continuous on $[a, b]$; furthermore, if f is continuous at $x_0 \in [a, b]$, then F is differentiable at x_0, and

$$F'(x_0) = f(x_0).$$

Proof. Suppose $f \in \mathcal{R}(\alpha)$. Since f is bounded, let $|f(t)| \le M$ for $t \in [a, b]$. If $a \le x < y \le b$, then

$$\begin{aligned} |F(y) - F(x)| &= \left| \int_a^y f(t)\,\mathrm{d}t - \int_a^x f(t)\,\mathrm{d}t \right| \\ &= \left| \int_x^y f(t)\,\mathrm{d}t \right| \\ &\le \int_x^y |f(t)|\,\mathrm{d}t \\ &\le M(y - x). \end{aligned}$$

Hence F is Lipschitz continuous, so F is uniformly continuous on $[a, b]$.

Now suppose f is continuous at x_0. Fix $\varepsilon > 0$, choose $\delta > 0$ such that for $a \le t \le b$,

$$|t - x_0| < \delta \implies |f(t) - f(x_0)| < \varepsilon.$$

Hence, if s, t are such that

$$x_0 - \delta < s \le x_0 \le t < x_0 + \delta \quad \text{and} \quad a \le x < t \le b,$$

we have, by 18.11(v),

$$\begin{aligned} \left| \frac{F(t) - F(s)}{t - s} - f(x_0) \right| &= \left| \frac{\int_a^s f(u)\,\mathrm{d}u - \int_a^s f(u)\,\mathrm{d}u}{t - s} - f(x_0) \right| \\ &= \left| \frac{1}{t - s} \int_s^t (f(u) - f(x_0))\,\mathrm{d}u \right| \\ &= \frac{1}{t - s} \left| \int_s^t (f(u) - f(x_0))\,\mathrm{d}u \right| \\ &\le \frac{1}{t - s} \int_s^t |f(u) - f(x_0)|\,\mathrm{d}u \\ &< \frac{1}{t - s} \varepsilon(t - s) = \varepsilon \end{aligned}$$

so it follows that $F'(x_0) = f(x_0)$. ∎

Theorem 18.17 (Fundamental theorem of calculus)**.** *Suppose $f \in \mathcal{R}(\alpha)$, and there exists a differentiable function F on $[a, b]$ such that $F' = f$. Then*

$$\int_a^b f(x)\,\mathrm{d}x = F(b) - F(a). \tag{18.3}$$

Proof. Let $\varepsilon > 0$ be given. Choose a partition $P = \{x_0, \dots, x_n\}$ of $[a, b]$ such that $U(f; P) - L(f; P) < \varepsilon$. By the mean value theorem, there exist $t_i \in [x_{i-1}, x_i]$ such that

$$\begin{aligned} F(x_i) - F(x_{i-1}) &= F'(t_i)\Delta x_i \\ &= f(t_i)\Delta x_i. \end{aligned}$$

Thus

$$\sum_{i=1}^{n} f(t_i)\Delta x_i = F(b) - F(a).$$

Then by 18.6,

$$\left| F(b) - F(a) - \int_a^b f(x)\,\mathrm{d}x \right| = \left| \sum_{i=1}^{n} f(t_i)\Delta x_i - \int_a^b f(x)\,\mathrm{d}x \right| < \varepsilon.$$

Since this holds for all $\varepsilon > 0$, the proof is complete. □

Lemma 18.18 (Integration by parts)**.** *Suppose F and G are differentiable on $[a, b]$, $F' = f \in \mathcal{R}$ and $G' = g \in \mathcal{R}$. Then*

$$\int_a^b F(x)g(x)\,\mathrm{d}x = F(b)G(b) - F(a)G(a) - \int_a^b f(x)G(x)\,\mathrm{d}x\,. \tag{18.4}$$

Proof. Let $H(x) = F(x)G(x)$. Then apply the fundamental theorem of calculus to H and its derivative. □

18.4 Integration of Vector-valued Functions

Let $f_1, \dots, f_k \colon [a, b] \to \mathbb{R}$, and let $\mathbf{f} = (f_1, \dots, f_k)$ where $\mathbf{f} \colon [a, b] \to \mathbb{R}^k$. We say that $\mathbf{f} \in \mathcal{R}(\alpha)$ if $f_1, \dots, f_k \in \mathcal{R}(\alpha)$. If this is the case, we define

$$\int_a^b \mathbf{f} \, \mathrm{d}\alpha := \left(\int_a^b f_1 \, \mathrm{d}\alpha \,, \dots, \int_a^b f_k \, \mathrm{d}\alpha \right).$$

In other words, we "integrate componentwise", so that $\int \mathbf{f} \, \mathrm{d}\alpha$ is the point in $\mathbb{R}^k$ whose i-th coordinate is $\int f_i \, \mathrm{d}\alpha$.

It is clear that parts (a), (c), and (e) of Theorem 6.12 are valid for these vector-valued integrals; we simply apply the earlier results to each coordinate. The same is true of Theorems 6.17, 6.20, and 6.21. To illustrate, we state the analogue of the fundamental theorem of calculus.

Theorem 18.19. *If $\mathbf{f}, \mathbf{F} \colon [a, b] \to \mathbb{R}^k$, $\mathbf{f} \in \mathcal{R}(\alpha)$, and $\mathbf{F}' = \mathbf{f}$. Then*

$$\int_a^b \mathbf{f}(t) \, \mathrm{d}t = \mathbf{F}(b) - \mathbf{F}(a). \tag{18.5}$$

The analogue of Theorem 6.13(b) offers some new features, however, at least in its proof.

Lemma 18.20 (Triangle inequality)**.** *Let $\mathbf{f} : [a, b] \to \mathbb{R}^k$, $\mathbf{f} \in \mathcal{R}(\alpha)$ where α is monotonically increasing on $[a, b]$. Then $|\mathbf{f}| \in \mathcal{R}(\alpha)$, and*

$$\left\| \int_a^b \mathbf{f} \, \mathrm{d}\alpha \right\| \le \int_a^b \|\mathbf{f}\| \, \mathrm{d}\alpha \,.$$

Proof. If $f_1, \dots, f_k$ are the components of $\mathbf{f}$, then

$$\|\mathbf{f}\| = \left({f_1}^2 + \cdots + {f_k}^2 \right)^{1/2}.$$

By 18.10, each of the functions ${f_i}^2 \in \mathcal{R}(\alpha)$, so their sum ${f_1}^2 + \cdots + {f_k}^2 \in \mathcal{R}(\alpha)$.

Since x^2 is a continuous function of x, Theorem 4.17 shows that the square-root function is continuous on $[0, M]$, for every real M. If we apply Theorem 6.11 once more, (41) shows that $\|\mathbf{f}\| \in \mathcal{R}(\alpha)$.

Let $\mathbf{y} = (y_1, \dots, y_k)$, where $y_i = \int f_i \, \mathrm{d}\alpha$. Then we have $\mathbf{y} = \int \mathbf{f} \, \mathrm{d}\alpha$, and

$$\|\mathbf{y}\|^2 = \sum_{i=1}^k {y_i}^2 = \sum_{i=1}^k \left(y_i \int f_i \, \mathrm{d}\alpha \right) = \int \left(\sum_{i=1}^k y_i f_i \right) \mathrm{d}\alpha \,.$$

By the Cauchy–Schwarz inequality,

$$\sum_{i=1}^k y_i f_i(t) \le \|\mathbf{y}\| \|\mathbf{f}(t)\| \quad (a \le t \le b);$$

hence Theorem 6.12(b) implies

$$\|\mathbf{f}\|^2 \le \|\mathbf{y}\| \int \|\mathbf{f}\| \, \mathrm{d}\alpha \,.$$

If $\mathbf{y} = \mathbf{0}$, (40) is trivial. If $\mathbf{y} \neq \mathbf{0}$, division of (43) by $\|\mathbf{y}\|$ gives (40). $\square$

to do

18.5 Rectifiable Curves

Definition 18.21 (Curve)**.** A ***curve*** in $\mathbb{R}^k$ is a continuous mapping $\gamma\colon [a,b] \to \mathbb{R}^k$. If γ is bijective, γ is called an *arc*. If $\gamma(a) = \gamma(b)$, γ is said to be a *closed curve*.

The case $k = 2$ (i.e., the case of plane curves) is of considerable importance in the study of analytic functions of a complex variable.

Remark. Note that we define a curve to be a mapping, not a point set. Of course, with each curve γ in $\mathbb{R}^k$ there is associated a subset of $\mathbb{R}^k$, namely the range of γ, but different curves may have the same range.

For each partition $P = \{x_0, \dots, x_n\}$ of $[a,b]$ and each curve γ on $[a,b]$, define

$$\Lambda(\gamma; P) := \sum_{i=1}^{n} |\gamma(x_i) - \gamma(x_{i-1})|.$$

The i-th term in this sum is the distance (in $\mathbb{R}^k$) between the points $\gamma(x_{i-1})$ and $\gamma(x_i)$, Hence $\Lambda(\gamma; P)$ is the length of a polygonal path with vertices at $\gamma(x_0), \gamma(x_1), \dots, \gamma(x_n)$, in this order. As our partition becomes finer and finer, this polygon approaches the range of γ more and more closely.

insert figure

Definition 18.22. The ***total variation*** (or *length*) of γ is

$$\Lambda(\gamma) := \sup_{P \in \mathcal{P}[a,b]} \Lambda(\gamma; P).$$

We say γ is ***rectifiable*** if $\Lambda(\gamma) < \infty$.

The next result gives a formula for calculating the length of a rectifiable curve that is continuously differentiable.

Proposition 18.23. *If γ is a continuously differentiable curve on $[a,b]$, then γ is rectifiable, and*

$$\Lambda(\gamma) = \int_a^b |\gamma'(t)| \, \mathrm{d}t\,. \tag{18.6}$$

Proof. If $a \le x_{i-1} < x_i \le b$, then

$$|\gamma(x_i) - \gamma(x_{i-1})| = \left| \int_{x_{i-1}}^{x_i} \gamma'(t) \, \mathrm{d}t \right| \le \int_{x_{i-1}}^{x_i} |\gamma'(t)| \, \mathrm{d}t\,.$$

Hence, for every partition P of $[a,b]$, taking the sum on both sides gives

$$\Lambda(\gamma; P) \le \int_a^b |\gamma'(t)| \, \mathrm{d}t$$

and taking sup gives

$$\Lambda(\gamma) \le \int_a^b |\gamma'(t)| \, \mathrm{d}t\,.$$

We now prove the opposite inequality. Since γ' is (continuous and thus) uniformly continuous on $[a, b]$, fix $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|s - t| < \delta \implies |\gamma'(s) - \gamma'(t)| < \varepsilon.$$

Let $P = \{x_0, \ldots, x_n\}$ be a partition of $[a, b]$, with $\Delta x_i < \delta$ for all i. If $t \in [x_{i-1}, x_i]$, it follows that

$$|\gamma'(t)| \le |\gamma'(x_i)| + \varepsilon.$$

Hence

$$\begin{aligned}
\int_{x_{i-1}}^{x_i} |\gamma'(t)| \,\mathrm{d}t &\le |\gamma'(x_i)|\Delta x_i + \varepsilon\Delta x_i \\
&= \left| \int_{x_{i-1}}^{x_i} \big(\gamma'(t) + \gamma'(x_i) - \gamma'(t)\big) \,\mathrm{d}t \right| + \varepsilon\Delta x_i \\
&\le \left| \int_{x_{i-1}}^{x_i} \gamma'(t) \,\mathrm{d}t \right| + \left| \int_{x_{i-1}}^{x_i} \big(\gamma'(x_i) - \gamma'(t)\big) \,\mathrm{d}t \right| + \varepsilon\Delta x_i \\
&\le |\gamma(x_i) - \gamma(x_{i-1})| + 2\varepsilon\Delta x_i.
\end{aligned}$$

If we add these inequalities, we obtain

$$\begin{aligned}
\int_a^b |\gamma'(t)| \,\mathrm{d}t &\le \Lambda(\gamma; P) + 2\varepsilon(b - a) \\
&\le \Lambda(\gamma) + 2\varepsilon(b - a).
\end{aligned}$$

Since ε was arbitrary, we must have

$$\int_a^b |\gamma'(t)| \le \Lambda(\gamma).$$

This completes the proof. □

Exercises

Chapter 19

Sequences and Series of Functions

Suppose $f_n : E \subset X \to Y$ is a sequence of functions. In some cases, we shall restrict ourselves to complex-valued functions (take $Y = \mathbb{C}$).

19.1 Pointwise Convergence

A natural extension of convergence of sequences of numbers to sequences of functions is to fix a point $x \in E$, and consider the behaviour of the sequence $(f_n(x))$.

Definition 19.1 (Pointwise convergence)**.** Suppose (f_n) is a sequence of functions, and $(f_n(x))$ converges for every $x \in E$. We say (f_n) ***converges pointwise*** to f on E, denoted by $f_n \to f$, if

$$f(x) = \lim_{n\to\infty} f_n(x) \quad (\forall x \in E).$$

That is, for all $x \in E$,

$$\forall \varepsilon > 0, \quad \exists N \in \mathbb{N}, \quad \forall n \geq N, \quad d\left(f_n(x) - f(x)\right) < \varepsilon.$$

f is called the *limit* (or *limit function*) of (f_n).

Similarly, if $\sum f_n(x)$ converges for every $x \in E$, and if we define

$$f(x) = \sum_{n=1}^{\infty} f_n(x) \quad (\forall x \in E)$$

the function f is called the *sum of the series* $\sum f_n$.

Example 19.2. The sequence of functions $f_n(x) = \frac{x}{n}$ converges pointwise to the zero function $f(x) = 0$.

The main problem which arises is to determine whether important properties of functions are preserved by pointwise convergence. For instance, if f_n are continuous, or differentiable, or integrable, is the same true of the limit function? What are the relations between f_n' and f', say, or between $\int f_n$ and $\int f$?

Example 19.3 (Continuity)**.** For $0 < x < 1$, the sequence of functions $f_n(x) = x^n$ converges pointwise to the function

$$f(x)=\begin{cases}1 & (x=1)\\ 0 & (0\le x<1)\end{cases}$$

Evidently f_n are continuous, but f is discontinuous. Hence

$$\lim_{x\to x_0}\lim_{n\to\infty} f_n(x)\neq\lim_{n\to\infty}\lim_{x\to x_0} f_n(x).$$

Example 19.4 (Differentiability)**.** For $x\in\mathbb{R}$, let

$$f_n(x)=\frac{\sin nx}{\sqrt{n}}\quad (n=1,2,\dots)$$

so

$$f(x)=\lim_{n\to\infty} f_n(x)=0.$$

Then $f'(x)=0$, and

$$f_n'(x)=\sqrt{n}\cos nx,$$

so (f_n') does not converge to f'.

This shows that the limit of the derivative does not equal the derivative of the limit.

Example 19.5 (Integrability)**.** Let

$$f_n(x)=\chi_{[n,n+1]}(x),$$

Then $\displaystyle\int_{\mathbb{R}} f_n(x)\,\mathrm{d}x=1$, so

$$\lim_{n\to\infty}\int_{\mathbb{R}} f_n(x)\,\mathrm{d}x=1.$$

However

$$\int_{\mathbb{R}}\lim_{n\to\infty} f_n(x)\,\mathrm{d}x=\int 0\,\mathrm{d}x=0.$$

This shows that the limit of the integral does not equal the integral of the limit. Thus we may not switch the order of limits.

Pointwise convergence does not preserve many nice properties of functions. Hence, we need a stronger notion of convergence for sequences and series of functions.

19.2 Uniform Convergence

Definition 19.6 (Uniform convergence)**.** We say (f_n) ***converges uniformly*** to f on E, denoted by $f_n \rightrightarrows f$, if

$$\forall \varepsilon > 0, \quad \exists N \in \mathbb{N}, \quad \forall x \in E, \quad \forall n \geq N, \quad d\left(f_n(x) - f(x)\right) < \varepsilon.$$

Similarly, a series of functions $\sum f_n(x)$ converges uniformly on E if the sequence of partial sums (s_n) defined by

$$s_n(x) = \sum_{k=1}^{n} f_k(x)$$

converges uniformly on E.

Intuitively, uniform convergence can be visualised as the sequence of functions (f_n) eventually contained in an ε-tube around f, for sufficiently large n. insert figure

Remark. Uniform convergence is stronger than pointwise convergence, since N is uniform (or "fixed") for all $x \in E$; for pointwise convergence, the choice of N is determined by x.

Uniform convergence implies pointwise convergence, but not the other way around.

Example 19.7. Consider the sequence of functions $f_n(x) = x^n$ defined on $(0,1)$. Then $f_n \to 0$. But $f_n \not\rightrightarrows 0$.

Proof. □

From now on, we shall restrict our focus to sequences of complex-valued functions defined on $E \subset X$, unless stated otherwise.

We say that (f_n) is *uniformly Cauchy* if

$$\forall \varepsilon > 0, \quad \exists N \in \mathbb{N}, \quad \forall x \in E, \quad \forall n, m \geq N, \quad |f_n(x) - f_m(x)| < \varepsilon.$$

The Cauchy criterion for uniform convergence is as follows.

Lemma 19.8 (Cauchy criterion)**.** *$f_n \rightrightarrows f$ on E if and only if (f_n) is uniformly Cauchy.*

Proof.

$\boxed{\implies}$ Suppose $f_n \rightrightarrows f$ on E. Let $\varepsilon > 0$ be given. Then there exists $N \in \mathbb{N}$ such that for all $x \in E$, for all $n \geq N$,

$$|f_n(x) - f(x)| < \frac{\varepsilon}{2}.$$

Then for all $n, m \geq N$,

$$\begin{aligned} |f_n(x) - f_m(x)| &= |(f_n(x) - f(x)) + (f(x) - f_m(x))| \\ &\leq |f_n(x) - f(x)| + |f_m(x) - f(x)| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

$\boxed{\Longleftarrow}$ Suppose that (f_n) is uniformly Cauchy.

Then for every $x \in E$, the sequence $(f_n(x))$ is a Cauchy sequence and thus converges to a limit $f(x)$. Hence by definition, $f_n \to f$ on E. We are left to prove that the convergence is uniform.

Let $\varepsilon > 0$ be given. There exists $N \in \mathbb{N}$ such that for all $n, m \geq N$ and for all $x \in E$,

$$|f_n(x) - f_m(x)| < \varepsilon.$$

Fix n, and let $m \to \infty$. Since $\lim_{m\to\infty} f_m(x) = f(x)$, thus for all $n \geq N$ and for all $x \in E$,

$$|f_n(x) - f(x)| < \varepsilon,$$

which completes the proof. □

Definition 19.9. If $f \in \mathcal{C}(X, \mathbb{C})$, we define the ***suprenum norm*** of f as

$$\|f\| := \sup_{x\in X} |f(x)|.$$

Lemma 19.10. *$\|f\|$ gives a norm on $\mathcal{C}(X, \mathbb{C})$. Then $\mathcal{C}(X, \mathbb{C})$ is a metric space, with metric $d(f, g) = \|f - g\|$.*

Proof. Check that $\|f\|$ satisfies the conditions for a norm:

(i) $|f(x)| \geq 0$ for all $x \in X$, so $\|f\| \geq 0$. It is clear that $\|f\| = 0$ if and only if $f(x) = 0$ for every $x \in X$, that is, only if $f = 0$.

(ii) For all $\lambda \in \mathbb{C}$,
$$\|\lambda f\| = \sup_{x\in X} |\lambda f(x)| = |\lambda| \sup_{x\in X} |f(x)| = |\lambda|\|f\|.$$

(iii) If $h = f + g$, then for all $x \in X$,
$$|h(x)| \leq |f(x)| + |g(x)| \leq \|f\| + \|g\|.$$
Hence taking sup on the left gives $\|f + g\| \leq \|f\| + \|g\|$.

Check conditions for metric space. □

The following result provides another equivalent way to determine uniform convergence.

Lemma 19.11. *$f_n \rightrightarrows f$ on E if and only if $f_n \to f$ on E with respect to the metric of $\mathcal{C}(E, \mathbb{C})$.*

Proof.

$$\begin{aligned}
f_n \to f &\iff \lim_{n\to\infty} \|f_n - f\| = 0 \\
&\iff \lim_{n\to\infty} \left(\sup_{x\in E} |f_n(x) - f(x)|\right) = 0 \\
&\iff \forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, \sup_{x\in E} |f_n(x) - f(x)| < \varepsilon \\
&\iff \forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, \forall x \in E, |f_n(x) - f(x)| < \varepsilon
\end{aligned}$$

which precisely means that $f_n \rightrightarrows f$ on E, by definition.

Note that for the last step, the $\boxed{\Longleftarrow}$ direction is tricky, since the limit can equal ε, so we take $\frac{\varepsilon}{2}$ instead. □

For series, there is a very convenient test for uniform convergence, due to Weierstrass.

Lemma 19.12 (Weierstrass M-test). *Suppose (f_n) is a sequence of complex-valued functions defined on E, and*

$$|f_n(x)| \leq M_n \quad (n = 1, 2, \dots, \; x \in E)$$

If $\sum M_n$ converges, then $\sum f_n$ converges uniformly on E.

Proof. Suppose $\sum M_n$ converges. Let $\varepsilon > 0$ be given, the partial sums of $\sum M_n$ form a Cauchy sequence, so there exists $N \in \mathbb{N}$ such that for all $n \geq m \geq N$,

$$\sum_{k=m}^{n} M_k < \varepsilon.$$

Then considering the partial sums of the series of functions,

$$\left|\sum_{k=m}^{n} f_k(x)\right| \leq \sum_{k=m}^{n} |f_k(x)| \leq \sum_{k=m}^{n} M_k < \varepsilon.$$

By the Cauchy criterion (19.8), we are done. □

Example 19.13.

- The series $\displaystyle\sum_{n=1}^{\infty} \frac{\sin nx}{n^2}$ converges uniformly on $\mathbb{R}$. (Note: this is a Fourier series, we'll see more of these later). That is because

$$\left|\frac{\sin nx}{n^2}\right| \leq \frac{1}{n^2} \quad \text{and} \quad \sum_{n=1}^{\infty} \frac{1}{n^2} \text{ converges.}$$

- The series $\displaystyle\sum_{n=0}^{\infty} \frac{x^n}{n!}$ converges uniformly on any bounded interval. For example take the interval $[-r, r] \subset \mathbb{R}$,

$$\left|\frac{x^n}{n!}\right| \leq \frac{r^n}{n!} \quad \text{and} \quad \sum_{n=1}^{\infty} \frac{r^n}{n!} \text{ converges by the ratio test.}$$

19.3 Properties of Uniform Convergence

We now consider properties preserved by uniform convergence.

19.3.1 Uniform Convergence and Continuity

We prove a more general result.

Proposition 19.14. *Suppose $f_n \rightrightarrows f$ on E. Let $x \in X$ be a limit point of E, and suppose that*

$$\lim_{t \to x} f_n(t) = A_n \quad (n = 1, 2, \dots).$$

Then (A_n) converges, and $\lim_{t \to x} f(t) = \lim_{n \to \infty} A_n$.

In other words, the conclusion is that

$$\lim_{t \to x} \lim_{n \to \infty} f_n(t) = \lim_{n \to \infty} \lim_{t \to x} f_n(t).$$

Proof.

1. We first show that (A_n) converges. Since (f_n) uniformly converges on E, by the Cauchy criterion (19.8), fix $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $n, m \geq N$, $t \in E$,

$$|f_n(t) - f_m(t)| < \varepsilon.$$

Letting $t \to x$, since $\lim_{t \to x} f_n(t) = A_n$, we have that for all $n, m \geq N$,

$$|A_n - A_m| < \varepsilon.$$

Thus (A_n) is a Cauchy sequence and therefore converges, say to A.

2. Next we will show that $\lim_{t \to x} f(t) = A$.

Idea. We want to bound the term $|f(t) - A|$, using terms of known values.

Write

$$|f(t) - A| \leq |f(t) - f_n(t)| + |f_n(t) - A_n| + |A_n - A|. \tag{1}$$

By the uniform convergence of (f_n), there exists $N_1 \in \mathbb{N}$ such that for all $n \geq N_1$,

$$|f(t) - f_n(t)| < \frac{\varepsilon}{3} \quad (t \in E).$$

By the convergence of (A_n), there exists $N_2 \in \mathbb{N}$ such that for all $n \geq N_2$,

$$|A_n - A| < \frac{\varepsilon}{3}.$$

Choose $N = \max\{N_1, N_2\}$ such that the above two inequalities hold simultaneously. Then for this n, since $\lim_{t\to x} f_n(t) = A_n$, we choose an open ball B of x such that if $t \in B \cap E$, $t \neq x$, then

$$|f_n(t) - A_n| < \frac{\varepsilon}{3}.$$

Substituting the above inequalities into (1) gives

$$|f(t) - A| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

provided $t \in B \cap E$, $t \neq x$. This is equivalent to $\lim_{t\to x} f(t) = A$.

□

An immediate important corollary is that uniform convergence preserves continuity.

Corollary 19.15. *Suppose (f_n) are continuous on E, and $f_n \rightrightarrows f$ on E. Then f is continuous on E.*

Proof. By continuity of f_n,

$$\lim_{t\to x} f_n(t) = f_n(x).$$

Then

$$\lim_{t\to x} f(t) = \lim_{t\to x}\left(\lim_{n\to\infty} f_n(t)\right) = \lim_{n\to\infty}\left(\lim_{t\to x} f_n(t)\right) = \lim_{n\to\infty} f_n(x) = f(x),$$

which precisely means that f is continuous on E. □

Remark. The converse is not true; for instance, the sequence of functions $f_n : (0,1) \to \mathbb{R}$ defined by $f_n(x) = x^n$ converges to the zero function, which is continuous, but the convergence is not uniform.

Let us see that we can have extra conditions such that the converse of the previous result is true.

Proposition 19.16 (Dini's theorem)**.** *Suppose K is compact, and (f_n) is a sequence of continuous functions on K, $f_n \to f$ on K, and (f_n) is monotonically decreasing:*

$$f_n(x) \geq f_{n+1}(x) \quad (n = 1, 2, \dots).$$

Then $f_n \rightrightarrows f$ on K.

Remark. The compactness in the hypotheses is necessary; for instance, on $(0,1)$ define $f_n(x) = \dfrac{1}{nx+1}$. Then $f_n(x) \to 0$ monotonically in $(0,1)$, but the convergence is not uniform.

Proof. Let $g_n = f_n - f$. Then g_n is continuous, $g_n \to 0$, and $g_n \geq g_{n+1} \geq 0$. We have to prove that $g_n \rightrightarrows 0$ on K.

Let $\varepsilon > 0$ be given. For $n = 1, 2, \dots$, let

$$K_n = \{x \in K \mid g_n(x) \geq \varepsilon\}.$$

Since g_n is continuous, and $\{g_n(x) \mid g_n(x) \geq \varepsilon\}$ is closed, by 16.13, its pre-image K_n is closed. Since K_n is a closed subset of a compact set K, by 14.39, K_n is compact.

Since $g_n \geq g_{n+1}$, we have $K_n \supset K_{n+1}$. Fix $x \in K$. Since $g_n(x) \to 0$, we see that $x \notin K_n$ if n is sufficiently large. Thus $x \notin \bigcap_{n=1}^{\infty} K_n$. In other words, $\bigcap_{n=1}^{\infty} K_n = \emptyset$. Hence $K_N = \emptyset$ for some N (by the converse of Cantor's intersection theorem). It follows that for all $x \in K$ and for all $n \geq N$,

$$0 \leq g_n(x) < \varepsilon.$$

Therefore $g_n \rightrightarrows 0$ on K, as desired. □

Lemma 19.17. *$\mathcal{C}(X, \mathbb{C})$ is a complete metric space.*

Proof. Let (f_n) be a Cauchy sequence in $\mathcal{C}(X, \mathbb{C})$. Then fix $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $n, m \geq N$,

$$\|f_n - f_m\| < \varepsilon.$$

By the Cauchy criterion (19.8), $f_n \rightrightarrows f$ for some $f : X \to \mathbb{C}$. We now need to show that $f \in \mathcal{C}(X, \mathbb{C})$; that is, f is continuous and bounded.

- f is continuous by 19.15.
- f is bounded, since there is an n such that $|f(x) - f_n(x)| < 1$ for all $x \in X$, and f_n is bounded.

Hence $f \in \mathcal{C}(X, \mathbb{C})$, and since $f_n \rightrightarrows f$ on X, we have $\|f - f_n\| \to 0$ as $n \to \infty$. □

19.3.2 Uniform Convergence and Integration

The next result states that the limit and integral can be interchanged.

Proposition 19.18. *Suppose $f_n \rightrightarrows f$ on $[a, b]$, $f_n \in \mathcal{R}(\alpha)$ and $\alpha \nearrow$. Then $f \in \mathcal{R}(\alpha)$, and*

$$\lim_{n\to\infty} \int_a^b f_n \, \mathrm{d}\alpha = \int_a^b f \, \mathrm{d}\alpha \,. \tag{19.1}$$

Proof. It suffices to prove this for real-valued f_n. Let

$$\varepsilon_n = \sup_{x\in[a,b]} |f_n(x) - f(x)|.$$

Then $|f_n - f| \le \varepsilon$, so

$$f_n - \varepsilon_n \le f \le f_n + \varepsilon_n,$$

so that the upper and lower integrals of f satisfy

$$\int_a^b (f_n - \varepsilon_n) \, \mathrm{d}\alpha \le \underline{\int_a^b} f \, \mathrm{d}\alpha \le \overline{\int_a^b} f \, \mathrm{d}\alpha \le \int_a^b (f_n + \varepsilon_n) \, \mathrm{d}\alpha \,.$$

Hence

$$0 \le \overline{\int_a^b} f \, \mathrm{d}\alpha - \underline{\int_a^b} f \, \mathrm{d}\alpha \le 2\varepsilon_n[\alpha(b) - \alpha(a)].$$

Since $f_n \rightrightarrows f$, we see that $\varepsilon_n \to 0$ as $n \to \infty$, the upper and lower integrals of f are equal. Hence $f \in \mathcal{R}(\alpha)$.

We have

$$\begin{aligned}
\left| \int_a^b f_n \, \mathrm{d}\alpha - \int_a^b f \, \mathrm{d}\alpha \right| &= \left| \int_a^b f_n - f \, \mathrm{d}\alpha \right| \\
&\le \int_a^b |f_n - f| \, \mathrm{d}\alpha \\
&\le [\alpha(b) - \alpha(a)] \sup_{x\in[a,b]} |f_n(x) - f(x)| \\
&= \varepsilon_n[\alpha(b) - \alpha(a)].
\end{aligned}$$

This implies

$$\lim_{n\to\infty} \int_a^b f_n \, \mathrm{d}\alpha = \int_a^b f \, \mathrm{d}\alpha \,.$$

$\square$

Corollary 19.19. *Suppose $f_n \in \mathcal{R}(\alpha)$ and*

$$f(x) = \sum_{n=1}^{\infty} f_n(x)$$

converges uniformly on $[a,b]$. *Then*

$$\int_a^b f \,\mathrm{d}\alpha = \sum_{n=1}^{\infty} \int_a^b f_n \,\mathrm{d}\alpha \,.$$

In other words, we can swap the integral and sum, such that the series may be integrated term by term.

Proof. Consider the sequence of partial sums

$$f_n(x) = \sum_{k=1}^{n} f_k(x) \quad (n = 1, 2, \dots).$$

It follows $f_n \in \mathcal{R}(\alpha)$ and $f_n \rightrightarrows f$. Apply above theorem to (f_n) and the conclusion follows. $\square$

Example 19.20. Let us show how to integrate a Fourier series:

$$\int_0^x \sum_{n=1}^{\infty} \frac{\cos nt}{n^2} \,\mathrm{d}t = \sum_{n=1}^{\infty} \int_0^x \frac{\cos nt}{n^2} \,\mathrm{d}t = \sum_{n=1}^{\infty} \frac{\sin nx}{n^3}.$$

19.3.3 Uniform Convergence and Differentiation

The next result shows that the process of limit and differentiation can be interchanged.

Proposition 19.21. *Suppose (f_n) are differentiable on $[a,b]$, and $(f_n(x_0))$ converges for some $x_0 \in [a,b]$. If f_n' converges uniformly on $[a,b]$, then there exists a differentiable f such that $f_n \rightrightarrows f$ on $[a,b]$, and*

$$f'(x) = \lim_{n\to\infty} f_n'(x) \quad (a \le x \le b). \tag{19.2}$$

Proof. We first show that (f_n) converges uniformly on $[a,b]$, then show that the limit f is differentiable, and finally show that (19.2) holds.

1. Let $\varepsilon > 0$ be given. Since $(f_n(x_0))$ converges, $(f_n(x_0))$ is a Cauchy sequence; choose $N \in \mathbb{N}$ such that for all $n, m \ge N$,
$$|f_n(x_0) - f_m(x_0)| < \frac{\varepsilon}{2}.$$
Since (f_n') converges uniformly on $[a,b]$, by 19.8, (f_n') is uniformly Cauchy. Thus
$$|f_n'(x) - f_m'(x)| < \frac{\varepsilon}{2(b-a)} \quad (a \le x \le b).$$
We now apply the mean value theorem (17.20) to the function $f_n - f_m$: for any $x, t \in [a,b]$, if $n, m \ge N$, then
$$|(f_n(x) - f_m(x)) - (f_n(t) - f_m(t))| < \frac{\varepsilon}{2(b-a)}|x-t| \le \frac{\varepsilon}{2} \tag{1}$$
Finally, by the triangle inequality,
$$\begin{aligned}|f_n(x) - f_m(x)| &\le |f_n(x) - f_m(x) - f_n(x_0) + f_m(x_0)| + |f_n(x_0) - f_m(x_0)| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.\end{aligned}$$
This holds true for all $x \in [a,b]$. Hence by 19.8, (f_n) converges uniformly on $[a,b]$.

2. Let
$$f(x) = \lim_{n\to\infty} f_n(x) \quad (a \le x \le b).$$
Fix a point $x \in [a,b]$, and let
$$\phi_n(t) = \frac{f_n(t) - f_n(x)}{t-x}, \quad \phi(t) = \frac{f(t) - f(x)}{t-x} \quad (a \le t \le b,\ t \ne x).$$

Idea. To show that f is differentiable, we need to show that $\lim_{t\to x} \phi(t)$ exists.

Note that since f_n are differentiable, we have
$$\lim_{t\to x} \phi_n(t) = f_n'(x) \quad (n = 1, 2, \dots).$$

By (1), for all $n, m \geq N$,

$$\begin{aligned}
|\phi_n(t) - \phi_m(t)| &= \frac{1}{|t-x|} \left|(f_n(t) - f_n(x)) - (f_m(t) - f_m(x))\right| \\
&= \frac{1}{|x-t|} \left|(f_n(x) - f_m(x)) - (f_n(t) - f_m(t))\right| \\
&< \frac{1}{\cancel{|x-t|}} \cdot \frac{\varepsilon}{2(b-a)} \cancel{|x-t|} = \frac{\varepsilon}{2(b-a)},
\end{aligned}$$

so (ϕ_n) converges uniformly, for $t \neq x$. Since (f_n) converges to f, we conclude that

$$\lim_{n\to\infty} \phi_n(t) = \lim_{n\to\infty} \frac{f_n(t) - f_n(x)}{t-x} = \frac{f(t) - f(x)}{t-x} = \phi(t)$$

uniformly for $a \leq t \leq b$, $t \neq x$.

Applying 19.14 to (ϕ_n), we obtain

$$\lim_{t\to x} \phi(t) = \lim_{t\to x} \lim_{n\to\infty} \phi_n(t) = \lim_{n\to\infty} \lim_{t\to x} \phi_n(t) = \lim_{n\to\infty} f_n'(x),$$

which is precisely (19.2).

□

Example 19.22 (Weierstrass function)**.** We will now construct a continuous nowhere differentiable function on $\mathbb{R}$. Define

$$\phi(x) = |x| \quad (-1 \leq x \leq 1).$$

We extend the definition of $\phi(x)$ to all of $\mathbb{R}$ by making ϕ 2-periodic: $\phi(x) = \phi(x+2)$. Then $\phi : \mathbb{R} \to \mathbb{R}$ is continuous as $|\phi(x) - \phi(y)| \leq |x-y|$ (not hard to prove).

Let the *Weierstrass function* be defined as

$$f(x) = \sum_{n=0}^{\infty} \left(\frac{3}{4}\right)^n \phi(4^n x).$$

Claim. The Weierstrass function is continuous and nowhere differentiable on $\mathbb{R}$.

- Since $\sum \left(\frac{3}{4}\right)^n$ converges, and $|\phi(x)| \leq 1$ for all $x \in \mathbb{R}$, by the Weierstrass M-test, $f(x)$ converges uniformly and hence is continuous.

- Fix $x \in \mathbb{R}$ and $m \in \mathbb{Z}^+$, and define

$$\delta_m = \pm\frac{1}{2} \cdot 4^{-m},$$

where the sign is chosen in such a way so that there is no integer between $4^m x$ and $4^m(x+\delta_m)$, which can be done since $4^m|\delta_m| = \frac{1}{2}$. Define

$$\gamma_n = \frac{\phi\left(4^n(x+\delta_m)\right) - \phi(4^n x)}{\delta_m}.$$

If $n > m$, then as $4^n\delta_m$ is an even integer. Then as ϕ is 2-periodic we get that $\gamma_n = 0$.

Furthermore, since there is no integer between $4^m x \pm \frac{1}{2}$ and $4^m x$, we have that

$$\left|\phi\left(4^m x \pm \frac{1}{2}\right) - \phi(4^m x)\right| = \left|\left(4^m x \pm \frac{1}{2}\right) - 4^m x\right| = \frac{1}{2}.$$

Therefore

$$|\gamma_n| = \left|\frac{\phi\left(4^m x \pm \frac{1}{2}\right) - \phi(4^m x)}{\pm\frac{1}{2}\cdot 4^{-m}}\right| = 4^m.$$

Similarly, if $n < m$, since $|\phi(s) - \phi(t)| \leq |s - t|$,

$$|\gamma_n| = \left|\frac{\phi\left(4^n x \pm \frac{1}{2}\cdot 4^{n-m}\right) - \phi(4^n x)}{\pm\frac{1}{2}\cdot 4^{-m}}\right| \leq \left|\frac{\pm\frac{1}{2}\cdot 4^{n-m}}{\pm\frac{1}{2}\cdot 4^{-m}}\right| = 4^n.$$

Finally,

$$\begin{aligned}
\left|\frac{f(x+\delta_m) - f(x)}{\delta_m}\right| &= \left|\sum_{n=0}^{\infty}\left(\frac{3}{4}\right)^n \frac{\phi(4^n(x+\delta_m)) - \phi(4^n x)}{\delta_m}\right| = \left|\sum_{n=0}^{\infty}\left(\frac{3}{4}\right)^n \gamma_n\right| \\
&= \left|\sum_{n=0}^{m}\left(\frac{3}{4}\right)^n \gamma_n\right| \\
&\geq \left|\frac{3}{4}^m \gamma_m\right| - \left|\sum_{n=0}^{m-1}\left(\frac{3}{4}\right)^n \gamma_n\right| \\
&\geq 3^m - \sum_{n=0}^{m-1} 3^n = 3^m - \frac{3^m - 1}{3 - 1} = \frac{3^m + 1}{2}.
\end{aligned}$$

It is obvious that $\delta_m \to 0$ as $m \to \infty$, but $\frac{3^m+1}{2}$ goes to infinity. Hence f cannot be differentiable at x.

19.4 Equicontinuous Families of Functions

We would like an analogue of Bolzano–Weierstrass; that is, every bounded sequence of functions has a convergent subsequence.

Definition 19.23. Suppose (f_n) is a sequence of functions. We say (f_n) is ***pointwise bounded*** on E if for every $x \in E$, the sequence $(f_n(x))$ is bounded; that is,

$$\forall x \in E, \quad \exists M \in \mathbb{R}, \quad \forall n \in \mathbb{N}, \quad |f_n(x)| \leq M.$$

We say (f_n) is ***uniformly bounded*** on E if

$$\exists M \in \mathbb{R}, \quad \forall x \in E, n \in \mathbb{N}, \quad |f_n(x)| \leq M.$$

Lemma 19.24. *Suppose (f_n) is a pointwise bounded sequence of complex-valued functions on a countable set E. Then (f_n) has a subsequence (f_{n_k}) such that $f_{n_k}(x)$ converges for every $x \in E$.*

Proof. We will use a very common and useful diagonal argument.

Arrange the points of E in a sequence (x_i), where $i = 1, 2, \ldots$.

Since (f_n) is pointwise bounded on E, the sequence $(f_n(x_1))_{n=1}^{\infty}$ is bounded. By the Bolzano–Weierstrass theorem, there exists a subsequence, which we denote by $(f_{1,k})_{k=1}^{\infty}$, such that $(f_{1,k}(x_1))_{k=1}^{\infty}$ converges.

Consider the array formed by the sequences $S_1, S_2, \ldots$:

$$\begin{array}{lllll} S_1: & f_{1,1} & f_{1,2} & f_{1,3} & \cdots \\ S_2: & f_{2,1} & f_{2,2} & f_{2,3} & \cdots \\ S_3: & f_{3,1} & f_{3,2} & f_{3,3} & \cdots \\ \vdots & & & & \end{array}$$

and which have the following properties:

(i) S_n is a subsequence of S_{n-1}, for $n = 2, 3, \ldots$

(ii) $(f_{n,k}(x_n))$ converges, as $k \to \infty$ (the boundedness of $(f_n(x_n))$ makes it possible to choose S_n in this way);

(iii) The order in which the functions appear is the same in each sequence; i.e., if one function precedes another in S_1, they are in the same relation in every S_n, until one or the other is deleted. Hence, when going from one row in the above array to the next below, functions may move to the left but never to the right.

We now go down the diagonal of the array; i.e., we consider the sequence

$$S: \quad f_{1,1} \quad f_{2,2} \quad f_{3,3} \quad \cdots$$

By (iii), the sequence S (except possibly its first $n-1$ terms) is a subsequence of S_n, for $n = 1, 2, \ldots$. Hence (ii) implies that $(f_{n,n}(x_i))$ converges, as $n \to \infty$, for every $x_i \in E$. □ to do

Definition 19.25. A family $\mathscr{F}$ of functions $f : E \subset X \to \mathbb{C}$ is ***equicontinuous*** on E if

$$\forall \varepsilon > 0, \quad \exists \delta > 0, \quad \forall x, y \in E, f \in \mathscr{F}, \quad d(x,y) < \delta \implies |f(x) - f(y)| < \varepsilon.$$

Proposition 19.26. *Suppose X is a compact metric space, $f_n \in \mathcal{C}(X, \mathbb{C})$, and (f_n) converges uniformly on X. Then (f_n) is equicontinuous on X.*

Proof. Let $\varepsilon > 0$ be given. Since (f_n) converges uniformly on X, $f_n \to f$ on X with respect to the metric of $\mathcal{C}(X, \mathbb{C})$. Then

$$\lim_{n\to\infty} \|f_n - f\| = 0,$$

i.e., there exists $N \in \mathbb{N}$ such that for all $n \geq N$,

$$\|f_n - f_N\| < \frac{\varepsilon}{3}.$$

Since continuous functions are uniformly continuous on compact sets, f_n are uniformly continuous on K, so there exists $\delta > 0$ such that

$$d(x,y) < \delta \implies |f_i(x) - f_i(y)| < \frac{\varepsilon}{3}$$

for $i = 1, \ldots, N$. If $n \geq N$ and $d(x,y) < \delta$,

$$\begin{aligned} |f_n(x) - f_n(y)| &\leq |f_n(x) - f_N(x)| + |f_N(x) - f_N(y)| + |f_N(y) - f_n(y)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

In conjunction with (43), this proves the theorem. □

We first need the following lemma.

Lemma 19.27. *A compact metric space X contains a countable dense subset.*

Proof. For each $n \in \mathbb{N}$, there exist finitely many balls of radius $\frac{1}{n}$ that cover X (by compactness of X). That is, for every n, there exist finitely many points $x_{n,1}, \ldots, x_{n,k_n}$ such that

$$X = \bigcup_{i=1}^{k_n} B_{\frac{1}{n}}(x_{n,i}).$$

Claim. $S = \{x_{n,i} \mid i = 1, \ldots, k_n\}$ is a countable dense subset of X.

- Since S is a countable union of finite sets, S is countable.

- For every $x \in X$ and every $\varepsilon > 0$, there exists $n \in \mathbb{N}$ such that $\frac{1}{n} < \varepsilon$ and an $x_{n,i} \in S$ such that

$$x \in B_{\frac{1}{n}}(x_{n,i}) \subset B_\varepsilon(x_{n,i}).$$

  Hence $x \in \overline{S}$, so $\overline{S} = X$ and therefore S is dense.

□

We can now prove the very useful Arzelà–Ascoli theorem about existence of convergent subsequences.

Theorem 19.28 (Arzelà–Ascoli theorem)**.** *Suppose X is compact, $f_n \in \mathcal{C}(X, \mathbb{C})$, and (f_n) is pointwise bounded and equicontinuous on X. Then (f_n) is uniformly bounded on X, and contains a uniformly convergent subsequence.*

Proof. Let us first show that the sequence is uniformly bounded. By equicontinuity, there exists $\delta > 0$ such that

$$B_\delta(x) \subset f_n^{-1}\left(B_1(f_n(x))\right) \quad (x \in X).$$

Since X is compact, there exist finitely many points $x_1, \ldots, x_k$ such that

$$X = \bigcup_{j=1}^{k} B_\delta(x_j).$$

Since (f_n) is pointwise bounded, there exist $M_1, \ldots, M_k$ such that

$$|f_n(x_j)| \leq M_j \quad (j = 1, \ldots, k)$$

for all n. Let $M = 1 + \max\{M_1, \ldots, M_k\}$. Now given any $x \in X$, $x \in B_\delta(x_j)$ for some $1 \leq j \leq k$. Therefore, for all n we have $x \in f_n^{-1}\left(B_1(f_n(x_j))\right)$ or in other words

$$|f_n(x) - fn(x_j)| < 1.$$

By reverse triangle inequality,

$$|f_n(x)| < 1 + |fn(x_j)| \leq 1 + M_j \leq M$$

Since x was arbitrary, (f_n) is uniformly bounded.

Next, pick a countable dense set S. By Theorem 7.23, there exists a subsequence (f_{n_j}) that converges pointwise on S. Write $g_j = f_{n_j}$ for simplicity. Note that (g_n) is equicontinuous.

Let $\varepsilon > 0$ be given, then pick $\delta > 0$ such that for all $x \in X$,

$$B_\delta(x) \subset g_n^{-1}\left(B_{\frac{\varepsilon}{3}}(g_n(x))\right).$$

By density of S, every $x \in X$ is in some $B_\delta(y)$ for some $y \in S$, and by compactness of X, there is a finite subset $\{x_1, \ldots, x_k\}$ of S such that

$$X = \bigcup_{j=1}^{k} B_\delta(x_j).$$

Now as there are finitely many points and we know that (g_n) converges pointwise on S, there exists $N \in \mathbb{N}$ such that for all $n, m \geq N$,

$$|g_n(x_j) - g_m(x_j)| < \frac{\varepsilon}{3} \quad (j = 1, \ldots, k).$$

Let $x \in X$ be arbitrary. There is some i such that $x \in B_\delta(x_i)$ and so we have for all $i \in \mathbb{N}$,

$$|g_i(x) - g_i(x_j)| < \frac{\varepsilon}{3}$$

and so $n, m \geq N$ that

$$\begin{aligned}|g_n(x) - g_m(x)| &\leq |g_n(x) - g_n(x_j)| + |g_n(x_j) - g_m(x_j)| + |g_m(x_j) - g_m(x)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.\end{aligned}$$

□

Corollary 19.29. *Suppose X is a compact metric space. Let $S \subset \mathcal{C}(X, \mathbb{C})$ be a closed, bounded and equicontinuous set. Then S is compact.*

Corollary 19.30. *Suppose (f_n) is a sequence of differentiable functions on $[a, b]$, (f_n') is uniformly bounded, and there exists $x_0 \in [a, b]$ such that $(f_n(x_0))$ is bounded. Then there exists a uniformly convergent subsequence (f_{n_k}).*

19.5 Stone–Weierstrass Approximation Theorem

Perhaps surprisingly, even a very badly behaving continuous function is really just a uniform limit of polynomials. We cannot really get any "nicer" as a function than a polynomial.

19.5.1 Weierstrass's Version

Theorem 19.31 (Weierstrass approximation theorem)**.** *If $f :: [a,b] \to \mathbb{C}$ is continuous, there exists a sequence of polynomials (P_n) such that $P_n \rightrightarrows f$ on $[a,b]$.*
If f is real, then P_n may be taken real.

Proof. WLOG assume that $[a,b]=[0,1]$. We may also assume that $f(0)=f(1)=0$. For if the theorem is proved for this case, consider

$$g(x)=f(x)-f(0)-x[f(1)-f(0)] \quad (0\le x\le 1).$$

Here $g(0)=g(1)=0$, and if g can be obtained as the limit of a uniformly convergent sequence of polynomials, it is clear that the same is true for f, since $f-g$ is a polynomial.

Furthermore, we define $f(x)$ to be zero for x outside $[0,1]$. Then f is uniformly continuous on the whole line.

Let

$$Q_n(x)=c_n(1-x^2)^n \quad (n=1,2,\dots),$$

where c_n is chosen such that

$$\int_{-1}^{1} Q_n(x)\,\mathrm{d}x=1 \quad (n=1,2,\dots).$$

We need some information about the order of magnitude of c_n. Since

$$\begin{aligned}\int_{-1}^{1}(1-x^2)^n\,\mathrm{d}x &= 2\int_0^1 (1-x^2)^n\,\mathrm{d}x\\ &\ge 2\int_0^{\frac{1}{\sqrt{n}}}(1-x^2)^n\,\mathrm{d}x\\ &\ge 2\int_0^{\frac{1}{\sqrt{n}}}(1-nx^2)\,\mathrm{d}x\\ &=\frac{4}{3\sqrt{n}}\\ &>\frac{1}{\sqrt{n}},\end{aligned}$$

it follows from (48) that

$$c_n<\sqrt{n}.$$

The inequality $(1-x^2)^n\ge 1-nx^2$ which we used above is easily shown to be true by considering the function

$$(1-x^2)^n-1+nx^2$$

which is zero at $x = 0$ and whose derivative is positive in $(0, 1)$.

For any $\delta > 0$, (49) implies

$$Q_n(x) \leq \sqrt{n}(1 - \delta^2)^n \quad (\delta \leq |x| \leq 1),$$

so that $Q_n \rightrightarrows 0$ in $\delta \leq |x| \leq 1$.

Now let

$$P_n(x) = \int_{-1}^{1} f(x + t) Q_n(t)\,\mathrm{d}t \quad (0 \leq x \leq 1).$$

Our assumptions about f show, by a simple change of variable, that

$$P_n(x) = \int_{-x}^{1-x} f(x + t) Q_n(t)\,\mathrm{d}t = \int_0^1 f(t) Q_n(t - x)\,\mathrm{d}t\,,$$

and the last integral is clearly a polynomial in x. Thus (P_n) is a sequence of polynomials, which are real if f is real.

Given $\varepsilon > 0$, we choose $\delta > 0$ such that

$$|y - x| < \delta \implies |f(y) - f(x)| < \frac{\varepsilon}{2}.$$

Let $M = \sup |f(x)|$, Using (48), (50), and the fact that $Q_n(x) \geq 0$, we see that for $0 \leq x \leq 1$,

$$\begin{aligned}
|P_n(x) - f(x)| &= \left| \int_{-1}^{1} [f(x + t) - f(x)] Q_n(t)\,\mathrm{d}t \right| \\
&\leq \int_{-1}^{1} |f(x + t) - f(x)| Q_n(t)\,\mathrm{d}t \\
&\leq 2M \int_{-1}^{-\delta} Q_n(t)\,\mathrm{d}t + \frac{\varepsilon}{2} \int_{-\delta}^{\delta} Q_n(t)\,\mathrm{d}t + 2M \int_{\delta}^{1} Q_n(t)\,\mathrm{d}t \\
&\leq 4M\sqrt{n}(1 - \delta^2)^n + \frac{\varepsilon}{2} \\
&< \varepsilon
\end{aligned}$$

for all large enough n, which proves the theorem. □

Think about the consequences of the theorem. If you have any property that gets preserved under uniform convergence and it is true for polynomials, then it must be true for all continuous functions.

Let us note an immediate application of the Weierstrass theorem. We have already seen that countable dense subsets can be very useful.

Corollary 19.32. *The metric space $C([a, b], \mathbb{C})$ contains a countable dense subset.*

Corollary 19.33. *For every interval $[-a, a]$, there exists a sequence of real polynomials P_n such that $P_n(0) = 0$ and*

$$\lim_{n\to\infty} P_n(x) = |x|$$

uniformly on $[-a, a]$.

19.5.2 Algebra of Functions

We shall now isolate those properties of the polynomials which make the Weierstrass theorem possible.

Definition 19.34. A family $\mathscr{A}$ of complex-valued functions $f : X \to \mathbb{C}$ is an ***algebra*** if, for all $f, g \in \mathscr{A}, c \in \mathbb{C}$,

(i) $f + g \in \mathscr{A}$; (closed under addition)

(ii) $fg \in \mathscr{A}$; (closed under multiplication)

(iii) $cf \in \mathscr{A}$. (closed under scalar multiplication)

If we talk of an algebra of real-valued functions, then of course we only need the above to hold for $c \in \mathbb{R}$.
$\mathscr{A}$ is *uniformly closed* if the limit of every uniformly convergent sequence in $\mathscr{A}$ is also in $\mathscr{A}$.
Let $\mathscr{B}$ be the set of all limits of uniformly convergent sequences in $\mathscr{A}$. Then $\mathscr{B}$ is the *uniform closure* of $\mathscr{A}$.

Example 19.35.

- $\mathcal{C}(X, Y)$ is an algebra of functions.

Proposition 19.36. *Let $\mathscr{B}$ be the uniform closure of an algebra $\mathscr{A}$ of bounded functions. Then $\mathscr{B}$ is a uniformly closed algebra.*

Now let us distill the right properties of polynomials that were sufficient for an approximation theorem.

Definition 19.37. Let $\mathscr{A}$ be a family of functions defined on X.
We say $\mathscr{A}$ *separates points* if for every $x, y \in X$, with $x \neq y$ there exists $f \in \mathscr{A}$ such that $f(x) \neq f(y)$.
We say $\mathscr{A}$ *vanishes at no point* if for every $x \in X$ there exists $f \in \mathscr{A}$ such that $f(x) \neq 0$.

Example 19.38.

Proposition 19.39. *Suppose $\mathscr{A}$ is an algebra of functions on X, that separates points and vanishes at no point. Suppose x, y are distinct points of X and $c, d \in \mathbb{C}$. Then there exists $f \in \mathscr{A}$ such that*

$$f(x) = c, \quad f(y) = d.$$

19.5.3 The Theorem

We now have all the material needed for Stone's generalisation of the Weierstrass theorem.

Theorem 19.40 (Stone–Weierstrass approximation theorem)**.** *Let X be a compact metric space and $\mathscr{A}$ an algebra of real-valued continuous functions on X, such that $\mathscr{A}$ separates points and vanishes at no point. Then the uniform closure of $\mathscr{A}$ is all of $\mathcal{C}(X, \mathbb{R})$.*

Exercises

Chapter 20

Some Special Functions

20.1 Power Series

Definition 20.1. Given a sequence (c_n) of complex numbers, a ***power series*** takes the form

$$\sum_{n=0}^{\infty} c_n z^n,$$

where $z \in \mathbb{C}$; the numbers c_n are called the *coefficients* of the series.

The convergence of $\sum c_n z^n$ depends on the choice of z (we would expect that a power series will be more likely to converge for small $|z|$ than for large $|z|$). More specifically, there is a "circle of convergence", where $\sum c_n z^n$ converges if z is in the interior of the circle, and diverges if z is in the exterior.

Lemma 20.2 (Cauchy–Hadamard theorem)**.** *Given the power series $\sum c_n z^n$, let*

$$\alpha = \limsup_{n\to\infty} \sqrt[n]{|c_n|}, \qquad R = \frac{1}{\alpha}. \tag{20.1}$$

(If $\alpha = 0$, $R = +\infty$; if $\alpha = +\infty$, $R = 0$.) Then

(i) If $|z| < R$, $\sum c_n z^n$ converges.

(ii) If $|z| > R$, $\sum c_n z^n$ diverges.

R is called the ***radius of convergence*** of $\sum c_n (z-a)^n$; the ***disk of convergence*** for the power series is

$$D_R(a) := \{z \in \mathbb{C} : |z| < R\}.$$

Proof. Let $a_n = c_n z^n$. We apply the root test:

$$\limsup_{n\to\infty} \sqrt[n]{|a_n|} = \limsup_{n\to\infty} \sqrt[n]{|c_n z^n|} = |z| \limsup_{n\to\infty} \sqrt[n]{|c_n|} = \frac{|z|}{R}.$$

(i) If $|z| < R$, then $\limsup_{n\to\infty} \sqrt[n]{|a_n|} < 1$. By the root test, $\sum c_n z^n$ converges absolutely and thus 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.

□

Remark. On the boundary of the disc of convergence, $|z| = R$, the situation is more delicate as one can have either convergence or divergence.

In the previous result, we have shown that the radius of convergence can be found by using the root test. We can also find it using the ratio test (which is easier to compute).

Lemma 20.3. *If $\sum c_n z^n$ has radius of convergence R, then*

$$R = \lim_{n\to\infty} \left| \frac{c_n}{c_{n+1}} \right|,$$

if this limit exists.

Proof. By the ratio test, $\sum c_n z^n$ converges if

$$\lim_{n\to\infty} \left| \frac{c_{n+1} z^{n+1}}{c_n z^n} \right| < 1.$$

This is equivalent to

$$|z| < \frac{1}{\lim_{n\to\infty} \left| \frac{c_{n+1}}{c_n} \right|} = \lim_{n\to\infty} \left| \frac{c_n}{c_{n+1}} \right|.$$

□

Proposition 20.4. *Suppose the radius of convergence of $\sum c_n z^n$ is 1, and suppose $c_0 \geq c_1 \geq c_2 \geq \cdots$, $c_n \to 0$. Then $\sum c_n z^n$ converges at every point on the circle $|z| = 1$, except possibly at $z = 1$.*

Proof. Let

$$a_n = z^n, \quad b_n = c_n.$$

Then the hypothesis of Lemma 15.45 are satisfied, since

$$|A_n| = \left| \sum_{k=0}^{n} z^k \right| = \left| \frac{1 - z^{n+1}}{1 - z} \right| \leq \frac{2}{|1 - z|}$$

if $|z| = 1$, $|z| \neq 1$. □

Definition 20.5. An ***analytic function*** is a function that can be represented by a power series; that is, functions of the form

$$f(x) = \sum_{n=0}^{\infty} c_n x^n$$

or, more generally,

$$f(x) = \sum_{n=0}^{\infty} c_n (x - a)^n.$$

We shall restrict ourselves to real values of x (since we have yet to define complex differentiation). Instead of circles of convergence we shall therefore encounter intervals of convergence.

As a matter of convenience, we shall often take $a = 0$ without any loss of generality. If $\sum c_n x^n$ converges for all $x \in (-R, R)$, for some $R > 0$, we say that f is *expanded in a power series* about the point $x = 0$.

Proposition 20.6. *Suppose $\sum c_n x^n$ converges for $|x| < R$. Let*

$$f(x) = \sum_{n=0}^{\infty} c_n x^n \quad (|x| < R).$$

Then

(i) $\sum c_n x^n$ converges absolutely and uniformly on $(-r, r)$ where $r < R$;

(ii) $f(x)$ is continuous and differentiable on $(-R, R)$, and

$$f'(x) = \sum_{n=1}^{\infty} n c_n x^{n-1} \quad (|x| < R).$$

Proof.

(i) We will show that $\sum c_n x^n$ converges absolutely and uniformly on $[-R+\varepsilon, R-\varepsilon]$ for all $\varepsilon > 0$.

Idea. Weierstrass M-test.

Let $\varepsilon > 0$ be given. For $|x| \le R - \varepsilon$, notice that we have

$$|c_n x^n| \le |c_n|(R-\varepsilon)^n \quad (n = 1, 2, \dots).$$

Consider the series

$$\sum |c_n|(R-\varepsilon)^n.$$

By Lemma 20.2, which states that every power series converges (absolutely) in the interior of its internal of convergence, we have that $\sum |c_n|(R-\varepsilon)^n$ converges.

By the Weierstrass M-test, $\sum c_n x^n$ uniformly converges on $[-R+\varepsilon, R-\varepsilon]$.

(ii) Since $\lim_{n\to\infty} \sqrt[n]{n} = 1$, we have

$$\limsup_{n\to\infty} \sqrt[n]{n|c_n|} = \limsup_{n\to\infty} \sqrt[n]{|c_n|},$$

so the series $\sum_{n=0}^{\infty} c_n x^n$ and $\sum_{n=1}^{\infty} n c_n x^{n-1}$ have the same radius of convergence; thus $\sum_{n=1}^{\infty} n c_n x^{n-1}$ has radius of convergence R.

Idea. Interchange limits and derivatives.

Since $\sum_{n=1}^{\infty} n c_n x^{n-1}$ is a power series, by (i), it converges uniformly in $[-R+\varepsilon, R-\varepsilon]$, for every $\varepsilon > 0$.

Consider the partial sums $f_n(x) = \sum_{k=0}^{n} c_k x^k$; evidently f_n are differentiable on $[-R+\varepsilon, R-\varepsilon]$, and $(f_n(x))$ converges whenever $|x| < R$. Also,

$$f_n'(x) = \sum_{k=1}^{n} k c_k x^{k-1},$$

converge uniformly on $[-R+\varepsilon, R-\varepsilon]$. By Proposition 19.21, for all $|x| < R$,

$$f'(x) = \lim_{n\to\infty} f_n'(x) = \lim_{n\to\infty} \sum_{k=1}^{n} k c_k x^{k-1} = \sum_{n=1}^{\infty} n c_n x^{n-1}.$$

Since f is differentiable on $(-R, R)$, by Lemma 17.2, f is continuous on $(-R, R)$.

□

Corollary 20.7. *f is infinitely differentiable in $(-R, R)$; its derivatives are given by*

$$f^{(k)}(x) = \sum_{n=k}^{\infty} n(n-1)\cdots(n-k+1)c_n x^{n-k}. \tag{20.2}$$

In particular,

$$f^{(k)}(0) = k!c_k \quad (k = 0, 1, 2, \dots). \tag{20.3}$$

Proof. Apply the previous result successively to $f, f', f'', \dots$.

Then plug in $x = 0$. □

Remark. (20.3) is very interesting.

- It shows, on the one hand, that the coefficients of the power series development of f are determined by the values of f and its derivatives at a single point.
- On the other hand, if the coefficients are given, the values of the derivatives of f at the center of the interval of convergence can be read off immediately from the power series.

If the series (3) converges at an endpoint, say at $x = R$, then f is continuous not only in $(-R, R)$, but also at $x = R$, as shown by the following result (for simplicity of notation, we take $R = 1$).

Proposition 20.8 (Abel's theorem)**.** *Suppose $\sum c_n$ converges. Let*

$$f(x) = \sum_{n=0}^{\infty} c_n x^n \quad (-1 < x < 1).$$

Then $f(x)$ is continuous at $x = 1$.

Proof. We want to show that $\lim_{x\to 1} f(x) = \sum_{n=0}^{\infty} c_n$.

Let

$$s_n = c_0 + \cdots + c_n, \quad s_{-1} = 0.$$

Then we write

$$\begin{aligned}\sum_{n=0}^{m} c_n x^n &= \sum_{n=0}^{m}(s_n - s_{n-1})x^n \\ &= \sum_{n=0}^{m} s_n x^n - \sum_{n=1}^{m} s_{n-1}x^n \\ &= \sum_{n=0}^{m} s_n x^n - \sum_{n=0}^{m-1} s_n x^{n+1} \\ &= (1-x)\sum_{n=0}^{m-1} s_n x^n + s_m x^m.\end{aligned}$$

For $|x| < 1$, we let $m \to \infty$ and obtain

$$f(x) = (1-x)\sum_{n=0}^{\infty} s_n x^n.$$

Suppose $s_n \to s$. We will show that $\lim_{x\to 1} f(x) = s$. Fix $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that

$$n \geq N \implies |s_n - s| < \frac{\varepsilon}{2}.$$

Note that for $|x| < 1$, since $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$, we can write

$$(1-x)\sum_{n=0}^{\infty} s x^n = s.$$

If $x > 1 - \delta$, for some suitably chosen $\delta > 0$, we have

$$\begin{aligned}|f(x) - s| &= \left|(1-x)\sum_{n=0}^{\infty}(s_n - s)x^n\right| \\ &= (1-x)\left|\sum_{n=0}^{N}(s_n - s)x^n + \sum_{n=N+1}^{\infty}(s_n - s)x^n\right| \\ &\leq (1-x)\left|\sum_{n=0}^{N}(s_n - s)x^n\right| + (1-x)\sum_{n=N+1}^{\infty}(s_n - s)x^n.\end{aligned}$$

Note that

$$\begin{aligned}(1-x)\left|\sum_{n=N+1}^{\infty}(s_n - s)x^n\right| &\leq (1-x)\sum_{n=N+1}^{\infty}|s_n - s|\,|x|^n \\ &< \frac{\varepsilon}{2}(1-x)\sum_{n=N+1}^{\infty} x^n \\ &= \frac{\varepsilon}{2}(1-x)\frac{x^{N+1}}{1-x} < \frac{\varepsilon}{2}\end{aligned}$$

and

$$\begin{aligned}(1-x)\left|\sum_{n=0}^{N}(s_n-s)x^n\right| &\le (1-x)\sum_{n=0}^{N}|s_n-s|\,|x|^n \\ &< (1-x)\sum_{n=0}^{N}|s_n-s|\end{aligned}$$

which can be bounded by, say, M because there are only finitely many terms in the sum. Choosing $\delta < \frac{\varepsilon}{2M}$ gives

$$(1-x)\sum_{n=0}^{N}|s_n-s| < (1-(1-\delta))\sum_{n=0}^{N}|s_n-s| < \delta M < \frac{\varepsilon}{2}.$$

Hence

$$|f(x)-s| < \frac{\varepsilon}{2}+\frac{\varepsilon}{2} = \varepsilon,$$

and therefore $\lim\limits_{x\to 1} f(x) = s$, as desired. □

We now require a result concerning an inversion in the order of summation.

Proposition 20.9 (Fubini's theorem for sums)**.** *Given a double sequence* (a_{ij}), $i = 1,2,\dots$, $j = 1,2,\dots$, *suppose that*

$$\sum_{j=1}^{\infty}|a_{ij}| = b_i \quad (i=1,2,\dots)$$

and $\sum b_i$ *converges. Then*

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}a_{ij} = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty}a_{ij}. \tag{20.4}$$

Remark. This is analogous to Fubini's theorem for the swapping of double integrals.

Proof. Let E be a countable set:

$$E = \{x_0, x_1, x_2, \dots\},$$

and suppose $x_n \to x_0$ as $n \to \infty$. Define the sequence of functions $f_i : E \to \mathbb{C}$ by

$$\begin{aligned} f_i(x_0) &= \sum_{j=1}^{\infty} a_{ij} \quad (i=1,2,\dots) \\ f_i(x_n) &= \sum_{j=1}^{n} a_{ij} \quad (i,n=1,2,\dots)\end{aligned}$$

See that each f_i is continuous at x_0.

Since $|f_i(x)| \le b_i$ for $x \in E$ (by triangle inequality), and $\sum b_i$ converges, by the Weierstrass M-test, $\sum\limits_{i=1}^{n} f_i(x)$ converges uniformly. Let

$$g(x) = \sum_{i=1}^{\infty} f_i(x) \quad (x \in E).$$

By 7.11, g is continuous at x_0, so

$$\lim_{n\to\infty} g(x_n) = g(x_0).$$

It follows that

$$\begin{aligned}
\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij} &= \sum_{i=1}^{\infty} f_i(x_0) = g(x_0) = \lim_{n\to\infty} g(x_n) \\
&= \lim_{n\to\infty} \sum_{i=1}^{\infty} f_i(x_n) = \lim_{n\to\infty} \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij} \\
&= \lim_{n\to\infty} \sum_{j=1}^{n}\sum_{i=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} a_{ij}.
\end{aligned}$$

□

Theorem 20.10 (Taylor's theorem)**.** *Suppose $\sum c_n x^n$ converges in $|x| < R$, let*

$$f(x) = \sum_{n=0}^{\infty} c_n x^n.$$

If $a \in (-R, R)$, then f can be expanded in a power series about the point $x = a$ which converges in $|x-a| < R - |a|$, and

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n \quad (|x-a| < R - |a|). \tag{20.5}$$

Proof. We have

$$\begin{aligned}
f(x) &= \sum_{n=0}^{\infty} c_n x^n = \sum_{n=0}^{\infty} c_n \left[(x-a)+a\right]^n \\
&= \sum_{n=0}^{\infty} c_n \sum_{m=0}^{n} \binom{n}{m} a^{n-m}(x-a)^m \\
&= \sum_{m=0}^{\infty} \left[\sum_{n=m}^{\infty} \binom{n}{m} c_n a^{n-m}\right](x-a)^m \qquad (1)
\end{aligned}$$

This is the desired expansion about the point $x = a$. We need to show that the swapping of summations in (1) is valid, which is applicable only if $\binom{n}{m} c_n a^{n-m}(x-a)^m$ satisfies Theorem 8.3, i.e.

$$\sum_{n=0}^{\infty}\sum_{m=0}^{n} \left| c_n \binom{n}{m} a^{n-m}(x-a)^m \right|$$

converges. Write

$$\begin{aligned}
&\sum_{n=0}^{\infty}\sum_{m=0}^{n}\left|c_n\binom{n}{m}a^{n-m}(x-a)^m\right| \\
&=\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\binom{n}{m}|c_n||a|^{n-m}|x-a|^m \\
&=\sum_{n=0}^{\infty}|c_n|\left(|x-a|+|a|\right)^n,
\end{aligned}$$

which converges if and only if $|x-a|+|a|<R$.

Finally, the form of the coefficients in (20.5) follows from (20.3): differentiate $f(x)$ repeatedly to obtain

$$\begin{aligned}
f^{(m)}(x) &= \sum_{n=m}^{\infty} c_n n(n-1)\cdots(n-m-1)x^{n-m} \\
&= \sum_{n=m}^{\infty} c_n m!\binom{n}{m}x^{n-m}
\end{aligned}$$

and then plug in $x=a$. □

If two power series converge to the same function in $(-R,R)$, (7) shows that the two series must be identical, i.e., they must have the same coefficients. It is interesting that the same conclusion can be deduced from much weaker hypotheses:

Proposition 20.11. *Suppose $\sum a_n x^n$ and $\sum b_n x^n$ converge in $S=(-R,R)$. Let E be the set of all $x\in S$ such that*

$$\sum_{n=0}^{\infty} a_n x^n = \sum_{n=0}^{\infty} b_n x^n.$$

If E has a limit point in S, then $a_n=b_n$ for $n=0,1,2,\ldots$. Hence (20) holds for all $x\in S$.

Proof. Let $c_n=a_n-b_n$, and let

$$f(x)=\sum_{n=0}^{\infty} c_n x^n \quad (x\in S)$$

We will show that $c_n=0$, so that $f(x)=0$ on E.

Let A be the set of all limit points of E in S, and let B consist of all other points of S. It is clear from the definition of ”limit point” that B is open. Suppose we can prove that A is open. Then A and B are disjoint open sets. Hence they are separated (Definition 2.45). Since $S=A\cup B$, and S is connected, one of A and B must be empty. By hypothesis, A is not empty. Hence B is empty, and $A=S$. Since f is continuous in S, $A\subset E$. Thus $E=S$, and (7) shows that $c_n=0$ for n = 0, 1, 2, ... , which is the desired conclusion.

Thus we have to prove that A is open. If $x_0\in A$, Theorem 8.4 shows that

$$f(x)=\sum_{n=0}^{\infty} d_n(x-x_0)^n \quad (|x-x_0|<R-|x_0|).$$

We claim that $d_n = 0$ for all n. Otherwise, let k be the smallest nonnegative integer such that $d_k \neq 0$. Then

$$f(x) = (x - x_0)^k g(x) \quad (|x - x_0| < R - |x_0|)\,,$$

where

$$g(x) = \sum_{m=0}^{\infty} d_{k+m}(x - x_0)^m.$$

Since g is continuous at x_0 and

$$g(x_0) = d_k \neq 0,$$

there exists $\delta > 0$ such that $g(x) \neq 0$ if $|x - x_0| < \delta$. It follows from (23) that $f(x) \neq 0$ if $0 < |x - x_0| < \delta$. But this contradicts the fact that x_0 is a limit point of E.

Thus $d_n = 0$ for all n, so that $f(x) = 0$ for all x for which (22) holds, i.e., in a neighborhood of x_0. This shows that A is open, and completes the proof.

to do

$\square$

20.1.1 Exponential and Logarithmic Functions

Definition 20.12 (Exponential function). For $z \in \mathbb{C}$, define

$$\exp(z) := \sum_{n=0}^{\infty} \frac{z^n}{n!}. \tag{20.6}$$

Notation. We shall usually replace $\exp(z)$ by the customary shorter expression e^z.

Lemma 20.13. $\exp(z)$ *converges for every* $z \in \mathbb{C}$.

Proof. Let $a_n = \frac{z^n}{n!}$. Then

$$\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| = \lim_{n\to\infty}\left|\frac{z}{n+1}\right| = |z| \lim_{n\to\infty}\frac{1}{n+1} = 0 < 1.$$

By the ratio test, the series converges absolutely for all $z \in \mathbb{C}$, and thus converges for all $z \in \mathbb{C}$. □

Thus $\exp(z)$ has infinite radius of convergence.

The series converges uniformly on every bounded subset of the complex plane. Thus exp is a continuous function.

Lemma 20.14 (Addition formula). *For* $z, w \in \mathbb{C}$,

$$\exp(z + w) = \exp(z)\exp(w). \tag{20.7}$$

Proof. By multiplication of absolutely convergent series,

$$\begin{aligned}
\exp(z)\exp(w) &= \sum_{n=0}^{\infty}\frac{z^n}{n!}\sum_{m=0}^{\infty}\frac{w^m}{m!} \\
&= \sum_{k=0}^{\infty}\left(\frac{z^k}{k!} + \frac{z^{k-1}}{(k-1)!}\frac{w}{1!} + \cdots + \frac{w^k}{k!}\right) \\
&= \sum_{k=0}^{\infty}\frac{1}{k!}\sum_{m+n=k}\binom{k}{n} z^n w^{k-n} \\
&= \sum_{k=0}^{\infty}\frac{1}{k!}(z+w)^k \\
&= \exp(z+w)
\end{aligned}$$

□

An immediate consequence is

$$\exp(z)\exp(-z) = 1 \tag{20.8}$$

for all $z \in \mathbb{C}$. This shows that $\exp(z) \neq 0$ for all $z \in \mathbb{C}$.

Lemma 20.15. *The restriction of* $\exp$ *to* $\mathbb{R}$ *is a monotonically increasing positive function, and*

$$\lim_{x\to\infty} e^x = \infty, \qquad \lim_{x\to-\infty} e^x = 0.$$

Proof. By (20.6), $\exp(x) > 0$ if $x > 0$ (since all the terms are positive); hence (20.8) shows that $\exp(x) > 0$ for all real x.

By (20.6), $\exp(x) \to +\infty$ as $x \to +\infty$; hence (20.8) shows that $\exp(x) \to 0$ as $x \to -\infty$ along the real axis.

By (20.6), $0 < x < y$ implies that $\exp(x) < \exp(y)$; by (20.8), it follows that $\exp(-y) < \exp(-x)$. □

Lemma 20.16. $\exp'(x) = \exp(x)$ *for all* $x \in \mathbb{R}$.

Proof. By the addition formula,

$$\begin{aligned}
\exp'(x) &= \lim_{h\to 0} \frac{\exp(x+h) - \exp(x)}{h} \\
&= \lim_{h\to 0} \frac{\exp(x)\exp(h) - \exp(x)}{h} \\
&= \exp(x) \lim_{h\to 0} \frac{\exp(h) - 1}{h} \\
&= \exp(x).
\end{aligned}$$

□

Since $\exp$ is strictly increasing and differentiable on $\mathbb{R}$, it has an inverse function $\log$ which is also strictly increasing and differentiable and whose domain is $\exp(\mathbb{R}) = \mathbb{R}^+$. Define the ***logarithm function*** $\log$ by

$$\exp(\log(y)) = y \quad (y > 0),$$

or, equivalently, by

$$\log(\exp(x)) = x \quad (x \in \mathbb{R}).$$

Differentiating this using the chain rule, we obtain

$$\log'(\exp(x))\exp(x) = 1.$$

Writing $y = \exp(x)$, this gives us

$$\log'(y) = \frac{1}{y} \quad (y > 0).$$

Taking $x = 0$, we see that $\log(1) = 0$. Hence

$$\log y = \int_1^y \frac{1}{x}\,\mathrm{d}x\,.$$

Lemma 20.17 (Addition formula)**.** *If* $u, v > 0$ *then*

$$\log(uv) = \log(u) + \log(v). \tag{20.9}$$

Proof. Write $u = \exp(x)$, $v = \exp(y)$. Then

$$\begin{aligned} \log(uv) &= \log\left(\exp(x)\exp(y)\right) \\ &= \log\left(\exp(x+y)\right) \\ &= x + y \\ &= \log(u) + \log(v). \end{aligned}$$

$\square$

20.1.2 Trigonometric Functions

Definition 20.18. For $z \in \mathbb{C}$, define

$$\cos z := \frac{e^{iz} + e^{-iz}}{2}, \quad \sin z := \frac{e^{iz} - e^{-iz}}{2i}. \tag{20.10}$$

By (20.6), we obtain the power series

$$\cos z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!}$$
$$\sin z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+1)!}$$

Lemma 20.19 (Euler's identity)**.** *For $z \in \mathbb{C}$,*

$$e^{iz} = \cos z + i \sin z.$$

Proof. This is immediate from (20.10). □

Let us now restrict ourselves to considering real values $x \in \mathbb{R}$. Then $\cos x$ and $\sin x$ are real for real x. By Euler's identity, $\cos x$ and $\sin x$ are the real and imaginary parts, respectively, of e^{ix}.

By (20.8),

$$|e^{ix}|^2 = e^{ix}\overline{e^{ix}} = e^{ix}e^{-ix} = 1,$$

so that

$$|e^{ix}| = 1 \quad (x \in \mathbb{R}).$$

Lemma 20.20 (Derivative of trigonometric functions)**.** *For $x \in \mathbb{R}$,*

$$\cos' x = -\sin x, \quad \sin' x = \cos x.$$

Let x_0 be the smallest positive number such that $\cos x_0 = 0$. This exists, since the set of zeros of a continuous function is closed, and $\cos 0 \neq 0$. We define the number π by

$$\pi := 2x_0.$$

Then $\cos \frac{\pi}{2} = 0$, and (48) shows that $\sin \frac{\pi}{2} = \pm 1$. Since $\cos x > 0$ in $\left(0, \frac{\pi}{2}\right)$, sin is increasing in $\left(0, \frac{\pi}{2}\right)$; hence $\sin \frac{\pi}{2} = 1$.

Lemma 20.21. $\exp$ *is periodic, with period $2\pi i$.*

Proof.

$$\exp\left(\frac{\pi i}{2}\right) = i,$$

and the addition formula gives

$$\exp(\pi i) = -1, \quad \exp(2\pi i) = 1.$$

Hence

$$\exp(z + 2\pi i) = \exp(z) \quad (z \in \mathbb{C}).$$

□

Lemma 20.22.

(i) *C and S are periodic, with period 2π.*

(ii) *If $0 < t < 2\pi$, then $\exp(it) \neq 1$.*

(iii) *If $z \in \mathbb{C}$, $|z| = 1$, there exists a unique $t \in [0, 2\pi)$ such that $\exp(it) = z$.*

20.2 Algebraic Completeness of the Complex Field

We now prove that the complex field is *algebraically complete*; that is, every non-constant polynomial with complex coefficients has a complex root.

Theorem 20.23 (Fundamental Theorem of Algebra)**.** *For $a_i \in \mathbb{C}$, let*

$$P(z) = \sum_{k=0}^{n} a_k z^k$$

where $n \geq 1$, $a_n \neq 0$. Then $P(z) = 0$ for some $z \in \mathbb{C}$.

Proof. WLOG assume $a_n = 1$. Let

$$\mu = \inf |P(z)| \quad (z \in \mathbb{C}).$$

If $|z| = R$, then

$$|P(z)| \geq R^n \left(1 - |a_{n-1}|R^{-1} - \cdots - |a_0|R^{-n}\right).$$

The RHS tends to ∞ as $R \to \infty$. Hence there exists R_0 such that $|P(z)| > \mu$ if $|z| > R_0$. Since $|P|$ is continuous on the closed disk $\overline{D}_{R_0}(0)$, Theorem 4.16 shows that $|P(z_0)| = \mu$ for some z_0.

Claim. $\mu = 0$.

If not, let $Q(z) = \dfrac{P(z+z_0)}{P(z_0)}$. Then Q is a non-constant polynomial, $Q(0) = 1$, and $|Q(z)| \geq 1$ for all z. There is a smallest integer k, $1 \leq k \leq n$ such that

$$Q(z) = 1 + b_k z^k + \cdots + b_n z^n \quad (b_k \neq 0).$$

By Theorem 8.7(d) there is a real θ such that

$$e^{ik\theta} b_k = -|b_k|.$$

If $r > 0$ and $r^k |b_k| < 1$, the above equation implies

$$|1 + b_k r^k e^{ik\theta}| = 1 - r^k |b_k|,$$

so that

$$\left|Q\left(re^{i\theta}\right)\right| \leq 1 - r^k \left(|b_k| - r|b_{k+1}| - \cdots - r^{n-k}|b_n|\right).$$

For sufficiently small r, the expression in braces is positive; hence $|Q(re^{i\theta})| < 1$, a contradiction.

Thus $\mu = 0$, that is, $P(z_0) = 0$. ∎

20.3 Fourier Series

Definition 20.24. A ***trigonometric polynomial*** is a finite sum of the form

$$f(x) = a_0 + \sum_{n=1}^{N} (a_n \cos nx + b_n \sin nx) \quad (x \in \mathbb{R})$$

where $a_0, a_1 \dots, a_N, b_1, \dots, b_N \in \mathbb{C}$.

Using (20.10), we can write the above in the form

$$f(x) = \sum_{n=-N}^{N} c_n e^{inx}$$

for some constants $c_n \in \mathbb{C}$. This is a more convenient form of trigonometric polynomials, which we shall work with.

It is clear that every trigonometric polynomial is periodic, with period 2π.

For non-zero integer n, e^{inx} is the derivative of $\frac{1}{in}e^{inx}$, which also has period 2π. Hence

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{inx}\,\mathrm{d}x = \begin{cases} 1 & (n = 0) \\ 0 & (n = \pm 1, \pm 2, \dots) \end{cases}$$

Definition 20.25. Let $f \in \mathcal{R}[-\pi, \pi]$. The ***Fourier coefficients*** of f are the numbers c_n, defined by

$$c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x) e^{-inx}\,\mathrm{d}x\,.$$

The series

$$\sum_{n=-\infty}^{\infty} c_n e^{inx}$$

formed with the Fourier coefficients is called the ***Fourier series*** of f; in this case we write

$$f \sim \sum_{n=-\infty}^{\infty} c_n e^{inx}.$$

We say f is an L^2 function if $|f|^2$ is Lebesgue integrable. The space of L^2 functions on a set E is denoted by $L^2(E)$. For all $f, g \in L^2(E)$, define the inner product

$$\langle f, g\rangle = \int_E f(x)\overline{g(x)}\,\mathrm{d}x\,.$$

Then the norm of f squared is defined as

$$\|f\|^2 := \langle f, f\rangle = \int_E |f(x)|^2\,\mathrm{d}x\,.$$

We say that f and g are *orthogonal* if $\langle f, g\rangle = 0$.

Definition 20.26. Let (ϕ_n) be a sequence of complex functions on $[a, b]$.

(i) We say (ϕ_n) is an ***orthogonal system*** of functions on $[a, b]$ if $\langle \phi_n, \phi_m \rangle = 0$ for all $n \neq m$.

(ii) We say (ϕ_n) is an ***orthonormal system*** of functions on $[a, b]$ if (ϕ_n) is an orthogonal system, and $\|\phi_n\| = 1$ for all n.

Example 20.27.

- $\left\{ \dfrac{1}{\sqrt{2\pi}} e^{inx} \right\}$ is an orthonormal system on $[-\pi, \pi]$.
- $\left\{ \dfrac{1}{\sqrt{2\pi}}, \dfrac{1}{\sqrt{\pi}} \cos nx, \dfrac{1}{\sqrt{\pi}} \sin nx \right\}$ is an orthonormal system on $[-\pi, \pi]$.

  Proof. We have

$$\begin{aligned}&\int_{-\pi}^{\pi} \cos nx \sin mx \, \mathrm{d}x \\ &= \int_{-\pi}^{\pi} \frac{e^{inx} + e^{-inx}}{2} \frac{e^{imx} - e^{-imx}}{2i} \, \mathrm{d}x \\ &= \int_{-\pi}^{\pi} \frac{e^{i(n+m)x} - e^{i(n-m)x} + e^{-i(n-m)x} - e^{-i(n+m)x}}{4i} \, \mathrm{d}x = 0\end{aligned}$$

  and similarly

$$\begin{aligned}&\int_{-\pi}^{\pi} \cos nx \cos mx \, \mathrm{d}x \\ &= \int_{-\pi}^{\pi} \frac{e^{inx} + e^{-inx}}{2} \frac{e^{imx} + e^{-imx}}{2i} \, \mathrm{d}x \\ &= \int_{-\pi}^{\pi} \frac{e^{i(n+m)x} + e^{i(n-m)x} + e^{-i(n-m)x} + e^{-i(n+m)x}}{4} \, \mathrm{d}x \\ &= \begin{cases} \frac{1}{2} \cdot 2\pi = \pi & (n = m) \\ 0 & (n \neq m) \end{cases}\end{aligned}$$

  $\square$

If (ϕ_n) is an orthonormal system of functions on $[a, b]$, then

$$f \sim \sum_{n=1}^{\infty} c_n \phi_n$$

where $c_n = \langle f, \phi_n \rangle$; we call c_n the n-th Fourier coefficient of f relative to (ϕ_n).

Example 20.28. In $\mathbb{R}^3$, let

$$\phi_1 = (1, 0, 0), \quad \phi_2 = (0, 1, 0), \quad \phi_3 = (0, 0, 1).$$

Suppose $f = (2, -1, 3)$. Then

$$\langle f, \phi_1 \rangle = 2, \quad \langle f, \phi_2 \rangle = -1, \quad \langle f, \phi_3 \rangle = 3.$$

Hence

$$f \sim 2\phi_1 - \phi_2 + 3\phi_3.$$

The following theorems show that the partial sums of the Fourier series of f have a certain minimum property. We shall assume here and in the rest of this chapter that $f \in \mathcal{R}$, although this hypothesis can be weakened.

Proposition 20.29. *Let (ϕ_n) be an orthonormal system of functions on $[a, b]$. Let*

$$s_n(x) = \sum_{k=1}^{n} c_k \phi_k(x)$$

be the n-th partial sum of the Fourier series of f, and let

$$t_n(x) = \sum_{k=1}^{n} \gamma_k \phi_k(x).$$

Then

$$\|f - s_n\| \leq \|f - t_n\|, \tag{20.11}$$

where equality holds if and only if $\gamma_k = c_k$ for $k = 1, \ldots, n$.

That is to say, among all functions t_n, s_n gives the best possible mean square approximation to f.

Proof. We want to show that

$$\langle f - s_n, f - s_n \rangle \leq \langle f - t_n, f - t_n \rangle .$$

Note that

$$\langle f, s_n \rangle = \left\langle f, \sum_{k=1}^{n} c_k \phi_k \right\rangle = \sum_{k=1}^{n} \overline{c_k} \langle f, \phi_k \rangle = \sum_{k=1}^{n} \overline{c_k} c_k = \sum_{k=1}^{n} |c_k|^2$$

$$\langle s_n, s_n \rangle = \left\langle \sum_{k=1}^{n} c_k \phi_k, \sum_{k=1}^{n} c_k \phi_k \right\rangle = \sum_{k=1}^{n} \langle c_k \phi_k, c_k \phi_k \rangle = \sum_{k=1}^{n} |c_k|^2$$

$$\langle f, t_n \rangle = \sum_{k=1}^{n} c_k \overline{\gamma_k}$$

$$\langle t_n, f \rangle = \sum_{k=1}^{n} \gamma_k \overline{c_k}$$

$$\langle t_n, t_n \rangle = \sum_{k=1}^{n} |\gamma_k|^2$$

Hence we rewrite the desired inequality as

$$\begin{aligned}
&\iff \langle f, f\rangle - \sum_{k=1}^{n} |c_k|^2 \le \langle f, f\rangle - \sum_{k=1}^{n} c_k\overline{\gamma_k} - \sum_{k=1}^{n} \gamma_k\overline{c_k} + \sum_{k=1}^{n} |\gamma_k|^2 \\
&\iff \sum_{k=1}^{n} (c_k\overline{c_k} - c_k\overline{\gamma_k} - \gamma_k\overline{c_k} + \gamma_k\overline{\gamma_k}) \ge 0 \\
&\iff \sum_{k=1}^{n} (c_k - \gamma_k)(\overline{c_k} - \overline{\gamma_k}) \ge 0 \\
&\iff \sum_{k=1}^{n} |c_k - \gamma_k|^2 \ge 0
\end{aligned}$$

which holds true. Then equality holds if and only if $|c_k - \gamma_k| = 0$, i.e.,

$$\gamma_k = c_k \quad (k = 1, \ldots, n).$$

□

Proposition 20.30 (Bessel inequality)**.** *Let (ϕ_n) be an orthonormal system of functions on $[a, b]$, and*

$$f(x) \sim \sum_{n=1}^{\infty} c_n\phi_n(x).$$

Then

$$\sum_{n=1}^{\infty} |c_n|^2 \le \|f\|. \tag{20.12}$$

In particular, $c_n \to 0$.

Proof. Letting $n \to \infty$ in (72), we obtain (73) □

the case where equality holds is called Parseval's identity

From now on we shall deal only with the trigonometric system. We shall consider functions f that have period 2π, and are Riemann-integrable on $[-\pi, \pi]$ (and hence on every bounded interval). The Fourier series of f is then the series (63) whose coefficients en are given by the integrals (62), and

$$s_N(x) = s_N(f; x) = \sum_{n=-N}^{N} c_n e^{inx}$$

is the N-th partial sum of the Fourier series of f. The inequality (72) now takes the form

In order to obtain an expression for sN that is more manageable than (75) we introduce the ***Dirichlet kernel***

$$D_N(x) := \sum_{n=-N}^{N} e^{inx}.$$

It follows that

$$\begin{aligned} D_N(x) &= \sum_{n=-N}^{N} e^{inx} \\ &= \frac{e^{-iNx}\left[(e^{ix})^{2N+1} - 1\right]}{e^{ix} - 1} \\ &= \frac{e^{i(N+1)x} - e^{iNx}}{e^{ix} - 1} \\ &= \frac{e^{i(N+\frac{1}{2})x} - e^{-i(N+\frac{1}{2})x}}{e^{\frac{ix}{2}} - e^{\frac{-ix}{2}}} \\ &= \frac{\sin\left(N + \frac{1}{2}\right)x}{\sin\frac{1}{2}x} \end{aligned}$$

Then, for some dummy variable t,

$$\begin{aligned} s_N(x) = \sum_{n=-N}^{N} c_n e^{inx} &= \sum_{n=-N}^{N} \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)e^{-int}\,\mathrm{d}t\right] e^{inx} \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \left[\sum_{n=-N}^{N} f(t)e^{in(x-t)}\right] \mathrm{d}t \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)\left[\sum_{n=-N}^{N} e^{in(x-t)}\right] \mathrm{d}t \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)D_N(x-t)\,\mathrm{d}t\,. \end{aligned}$$

Define the *convolution* of f and g as

$$(f * g)(t) := \int_E f(t)g(x-t)\,\mathrm{d}t\,.$$

The periodicity of all functions involved shows that it is immaterial over which interval we integrate, as long as its length is 2π. This shows that

$$s_N(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)D_N(x-t)\,\mathrm{d}t = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x-t)D_N(t)\,\mathrm{d}t\,.$$

We shall prove just one result about the pointwise convergence of Fourier series. Before that, we require the following result.

Proposition 20.31 (Riemann–Lebesgue lemma)**.** *Let $f \in \mathcal{R}[a,b]$. Then*

$$\lim_{n\to\infty}\int_a^b f(x)\sin nx\,\mathrm{d}x = 0. \tag{20.13}$$

Proof. □

Proposition 20.32 (Pointwise convergence of Fourier series)**.** *Suppose for some $x \in [-\pi, \pi]$ there exists $M > 0$, $\delta > 0$ such that*

$$\forall t \in (-\delta, \delta), \quad |f(x+t) - f(x)| \leq M|t|.$$

Then

$$\lim_{N\to\infty} s_N(f; x) = f(x).$$

Proof. Since

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} D_N(x)\,\mathrm{d}x = 1,$$

we can write

$$f(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x) D_N(t)\,\mathrm{d}t\,.$$

Then

$$\begin{aligned} s_N(x) - f(x) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} [f(x-t) - f(x)] D_N(t)\,\mathrm{d}t \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} [f(x-t) - f(x)] \frac{\sin\left(N + \frac{1}{2}\right)t}{\sin \frac{1}{2}t}\,\mathrm{d}t \end{aligned}$$

Let $g(t) = \dfrac{f(x-t) - f(x)}{\sin \frac{1}{2}t}$, then

$$s_N(x) - f(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(t) \sin\left(N + \frac{1}{2}\right) t\,\mathrm{d}t\,.$$

By the Riemann–Lebesgue lemma, we are done. □

Corollary 20.33.

Here is another formulation of this corollary:

This is usually called the localisation theorem. It shows that the behaviour of the sequence $(s_N(f; x))$, as far as convergence is concerned, depends only on the values of f in some (arbitrarily small) neighbourhood of x. Two Fourier series may thus have the same behavior in one interval, but may behave in entirely different ways in some other interval. We have here a very striking contrast between Fourier series and power series (Theorem 8.5).

We conclude with two other approximation theorems.

Theorem 20.34. *If f is continuous (with period 2π) and if $\varepsilon > 0$, then there exists a trigonometric polynomial P such that*

$$|P(x) - f(x)| < \varepsilon \quad (x \in \mathbb{R}).$$

Proof. □

Theorem 20.35 (Parseval's theorem)**.** *Suppose f and g are Riemann-integrable functions with period 2π, and*

$$f(x) \sim \sum_{n=-\infty}^{\infty} c_n e^{inx}, \quad g(x) \sim \sum_{n=-\infty}^{\infty} \gamma_n e^{inx}.$$

Then

(i)

$$\lim_{N\to\infty} \|f - s_N(f)\|^2 = 0.$$

(ii)

$$\frac{1}{2\pi} \langle f, g \rangle = \sum_{n=-\infty}^{\infty} c_n \overline{\gamma_n}.$$

(iii)

$$\|f\|^2 = \sum_{n=-\infty}^{\infty} |c_n|^2.$$

Proof.

(i)

(ii)

(iii)

□

20.4 Gamma Function

The *Gamma function* simulates the factorial.

Definition 20.36 (Gamma function)**.** For $0 < x < \infty$, the ***Gamma function*** is defined as

$$\Gamma(x) := \int_0^\infty t^{x-1} e^{-t} \, \mathrm{d}t \, . \tag{20.14}$$

The integral converges for these x. (When $x < 1$, both 0 and ∞ have to be looked at.)

Lemma 20.37.

(i) The functional equation

$$\Gamma(x+1) = x\Gamma(x)$$

holds for $0 < x < \infty$.

(ii) $\Gamma(n+1) = n!$ *for* $n = 1, 2, 3, \ldots$

(iii) $\log \Gamma$ *is convex on* $(0, \infty)$.

Proof.

(i) Integrate by parts:

$$\begin{aligned}\Gamma(x+1) &= \int_0^\infty t^x e^{-t} \, \mathrm{d}t \\ &= \left[-t^x e^{-t}\right]_0^\infty + \int_0^\infty x t^{x-1} e^{-t} \, \mathrm{d}t \\ &= 0 + x\Gamma(x) = x\Gamma(x).\end{aligned}$$

(ii) We have

$$\Gamma(1) = \int_0^\infty e^{-t} \, \mathrm{d}t = \left[-e^{-t}\right]_0^\infty = 1.$$

Since $\Gamma(1) = 1$, (i) implies (ii) by induction.

(iii) To show that $\log \Gamma(x)$ is convex, we need to show that for all $p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$,

$$\log \Gamma\left(\frac{x}{p} + \frac{y}{q}\right) \geq \frac{1}{p} \log \Gamma(x) + \frac{1}{q} \log \Gamma(y).$$

This is equivalent to showing

$$\Gamma\left(\frac{x}{p} + \frac{y}{q}\right) \geq \Gamma(x)^{\frac{1}{p}} + \Gamma(y)^{\frac{1}{q}}.$$

We have

$$
\begin{aligned}
\Gamma\left(\frac{x}{p}+\frac{y}{q}\right) &= \int_0^\infty t^{\frac{x}{p}+\frac{y}{q}-1} e^{-t}\,\mathrm{d}t \\
&= \int_0^\infty t^{\frac{x-1}{p}+\frac{y-1}{q}} + e^{-t\left(\frac{1}{p}+\frac{1}{q}\right)}\,\mathrm{d}t \\
&= \int_0^\infty \left(t^{x-1}e^{-t}\right)^{\frac{1}{p}} \left(t^{y-1}e^{-t}\right)^{\frac{1}{q}}\,\mathrm{d}t \\
&\leq \left[\int_0^\infty \left(t^{\frac{x-1}{p}} e^{-\frac{t}{p}}\right)^p \mathrm{d}t\right]^{\frac{1}{p}} \left[\int_0^\infty \left(t^{\frac{y-1}{q}} e^{-\frac{t}{q}}\right)^q \mathrm{d}t\right]^{\frac{1}{q}} \\
&= \Gamma(x)^{\frac{1}{p}}\Gamma(y)^{\frac{1}{q}}
\end{aligned}
$$

where the penultimate line holds as a result of Holder's inequality.

□

In fact, these three properties characterise Γ completely.

Lemma 20.38 (Characterisation of Γ)**.** *If f is a positive function on $(0,\infty)$ such that*

(i) $f(x+1) = xf(x)$,

(ii) $f(1) = 1$,

(iii) $\log f$ *is convex,*

then $f(x) = \Gamma(x)$.

Proof. □

Definition 20.39 (Beta function)**.** For $x > 0$ and $y > 0$, the ***beta function*** is defined as

$$B(x,y) := \int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t\,.$$

Lemma 20.40.

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

Proof. Let $f(x) = \dfrac{\Gamma(x+y)}{\Gamma(y)}B(x,y)$. We want to prove that $f(x) = \Gamma(x)$, using Lemma 20.38.

(i)

$$B(x+1,y) = \int_0^1 t^x(1-t)^{y-1}\,\mathrm{d}t\,.$$

Integrating by parts gives

$$
\begin{aligned}
B(x+1,y) &= \underbrace{\left[t^x \cdot \frac{(1-t)^y}{y}(-1)\right]_0^1}_{0} + \int_0^1 x t^{x-1} \frac{(1-t)^y}{y}\,\mathrm{d}t \\
&= \frac{x}{y}\int_0^1 t^{x-1}(1-t)^{y-1}(1-t)\,\mathrm{d}t \\
&= \frac{x}{y}\left(\int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t - \int_0^1 t^{x}(1-t)^{y-1}\,\mathrm{d}t\right) \\
&= \frac{x}{y}\left(B(x,y) - B(x+1,y)\right)
\end{aligned}
$$

which gives $B(x+1,y) = \dfrac{x}{x+y}B(x,y)$. Thus

$$
\begin{aligned}
f(x+1) &= \frac{\Gamma(x+1+y)}{\Gamma(y)}B(x+1,y) \\
&= \frac{(x+y)B(x+y)}{\Gamma(y)} \cdot \frac{x}{x+y}B(x,y) \\
&= xf(x).
\end{aligned}
$$

(ii)

$$
B(1,y) = \int_0^1 (1-t)^{y-1}\,\mathrm{d}t = \left[-\frac{(1-t)^y}{y}\right]_0^1 = \frac{1}{y}
$$

and thus

$$
f(1) = \frac{\Gamma(1+y)}{\Gamma(y)}B(1,y) = \frac{y\Gamma(y)}{\Gamma(y)}\frac{1}{y} = 1.
$$

(iii) We now show that $\log B(x,y)$ is convex, so that

$$
\log f(x) = \underbrace{\log\Gamma(x+y)}_{\text{convex}} + \log B(x,y) - \underbrace{\log\Gamma(y)}_{\text{constant}}
$$

is convex with respect to x.

$$
B(x_1,y)^{\frac{1}{p}}B(x_2,y)^{\frac{1}{q}} = \left(\int_0^1 t^{x_1-1}(1-t)^{y-1}\,\mathrm{d}t\right)^{\frac{1}{p}}\left(\int_0^1 t^{x_2-1}(1-t)^{y-1}\,\mathrm{d}t\right)^{\frac{1}{q}}
$$

By Hölder's inequality,

$$
\begin{aligned}
B(x_1,y)^{\frac{1}{p}}B(x_2,y)^{\frac{1}{q}} &= \int_0^1 \left[t^{x_1-1}(1-t)^{y-1}\right]^{\frac{1}{p}}\left[t^{x_2-1}(1-t)^{y-1}\right]^{\frac{1}{q}}\,\mathrm{d}t \\
&= \int_0^1 t^{\frac{x_1}{p}+\frac{x_2}{q}-1}(1-t)^{y-1}\,\mathrm{d}t \\
&= B\left(\frac{x_1}{p}+\frac{x_2}{q},y\right).
\end{aligned}
$$

Taking log on both sides gives

$$
\log B(x,y)^{\frac{1}{p}}B(x_2,y)^{\frac{1}{q}} \geq \log B\left(\frac{x_1}{p}+\frac{x_2}{q},y\right)
$$

or

$$\frac{1}{p}\log B(x,y)+\frac{1}{q}\log B(x_2,y)\geq \log B\left(\frac{x_1}{p}+\frac{x_2}{q},y\right).$$

Hence $\log B(x,y)$ is convex, so $\log f(x)$ is convex.

Therefore $f(x)=\Gamma(x)$ which implies $B(x,y)=\dfrac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$. $\square$

An alternative form of Γ is as follows:

$$\Gamma(x)=2\int_0^{+\infty} t^{2x-1}e^{-t^2}\,\mathrm{d}t\,.$$

Using this form of Γ, we present an alternative proof.

Proof.

$$\begin{aligned}\Gamma(x)\Gamma(y)&=\left(2\int_0^{+\infty} t^{2x-1}e^{-t^2}\,\mathrm{d}t\right)\left(2\int_0^{+\infty} s^{2y-1}e^{-s^2}\,\mathrm{d}s\right)\\&=4\iint_{[0,+\infty)\times[0,+\infty)} t^{2x-1}s^{2y-1}e^{-\left(t^2+s^2\right)}\,\mathrm{d}t\,\mathrm{d}s\end{aligned}$$

Using polar coordinates transformation, let $t=r\cos\theta$, $s=r\sin\theta$. Then $\mathrm{d}t\,\mathrm{d}s=r\,\mathrm{d}r\,\mathrm{d}\theta$. Thus

$$\begin{aligned}\Gamma(x)\Gamma(y)&=4\int_0^{\frac{\pi}{2}}\left[\int_0^{+\infty} r^{2x-1}\cos^{2x-1}\theta\cdot r^{2y-1}\sin^{2y-1}\theta\cdot e^{-r^2}\cdot r\,\mathrm{d}r\right]\mathrm{d}\theta\\&=\underbrace{2\int_0^{\frac{\pi}{2}}\cos^{2x-1}\theta\sin^{2y-1}\theta\,\mathrm{d}\theta}_{B(x,y)}\cdot\underbrace{2\int_0^{+\infty} r^{2(x+y)-1}e^{-r^2}\,\mathrm{d}r}_{\Gamma(x+y)}\end{aligned}$$

since

$$\begin{aligned}B(x,y)&=\int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t\qquad t=\cos^2\theta\\&=\int_{\frac{\pi}{2}}^0\cos^{2(x-1)}\theta\sin^{2(y-1)}\theta\cdot 2\cos\theta(-\sin\theta)\,\mathrm{d}\theta\\&=2\int_0^{\frac{\pi}{2}}\cos^{2x-1}\theta\sin^{2y-1}\theta\,\mathrm{d}\theta\,.\end{aligned}$$

Hence $B(x,y)=\dfrac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$. $\square$

More on polar coordinates:

$$I=\int_{-\infty}^{+\infty} e^{-x^2}\,\mathrm{d}x=\sqrt{\pi}. \tag{20.15}$$

Proof.

$$\begin{aligned}
I^2 &= \int_{-\infty}^{+\infty} e^{-x^2}\,\mathrm{d}x \int_{-\infty}^{+\infty} e^{-y^2}\,\mathrm{d}y \\
&= \iint_{\mathbb{R}^2} e^{-(x^2+y^2)}\,\mathrm{d}x\,\mathrm{d}y \qquad x = r\cos\theta, y = r\sin\theta \\
&= \int_0^{2\pi} \underbrace{\int_0^{+\infty} e^{-r^2} r\,\mathrm{d}r}_{\text{constant w.r.t. } \theta}\,\mathrm{d}\theta \qquad s = r^2, \mathrm{d}s = 2r\,\mathrm{d}r \\
&= 2\pi \int_0^{+\infty} e^{-s} \cdot \frac{1}{2}\,\mathrm{d}s \\
&= 2\pi \left[\frac{1}{2} e^{-s}(-1)\right]_0^{\infty} = \pi
\end{aligned}$$

and thus

$$I = \int_{-\infty}^{+\infty} e^{-x^2}\,\mathrm{d}x = \sqrt{\pi}.$$

□

From this, we have

$$\Gamma\left(\frac{1}{2}\right) = 2\int_0^{\infty} e^{-t^2}\,\mathrm{d}t = \sqrt{\pi}.$$

Lemma 20.41.

$$\Gamma(x) = \frac{2^{x-1}}{\sqrt{\pi}}\Gamma\left(\frac{x}{2}\right)\Gamma\left(\frac{x+1}{2}\right).$$

Proof. Let $f(x) = \frac{2^{x-1}}{\sqrt{\pi}}\Gamma\left(\frac{x}{2}\right)\Gamma\left(\frac{x+1}{2}\right)$. We want to prove that $f(x) = \Gamma(x)$.

(i)

$$\begin{aligned}
f(x+1) &= \frac{2^x}{\sqrt{\pi}}\Gamma\left(\frac{x+1}{2}\right)\Gamma\left(\frac{x}{2}+1\right) \\
&= \frac{2^x}{\sqrt{\pi}}\Gamma\left(\frac{x+1}{2}\right)\frac{x}{2}\Gamma\left(\frac{x}{2}\right) \\
&= xf(x)
\end{aligned}$$

(ii) $f(1) = \frac{1}{\sqrt{\pi}}\Gamma\left(\frac{1}{2}\right)\Gamma(1) = 1$ since $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

(iii)

$$\log f(x) = \underbrace{(x-1)\log 2}_{\text{linear}} + \underbrace{\log\Gamma\left(\frac{x}{2}\right)}_{\text{convex}} + \underbrace{\log\Gamma\left(\frac{x+1}{2}\right)}_{\text{convex}} - \underbrace{\log\sqrt{\pi}}_{\text{constant}}$$

and hence $\log f(x)$ is convex.

Therefore $f(x) = \Gamma(x)$. □

Theorem 20.42 (Stirling's formula)**.** *This provides a simple approximate expression for $\Gamma(x+1)$ when x is large (hence for $n!$ when n is large). The formula is*

$$\lim_{x\to\infty} \frac{\Gamma(x+1)}{(x/e)^x\sqrt{2\pi x}} = 1. \tag{20.16}$$

Proof. □

Lemma 20.43.

$$B(p, 1-p) = \Gamma(p)\Gamma(1-p) = \frac{\pi}{\sin p\pi}.$$

Proof. We have

$$\begin{aligned}
B(p,1-p) &= \int_0^1 t^{p-1}(1-t)^{-p}\,\mathrm{d}t \\
&= \int_0^\infty \left(\frac{x}{1+x}\right)^{p-1}\left(\frac{1}{1+x}\right)^{-p}\frac{1}{(1+x)^2}\,\mathrm{d}x \quad [x = \frac{t}{1-t}] \\
&= \int_0^\infty \frac{x^{p-1}}{1+x}\,\mathrm{d}x \\
&= \int_0^1 \frac{x^{p-1}}{1+x}\,\mathrm{d}x + \int_1^\infty \frac{x^{p-1}}{1+x}\,\mathrm{d}x
\end{aligned}$$

See that

$$\begin{aligned}
\int_1^\infty \frac{x^{p-1}}{1+x}\,\mathrm{d}x &= \int_1^0 \frac{y^{1-p}}{1+\frac{1}{y}}\left(-\frac{1}{y^2}\right)\mathrm{d}y \quad [x = \frac{1}{y}] \\
&= \int_0^1 \frac{y^{-p}}{1+y}\,\mathrm{d}y = \int_0^1 \frac{x^{-p}}{1+x}\,\mathrm{d}x
\end{aligned}$$

so

$$
\begin{aligned}
B(p, 1-p) &= \int_0^1 \frac{x^{p-1} + x^{-p}}{1+x} \, \mathrm{d}x \\
&= \lim_{r \to 1^-} \int_0^r (x^{p-1} + x^{-p}) \sum_{k=0}^{\infty} (-1)^k x^k \, \mathrm{d}x \\
&= \lim_{r \to 1^-} \int_0^r \left(\sum_{k=0}^{\infty} (-1)^k x^{k+p-1} + \sum_{k=0}^{\infty} (-1)^k x^{k-p} \right) \mathrm{d}x \\
&= \lim_{r \to 1^-} \left[\sum_{k=0}^{\infty} (-1)^k \frac{x^{k+p}}{k+p} + \sum_{k=0}^{\infty} (-1)^k \frac{x^{k-p+1}}{k-p+1} \right]_0^r \\
&= \sum_{k=0}^{\infty} (-1)^k \frac{1}{k+p} + \sum_{k=0}^{\infty} (-1)^k \frac{1}{k-p+1} \\
&= \frac{1}{p} + \sum_{k=1}^{\infty} (-1)^k \frac{1}{k+p} + \sum_{k=1}^{\infty} (-1)^{k-1} \frac{1}{k+p} \\
&= \frac{1}{p} + \sum_{k=1}^{\infty} \frac{(-1)^k 2p}{p^2 - k^2}
\end{aligned}
$$

□

Exercises

V

Multivariable Analysis

Chapter 21

Differentiation

We shall now switch to a different topic, namely that of differentiation in several variable calculus. More precisely, we shall be dealing with maps $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ from one Euclidean space to another, and trying to understand what the derivative of such a map is.

21.1 Basic Definitions

Recall that for $f\colon \mathbb{R} \to \mathbb{R}$, we defined the derivative at x as

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

This equation certainly makes no sense in the general case of a function $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$, but can reformulated in a way that does. If $A\colon \mathbb{R} \to \mathbb{R}$ is the linear transformation defined by $A(h) = f'(x) \cdot h$, then we can rewrite

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

Thus we reformulate the definition of differentiability as follows:

> $f\colon \mathbb{R} \to \mathbb{R}$ is differentiable at $x \in \mathbb{R}$ if there exists a linear transformation $A \in \mathcal{L}(\mathbb{R}, \mathbb{R})$ such that
> $$\lim_{h\to 0} \frac{f(x+h) - f(a) - Ah}{h} = 0.$$

In this form, the definition has a simple generalisation to higher dimensions:

Definition 21.1 (Differentiability)**.** Let $U \subset \mathbb{R}^n$ be open, $\mathbf{f}\colon U \to \mathbb{R}^m$. We say $\mathbf{f}$ is ***differentiable*** at $\mathbf{x} \in U$ if there exists $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ such that

$$\lim_{\mathbf{h}\to \mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\|}{\|\mathbf{h}\|} = 0.$$

Remark. $\mathbf{h}$ is a point of $\mathbb{R}^n$, and $\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}$ is a point of $\mathbb{R}^m$, so the norm signs are essential.

The linear transformation A is denoted as $D\mathbf{f}(\mathbf{x})$, and called the ***derivative*** of $\mathbf{f}$ at $\mathbf{x}$.

Remark. If $n = m = 1$, then $D\mathbf{f}(\mathbf{x})$ coincides with the familiar $f'(x)$ in single-variable calculus.

Notation. Some authors also use $\mathbf{f}'(\mathbf{x})$ to denote the derivative.

If $\mathbf{f}$ is differentiable at every $\mathbf{x} \in U$, we say $\mathbf{f}$ is *differentiable on* U.

The justification for the phrase "*the* linear transformation" is as follows:

Lemma 21.2 (Uniqueness of derivative)**.** *Let $U \subset \mathbb{R}^n$ be open, $\mathbf{f}\colon U \to \mathbb{R}^m$. Suppose $\mathbf{x} \in U$, and there exist $A, B \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ such that*

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\|}{\|\mathbf{h}\|} = 0 \quad \textit{and} \quad \lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - B\mathbf{h}\|}{\|\mathbf{h}\|} = 0.$$

Then $A = B$.

Proof. Suppose $\mathbf{h} \neq \mathbf{0}$. Then

$$\begin{aligned}
\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|(A-B)\mathbf{h}\|}{\|\mathbf{h}\|} &= \lim_{\mathbf{h}\to\mathbf{0}} \frac{\|(\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}) - (\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - B\mathbf{h})\|}{\|\mathbf{h}\|} \\
&\le \lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\|}{\|\mathbf{h}\|} + \lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - B\mathbf{h}\|}{\|\mathbf{h}\|} \\
&= 0 + 0 = 0.
\end{aligned}$$

If $\mathbf{x} \in \mathbb{R}^n$, then $t\mathbf{x} \to 0$ as $t \to 0$. Hence for $\mathbf{x} \neq \mathbf{0}$ we have

$$0 = \lim_{t\to 0} \frac{\|A(t\mathbf{x}) - B(t\mathbf{x})\|}{\|t\mathbf{x}\|} = \frac{\|A\mathbf{x} - B\mathbf{x}\|}{\|\mathbf{x}\|}.$$

Therefore $A\mathbf{x} = B\mathbf{x}$, which implies $A = B$. □

Example 21.3. Consider the function $f\colon \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x,y) = \sin x$. Then $f'(a,b) = (\cos a)\cdot x$. To prove this, note that

$$\lim_{(h,k)\to(0,0)} \frac{\|f(a+h,b+k) - f(a,b) - f'(h,k)\|}{\|(h,k)\|} = \lim_{(h,k)\to(0,0)} \frac{|\sin(a+h) - \sin a - (\cos a)\cdot h|}{\|(h,k)\|}.$$

Since $\sin'(a) = \cos a$, we have

$$\lim_{h\to 0} \frac{|\sin(a+h) - \sin a - (\cos a)\cdot h|}{|h|} = 0.$$

Since $\|(h,k)\| \ge |h|$, it is also true that

$$\lim_{(h,k)\to(0,0)} \frac{|\sin(a+h) - \sin a - (\cos a)\cdot h|}{\|(h,k)\|} = 0.$$

21.2 Basic Theorems

Lemma 21.4 (Differentiability implies continuity). *Let $U \subset \mathbb{R}^n$ be open, suppose $\mathbf{f}\colon U \to \mathbb{R}^m$ is differentiable at $\mathbf{x} \in U$. Then $\mathbf{f}$ is continuous at $\mathbf{x}$.*

Proof. Another way to write the differentiability of $\mathbf{f}$ at $\mathbf{x}$ is to consider the *remainder*:

$$\mathbf{r}(\mathbf{h}) := \mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - D\mathbf{f}(\mathbf{x})\mathbf{h}.$$

By definition, $\mathbf{f}$ is differentiable at $\mathbf{x}$ if $\lim_{\mathbf{h}\to\mathbf{0}} \|\mathbf{r}(\mathbf{h})\|/\|\mathbf{h}\| = 0$. Thus $\lim_{\mathbf{h}\to\mathbf{0}} \mathbf{r}(\mathbf{h}) = \mathbf{0}$.
The mapping $\mathbf{h} \mapsto D\mathbf{f}(\mathbf{x})\mathbf{h}$ is a linear mapping between finite-dimensional spaces, hence continuous and $\lim_{\mathbf{h}\to\mathbf{0}} D\mathbf{f}(\mathbf{x})\mathbf{h} = \mathbf{0}$. Hence

$$\lim_{\mathbf{h}\to\mathbf{0}} \mathbf{f}(\mathbf{x}+\mathbf{h}) = \mathbf{f}(\mathbf{x}).$$

This precisely means that $\mathbf{f}$ is continuous at $\mathbf{x}$. □

The next result implies that differentiation is a linear map on the space of differentiable functions.

Lemma 21.5. *Let $U \subset \mathbb{R}^n$ be open, suppose $\mathbf{f}, \mathbf{g}\colon U \to \mathbb{R}^m$ are differentiable at $\mathbf{x} \in U$, let $\alpha \in \mathbb{R}$. Then*

(i) $\mathbf{f}+\mathbf{g}$ is differentiable at $\mathbf{x}$, and *(addition)*

$$D(\mathbf{f}+\mathbf{g})(\mathbf{x}) = D\mathbf{f}(\mathbf{x}) + D\mathbf{g}(\mathbf{x}).$$

(ii) $\alpha\mathbf{f}$ is differentiable at $\mathbf{x}$, and *(scalar multiplication)*

$$D(\alpha\mathbf{f})(\mathbf{x}) = \alpha D\mathbf{f}(\mathbf{x}).$$

Proof. Let $\mathbf{h} \in \mathbb{R}^n$, $\mathbf{h} \neq \mathbf{0}$.

(i) We have

$$\begin{aligned}&\frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) + \mathbf{g}(\mathbf{x}+\mathbf{h}) - (\mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})) - (D\mathbf{f}(\mathbf{x}) + D\mathbf{g}(\mathbf{x}))\mathbf{h}\|}{\|\mathbf{h}\|} \\ &\leq \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - D\mathbf{f}(\mathbf{x})\mathbf{h}\|}{\|\mathbf{h}\|} + \frac{\|\mathbf{g}(\mathbf{x}+\mathbf{h}) - \mathbf{g}(\mathbf{x}) - D\mathbf{g}(\mathbf{x})\mathbf{h}\|}{\|\mathbf{h}\|}\end{aligned}$$

Then take limits $\mathbf{h} \to \mathbf{0}$ on both sides of the equation.

(ii) Write

$$\frac{\|\alpha\mathbf{f}(\mathbf{x}+\mathbf{h}) - \alpha\mathbf{f}(\mathbf{x}) - \alpha D\mathbf{f}(\mathbf{x})\mathbf{h}\|}{\|\mathbf{h}\|} = |\alpha| \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - D\mathbf{f}(\mathbf{x})\mathbf{h}\|}{\|\mathbf{h}\|}.$$

Then take limits $\mathbf{h} \to \mathbf{0}$ on both sides of the equation.

□

We now extend the chain rule to the present situation.

Lemma 21.6 (Chain rule)**.** *Let $U \subset \mathbb{R}^n$, $V \subset \mathbb{R}^m$ be open. Suppose $\mathbf{f} \colon U \to \mathbb{R}^m$ is differentiable at $\mathbf{x} \in U$, $\mathbf{f}(U) \subset V$, and $g : V \to \mathbb{R}^k$ is differentiable at $\mathbf{f}(\mathbf{x})$.*
Then $\mathbf{g} \circ \mathbf{f}$ is differentiable at $\mathbf{x}$, and

$$D(\mathbf{g} \circ \mathbf{f})(\mathbf{x}) = D\mathbf{g}\left(\mathbf{f}(\mathbf{x})\right) D\mathbf{f}(\mathbf{x}).$$

Proof. Let $\mathbf{F} = \mathbf{g} \circ \mathbf{f}$. Let $A = D\mathbf{f}(\mathbf{x})$ and $B = D\mathbf{g}\left(\mathbf{f}(\mathbf{x})\right)$. We will show that $D\mathbf{F}(\mathbf{x}) = BA$.

Take a non-zero $\mathbf{h} \in \mathbb{R}^n$ and write $\mathbf{y} = \mathbf{f}(\mathbf{x})$, $\mathbf{k} = \mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x})$. Let

$$\mathbf{r}(\mathbf{h}) = \mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}.$$

Then $\mathbf{r}(\mathbf{h}) = \mathbf{k} - A\mathbf{h}$ or $A\mathbf{h} = \mathbf{k} - \mathbf{r}(\mathbf{h})$, and $\mathbf{f}(\mathbf{x}+\mathbf{h}) = \mathbf{y} + \mathbf{k}$. We look at the quantity we need to go to zero:

$$\begin{aligned}
\frac{\|\mathbf{F}(\mathbf{x}+\mathbf{h}) - \mathbf{F}(\mathbf{x}) - BA\mathbf{h}\|}{\|\mathbf{h}\|} &= \frac{\|\mathbf{g}(\mathbf{f}(\mathbf{x}+\mathbf{h})) - \mathbf{g}(\mathbf{f}(\mathbf{x})) - BA\mathbf{h}\|}{\|\mathbf{h}\|} \\
&= \frac{\|\mathbf{g}(\mathbf{y}+\mathbf{k}) - \mathbf{g}(\mathbf{y}) - B(\mathbf{k} - \mathbf{r}(\mathbf{h}))\|}{\|\mathbf{h}\|} \\
&\leq \frac{\|\mathbf{g}(\mathbf{y}+\mathbf{k}) - \mathbf{g}(\mathbf{y}) - B\mathbf{k}\|}{\|\mathbf{h}\|} + \|B\| \frac{\|\mathbf{r}(\mathbf{h})\|}{\|\mathbf{h}\|} \\
&= \frac{\|\mathbf{g}(\mathbf{y}+\mathbf{k}) - \mathbf{g}(\mathbf{y}) - B\mathbf{k}\|}{\|\mathbf{k}\|} \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x})\|}{\|\mathbf{h}\|} + \|B\| \frac{\|\mathbf{r}(\mathbf{h})\|}{\|\mathbf{h}\|}.
\end{aligned}$$

Take the limit $\mathbf{h} \to \mathbf{0}$. We examine the three terms:

- Since $\mathbf{f}$ is differentiable at $\mathbf{x}$, $\displaystyle\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{r}(\mathbf{h})\|}{\|\mathbf{h}\|} = 0$.
- Since $\mathbf{f}$ is continuous at $\mathbf{x}$, $\mathbf{k} \to \mathbf{0}$ as $\mathbf{h} \to \mathbf{0}$. Thus since $\mathbf{g}$ is differentiable at $\mathbf{y}$,

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{g}(\mathbf{y}+\mathbf{k}) - \mathbf{g}(\mathbf{y}) - B\mathbf{k}\|}{\|\mathbf{k}\|} = 0.$$

- We have

$$\begin{aligned}
\frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x})\|}{\|\mathbf{h}\|} &\leq \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\|}{\|\mathbf{h}\|} + \frac{\|A\mathbf{h}\|}{\|\mathbf{h}\|} \\
&\leq \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\|}{\|\mathbf{h}\|} + \|A\|.
\end{aligned}$$

  Since $\mathbf{f}$ is differentiable at $\mathbf{x}$, for small enough $\mathbf{h}$, the quantity $\frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\|}{\|\mathbf{h}\|}$ is bounded. Thus the term $\frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x})\|}{\|\mathbf{h}\|}$ stays bounded as $\mathbf{h} \to \mathbf{0}$.

Therefore

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{F}(\mathbf{x}+\mathbf{h}) - \mathbf{F}(\mathbf{x}) - BA\mathbf{h}\|}{\|\mathbf{h}\|} = 0,$$

so $D\mathbf{F}(\mathbf{x}) = BA$ as desired. □

We now show the derivatives of several basic functions.

Lemma 21.7.

(i) If $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ *is a constant function, then*

$$D\mathbf{f}(\mathbf{x}) = \mathbf{0}.$$

(ii) If $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ *is a linear transformation, then*

$$D\mathbf{f}(\mathbf{x}) = \mathbf{f}.$$

Proof.

(i) Suppose $\mathbf{f}(\mathbf{x}) = \mathbf{y}$ for all $\mathbf{x} \in \mathbb{R}^n$. Then

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{0}\|}{\|\mathbf{h}\|} = \lim_{\mathbf{h}\to\mathbf{0}} \frac{\|\mathbf{y} - \mathbf{y} - \mathbf{0}\|}{\|\mathbf{h}\|} = 0.$$

(ii) We have

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

□

We are now assured of the differentiability of those functions $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$, whose component functions are obtained by addition, multiplication, division, and composition, from the functions π_i (which are linear maps) and the functions which we can already differentiate by elementary calculus.

Finding $D\mathbf{f}(\mathbf{x})$, however, may be a fairly formidable task.

Example 21.8. Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(x,y) = \sin\left(xy^2\right)$. Since $f = \sin\circ(\pi_1 \cdot (\pi_2)^2)$, we have

$$\begin{aligned} f'(x,y) &= \sin'(xy^2) \cdot [y^2(\pi_1)'(x,y) + x((\pi_2)^2)'(x,y)] \\ &= \sin'(xy^2) \cdot [y^2(\pi_1)'(x,y) + 2xy(\pi_2)'(x,y)] \\ &= \left(\cos\left(xy^2\right)\right) \cdot [y^2(1,0) + 2xy(0,1)] \\ &= \left(y^2\cos\left(xy^2\right), 2xy\cos\left(xy^2\right)\right). \end{aligned}$$

Fortunately, we will soon discover a much simpler method of computing $D\mathbf{f}$.

21.3 Partial Derivatives

We begin the attack on the problem of finding derivatives "one variable at a time"; that is, we hold all but one variables constant and take the regular derivative. This is known as a *partial derivative*.

Definition 21.9 (Partial derivative)**.** Let $U \subset \mathbb{R}^n$ be open, $\mathbf{f}\colon U \to \mathbb{R}^m$. The j-th partial derivative at $\mathbf{x} = (x_1, \dots, x_n) \in U$ is

$$\begin{aligned} D_j\mathbf{f}(\mathbf{x}) &= \lim_{h\to 0} \frac{\mathbf{f}(x_1, \dots, x_j + h, \dots, x_n) - \mathbf{f}(x_1, \dots, x_n)}{h} \\ &= \lim_{t\to 0} \frac{\mathbf{f}(\mathbf{x} + t\mathbf{e}_j) - \mathbf{f}(\mathbf{x})}{t} \qquad (j = 1, \dots, n) \end{aligned}$$

provided the limit exists.

Notation. Some authors also denote the j-th partial derivative at $\mathbf{x}$ by $\dfrac{\partial \mathbf{f}}{\partial x_j}(\mathbf{x})$.

Note that $D_j\mathbf{f}(\mathbf{x})$ is the ordinary derivative of a certain function; if $g\colon \mathbb{R} \to \mathbb{R}^m$ is defined by

$$\mathbf{g}(x) = \mathbf{f}(x_1, \dots, x, \dots, x_n),$$

then $D_j\mathbf{f}(\mathbf{x}) = g'(x_j)$. This implies that the computation of $D_j\mathbf{f}(\mathbf{x})$ is a problem we can already solve. If $\mathbf{f}(x_1, \dots, x_n)$ is given by some formula involving $x_1, \dots, x_n$, then we find $D_j\mathbf{f}(x_1, \dots, x_n)$ by differentiating the function whose value at x_j is given by the formula when all x_i, $i \neq j$, are thought of as constants.

Example 21.10. If $f(x, y) = \sin\left(xy^2\right)$, then

$$\begin{aligned} D_1 f(x, y) &= \frac{\partial f}{\partial x} = y^2 \cos\left(xy^2\right), \\ D_2 f(x, y) &= \frac{\partial f}{\partial y} = 2xy \cos\left(xy^2\right). \end{aligned}$$

If $f(x, y) = x^y$, then

$$\begin{aligned} D_1 f(x, y) &= \frac{\partial f}{\partial x} = yx^{y-1}, \\ D_2 f(x, y) &= \frac{\partial f}{\partial y} = x^y \log x. \end{aligned}$$

Partial derivatives can be used to find the maxima and minima of functions.

Proposition 21.11. *Let $U \subset \mathbb{R}^n$ be open. If the maximum (or minimum) of $\mathbf{f}\colon U \to \mathbb{R}$ is at $\mathbf{x} \in U$, then*

$$D_j\mathbf{f}(\mathbf{x}) = 0,$$

provided the partial derivatives exists.

Proof. For each $j = 1, \dots, n$, define $g_j(x)\colon \mathbb{R} \to \mathbb{R}^m$ by

$$g_j(\mathbf{x}) = \mathbf{f}(x_1, \dots, x, \dots, x_n).$$

Then $g_j'(x_j) = D_j\mathbf{f}(\mathbf{x})$.

Clearly g_j has a maximum (or minimum) at x_j, and g_j is defined in an open interval containing x_j. Hence $D_j\mathbf{f}(\mathbf{x}) = g_j'(x_j) = 0$. $\square$

Remark. The converse is false even if $n = 1$ (if $f\colon \mathbb{R} \to \mathbb{R}$ is defined by $f(x) = x^3$, then $f'(0) = 0$, but 0 is not even a local maximum or minimum).

If $n > 1$, the converse may fail to be true in a rather spectacular way. Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(x, y) = x^2 - y^2$. Since g_1 has a minimum at 0, and g_2 has a maximum at 0,

$$\begin{aligned} D_1 f(0,0) &= 0, \\ D_2 f(0,0) &= 0, \end{aligned}$$

but clearly $(0, 0)$ is neither a local maximum or minimum.

Let $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ and $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$ be the standard bases of $\mathbb{R}^n$ and $\mathbb{R}^m$. Let $U \subset \mathbb{R}^n$ be open, $\mathbf{f}\colon U \to \mathbb{R}^m$. The *components* of $\mathbf{f}$ are the real-valued functions $f_1, \dots, f_m\colon U \to \mathbb{R}$ defined by

$$\mathbf{f}(\mathbf{x}) = \sum_{i=1}^{m} f_i(\mathbf{x})\mathbf{u}_i \quad (\mathbf{x} \in U).$$

Hence with respect to the standard bases of $\mathbb{R}^m$, we can write

$$[\mathbf{f}(\mathbf{x})] = \begin{pmatrix} f_1(\mathbf{x}) & \cdots & f_m(\mathbf{x}) \end{pmatrix},$$

which we can identify with as a vector in $\mathbb{R}^m$:

$$(f_1(\mathbf{x}), \dots, f_m(\mathbf{x})).$$

The next result states that in order to differentiate a function $\mathbf{f}$, we can differentiate each of its components.

Lemma 21.12. *If $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$, then $\mathbf{f}$ is differentiable at $\mathbf{x} \in \mathbb{R}^n$ if and only if each component f_i is differentiable at $\mathbf{x}$, and*

$$D\mathbf{f}(\mathbf{x}) = \left(f_1'(\mathbf{x}), \dots, f_m'(\mathbf{x})\right).$$

Thus $D\mathbf{f}(\mathbf{x})$ is the $m \times n$ matrix whose i-th row is $(f_i)'(\mathbf{x})$.

Proof. Suppose each f_i is differentiable at $\mathbf{x}$. Let

$$A = \left(f_1'(\mathbf{x}), \dots, f_m'(\mathbf{x})\right).$$

Then

$$\begin{aligned} &\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h} \\ &= (f_1(\mathbf{x}+\mathbf{h}), \dots, f_m(\mathbf{x}+\mathbf{h})) - (f_1(\mathbf{x}), \dots, f_m(\mathbf{x})) - \left(f_1'(\mathbf{x})\mathbf{h}, \dots, f_m'(\mathbf{x})\mathbf{h}\right) \\ &= \left(f_1(\mathbf{x}+\mathbf{h}) - f_1(\mathbf{x}) - f_1'(\mathbf{x})\mathbf{h}, \dots, f_m(\mathbf{x}+\mathbf{h}) - f_m(\mathbf{x}) - f_m'(\mathbf{x})\mathbf{h}\right). \end{aligned}$$

Therefore

$$\lim_{h \to 0} \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - A\mathbf{h}\|}{\|\mathbf{h}\|} \leq \lim_{h \to 0} \sum_{i=1}^{m} \frac{\|f_i(\mathbf{x}+\mathbf{h}) - f_i(\mathbf{x}) - f_i'(\mathbf{x})\mathbf{h}\|}{\|\mathbf{h}\|} = 0.$$

$\square$

21.4 Derivatives

Likewise we can also take partial derivatives of each component: the j-th partial derivative of the i-th component of $\mathbf{f}$ at $\mathbf{x} \in U$ is

$$D_j f_i(\mathbf{x}) := \lim_{t \to 0} \frac{f_i(\mathbf{x} + t\mathbf{e}_j) - f_i(\mathbf{x})}{t} \quad (1 \le i \le m,\ 1 \le j \le n),$$

provided the limit exists.

Partial derivatives are easier to compute with all the machinery of calculus, and they provide a way to compute the total derivative of a function.

Theorem 21.13. *Let $U \subset \mathbb{R}^n$ be open, suppose $\mathbf{f}\colon U \to \mathbb{R}^m$ is differentiable at $\mathbf{x} \in U$. Then all the partial derivatives at $\mathbf{x}$ exist, and the matrix of $D\mathbf{f}(\mathbf{x})$ with respect to the standard bases of $\mathbb{R}^n$ and $\mathbb{R}^m$ is*

$$[D\mathbf{f}(\mathbf{x})]_{ij} = D_j f_i(\mathbf{x}).$$

That is,

$$[D\mathbf{f}(\mathbf{x})] = \begin{pmatrix} D_1 f_1(\mathbf{x}) & D_2 f_1(\mathbf{x}) & \cdots & D_n f_1(\mathbf{x}) \\ D_1 f_2(\mathbf{x}) & D_2 f_2(\mathbf{x}) & \cdots & D_n f_2(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ D_1 f_m(\mathbf{x}) & D_2 f_m(\mathbf{x}) & \cdots & D_n f_m(\mathbf{x}) \end{pmatrix}.$$

We call $[D\mathbf{f}(\mathbf{x})]$ the ***Jacobian matrix*** of $\mathbf{f}$ at $\mathbf{x}$.

Remark. If $f\colon \mathbb{R} \to \mathbb{R}$, then $f'(x)$ is a 1×1 matrix whose single entry is the number denoted as $f'(x)$ in single-variable calculus.

Proof. Fix $j \in \{1, \dots, n\}$. We want to show that

$$D\mathbf{f}(\mathbf{x})\mathbf{e}_j = \sum_{i=1}^{m} D_j f_i(\mathbf{x})\mathbf{u}_i.$$

Since $\mathbf{f}$ is differentiable at $\mathbf{x}$, writing differentiability in terms of the remainder gives us

$$\mathbf{f}(\mathbf{x} + t\mathbf{e}_j) - \mathbf{f}(\mathbf{x}) = D\mathbf{f}(\mathbf{x})(t\mathbf{e}_j) + \mathbf{r}(t\mathbf{e}_j)$$

where $\|\mathbf{r}(t\mathbf{e}_j)\|/t \to 0$ as $t \to 0$. Taking the limit $t \to 0$ on both sides, the linearity of $D\mathbf{f}(\mathbf{x})$ shows that

$$\lim_{t \to 0} \frac{\mathbf{f}(\mathbf{x} + t\mathbf{e}_j) - \mathbf{f}(\mathbf{x})}{t} = D\mathbf{f}(\mathbf{x})\mathbf{e}_j.$$

If we now represent $\mathbf{f}$ in terms of its components, the above equation becomes

$$\lim_{t \to 0} \sum_{i=1}^{m} \frac{f_i(\mathbf{x} + t\mathbf{e}_j) - f_i(\mathbf{x})}{t} \mathbf{u}_i = D\mathbf{f}(\mathbf{x})\mathbf{e}_j.$$

It follows that each quotient in this sum has a limit as $t \to 0$ (see Theorem 4.10), so that each partial derivative $D_j f_i$ exists. Hence

$$D\mathbf{f}(\mathbf{x})\mathbf{e}_j = \sum_{i=1}^{m} D_j f_i(\mathbf{x})\mathbf{u}_i \quad (j = 1, \dots, n).$$

□

The converse is true if one hypothesis is added.

Let $U \subset \mathbb{R}^n$ be open. We say $\mathbf{f}\colon U \to \mathbb{R}^m$ is ***continuously differentiable*** if $\mathbf{f}$ is differentiable, and $D\mathbf{f}$ is continuous; we also say that $\mathbf{f}$ is a $\mathcal{C}'$-mapping, or that $\mathbf{f} \in \mathcal{C}'(U)$.

More explicitly, it is required that

$$\forall \varepsilon > 0, \quad \exists \delta > 0, \quad \forall x \in U, \quad \|\mathbf{x} - \mathbf{y}\| < \delta \implies \|D\mathbf{f}(\mathbf{y}) - D\mathbf{f}(\mathbf{x})\| < \varepsilon.$$

Theorem 21.14. *Let $U \subset \mathbb{R}^n$ be open, $\mathbf{f}\colon U \to \mathbb{R}^m$. Then $\mathbf{f}$ is continuously differentiable if and only if all the partial derivatives $D_j f_i$ exist and are continuous on U.*

Proof.

$\boxed{\implies}$ Suppose $\mathbf{f} \in \mathcal{C}'(U)$. Then $\mathbf{f}$ is differentiable on U, so

$$(D_j f_i)(\mathbf{x}) = (D\mathbf{f}(x)\mathbf{e}_j) \cdot \mathbf{u}_i$$

for all i, j, and for all $\mathbf{x} \in U$. Hence

$$(D_j f_i)(\mathbf{y}) - (D_j f_i)(\mathbf{x}) = ((D\mathbf{f}(\mathbf{y}) - D\mathbf{f}(\mathbf{x}))\mathbf{e}_j) \cdot \mathbf{u}_i.$$

Since $\|\mathbf{u}_i\| = \|\mathbf{e}_j\| = 1$, it follows that

$$\begin{aligned} \|(D_j f_i)(\mathbf{y}) - (D_j f_i)(\mathbf{x})\| &\le \|(D\mathbf{f}(\mathbf{y}) - D\mathbf{f}(\mathbf{x}))\mathbf{e}_j\| \\ &\le \|D\mathbf{f}(\mathbf{y}) - D\mathbf{f}(\mathbf{x})\|. \end{aligned}$$

Hence $D_j f_i$ is continuous.

$\boxed{\impliedby}$ □

21.4.1 Gradients, Curves, and Directional Derivatives

Let γ be a differentiable mapping of $(a, b) \subset \mathbb{R}$ into an open set $U \subset \mathbb{R}^n$; that is, γ is a differentiable curve in U. Let $f\colon U \to \mathbb{R}$ be differentiable.

For $t \in (a, b)$, define

$$g(t) = f(\gamma(t)).$$

By the chain rule,

$$g'(t) = f'(\gamma(t))\,\gamma'(t).$$

Since $\gamma'(t) \in \mathcal{L}(\mathbb{R}, \mathbb{R}^n)$ and $f'(\gamma(t)) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R})$, $g'(t)$ is a linear operator on $\mathbb{R}$; thus, we can regard $g'(t)$ as a real number. This number can be computed in terms of the partial derivatives of f and the derivatives of the components of γ_i, as we shall now see.

With respect to the standard basis $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ of $\mathbb{R}^n$, the matrix of $\gamma'(t)$ is the $n \times 1$ matrix which has $\gamma_i'(t)$ in the i-th row, where $\gamma_1, \dots, \gamma_n$ are the components of γ. For every $\mathbf{x} \in U$, the matrix of $f'(\mathbf{x})$ is the $1 \times n$ matrix which has $\frac{\partial f}{\partial x_j}$ in the j-th column. Hence the matrix of $g'(t)$ is the 1×1 matrix whose only entry is the real number

$$g'(t) = f'(\gamma(t))\gamma'(t) = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\gamma(t))\frac{\mathrm{d}\gamma_j}{\mathrm{d}t} = \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}\frac{\mathrm{d}\gamma_j}{\mathrm{d}t}.$$

Definition 21.15 (Gradient)**.** Let $U \subset \mathbb{R}^n$ be open, suppose $f\colon U \to \mathbb{R}$ is differentiable. The ***gradient*** at $\mathbf{x} \in U$ is defined as

$$(\nabla f)(\mathbf{x}) := \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\mathbf{x})\mathbf{e}_j. \tag{21.1}$$

Writing $\gamma'(t)$ as components

$$\gamma'(t) = \sum_{j=1}^{n} \gamma_i'(t)\mathbf{e}_j,$$

using the scalar product, we can rewrite $g'(t)$ as

$$g'(t) = (\nabla f)(\gamma(t)) \cdot \gamma'(t). \tag{21.2}$$

Let us now fix $\mathbf{x} \in U$, take a unit vector $\mathbf{u} \in \mathbb{R}^n$, and let γ be

$$\gamma(t) = \mathbf{x} + t\mathbf{u}.$$

Then $\gamma'(t) = \mathbf{u}$ for every t. Hence (21.2) shows that

$$g'(0) = (\nabla f)(\mathbf{x}) \cdot \mathbf{u}.$$

On the other hand, we have

$$g(t) - g(0) = f(\mathbf{x} + t\mathbf{u}) - f(\mathbf{x}).$$

Hence
$$\lim_{t\to 0}\frac{f(\mathbf{x}+t\mathbf{u})-f(x)}{t}=(\nabla f)(\mathbf{x})\cdot\mathbf{u}. \tag{21.3}$$

We call this limit the ***directional derivative*** of f at $\mathbf{x}$, in the direction of the unit vector $\mathbf{u}$, and may be denoted by $(D_{\mathbf{u}}f)(\mathbf{x})$.

If f and $\mathbf{x}$ are fixed, but $\mathbf{u}$ varies, then (21.3) shows that $(D_{\mathbf{u}}f)(\mathbf{x})$ attains it maximum when $\mathbf{u}$ is a positive scalar multiple of $(\nabla f)(\mathbf{x})$. [The case $(\nabla f)(\mathbf{x})=\mathbf{0}$ should be excluded here.]

If $\mathbf{u}=\sum_j u_j\mathbf{e}_j$, then (21.3) shows that $(D_{\mathbf{u}}f)(\mathbf{x})$ can be expressed in terms of the partial derivatives of f at $\mathbf{x}$:
$$(D_{\mathbf{u}}f)(\mathbf{x})=\sum_{j=1}^{n}\frac{\partial f}{\partial x_i}(\mathbf{x})\mathbf{u}_i. \tag{21.4}$$

Proposition 21.16. *Let $U\subset\mathbb{R}^n$ be open and convex, suppose $\mathbf{f}\colon U\to\mathbb{R}^m$ is differentiable on U, and there exists a real number M such that*
$$\|D\mathbf{f}(\mathbf{x})\|\le M \quad (\mathbf{x}\in U).$$
Then for all $\mathbf{a},\mathbf{b}\in U$,
$$\|\mathbf{f}(\mathbf{b})-\mathbf{f}(\mathbf{a})\|\le M\|\mathbf{b}-\mathbf{a}\|.$$

Proof. Fix $\mathbf{a},\mathbf{b}\in U$. Define
$$\gamma(t)=(1-t)\mathbf{a}+t\mathbf{b}$$
for all $t\in\mathbb{R}$ such that $\gamma(t)\in U$. Since U is convex, $\gamma(t)\in U$ if $0\le t\le 1$. Put
$$\mathbf{g}(t)=\mathbf{f}\left(\gamma(t)\right).$$
Then
$$\mathbf{g}'(t)=D\mathbf{f}\left(\gamma(t)\right)\gamma'(t)=D\mathbf{f}\left(\gamma(t)\right)(\mathbf{b}-\mathbf{a}),$$
so that
$$|\mathbf{g}'(t)|\le\|D\mathbf{f}\left(\gamma(t)\right)\||\mathbf{b}-\mathbf{a}|\le M|\mathbf{b}-\mathbf{a}|$$
for all $t\in[0,1]$. By Theorem 5.19,
$$\|\mathbf{g}(1)-\mathbf{g}(0)\|\le M\|\mathbf{b}-\mathbf{a}\|.$$
But $\mathbf{g}(0)=\mathbf{f}(\mathbf{a})$ and $\mathbf{g}(1)=\mathbf{f}(\mathbf{b})$. This completes the proof. □

Corollary 21.17. *If, in addition, $D\mathbf{f}(\mathbf{x})=\mathbf{0}$ for all $\mathbf{x}\in U$, then $\mathbf{f}$ is constant.*

Proof. To prove this, note that the hypotheses of the previous result hold now with $M=0$. □

21.4.2 The Jacobian

Definition 21.18 (Jacobian)**.** Let $U \subset \mathbb{R}^n$, suppose $\mathbf{f}\colon U \to \mathbb{R}^m$ is differentiable. Define the ***Jacobian*** of $\mathbf{f}$ at $\mathbf{x} \in U$ as

$$J_{\mathbf{f}}(\mathbf{x}) := \det[D\mathbf{f}(\mathbf{x})].$$

We shall also denote $J_{\mathbf{f}}$ as

$$\frac{\partial(f_1, \ldots, f_n)}{\partial(x_1, \ldots, x_n)}.$$

This last piece of notation may seem somewhat confusing, but it is quite useful when we need to specify the exact variables and function components used, as we will do, for example, in the implicit function theorem.

The Jacobian determinant $J_{\mathbf{f}}$ is a real-valued function, and when $n = 1$ it is simply the derivative. From the chain rule and $\det AB = \det A \det B$, it follows that

$$J_{\mathbf{f}\circ\mathbf{g}}(\mathbf{x}) = J_{\mathbf{f}}(\mathbf{g}(\mathbf{x}))J_{\mathbf{g}}(\mathbf{x}).$$

The determinant of a linear mapping tells us what happens to area/volume under the mapping. Similarly, the Jacobian determinant measures how much a differentiable mapping stretches things locally, and if it flips orientation. In particular, if the Jacobian determinant is non-zero than we would assume that locally the mapping is invertible (and we would be correct as we will later see).

21.4.3 Continuity and The Derivative

Let us prove a "mean value theorem" for vector-valued functions.

Theorem 21.19. *If $\phi : [a, b] \to \mathbb{R}^n$ is differentiable on (a, b) and continuous on $[a, b]$, then there exists $t \in [a, b]$ such that*

$$\|\phi(b) - \phi(a)\| \leq (b - a)\|\phi'(t)\|. \tag{21.5}$$

21.5 Inverse Functions

Suppose $f\colon \mathbb{R} \to \mathbb{R}$ is continuously differentiable in an open set containing a, and $f'(a) \neq 0$. If $f'(a) > 0$, there is an open interval V containing a such that $f'(x) > 0$ for all $x \in V$, and a similar statement holds if $f'(a) < 0$.

Thus f is increasing (or decreasing) on V, and is therefore bijective with an inverse function f^{-1} defined on some open interval W containing $f(a)$. Moreover it is not hard to show that f^{-1} is differentiable, and

$$(f^{-1})'(y) = \frac{1}{f'(f^{-1}(y))} \quad (y \in W).$$

An analogous discussion in higher dimensions is much more involved. The inverse function theorem states, roughly speaking, that a continuously differentiable mapping $\mathbf{f}$ is invertible in a neighbourhood of any point $\mathbf{x}$ at which the linear map $D\mathbf{f}(\mathbf{x})$ is invertible:

Theorem 21.20 (Inverse function theorem)**.** *Let $U \subset \mathbb{R}^n$ be open, suppose $\mathbf{f}\colon U \to \mathbb{R}^n$ is a $\mathcal{C}'$-mapping, and $D\mathbf{f}(\mathbf{a})$ is invertible for some $\mathbf{a} \in U$, and $\mathbf{b} = \mathbf{f}(\mathbf{a})$. Then*

(i) there exist open sets $U, V \subset \mathbb{R}^n$ such that $\mathbf{a} \in U$, $\mathbf{b} \in V$, $\mathbf{f}$ is bijective on U, and $\mathbf{f}(U) = V$;

(ii) if $\mathbf{g}$ is the inverse of $\mathbf{f}$ [which exists, by (i)], defined in V by

$$\mathbf{g}(\mathbf{f}(\mathbf{x})) = \mathbf{x} \quad (\mathbf{x} \in U),$$

then $\mathbf{g} \in \mathcal{C}'(V)$.

Writing the equation $\mathbf{y} = \mathbf{f}(\mathbf{x})$ in component form, we arrive at the following interpretation of the conclusion of the theorem: The system of n equations

$$y_i = f_i(x_1, \dots, x_n) \quad (i = 1, \dots, n)$$

can be solved for $x_1, \dots, x_n$ in terms of $y_1, \dots, y_n$, if we restrict x and y to small enough neighbourhoods of $\mathbf{a}$ and $\mathbf{b}$; the solutions are unique and continuously differentiable.

Corollary 21.21. *Let $U \subset \mathbb{R}^n$ be open, suppose $\mathbf{f}\colon U \to \mathbb{R}^m$ is continuously differentiable on U. If $D\mathbf{f}(\mathbf{x})$ is invertible for every $\mathbf{x} \in U$, then $f(W)$ is an open subset of $\mathbb{R}^n$ for every open set $W \subset U$.*

21.6 Implicit Functions

Theorem 21.22 (Implicit function theorem)**.** *Let $U \subset \mathbb{R}^n$ be open,*

21.7 Derivatives of Higher Order

Definition 21.23. Let $U \subset \mathbb{R}^n$ be open, suppose $f\colon U \to \mathbb{R}$, with partial derivatives $\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}$. If the functions $\frac{\partial f}{\partial x_j}$ are themselves differentiable, then the *second-order partial derivatives* of f are defined by

$$\frac{\partial^2 f}{\partial x_i\,\partial x_j} = \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right) \quad (i, j = 1, \dots, n).$$

If all these functions $\frac{\partial^2 f}{\partial x_i\,\partial x_j}$ are continuous on U, we say that f is of class $\mathcal{C}''$ in U, or that $f \in \mathcal{C}''(U)$.
$\mathbf{f}\colon U \subset \mathbb{R}^n \to \mathbb{R}^m$ is said to be of class $\mathcal{C}''$ if each component of $\mathbf{f}$ is of class $\mathcal{C}''$.

It can happen that

21.8 Differentiation of Integrals

Exercises

Chapter 22

Integration

22.1 Basic Definitions

Recall that a partition P of a closed interval $[a, b]$ is a set of points $t_0, \ldots, t_k$ where

$$a = t_0 \leq t_1 \leq \cdots \leq t_k = b.$$

The partition P divides the interval $[a, b]$ into k subintervals $[t_{i-1}, t_i]$.

More generally, a *partition* of a rectangle $[a_1, b_1] \times \cdots \times [a_n, b_n]$ is a collection

$$P = (P_1, \ldots, P_n),$$

where each P_i is a partition of the interval $[a_i, b_i]$.

Example 22.1. Suppose $P_1 = \{t_0, \ldots, t_k\}$ is a partition of $[a_1, b_1]$ and $P_2 = \{s_0, \ldots, s_l\}$ is a partition of $[a_2, b_2]$. Then the partition $P = (P_1, P_2)$ of $[a_1, b_1] \times [a_2, b_2]$ divides the closed rectangle into $k \cdot l$ subrectangles of the form $[t_{i-1}, t_i] \times [s_{j-1}, s_j]$.

In general, if P_i divides $[a_i, b_i]$ into N_i subintervals, then $P = (P_1, \ldots, P_n)$ divides $[a_1, b_1] \times \cdots \times [a_n, b_n]$ into $N = N_1 \cdots N_n$ subrectangles. These subrectangles are called *subrectangles of the partition* P.

Let $A \subset \mathbb{R}^n$ be a rectangle, $f \colon A \to \mathbb{R}$ be a bounded function, and P be a partition of A. For each subrectangle S of the partition, let

$$m_S(f) = \inf_{x \in S} f(x)$$
$$M_S(f) = \sup_{x \in S} f(x)$$

and let $v(S)$ be the volume of $S = [a_1, b_1] \times \cdots \times [a_n, b_n]$, defined by

$$v(S) = (b_1 - a_1) \cdots (b_n - a_n).$$

The *lower* and *upper sums* of f for P are defined by

$$L(f,P) = \sum_S m_S(f) \cdot v(S),$$
$$U(f,P) = \sum_S M_S(f) \cdot v(S).$$

Clearly $L(f,P) \le U(f,P)$.

We say a partition P' *refines* P if each subrectangle of P' is in a subrectangle of P.

Lemma 22.2. *Suppose the partition P' refines P. Then*

$$L(f,P) \le L(f,P') \quad \textit{and} \quad U(f,P') \le U(f,P).$$

Proof. Each subrectangle S of P is divided into several subrectangles $S_1, \ldots, S_\alpha$ of P', so $v(S) = v(S_1) + \cdots + v(S_\alpha)$. Now $m_S(f) \le m_{S_i}(f)$, since the values $f(x)$ for $x \in S$ include all values $f(x)$ for $x \in S_i$ (and possibly smaller ones). Thus

$$\begin{aligned} m_S(f) \cdot v(S) &= m_S(f) \cdot v(S_1) + \cdots + m_S(f) \cdot v(S_\alpha) \\ &\le m_{S_1}(f) \cdot v(S_1) + \cdots + m_{S_\alpha}(f) \cdot v(S_\alpha). \end{aligned}$$

The sum, for all S, of the terms on the LHS is $L(f,P)$, while the sum of all the terms on the RHS is $L(f,P')$. Hence $L(f,P) < L(f,P')$. The proof for upper sums is similar. □

Corollary 22.3. *If P and P' are any two partitions, then $L(f,P') \le U(f,P)$.*

Proof. Let P'' be a partition which refines both P and P'. Then

$$L(f,P') \le L(f,P'') \le U(f,P'') \le U(f,P).$$

□

Define the *upper* and *lower integrals* of f over A by

$$\overline{\int_A} f = \inf_{P \in \mathcal{P}(A)} U(f,P)$$
$$\underline{\int_A} f = \sup_{P \in \mathcal{P}(A)} L(f,P)$$

where $\mathcal{P}(A)$ denotes the set of all partitions of A.

The previous result implies that the sup of all lower sums for f is less than or equal to the inf of all upper sums for f; in other words,

$$\underline{\int_A} f \le \overline{\int_A} f.$$

If the two values coincide, we say f is integrable:

Definition 22.4. We say a bounded function $f\colon A \to \mathbb{R}$ is ***integrable*** on the rectangle A if

$$\underline{\int_A} f = \overline{\int_A} f.$$

This common number is called the *integral* of f over A, and denoted

$$\int_A f.$$

Often, the notation

$$\int_A f(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n$$

is used. If $f\colon [a, b] \to \mathbb{R}$, then this coincides with the Riemann integral: $\int_a^b f = \int_{[a,b]} f$.

A simple but useful criterion for integrability is provided by the next result.

Lemma 22.5 (Integrability criterion)**.** *A bounded function $f\colon A \to \mathbb{R}$ is integrable if and only if*

$$\forall \varepsilon > 0, \quad \exists P, \quad U(f, P) - L(f, P) < \varepsilon.$$

This means we can make the upper and lower sums arbitrarily close.

Proof.

$\boxed{\implies}$ Suppose f is integrable. Then

$$\sup_{P \in \mathcal{P}(A)} L(f, P) = \inf_{P \in \mathcal{P}(A)} U(f, P).$$

Thus for any $\varepsilon > 0$, there exists partitions P and P' such that

$$U(f, P) - L(f, P') < \varepsilon.$$

Let P'' be a common refinement of P and P'. Then

$$U(f, P'') - L(f, P'') \le U(f, P) - L(f, P') < \varepsilon.$$

$\boxed{\impliedby}$ Let $\varepsilon > 0$ be given. Suppose there exists a partition P such that $U(f, P) - L(f, P) < \varepsilon$. Then it is clear that

$$\sup_{P \in \mathcal{P}(A)} L(f, P) = \inf_{P \in \mathcal{P}(A)} U(f, P).$$

Hence f is integrable. $\square$

In the following sections we will characterize the integrable functions and discover a method of computing integrals. For the present we consider two functions, one integrable and one not.

Example 22.6 (Constant function)**.** Let $f\colon A \to \mathbb{R}$ be a constant function, $f(x) = c$. Then for any partition P and subrectangle S, we have

$$m_S(f) = M_S(f) = c,$$

so that

$$L(f,P) = U(f,P) = \sum_S c \cdot v(S) = c \cdot v(A).$$

Hence $\int_A f = c \cdot v(A)$.

Example 22.7 (Dirichlet's function)**.** Let $f\colon [0,1] \times [0,1] \to \mathbb{R}$ be defined by

$$f(x,y) = \begin{cases} 0 & (x \in \mathbb{Q}) \\ 1 & (x \in \mathbb{R} \setminus \mathbb{Q}) \end{cases}$$

If P is a partition, then every subrectangle S will contain points (x,y) with x rational, and also points (x,y) with x irrational. Hence $m_S(f) = 0$ and $M_S(f) = 1$, so

$$L(f,P) = \sum_S 0 \cdot v(S) = 0$$
$$U(f,P) = \sum_S 1 \cdot v(S) = v\left([0,1] \times [0,1]\right) = 1.$$

Therefore f is not integrable.

22.2 Measure Zero and Content Zero

Definition 22.8 (Measure zero)**.** We say $A \subset \mathbb{R}^n$ has ***measure*** 0, if for every $\varepsilon > 0$ there exists a cover $\{U_1, U_2, \dots\}$ of A by closed rectangles such that

$$\sum_{n=1}^{\infty} v(U_n) < \varepsilon.$$

It is obvious (but nevertheless useful to remember) that if A has measure 0 and $B \subset A$, then B has measure 0. The reader may verify that open rectangles may be used instead of closed rectangles in the definition of measure 0.

Example 22.9. A set with only finitely many points clearly has measure 0.

Lemma 22.10. *If A has countably many points, then A also has measure 0.*

Lemma 22.11. *If $A = A_1 \cup A_2 \cup \cdots$ and each A_n has measure 0, then A has measure 0.*

Definition 22.12 (Content zero)**.**

22.3 Integrable Functions

22.4 Fubini's Theorem

22.5 Partitions of Unity

22.6 Change of Variables

VI
Complex Analysis

The starting point of our study is the idea of extending a function initially given for real values of the argument to one that is defined when the argument is complex. Thus, here the central objects are functions from the complex plane to itself

$$f\colon \mathbb{C} \to \mathbb{C},$$

or more generally, complex-valued functions defined on open subsets of $\mathbb{C}$.

Chapter 23

Complex Functions

23.1 The Complex Plane

$\mathbb{C}$ is a metric space, with metric $d(z, w) = |z - w|$. Hence all notions defined for general metric spaces, as outlined in Chapters 14 to 16 and 20, are applicable to $\mathbb{C}$.

If $z_0 \in \mathbb{C}$ and $r > 0$, we define the ***open disc*** of radius r centered at z_0 to be

$$D_r(z_0) = \{z \in \mathbb{C} \mid |z - z_0| < r\}.$$

The ***closed disc*** $D_r(z_0)$ of radius r centered at z_0 is defined by

$$D_r(z_0) = \{z \in \mathbb{C} \mid |z - z_0| \leq r\}.$$

The ***boundary*** of either the open or closed disc is the circle

$$C_r(z_0) = \{z \in \mathbb{C} \mid |z - z_0| = r\}.$$

Since the ***unit disc*** (the open disc centered at the origin and of radius 1) plays an important role, we will often denote it by $\mathbb{D}$:

$$\mathbb{D} = \{z \in \mathbb{C} \mid |z| < 1\}.$$

The last notion we need is that of connectedness. An open set $\Omega \subset \mathbb{C}$ is ***connected*** if it is not possible to find two disjoint non-empty open sets Ω_1 and Ω_2 such that

$$\Omega = \Omega_1 \cup \Omega_2.$$

A connected open set in $\mathbb{C}$ is called a ***region***. Similarly, a closed set F is connected if one cannot write $F = F_1 \cup F_2$ where F_1 and F_2 are disjoint non-empty closed sets.

There is an equivalent definition of connectedness for open sets in terms of curves, which is often useful in practice: an open set Ω is connected if and only if any two points in Ω can be joined by a curve γ entirely contained in Ω.

23.2 Functions on The Complex Plane

23.2.1 Holomorphic Functions

We present the complex analogue of differentiability, which, at first glance, seems no different from the real case.

Definition 23.1. Let $\Omega \subset \mathbb{C}$ be open. We say $f \colon \Omega \to \mathbb{C}$ is ***holomorphic*** at $z_0 \in \Omega$ if

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} \tag{23.1}$$

exists.

The value of the limit in (23.1) is known as the *derivative* of f at z_0; we write

$$f'(z_0) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}.$$

If $z = z_0 + h$ for some $h \in \mathbb{C}$, we can rewrite the above equation as

$$f'(z_0) = \lim_{h \to 0} \frac{f(z_0 + h) - f(z_0)}{h}.$$

It should be emphasised that h is a complex number that may approach 0 from any direction.

We can also rewrite

$$f(z_0 + h) - f(z_0) - f'(z_0)h = h\varepsilon(h), \tag{23.2}$$

where $\varepsilon(h) \to 0$ as $h \to 0$.

If f is holomorphic at every point of Ω, we say f is *holomorphic on* Ω. If $C \subset \mathbb{C}$ is closed, we say f is holomorphic on C if f is holomorphic in some open set containing C. Finally, we say f is ***entire*** if f is holomorphic in all of $\mathbb{C}$.

Lemma 23.2. *If f is holomorphic at z_0, then f is continuous at z_0.*

Proof. Suppose f is holomorphic at z_0. Then the limit $\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$ exists. Thus

$$\begin{aligned}
\lim_{z \to z_0} f(z) - f(z_0) &= \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}(z - z_0) \\
&= \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} \cdot \lim_{z \to z_0} (z - z_0) \\
&= f'(z_0) \cdot 0 = 0.
\end{aligned}$$

Hence $\lim_{z \to z_0} f(z) = f(z_0)$, so f is continuous at z_0. □

The following lemma collects the basic facts about holomorphic functions. We omit the proof, which is essentially identical to the real case.

Lemma 23.3. *Suppose $\Omega \subset \mathbb{C}$ is open, and $f, g\colon \Omega \to \mathbb{C}$ are holomorphic on Ω.*

(i) $f+g$ is holomorphic on Ω, and *(sums)*

$$(f+g)' = f' + g'.$$

(ii) fg is holomorphic on Ω, and *(products)*

$$(fg)' = f'g + fg'.$$

(iii) f/g is holomorphic on Ω (provided $g(z) \neq 0$), and *(quotients)*

$$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}.$$

Lemma 23.4 (Chain rule)**.** *Suppose Ω and U are open subsets of $\mathbb{C}$, and $f\colon \Omega \to U$ and $g\colon U \to \mathbb{C}$ are holomorphic. Then*

$$(g \circ f)'(z) = g'(f(z))f'(z) \quad (z \in \Omega). \tag{23.3}$$

Example 23.5 (Polynomials)**.** $f(z) = z$ is holomorphic on any open set in $\mathbb{C}$, and $f'(z) = 1$. In fact, any polynomial

$$p(z) = a_n z^n + \cdots + a_1 z + a_0$$

is holomorphic in the entire complex plane, and

$$p'(z) = n a_n z^{n-1} + \cdots + a_1.$$

Example 23.6. $f(z) = \frac{1}{z}$ is holomorphic on any open set in $\mathbb{C}$ that does not contain the origin, and $f'(z) = -\frac{1}{z^2}$.

Example 23.7. $f(z) = \overline{z}$ is not holomorphic. Indeed, we have

$$\frac{f(z_0+h) - f(z_0)}{h} = \frac{\overline{h}}{h}$$

which has no limit as $h \to 0$, as one can see by first taking h real and then h purely imaginary.

23.2.2 Cauchy–Riemann Equations

To each complex-valued function $f = u + iv$, we associate the mapping $F(x,y) = (u(x,y), v(x,y))$ from $\mathbb{R}^2$ to $\mathbb{R}^2$.

Recall from Chapter 21 that a function $F(x,y) = (u(x,y), v(x,y))$ is said to be *differentiable* at a point $P_0 = (x_0, y_0)$ if there exists a linear transformation $J\colon \mathbb{R}^2 \to \mathbb{R}^2$ such that

$$\lim_{|h| \to 0} \frac{F(P_0 + h) - F(P_0) - J(h)}{|h|} = 0.$$

Equivalently, we can write
$$F(P_0 + h) - F(P_0) = J(h) + |h|\varepsilon(h),$$
with $\|\varepsilon(h)\| \to 0$ as $|h| \to 0$. The linear transformation J is unique and is called the *derivative* of F at P_0.

If F is differentiable, the partial derivatives of u and v exist, and the linear transformation J can be described in the standard basis of $\mathbb{R}^2$ by the Jacobian matrix of F:

$$J = J_F(x, y) = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix}.$$

In the case of complex differentiation, the derivative is a complex number $f'(z_0)$; in the case of real derivatives, it is a matrix. There is, however, a connection between these two notions, which is given in terms of special relations that are satisfied by the entries of the Jacobian matrix, that is, the partials of u and v.

Theorem 23.8 (Cauchy–Riemann equations)**.** *Suppose f is holomorphic. Then its real and imaginary parts satisfy*
$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}. \tag{23.4}$$

Proof. Consider the limit
$$f'(z_0) = \lim_{h \to 0} \frac{f(z_0 + h) - f(z_0)}{h}$$
when h is first real, say $h = h_1 + ih_2$ with $h_2 = 0$. Then, if we write $z = x + iy$, $z_0 = x_0 + iy_0$, and $f(z) = f(x, y)$, we find that

$$\begin{aligned} f'(z_0) &= \lim_{h_1 \to 0} \frac{f(x_0 + h_1, y_0) - f(x_0, y_0)}{h_1} \\ &= \frac{\partial f}{\partial x}(z_0). \end{aligned}$$

Now taking h purely imaginary, say $h = ih_2$, a similar argument yields

$$\begin{aligned} f'(z_0) &= \lim_{h_2 \to 0} \frac{f(x_0, y_0 + h_2) - f(x_0, y_0)}{ih_2} \\ &= \frac{1}{i}\frac{\partial f}{\partial y}(z_0) = -i\frac{\partial f}{\partial y}(z_0). \end{aligned}$$

Hence
$$\frac{\partial f}{\partial x} = -i\frac{\partial f}{\partial y}.$$
Writing $f = u + iv$, we find after separating real and imaginary parts, that the partials of u and v exist, and they satisfy (23.4). □

Remark. This suggests that complex differentiability is a much more rigid property than one might think at first sight; if f is differentiable then these partial derivatives do exist, and moreover they are subject to a constraint.

We can clarify the situation further by defining two differential operators

$$\frac{\partial}{\partial z} = \frac{1}{2}\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right),$$
$$\frac{\partial}{\partial \bar{z}} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right).$$

Proposition 23.9. *If f is holomorphic at z_0, then*

$$\frac{\partial f}{\partial \bar{z}} = 0$$

and thus

$$f'(z_0) = \frac{\partial f}{\partial z}(z_0) = 2\frac{\partial u}{\partial z}(z_0).$$

If we write $F(x, y) = f(z)$, then F is differentiable in the sense of real variables, and

$$\det J_F(x_0, y_0) = |f'(z_0)|^2.$$

Proof. Taking real and imaginary parts, it is easy to see that the Cauchy–Riemann equations are equivalent to $\frac{\partial f}{\partial \bar{z}} = 0$.

Moreover, by our earlier observation,

$$f'(z_0) = \frac{1}{2}\left(\frac{\partial f}{\partial x}(z_0) - i\frac{\partial f}{\partial y}(z_0)\right) = \frac{\partial f}{\partial z}(z_0),$$

and the Cauchy–Riemann equations give

$$\frac{\partial f}{\partial z} = 2\frac{\partial u}{\partial z}.$$

To prove that F is differentiable it suffices to observe that if $h = (h_1, h_2)$ and $h = h_1 + ih_2$, then the Cauchy–Riemann equations imply

$$J_F(x_0, y_0)(h) = \left(\frac{\partial u}{\partial x} - i\frac{\partial u}{\partial y}\right)(h_1 + ih_2) = f'(z_0)h,$$

where we have identified a complex number with the pair of real and imaginary parts. After a final application of the Cauchy–Riemann equations, the above results imply that

$$\det J_F(x_0, y_0) = \frac{\partial u}{\partial x}\frac{\partial v}{\partial y} - \frac{\partial v}{\partial x}\frac{\partial u}{\partial y} = \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 = \left|2\frac{\partial u}{\partial z}\right|^2 = |f'(z_0)|^2.$$

□

So far, we have assumed that f is holomorphic and deduced relations satisfied by its real and imaginary parts. The next result contains an important converse, which completes the circle of ideas presented here.

Proposition 23.10. *Suppose $f = u + iv$ is a complex-valued function defined on an open set $\Omega \subset \mathbb{C}$. If $u, v\colon \mathbb{R}^2 \to \mathbb{R}$ are continuously differentiable and satisfy the Cauchy–Riemann* *equations on Ω, then f is holomorphic on Ω and $f'(z) = \frac{\partial f}{\partial z}$.*

Proof. Write
$$u(x+h_1, y+h_2) - u(x,y) = \frac{\partial u}{\partial x} h_1 + \frac{\partial u}{\partial y} h_2 + |h|\varepsilon_1(h)$$
and
$$v(x+h_1, y+h_2) - v(x,y) = \frac{\partial v}{\partial x} h_1 + \frac{\partial v}{\partial y} h_2 + |h|\varepsilon_2(h)$$
where the remainders $\varepsilon_1(h), \varepsilon_2(h) \to 0$ as $|h| \to 0$, and $h = h_1 + ih_2$. Using the Cauchy–Riemann equations we find that

$$f(z+h) - f(z) = \left(\frac{\partial u}{\partial x} - i\frac{\partial u}{\partial y}\right)(h_1 + ih_2) + |h|\varepsilon(h),$$

where $\varepsilon(h) = \varepsilon_1(h) + \varepsilon_2(h) \to 0$ as $|h| \to 0$. Hence f is holomorphic and

$$f'(z) = 2\frac{\partial u}{\partial z} = \frac{\partial f}{\partial z}.$$

□

Example 23.11. The function $f(z) = \bar{z}$ is not holomorphic.

Proof. Let $u, v : \mathbb{R}^2 \to \mathbb{R}$ be the components of f. Then $u(x,y) = x$, $v(x,y) = -y$, and so

$$\frac{\partial u}{\partial x} = 1, \quad \frac{\partial u}{\partial y} = 0, \quad \frac{\partial v}{\partial x} = 0, \quad \frac{\partial v}{\partial y} = -1.$$

Since u and v do not satisfy the Cauchy–Riemann equations, f is not holomorphic. □

We shall prove later that the derivative of an holomorphic function is itself holomorphic. By this fact, u and v have continuous second partial derivatives. Differentiating (23.4) again gives

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 v}{\partial x\,\partial y}, \quad \frac{\partial^2 u}{\partial y^2} = -\frac{\partial^2 v}{\partial y\,\partial x}.$$

Hence
$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0. \tag{23.5}$$

(23.5) is called *Laplace's equation*. Solutions to (23.5) are said to be ***harmonic***. If two harmonic functions u and v satisfy the Cauchy–Riemann equations (23.4), then v is called the *conjugate harmonic function* of u.

Proposition 23.12. *Let Ω be either $\mathbb{C}$ or some open disk. If $u \colon \Omega \to \mathbb{R}$ is a harmonic function, then u has a harmonic conjugate.*

Proof. Let $U = D_R(0)$ where $0 < R \leq \infty$, and let $u : U \to \mathbb{R}$ be a harmonic function. We will find a harmonic function v such that u and v satisfy the Cauchy–Riemann equations.

Let

$$v(x,y) = \int_0^y u_x(x,t)\,\mathrm{d}t + \phi(x)$$

and determine ϕ so that $u_x = -u_y$. Differentiating both sides of this equation with respect to x gives

$$\begin{aligned} v_x(x,y) &= \int_0^y u_{xx}(x,t)\,\mathrm{d}t + \phi'(x) \\ &= -\int_0^y u_{yy}(x,t)\,\mathrm{d}t + \phi'(x) \\ &= -u_y(x,y) + u_y(x,0) + \phi'(x) \end{aligned}$$

So it must be that $\phi'(x) = -u_y(x,0)$. It is easily checked that u and

$$v(x,y) = \int_0^y u_x(x,t)\,\mathrm{d}t - \int_0^x u_y(s,0)\,\mathrm{d}s$$

so satisfy the Cauchy–Riemann equations. □

23.2.3 Power Series

Recall our discussion of power series in Chapter 20, including Hadamard's formula (20.1) for the radius of convergence of a power series.

Having defined complex differentiation, we now prove the complex analogue of 20.6, concerning the derivative of complex power series.

Proposition 23.13. *The power series $f(z) = \sum_{n=0}^{\infty} a_n z^n$ defines a holomorphic function in its disc of convergence. Then the derivative of f is obtained by differentiating term-by-term the series of f:*

$$f'(z) = \sum_{n=0}^{\infty} n a_n z^{n-1}. \tag{23.6}$$

Moreover, f' has the same radius of convergence as f.

Corollary 23.14. *A power series is infinitely complex differentiable in its disc of convergence, and the higher derivatives are also power series obtained by termwise differentiation.*

Definition 23.15. We say $f\colon \Omega \subset \mathbb{C} \to \mathbb{C}$ is ***analytic*** (or have a *power series expansion*) at $z_0 \in \Omega$ if there exists a power series $\sum a_n(z-z_0)^n$ centered at z_0, with positive radius of convergence, such that

$$f(z) = \sum_{n=0}^{\infty} a_n(z-z_0)^n \quad \text{for all } z \text{ in a neighbourhood of } z_0.$$

If f has a power series expansion at every point in Ω, we say f is *analytic on* Ω.

Exponential Function

We are familiar with the exponential function e^x of a real variable, which has the property that $(e^x)' = e^x$. The complex exponential has the same property.

Lemma 23.16. $\exp'(z) = \exp(z)$ *for all* $z \in \mathbb{C}$.

Proof. Using 23.13, we calculate the derivative of $\exp(z)$ by differentiating term-by-term:

$$\exp'(z) = (1)' + \sum_{n=1}^{\infty} \left(\frac{z^n}{n!}\right)' = \sum_{n=1}^{\infty} \frac{nz^{n-1}}{n!} = \sum_{n=1}^{\infty} \frac{z^{n-1}}{(n-1)!} = \exp(z).$$

□

Lemma 23.17 (Basic properties of exp)**.**

(i) There exists a positive number π *such that* $e^{\frac{\pi i}{2}} = i$ *and such that* $e^z = 1$ *if and only if* $\frac{z}{2\pi i}$ *is an integer.*

(ii) $\exp$ *is a periodic function, with period* $2\pi i$*.*

(iii) The mapping $t \mapsto e^{it}$ *maps the real axis onto the unit circle.*

(iv) If $w \in \mathbb{C}$*,* $w \neq 0$*, then* $w = e^z$ *for some* z*.*

We shall encounter the integral of $(1+x^2)^{-1}$ over the real line. To evaluate it, put $\phi(t) = \frac{\sin t}{\cos t}$ in $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. By (6), $\phi' = 1 + \phi^2$. Hence ϕ is a monotonically increasing mapping of $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ onto $(-\infty, \infty)$, and we obtain

$$\int_{-\infty}^{\infty} \frac{1}{1+x^2}\,\mathrm{d}x = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{\phi'(t)}{1+\phi^2(t)}\,\mathrm{d}t = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \mathrm{d}t = \pi.$$

23.3 Integration Along Curves

Definition 23.18. A ***parametrised curve*** is a function $z(t)$ which maps a closed interval $[a,b] \subset \mathbb{R}$ to $\mathbb{C}$.
We say that the parametrised curve is ***smooth*** if $z'(t)$ exists and is continuous on $[a,b]$, and $z'(t) \neq 0$ for $t \in [a,b]$.

Remark. At the endpoints $t = a$ and $t = b$, the quantities $z'(a)$ and $z'(b)$ are interpreted as the one-sided limits

$$z'(a) = \lim_{h \to 0^+} \frac{z(a+h) - z(a)}{h}, \qquad z'(b) = \lim_{h \to 0^-} \frac{z(b+h) - z(b)}{h},$$

which are the right-hand derivative of $z(t)$ at a, and the left-hand derivative of $z(t)$ at b, respectively.

Similarly we say that the parametrised curve is ***piecewise-smooth*** if z is continuous on $[a,b]$ and if there exist points

$$a = a_0 < a_1 < \cdots < a_n = b,$$

where $z(t)$ is smooth in the intervals $[a_i, a_{i+1}]$.

We say two parametrisations

$$z\colon [a,b] \to \mathbb{C}, \quad \overline{z}\colon [c,d] \to \mathbb{C}$$

are *equivalent* if there exists a continuously differentiable bijection $s \mapsto t(s)$ from $[c,d]$ to $[a,b]$ so that $t'(s) > 0$ and

$$\overline{z}(s) = z(t(s)).$$

Remark. The condition $t'(s) > 0$ says precisely that the orientation is preserved: as s travels from c to d, then $t(s)$ travels from a to b.

The family of all parametrisations that are equivalent to $z(t)$ determines a smooth curve $\gamma \subset \mathbb{C}$, namely the image of $[a,b]$ under z with the orientation given by z as t travels from a to b.

The points $z(a)$ and $z(b)$ are called the *end-points* of the curve, and are independent on the parametrisation. Since γ carries an orientation, it is natural to say that γ begins at $z(a)$ and ends at $z(b)$.

A smooth or piecewise-smooth curve is ***closed*** if $z(a) = z(b)$ for any of its parametrisations.

Finally, a smooth or piecewise-smooth curve is ***simple*** if it is not self-intersecting, that is, $z(t) \neq z(s)$ unless $s = t$. (Of course, if the curve is closed to begin with, then we say that it is simple whenever $z(t) \neq z(s)$ unless $s = t$, or $s = a$ and $t = b$.)

For brevity, we shall call any piecewise-smooth curve a *curve*, since these will be the objects we shall be primarily concerned with.

A basic example consists of a circle. Consider the circle $C_r(z_0)$ centered at z_0 and of radius r:

$$C_r(z_0) = \{z \in \mathbb{C} \mid |z - z_0| = r\}.$$

Definition 23.19 (Orientation)**.** The positive orientation (counterclockwise) is the one that is given by the standard parametrisation

$$z(t) = z_0 + re^{it} \quad (0 \le t \le 2\pi).$$

The negative orientation (clockwise) is given by

$$z(t) = z_0 + re^{-it} \quad (0 \le t \le 2\pi).$$

In the following chapters, we shall denote by C a general positively oriented circle.

Definition 23.20 (Integral along curve)**.** Given a smooth curve $\gamma \subset \mathbb{C}$ parametrised by $z\colon [a,b] \to \mathbb{C}$, and f a continuous function on γ, define the ***integral of f along γ*** by

$$\int_\gamma f(z)\,\mathrm{d}z := \int_a^b f(z(t))\, z'(t)\,\mathrm{d}t\,. \tag{23.7}$$

In order for this definition to be meaningful, we must show that the integral on the RHS is independent of the parametrisation chosen for γ. Suppose $\overline{z}$ is an equivalent parametrisation as above. Then the change of variables formula and the chain rule imply that

$$\int_a^b f(z(t))\, z'(t)\,\mathrm{d}t = \int_c^d f(z(t(s)))\, z'(t(s))\, t'(s)\,\mathrm{d}s = \int_c^d f(\overline{z}(s))\, \overline{z}'(s)\,\mathrm{d}s\,.$$

This proves that the integral of f over γ is well defined.

If γ is piecewise smooth, then the integral of f over γ is simply the sum of the integrals of f over the smooth parts of γ, so if $z(t)$ is a piecewise-smooth parametrisation as before, then

$$\int_\gamma f(z)\,\mathrm{d}z = \sum_{i=1}^n \int_{a_{i-1}}^{a_i} f(z(t))\, z'(t)\,\mathrm{d}t\,.$$

By definition, the ***length*** of the smooth curve γ is

$$\Lambda(\gamma) = \int_a^b |z'(t)|\,\mathrm{d}t\,.$$

Arguing as we just did, it is clear that this definition is also independent of the parametrisation. Also, if γ is only piecewise-smooth, then its length is the sum of the lengths of its smooth parts.

Lemma 23.21 (Basic properties)**.**

(i) Linearity: if $\alpha, \beta \in \mathbb{C}$, then

$$\int_\gamma (\alpha f(z) + \beta g(z))\,\mathrm{d}z = \alpha \int_\gamma f(z)\,\mathrm{d}z + \beta \int_\gamma g(z)\,\mathrm{d}z\,.$$

(ii) If γ^- is γ with the reverse orientation, then

$$\int_\gamma f(z)\,\mathrm{d}z = -\int_{\gamma^-} f(z)\,\mathrm{d}z\,.$$

(iii) One has the inequality

$$\left|\int_\gamma f(z)\,\mathrm{d}z\right| \leq \sup_{z\in\gamma} |f(z)| \cdot \Lambda(\gamma).$$

A ***primitive*** for f on Ω is a function F that is holomorphic on Ω and such that $F'(z) = f(z)$ for all $z \in \Omega$.

Proposition 23.22. *If a continuous function f has a primitive F in Ω, and γ is a curve in Ω that begins at w_1 and ends at w_2, then*

$$\int_\gamma f(z)\,\mathrm{d}z = F(w_2) - F(w_1). \tag{23.8}$$

Corollary 23.23. *If γ is a closed curve in an open set Ω, and f is continuous and has a primitive in Ω, then*

$$\int_\gamma f(z)\,\mathrm{d}z = 0.$$

Proof. This is immediate since the end-points of a closed curve coincide. □

Corollary 23.24. *If f is holomorphic in a region Ω and $f' = 0$, then f is constant.*

VII
General Topology

The study of topology simultaneously simplifies and generalises the theory of metric spaces. By discarding the metric, and focusing solely on the more basic and fundamental notion of an open set, many arguments and proofs are simplified. And many constructions (such as the important concept of a quotient space) cannot be carried out in the setting of metric spaces: they need the more general framework of topological spaces.

Chapter 24

Topological Spaces and Continuous Functions

24.1 Topologies

24.1.1 Definitions and Examples

Definition 24.1 (Topological space)**.** A collection $\mathcal{T}$ of subsets of a set X is said to be a ***topology*** on X if

(i) $\emptyset, X \in \mathcal{T}$;

(ii) if $\{U_i \mid i \in I\}$ are in $\mathcal{T}$, then $\bigcup_{i \in I} U_i \in \mathcal{T}$; (closed under arbitrary unions)

(iii) if $U_1, \dots, U_n \in \mathcal{T}$, then $\bigcap_{i=1}^{n} U_i \in \mathcal{T}$. (closed under finite intersections)

If $\mathcal{T}$ is a topology on X, then $(X, \mathcal{T})$ is a ***topological space***, and the members of $\mathcal{T}$ are called *open sets* in X.

Notation. If the topology $\mathcal{T}$ is clear, we simply omit it and denote a topological space as X.

Example 24.2. Let X be any non-empty set.

- The *discrete topology* on X is the set of all subsets of X; that is, $\mathcal{T} = \mathcal{P}(X)$.
- The *indiscrete topology* (or *trivial topology*) on X is $\mathcal{T} = \{X, \emptyset\}$.
- The *co-finite topology* on X consists of the empty set together with every subset U of X such that $X \setminus U$ is finite.
- Any metric space (X, d), with $\mathcal{T}$ equal to the collection of all subsets of X that are open in the metric space sense. This topology is called the *metric topology* on X.

Definition 24.3. Suppose $\mathcal{T}$ and $\mathcal{T}'$ are two topologies on a given set X. We say that

(i) $\mathcal{T}$ is ***finer*** than $\mathcal{T}'$ if $\mathcal{T} \supset \mathcal{T}'$;

(ii) $\mathcal{T}$ is ***coarser*** than $\mathcal{T}'$ if $\mathcal{T} \subset \mathcal{T}'$;

(iii) $\mathcal{T}$ is ***comparable*** with $\mathcal{T}'$ if either $\mathcal{T} \supset \mathcal{T}'$ or $\mathcal{T} \subset \mathcal{T}'$.

Example 24.4. The indiscrete topology is the coarsest topology possible, while the discrete topology is the finest topology possible.

24.1.2 Bases

In linear algebra, every vector space is generated by a basis. In topology, we have a similar notion, as it is usually hard to define a topology by specifying all the open sets.

Definition 24.5 (Basis)**.** A ***basis*** for a topology on X is a collection $\mathcal{B}$ of subsets of X (called *basis elements*) if

(i) for all $x \in X$, there exists $B \in \mathcal{B}$ such that $x \in B$;

(ii) for all $B_1, B_2 \in \mathcal{B}$ and $x \in B_1 \cap B_2$, there exists $B_3 \in \mathcal{B}$ such that $x \in B_3 \subset B_1 \cap B_2$.

Property (i) states that the elements of $\mathcal{B}$ cover X. Property (ii) can be visualised as follows:

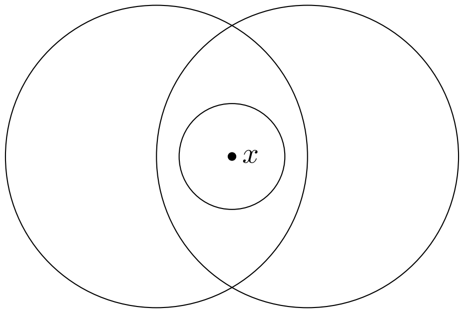


Figure 24.1: Property (ii) of Definition 24.5

Definition 24.6 (Topology generated by basis)**.** We define the *topology* $\mathcal{T}$ *generated by basis* $\mathcal{B}$ as follows:

$$U \in \mathcal{T} \iff \forall x \in U, \quad \exists B \in \mathcal{B}, \quad x \in B \subset U.$$

Lemma. *The collection* $\mathcal{T}$ *generated by the basis* $\mathcal{B}$ *is indeed a topology on* X*.*

Proof.

(i) $\emptyset$ satisfies the defining condition of openness vacuously, so $\emptyset \in \mathcal{T}$. $X \in \mathcal{T}$ follows from (i) of Definition 24.5.

(ii) Let $\{U_i \mid i \in I\}$ be a collection of elements of $\mathcal{T}$. We want to show that $U = \bigcup_{i \in I} U_i \in \mathcal{T}$.

Let $x \in U$. Then $x \in U_i$ for some $i \in I$. Since $U_i \in \mathcal{T}$, there exists $B \in \mathcal{B}$ such that $x \in B \subset U_i$. Thus $x \in B \subset U$, so $U \in \mathcal{T}$.

(iii) Let $U_1, U_2 \in \mathcal{T}$. We want to show that $U_1 \cap U_2 \in \mathcal{T}$.

Let $x \in U_1 \cap U_2$. Since $U_1 \in \mathcal{T}$, there exists $B_1 \in \mathcal{B}$ such that $x \in B_1 \subset U_1$; since $U_2 \in \mathcal{T}$, there exists $B_2 \in \mathcal{B}$ such that $x \in B_2 \subset U_2$. Then $x \in B_1 \cap B_2$.

Since $\mathcal{B}$ is a basis, by (ii) of Definition 24.5, there exists $B_3 \in \mathcal{B}$ such that $x \in B_3 \subset B_1 \cap B_2$. Thus $U_1 \cap U_2 \in \mathcal{T}$.

Finally, we show by induction that any finite intersection $U_1 \cap \cdots \cap U_n \in \mathcal{T}$. This is trivial for $n = 1$; suppose it true for $n - 1$ and prove it for n. Now

$$(U_1 \cap \cdots \cap U_n) = (U_1 \cap \cdots \cap U_{n-1}) \cap U_n.$$

By hypothesis, $U_1 \cap \cdots \cap U_{n-1} \in \mathcal{T}$. Thus the intersection of $U_1 \cap \cdots \cap U_{n-1}$ and U_n also belongs to $\mathcal{T}$.

□

Another way of describing the topology generated by a basis is given in the following result:

Lemma 24.7. *Let $\mathcal{T}$ be the topology on X generated by basis $\mathcal{B}$. Then $\mathcal{T}$ equals the collection of all unions of elements of $\mathcal{B}$.*

Proof. Let $\mathcal{B} = \{B_i \mid i \in I\}$.

$\boxed{\supset}$ If $B_i \in \mathcal{B}$, see that

$$\forall x \in B_i, \quad x \in B_i \subset B_i \implies B_i \in \mathcal{T}.$$

Since $\mathcal{T}$ is a topology, the arbitrary unions of B_i's must be in $\mathcal{T}$.

$\boxed{\subset}$ Let $U \in \mathcal{T}$. Then for each $x \in U$, there exists $B_x \in \mathcal{B}$ such that $x \in B_x \subset U$. Then $U = \bigcup_{x \in U} B_x$, so U is a union of elements of $\mathcal{B}$. □

Remark. The above result states that every $U \in \mathcal{T}$ can be expressed as a union of basis elements.

We have described in two different ways how to go from a basis to the topology it generates. Sometimes we need to go in the reverse direction, from a topology to a basis generating it. Here is one useful way of obtaining a basis for a given topology.

Lemma 24.8. *Let $(X, \mathcal{T})$ be a topological space. Suppose that $\mathcal{C}$ is a collection of open sets of X, such that*

$$\forall U \in \mathcal{T}, \quad \forall x \in U, \quad \exists C \in \mathcal{C}, \quad x \in C \subset U.$$

Then $\mathcal{C}$ is a basis for $\mathcal{T}$.

Proof. We first show that $\mathcal{C}$ is a basis.

(i) For all $x \in X$, since $X \in \mathcal{T}$, by hypothesis, there exists $C \in \mathcal{C}$ such that $x \in C \subset X$.

(ii) Let $x \in C_1 \cap C_2$, where $C_1, C_2 \in \mathcal{C} \subset \mathcal{T}$. Thus $C_1, C_2 \in \mathcal{T}$, so $C_1 \cap C_2 \in \mathcal{T}$. By hypothesis, there exists $C_3 \in \mathcal{C}$ such that $x \in C_3 \subset C_1 \cap C_2$.

Let $\mathcal{T}'$ be the topology generated by $\mathcal{C}$; that is,

$$U \in \mathcal{T}' \iff \forall x \in U, \quad \exists C \in \mathcal{C}, \quad x \in C \subset U.$$

We will show that $\mathcal{T} = \mathcal{T}'$.

$\boxed{\subset}$ Let $U \in \mathcal{T}$, $x \in U$. By hypothesis, there exists $C \in \mathcal{C}$ such that $x \in C \subset U$. By definition, $U \in \mathcal{T}'$. Hence $\mathcal{T} \subset \mathcal{T}'$.

$\boxed{\supset}$ Conversely, let $W \in \mathcal{T}'$. By 24.7, W is a union of elements of $\mathcal{C}$. Since each element of $\mathcal{C}$ is an element of $\mathcal{T}$ (and thus open), and a union of open sets is open, we have $W \in \mathcal{T}$. Hence $\mathcal{T}' \subset \mathcal{T}$. □

When topologies are given by bases, the next result is a criterion to determine whether one topology is finer than another.

Lemma 24.9. *Let $\mathcal{B}$ and $\mathcal{B}'$ be bases for the topologies $\mathcal{T}$ and $\mathcal{T}'$ respectively on X. Then the following are equivalent:*

(i) $\mathcal{T}'$ is finer than $\mathcal{T}$.

(ii) For all $x \in X$, and for all $B \in \mathcal{B}$ such that $x \in B$, there exists $B' \in \mathcal{B}'$ such that $x \in B' \subset B$.

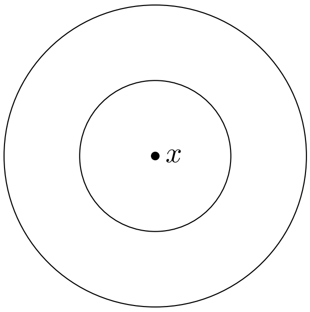


Figure 24.2: (ii) of 24.9

Proof.

$\boxed{\text{(ii)} \implies \text{(i)}}$ Let $U \in \mathcal{T}$. To show that $\mathcal{T} \subset \mathcal{T}'$, we want to show that $U \in \mathcal{T}'$.

Let $x \in U$. Since $\mathcal{B}$ generates $\mathcal{T}$, there exists $B \in \mathcal{B}$ such that $x \in B \subset U$. By (ii), there exists $B' \in \mathcal{B}'$ such that $x \in B' \subset B$. Then $x \in B' \subset U$, so $U \in \mathcal{T}'$, by definition.

$\boxed{\text{(i)} \implies \text{(ii)}}$ We are given $x \in X$ and $B \in \mathcal{B}$, with $x \in B$.

Now $B \in \mathcal{T}$ by definition, and $\mathcal{T} \subset \mathcal{T}'$ by (i); therefore, $B \in \mathcal{T}'$. Since $\mathcal{T}'$ is generated by $\mathcal{B}'$, there exists $B' \in \mathcal{B}'$ such that $x \in B' \subset B$. ☐

We now define three topologies on the real line $\mathbb{R}$, all of which are of interest.

Definition 24.10.

(i) Let $\mathcal{B}$ be the collection of all open intervals in $\mathbb{R}$. The topology generated by $\mathcal{B}$ is called the ***standard topology*** on $\mathbb{R}$.

Whenever we consider $\mathbb{R}$, we shall suppose it is given this topology unless stated otherwise.

(ii) Let $\mathcal{B}'$ be the collection of all half-open intervals of the form $[a, b)$. The topology generated by $\mathcal{B}'$ is called the ***lower limit topology*** on $\mathbb{R}$.

When $\mathbb{R}$ is given the lower limit topology, we denote it by $\mathbb{R}_\ell$.

(iii) Let $K = \{\frac{1}{n} \mid n \in \mathbb{Z}^+\}$, and let $\mathcal{B}''$ be the collection of all open intervals (a, b), along with all sets of the form $K \setminus (a, b)$. The topology generated by B'' is called the ***K-topology*** on $\mathbb{R}$.

When $\mathbb{R}$ is given this topology, we denote it by $\mathbb{R}_K$.

It is easy to see that all three of these collections are bases; in each case, the intersection of two basis elements is either another basis element or is empty. The relation between these topologies is the following:

Lemma 24.11. *The topologies of $\mathbb{R}_\ell$ and $\mathbb{R}_K$ are strictly finer than the standard topology on $\mathbb{R}$, but are not comparable with one another.*

Proof. Let $\mathcal{T}$, $\mathcal{T}'$, and $\mathcal{T}''$ be the topologies of $\mathbb{R}$, $\mathbb{R}_\ell$, and $\mathbb{R}_K$, respectively.

Given a basis element (a, b) for $\mathcal{T}$ and a point x of (a, b), the basis element $[x, b)$ for $\mathcal{T}'$ contains x and lies in (a, b). On the other hand, given the basis element $[x, d)$ for $\mathcal{T}'$, there is no open interval (a, b) that contains x and lies in $[x, d)$. Thus $\mathcal{T} \subset \mathcal{T}'$, so $\mathcal{T}'$ is strictly finer than $\mathcal{T}$.

A similar argument applies to $\mathbb{R}_K$. Given a basis element (a, b) for $\mathcal{T}$ and a point x of (a, b), this same interval is a basis element for $\mathcal{T}''$ that contains x. On the other hand, given the basis element $B = (-1, 1) \setminus K$ for $\mathcal{T}''$ and the point 0 of B, there is no open interval that contains 0 and lies in $\mathcal{B}$.

We leave it to you to show that the topologies of $\mathbb{R}_\ell$ and $\mathbb{R}_K$ are not comparable. □

Since the topology generated by a basis $\mathcal{B}$ may be described as the collection of arbitrary unions of elements of $\mathcal{B}$ (by 24.7), what happens if we start with a given collection of sets and take finite intersections of them as well as arbitrary unions? This leads to the notion of a *subbasis* for a topology.

Definition 24.12 (Subbasis). A ***subbasis*** $\mathcal{S}$ for a topology on X is a collection of subsets of X whose union equals X.

Definition 24.13 (Topology generated by subbasis). The *topology $\mathcal{T}$ generated by the subbasis $\mathcal{S}$* is defined as the collection of all unions of finite intersections of elements of $\mathcal{S}$:

$$U \in \mathcal{T} \iff U = \text{union of finite intersections in } \mathcal{S}.$$

Lemma. *The collection $\mathcal{T}$ generated by the subbasis $\mathcal{S}$ is a topology.*

Proof. Consider the collection

$$\mathcal{B} = \{\text{all finite intersections of elements of } \mathcal{S}\}.$$

It suffices to show that $\mathcal{B}$ is a basis, for then by 24.7, the collection $\mathcal{T}$ of all unions of elements of $\mathcal{B}$ is a topology.

(i) Let $x \in X$. Then x belongs to an element of $\mathcal{S}$, and thus belongs to an element of $\mathcal{B}$.

(ii) Let

$$B_1 = S_1 \cap \cdots \cap S_m, \quad B_2 = S_1' \cap \cdots \cap S_n'$$

be two elements of $\mathcal{B}$. Their intersection

$$B_1 \cap B_2 = (S_1 \cap \cdots \cap S_m) \cap (S_1' \cap \cdots \cap S_n')$$

is also a finite intersection of elements of $\mathcal{S}$, so it belongs to $\mathcal{B}$.

$\square$

24.2 Examples of Topologies

24.2.1 Order Topology

Definition 24.14 (Order topology). Let $(X, <)$, $|X| > 1$. Let $\mathcal{B}$ be the collection of all sets of the following types:

(i) All open intervals (a, b) in X.

(ii) All intervals of the form $[a_0, b)$, where a_0 is the smallest element (if any) of X.

(iii) All intervals of the form $(a, b_0]$, where b_0 is the largest element (if any) of X.

The topology generated by $\mathcal{B}$ is called the ***order topology***.

We need to check that $\mathcal{B}$ is a basis of X.

(i) Every $x \in X$ lies in some element of $\mathcal{B}$: the smallest element (if any) lies in all sets of type (ii), the largest element (if any) lies in all sets of type (iii), and every other element lies in a set of type (i).

(ii) The intersection of any two sets of the preceding types is a set of one of these types, or is empty. Several cases need to be checked; we leave it to you.

For instance, let $x \in (a, b) \cap (c, d)$. Let $p = \max\{a, c\}$, $q = \min\{b, d\}$. Then $x \in (p, q) \subset (a, b) \cap (c, d)$, where $(p, q) \in \mathcal{B}$.

Example 24.15.

- The standard topology on $\mathbb{R}$ is just the order topology derived from the usual order on $\mathbb{R}$.

Definition 24.16. Let $(X, <)$, $a \in X$. Then the following subsets of X are ***rays*** determined by a:

$$(a, +\infty) = \{x \in X \mid x > a\},$$
$$[a, +\infty) = \{x \in X \mid x \geq a\},$$
$$(-\infty, a) = \{x \in X \mid x < a\},$$
$$(-\infty, a] = \{x \in X \mid x \leq a\}.$$

$(a, +\infty)$ and $(-\infty, a)$ are called *open rays*, since they are open; for instance, $(a, +\infty) = \bigcup_{x>a}(a, x)$. Similarly, $[a, +\infty)$ and $(-\infty, a]$ are *closed rays*.

Lemma 24.17. *The collection of open rays form a subbasis for the order topology.*

Proof. Let $\mathcal{T}$ be the order topology on X, let $\mathcal{T}'$ be the topology generated by the subbasis of open rays. We will show that $\mathcal{T} = \mathcal{T}'$.

- Because the open rays are open in the order topology, the topology they generate is contained in the order topology. Hence $\mathcal{T}' \subset \mathcal{T}$.

- On the other hand, every basis element for the order topology equals a finite intersection of open rays; the interval (a, b) equals the intersection of $(-\infty, b)$ and $(a, +\infty)$, while $[a_0, b)$ and $(a, b_0]$, if they exist, are themselves open rays. Hence the topology generated by the open rays contains the order topology, so $\mathcal{T} \subset \mathcal{T}'$.

□

24.2.2 Product Topology

If X and Y are topological spaces, there is a standard way of defining a topology on the cartesian product $X \times Y$. We consider this topology now and study some of its properties.

Definition 24.18 (Product topology)**.** Let $(X, \mathcal{T}_X)$ and $(Y, \mathcal{T}_Y)$ be topological spaces. The ***product topology*** on $X \times Y$ is the topology $\mathcal{T}_{X\times Y}$ with basis

$$\mathcal{B} = \{U \times V \mid U \in \mathcal{T}_X, V \in \mathcal{T}_Y\}.$$

Lemma. *$\mathcal{B}$ is a basis.*

Proof.

(i) $X \times Y$ is a basis element, so every element of $X \times Y$ is contained in $X \times Y$.

(ii) Let $U_1 \times V_1, U_2 \times V_2 \in \mathcal{B}$. Then their intersection is

$$(U_1 \times V_1) \cap (U_2 \times V_2) = (U_1 \cap U_2) \times (V_1 \cap V_2).$$

Since $U_1 \cap U_2 \in \mathcal{T}_X$, $V_1 \cap V_2 \in \mathcal{T}_Y$, we have that $(U_1 \cap U_2) \times (V_1 \cap V_2) \in \mathcal{B}$.

□

What can one say if the topologies on X and Y are given by bases? The answer is as follows:

Lemma 24.19. *If $\mathcal{B}$ is a basis for the topology of X, $\mathcal{C}$ is a basis for the topology of Y, then the collection*

$$\mathcal{D} = \{B \times C \mid B \in \mathcal{B}, C \in \mathcal{C}\}$$

is a basis for the topology of $X \times Y$.

Proof. We apply 24.8.

Let W be an open set of $X \times Y$, and let $(x, y) \in W$. By definition of product topology there exists a basis element $U \times V$ such that $(x, y) \in U \times V \subset W$.

Since $\mathcal{B}$ and $\mathcal{C}$ are bases for X and Y respectively, there exists $B \in \mathcal{B}$ such that $x \in B \subset U$, and $C \in \mathcal{C}$ such that $y \in C \subset V$. Then $(x, y) \in B \times C \subset W$.

Thus the collection $\mathcal{D}$ meets the criterion of 24.8, so $\mathcal{D}$ is a basis for $X \times Y$. □

Definition 24.20 (Projection map)**.** Define the projection map of $X \times Y$ onto X as

$$\begin{aligned}\pi_1 \colon X \times Y &\to X \\ (x, y) &\mapsto x\end{aligned}$$

and the projection map of $X \times Y$ onto Y as

$$\begin{aligned}\pi_2 \colon X \times Y &\to Y \\ (x, y) &\mapsto y\end{aligned}$$

Remark. We use the word "onto" because π_1 and π_2 are surjective (unless one of the spaces X or Y happens to be empty, in which case $X \times Y$ is empty and our whole discussion is empty as well!).

If U is an open subset of X, then the set $\pi_1^{-1}(U)$ is precisely the set $U \times Y$, which is open in $X \times Y$. Similarly, if V is open in Y, then $\pi_2^{-1}(V) = X \times V$, which is open in $X \times Y$.

Lemma 24.21. *The collection*

$$\mathcal{S} = \{\pi_1^{-1}(U) \mid U \text{ open in } X\} \cup \{\pi_2^{-1}(V) \mid V \text{ open in } Y\}$$

is a subbasis for the product topology on $X \times Y$.

Proof. Let $\mathcal{T}$ denote the product topology on $X \times Y$; let $\mathcal{T}'$ be the topology generated by $\mathcal{S}$. We will show that $\mathcal{T} = \mathcal{T}'$.

$\boxed{\supset}$ Since every element of $\mathcal{S}$ belongs to $\mathcal{T}$, so do arbitrary unions of finite intersections of elements of $\mathcal{S}$. Thus $\mathcal{T}' \subset \mathcal{T}$.

$\boxed{\subset}$ Every basis element $U \times V$ for the topology $\mathcal{T}$ is a finite intersection of elements of $\mathcal{S}$, since

$$U \times V = \pi_1^{-1}(U) \cap \pi_2^{-1}(V).$$

Thus $U \times V$ belongs to $\mathcal{T}'$, so $\mathcal{T} \subset \mathcal{T}'$. □

24.2.3 Subspace Topology

Definition 24.22 (Subspace topology)**.** Let $(X, \mathcal{T})$ be a topological space. If $Y \subset X$, the collection

$$\mathcal{T}_Y := \{Y \cap U \mid U \in \mathcal{T}\}$$

is a topology on Y, called the ***subspace topology***. With this topology, Y is called a ***subspace*** of X; its open sets consist of all intersections of open sets of X with Y.

Lemma. *$\mathcal{T}_Y$ is a topology.*

Proof.

(i) $\emptyset = Y \cap \emptyset$ so $\emptyset \in \mathcal{T}_Y$. $Y = Y \cap X$ so $Y \in \mathcal{T}_Y$.

(ii) $\mathcal{T}_Y$ is closed under finite intersections, since

$$(U_1 \cap Y) \cap \cdots \cap (U_n \cap Y) = (U_1 \cap \cdots \cap U_n) \cap Y.$$

(iii) $\mathcal{T}_Y$ is closed under arbitrary unions, since

$$\bigcup_{i \in I} (U_i \cap Y) = \left(\bigcup_{i \in I} U_i \right) \cap Y.$$

□

Lemma 24.23. *If $\mathcal{B}$ is a basis for the topology of X, then*

$$\mathcal{B}_Y = \{B \cap Y \mid B \in \mathcal{B}\}$$

is a basis for the subspace topology on Y.

Proof. Let U be open in X, $y \in U \cap Y$. Since $\mathcal{B}$ is a basis for the topology of X, there exists $B \in \mathcal{B}$ such that $y \in B \subset U$. Then $y \in B \cap Y \subset U \cap Y$.

By 24.8, $\mathcal{B}_Y$ is a basis for the subspace topology on Y. □

When dealing with a space X and a subspace Y, one needs to be careful when one uses the term "open set". Does one mean an element of the topology of Y or an element of the topology of X?

We make the following definition: If Y is a subspace of X, we say that a set U is *open in Y* if it belongs to the topology of Y; this implies in particular that it is a subset of Y. We say that U is *open in X* if it belongs to the topology of X.

There is a special situation in which every set open in Y is also open in X:

Lemma 24.24. *Let Y be a subspace of X. If U is open in Y, and Y is open in X, then U is open in X.*

Proof. Since U is open in Y, $U = Y \cap V$ for some set V open in X. Since Y and V are both open in X, so is $Y \cap V$. □

Now let us explore the relation between the subspace topology and the product topology.

Lemma 24.25. *If A is a subspace of X, and B is a subspace of Y, then the product topology on $A \times B$ is the same as the topology $A \times B$ inherits as a subspace of $X \times Y$.*

Proof. The set $U \times V$ is the general basis element for $X \times Y$, where U is open in X and V is open in Y. Hence $(U \times V) \cap (A \times B)$ is the general basis element for the subspace topology on $A \times B$. Now

$$(U \times V) \cap (A \times B) = (U \cap A) \times (V \cap B).$$

Since $U \cap A$ and $V \cap B$ are the general open sets for the subspace topologies on A and B, respectively, the set $(U \cap A) \times (V \cap B)$ is the general basis element for the product topology on $A \times B$.

The conclusion we draw is that the bases for the subspace topology on $A \times B$ and for the product topology on $A \times B$ are the same. Hence the topologies are the same. $\square$

24.3 Closed Sets and Limit Points

Let X be a topological space. If E is an open set containing x, we say E is a ***neighbourhood*** of x.

24.3.1 Closed Sets

Definition 24.26 (Closed set)**.** We say $E \subset X$ is ***closed*** if its complement $E^c = X \setminus E$ is open.

The collection of closed subsets of a space X has properties similar to those satisfied by the collection of open subsets of X:

Lemma 24.27. *Let X be a topological space.*

(i) $\emptyset$ and X are closed.

(ii) Arbitrary intersections of closed sets are closed.

(iii) Finite unions of closed sets are closed.

Proof.

(i) $\emptyset$ and X are closed because they are the complements of the open sets X and $\emptyset$, respectively.

(ii) Suppose $\{A_i \mid i \in I\}$ is a collection of closed sets. By de Morgan's laws,

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

Since each ${A_i}^c$ is open, the RHS is open since it is an arbitrary union of open sets. Hence $\bigcap A_i$ is closed.

(iii) Suppose A_i is closed for $i = 1, \dots, n$. Then

$$\left(\bigcup_{i=1}^{n} A_i\right)^c = \bigcap_{i=1}^{n} {A_i}^c.$$

The RHS is a finite intersection of open sets and is thus open. Hence $\bigcup A_i$ is closed.

$\square$

Remark. Note that $\emptyset$ and X are both open and closed. This explains the statement "a door is not a set": a door must be either open or closed, and cannot be both, while a set can be open, or closed, or both, or neither!

Remark. Instead of using open sets, one could just as well specify a topology on a space by giving a collection of sets (to be called "closed sets") satisfying the three properties above. However this procedure has no particular advantage over the one we have adopted, so we shall use open sets to define topologies.

If Y is a subspace of X, we say E is *closed in* Y if $E \subset Y$ and E is closed in the subspace topology of Y (that is, $Y \setminus E$ is open in Y).

Proposition 24.28. *Let Y be a subspace of X. Then E is closed in Y if and only if it equals the intersection of a closed set of X with Y.*

Proof.

$\boxed{\Longleftarrow}$ Suppose $E = C \cap Y$, where C is closed in X.

Then $X \setminus C$ is open in X, so that $(X \setminus C) \cap Y$ is open in Y, by definition of subspace topology.

But $(X \setminus C) \cap Y = Y \setminus E$. Hence $Y \setminus E$ is open in Y, so that E is closed in Y.

$\boxed{\Longrightarrow}$ Suppose E is closed in Y. Then $Y \setminus E$ is open in Y.

By definition of subspace topology, $Y \setminus E$ is the intersection of an open set U of X with Y. Note that $X \setminus U$ is closed in X, and $E = Y \cap (X \setminus U)$. Thus E equals the intersection of a closed set of X with Y, as desired. □

A set E that is closed in the subspace Y may or may not be closed in the larger space X. As was the case with open sets, there is a criterion for E to be closed in E:

Proposition 24.29. *Let Y be a subspace of X. If E is closed in Y, and Y is closed in X, then E is closed in X.*

Proof. □

24.3.2 Closure and Interior

Definition 24.30. The ***interior*** $\operatorname{Int} E$ of $E \subset X$ is the union of all open sets contained in E. The ***closure*** $\overline{E}$ of E is the intersection of all closed sets containing E.

Obviously $\operatorname{Int} E$ is an open set and $\overline{E}$ is a closed set; furthermore,

$$\operatorname{Int} E \subset E \subset \overline{E}.$$

If E is open, $E = \operatorname{Int} E$; while if E is closed, $E = \overline{E}$.

We shall not make much use of the interior of a set, but the closure of a set will be quite important.

When dealing with a topological space X and a subspace Y, one needs to exercise care in taking closures of sets. If E is a subset of Y, the closure of E in Y and the closure of E in X will in general be different.

Notation. We reserve the notation $\overline{E}$ to mean the closure of E in X.

The next result shows that the closure of E in Y can be expressed in terms of $\overline{E}$.

Proposition 24.31. *Let Y be a subspace of X; let $E \subset Y$, let $\overline{E}$ denote the closure of E in X. Then the closure of E in Y equals $\overline{E} \cap Y$.*

Proof. Let F denote the closure of E in Y. We will show that $F = \overline{E} \cap Y$.

$\boxed{\subset}$ $\overline{E}$ is closed in X, so $\overline{E} \cap Y$ is closed in Y.

Since $\overline{E} \cap Y$ contains E, and since by definition F equals the intersection of all closed subsets of Y containing E, we must have $F \subset (\overline{E} \cap Y)$.

$\boxed{\supset}$ Since F is closed in Y, $F = C \cap Y$ for some set C closed in X. Then C is a closed set of X containing E; since $\overline{E}$ is the intersection of all such closed sets, we have $\overline{E} \subset C$. Then $(\overline{E} \cap Y) \subset (C \cap Y) = F$. $\square$

The next result provides a convenient characterisation of the closure of a set.

Lemma 24.32. *If $E \subset X$, then $x \in \overline{E}$ if and only if every neighbourhood of x intersects E.*

Proof.

$\boxed{\implies}$ We prove the contrapositive. Suppose $x \notin \overline{E}$.

Then $\overline{E}^c$ is an open set containing x which does not intersect E, as desired.

$\boxed{\impliedby}$ We prove the contrapositive. Suppose there exists a neighbourhood U of x which does not intersect E. Then U^c is a closed set containing E. By definition of closure $\overline{E}$, U^c must contain $\overline{E}$; hence x cannot be in $\overline{E}$. $\square$

Corollary 24.33. *Supposing the topology of X is given by a basis, then $x \in \overline{E}$ if and only if every basis element B containing x intersects E.*

Proof.

$\boxed{\Longrightarrow}$ If every open set containing x intersects E, so does every basis element B containing x, because B is an open set.

$\boxed{\Longleftarrow}$ If every basis element containing x intersects E, so does every open set U containing x, because U contains a basis element that contains x. $\square$

24.3.3 Limit Points

Definition 24.34. Suppose $E \subset X$. We say $x \in X$ is a ***limit point*** of E, if every neighbourhood of x intersects E in some point other than x itself.

Let E' denote the set of all limit points of E.

We shall now see that limit points provide another way to describe the closure of a set.

Proposition 24.35. *Let $E \subset X$. Then $\overline{E} = E \cup E'$.*

Proof.

$\boxed{\supset}$ Let $x \in E'$. Then every neighbourhood of x intersects E (in a point different from x).

By 24.32, $x \in \overline{E}$. Hence $E' \subset \overline{E}$. Since $E \subset \overline{E}$, it follows that $E \cup E' \subset \overline{E}$.

$\boxed{\subset}$ Let $x \in \overline{E}$.

- If $x \in E$, it is trivial that $x \in E \cup E'$.
- If $x \notin E$, since $x \in \overline{E}$, we know that every neighbourhood U of x intersects A. Since $x \notin E$, the set U must intersect E in a point different from x. Then $x \in E'$.

Thus $x \in E \cup E'$, so $\overline{E} \subset E \cup E'$. □

Corollary 24.36. *$E \subset X$ is closed if and only if it contains all its limit points.*

Proof. E is closed if and only if $E = \overline{E}$, and the latter holds if and only if $E' \subset E$. □

24.3.4 Hausdorff Spaces

Definition 24.37 (Hausdorff space)**.** A topological space X is a ***Hausdorff space*** if, for all distinct $x, y \in X$, there exist neighbourhoods U and V of x and y respectively that are disjoint.

Lemma 24.38. *Every finite point set in a Hausdorff space X is closed.*

Proof. It suffices to show that every one-point set $\{x_0\}$ is closed.

If x is a point of X different from x_0, then x and x_0 have disjoint neighbourhoods U and V, respectively. Since U does not intersect $\{x_0\}$, the point x cannot belong to the closure of the set $\{x_0\}$.

Hence the closure of $\{x_0\}$ is $\{x_0\}$ itself, so $\{x_0\}$ is closed. □

The condition that finite point sets be closed is in fact weaker than the Hausdorff condition. For example, $\mathbb{R}$ in the finite complement topology is not a Hausdorff space, but it is a space in which finite point sets are closed.

The condition that finite point sets be closed is called the T_1 ***axiom***.

Proposition 24.39. *Let X be a space satisfying the T_1 axiom; let $E \subset X$. Then x is a limit point of E if and only if every neighbourhood of x contains infinitely many points of E.*

Proof.

$\boxed{\Longleftarrow}$ If every neighbourhood of x intersects E in infinitely many points, it certainly intersects E in some point other than x itself, so that x is a limit point of E.

$\boxed{\Longrightarrow}$ Let x be a limit point of E. Suppose, for a contradiction, that there exists a neighbourhood U of x which intersects E at finitely many points.

Then U also intersects $E \setminus \{x\}$ at finitely many points, say $x_1, \dots, x_m$; that is,

$$U \cap (E \setminus \{x\}) = \{x_1, \dots, x_m\}.$$

Since the finite point set $\{x_1, \dots, x_m\}$ is closed, its complement $\{x_1, \dots, x_m\}^c$ is open. Then

$$U \cap \{x_1, \dots, x_m\}^c$$

is a neighbourhood of x that does not intersect $E \setminus \{x\}$. This contradicts the assumption that x is a limit point of E. □

Definition 24.40 (Limit)**.** If a sequence (x_n) in a Hausdorff space X converges to $x \in X$, we write $x_n \to x$, and say x is the ***limit*** of (x_n).

Lemma 24.41 (Uniqueness of limit)**.** *If X is a Hausdorff space, then a sequence of points of X converges to at most one point of X.*

Proof. Suppose (x_n) is a sequence in X, $x_n \to x$.

If $y \neq x$, let U and V be disjoint neighborhoods of x and y, respectively. Since U contains x_n for all but finitely many values of n, the set V cannot. Therefore, $x_n \not\to y$. $\square$

Proposition 24.42. *Every simply ordered set is a Hausdorff space in the order topology. The product of two Hausdorff spaces is a Hausdorff space. A subspace of a Hausdorff space is a Hausdorff space.*

24.3.5 Compactness

Definition 24.43. We say $K \subset X$ is ***compact*** if every open cover of K contains a finite subcover; that is, if $\{U_i \mid i \in I\}$ is a collection of open sets whose union contains K, then the union of some finite subcollection of $\{U_i\}$ also contains K.
In particular, if X is itself compact, then X is called a *compact space*.
X is *locally compact* if every point of X has a neighbourhood whose closure is compact.

Obviously, every compact space is locally compact.

We recall the Heine-Borel theorem: The compact subsets of a euclidean space $\mathbb{R}^n$ are precisely those that are closed and bounded. From this it follows easily that $\mathbb{R}^n$ is a locally compact Hausdorff space. Also, every metric space is a Hausdorff space.

The next result states that a closed subset of a compact set is compact.

Lemma 24.44. *Suppose K is compact and F is closed, in a topological space X. If $F \subset K$, then F is compact.*

Proof. If $\{U_i \mid i \in I\}$ is an open cover of F and $W = F^c$, then $W \cup \bigcup_{i \in I} U_i$ covers X; hence there is a finite collection $\{U_{i_k}\}$ such that

$$K \subset \bigcup_{k=1}^{n} U_{i_k}.$$

Then $F \subset \bigcup_{k=1}^{n} U_{i_k}$. □

Corollary 24.45. *Suppose $A \subset B$. If $\overline{B}$ is compact, then $\overline{A}$ is compact.*

Proposition 24.46. *Suppose X is a Hausdorff space, $K \subset X$ is compact, and $x \in K^c$. Then there exist open sets U and W such that $x \in U$, $K \subset W$, and $U \cap W = \emptyset$.*

Proof. Let $y \in K$. The Hausdorff separation axiom implies the existence of disjoint open sets U_y and V_y such that $x \in U_y$ and $y \in U_y$. Since K is compact, there exist points $y_1, \ldots, y_n \in K$ such that

$$K \subset V_{y_1} \cup \cdots \cup V_{y_n}.$$

Our requirements are then satisfied by the sets

$$U = U_{y_1} \cap \cdots \cap U_{y_n} \quad \text{and} \quad W = V_{y_1} \cup \cdots \cup V_{y_n}.$$

□

Corollary 24.47. *Compact subsets of Hausdorff spaces are closed.*

Corollary 24.48. *If F is closed and K is compact in a Hausdorff space, then $F \cap K$ is compact.*

Proposition 24.49. *If $\{K_i \mid i \in I\}$ is a collection of compact subsets of a Hausdorff space and if $\bigcap_{i \in I} K_i = \emptyset$, then some finite subcollection of $\{K_i\}$ also has empty intersection.*

Proposition 24.50. *Suppose U is open in a locally compact Hausdorff space X, $K \subset U$, and K is compact. Then there exists an open set V with compact closure such that*

$$K \subset V \subset \overline{V} \subset U.$$

24.4 Continuous Functions

24.4.1 Continuity of a Function

Definition 24.51. Let X and Y be topological spaces. We say $f\colon X \to Y$ is ***continuous*** if $f^{-1}(U)$ is an open set in X for every open set U in Y.

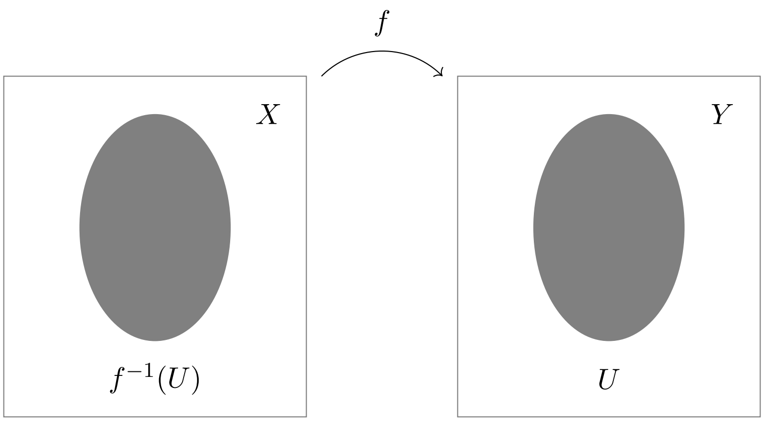


Figure 24.3: Pre-image of a set

The definition of continuity in Definition 24.51 is a global one. Frequently it is desirable to define continuity locally:

> $f\colon X \to Y$ is *continuous* at $x_0 \in X$ if, for every neighbourhood U of $f(x_0)$, there exists a neighbourhood V of x_0 such that $f(V) \subset U$.

When X and Y are metric spaces, the local and global definitions are equivalent. The following result relates the local and global definitions of continuity for topological spaces.

Lemma 24.52 (Equivalent definitions for continuity)**.** *Let X and Y be topological spaces. Then $f\colon X \to Y$ is continuous if and only if f is continuous at every point of X.*

Proof.

$\boxed{\implies}$ Suppose f is continuous, let $x_0 \in X$.

Let U be a neighbourhood of $f(x_0)$, then U is open in X. By continuity of f, $f^{-1}(U)$ is open in X. Since

- $f^{-1}(U)$ is a neighbourhood of x_0, and
- $f(f^{-1}(U)) \subset U$,

it follows that f is continuous at x_0.

$\boxed{\impliedby}$ Suppose f is continuous at every point of X.

Let U be a open set in Y. By (local) continuity of f, every point $x \in f^{-1}(U)$ has a neighbourhood V_x such that $f(V_x) \subset U$. Thus $V_x \subset f^{-1}(U)$. Since

$$f^{-1}(U) = \bigcup_x V_x,$$

$f^{-1}(U)$ is the union of open sets V_x, so $f^{-1}(U)$ is itself open. Hence f is continuous. □

Lemma 24.53. *Let X and Y be topological spaces; let $f\colon X \to Y$. Then the following are equivalent:*

(i) f is continuous.

(ii) $f(\overline{A}) \subset \overline{f(A)}$ for every $A \subset X$.

(iii) $f^{-1}(B)$ is closed in X for every closed set B of Y.

(iv) For each $x \in X$ and each neighbourhood V of $f(x)$, there is a neighbourhood U of x such that $f(U) \subset V$.

The next result states that continuous functions of continuous functions are continuous.

Lemma 24.54. *Let X, Y and Z be topological spaces. If $f\colon X \to Y$ and $g\colon Y \to Z$ are continuous, then $h = g \circ f$ is continuous.*

Proof. Let U be open in Z. By continuity of g, we have that $g^{-1}(U)$ is open in Y. Note that

$$h^{-1}(U) = f^{-1}\left(g^{-1}(U)\right).$$

If f is continuous, it follows that $h^{-1}(U)$ is open, so h is continuous. □

24.4.2 Homeomorphisms

Definition 24.55. Let X and Y be topological spaces; let $f\colon X \to Y$ be a bijection. We say f is a ***homeomorphism*** if both f and its inverse f^{-1} are continuous.

24.4.3 Constructing Continuous Functions

24.5 Metric Topology

One of the most important and frequently used ways of imposing a topology on a set is to define the topology in terms of a metric on the set.

Definition 24.56.

24.6 Quotient Topology

Definition 24.57 (Quotient map)**.** Let X and Y be topological spaces; let $p\colon X \to Y$ be a surjective map. The map p is said to be a ***quotient map*** if U is open in Y if and only if $p^{-1}(U)$ is open in X.

Definition 24.58 (Quotient topology)**.** If X is a space and A is a set and if $p\colon X \to A$ is a surjective map, then there exists exactly one topology $\mathcal{T}$ on A relative to which p is a quotient map; it is called the ***quotient topology*** induced by p.

Lemma. *The quotient topology $\mathcal{T}$ is a topology.*

Definition 24.59 (Quotient space)**.** Let X be a topological space, and let X^* be a partition of X into disjoint subsets whose union is X. Let $p\colon X \to X^*$ be the surjective map that carries each point of X to the element of X^* containing it. In the quotient topology induced by p, the space X^* is called a ***quotient space*** of X.

VIII
Measure Theory

In measure theory, the main idea is that we want to assign "sizes" to different sets. For example, we might think $[0, 2] \subset \mathbb{R}$ has size 2, while perhaps $\mathbb{Q} \subset \mathbb{R}$ has size 0. This is known as a *measure*.

One of the main applications of a measure is that we can use it to come up with a new definition of an integral, known as the *Lebesgue integral*. Instead of integrating functions $[a, b] \to \mathbb{R}$ only, we can replace the domain with any measure space, allowing us to integrate a much wider class of functions.

Chapter 25

Measures

25.1 Introduction

One of the most venerable problems in geometry is to determine the area of volume of a region in the plane or in 3-space.

Ideally, for $n \in \mathbb{N}$ we would like to have a function μ that assigns to each $E \subset \mathbb{R}^n$ a number $\mu(E) \in [0, \infty]$, the n-dimensional measure of E. We desire such a function μ to possess the following properties:

1. If $E_1, E_2, \ldots$ is a finite or infinite sequence of disjoint sets, then

$$\mu(E_1 \cup E_2 \cup \cdots) = \mu(E_1) + \mu(E_2) + \cdots .$$

2. If E is congruent to F (that is, if E can be transformed into F by translations, rotations, and reflections), then $\mu(E) = \mu(F)$.

3. $\mu(Q) = 1$, where Q is the unit cube

$$Q = \{x \in \mathbb{R}^n \mid 0 \leq x_i < 1,\ i = 1, \ldots, n\}.$$

Unfortunately, these conditions are mutually inconsistent.

Proposition 25.1. *There does not exist such a volume measure defined on $\mathcal{P}(\mathbb{R}^n)$.*

Proof. It suffices to prove the case when $n = 1$. Define an equivalence relation $\sim$ on $\mathbb{R}$ by

$$x \sim y \iff x - y \in \mathbb{Q}.$$

We check that this is an equivalence relation:

(i) $x - x = 0$ implies $x \sim x$.

(ii) $x - y \in \mathbb{Q}$ implies $y - x \in \mathbb{Q}$.

(iii) $x - y, y - z \in \mathbb{Q}$ implies $x - z = (x - y) + (y - z) \in \mathbb{Q}$.

Let $N \subset [0,1)$ be a subset containing exactly one representative from each equivalence class of $\sim$. (To find such an N, one must invoke the axiom of choice.)

Next, let $R = \mathbb{Q} \cap [0,1)$, and for each $r \in R$ let

$$N_r = \{x + r \mid x \in N \cap [0, 1-r]\} \cup \{x + r - 1 \mid x \in N \cap [1-r, 1]\}.$$

That is, to obtain N_r, shift N to the right by r units and then shift the part that sticks out beyond $[0,1)$ one unit to the left. Then $N_r \subset [0,1)$, and every $x \in [0,1)$ belongs to precisely one N_r.

Indeed, if y is the element of N that belongs to the equivalence class of x, then $x \in N_r$, where $r = x - y$ if $x \geq y$ or $r = x + y - 1$ if $x < y$; on the other hand, if $x \in N_r \cap N_s$, then $x - r$ (or $x - r + 1$) and $x - s$ (or $x - s + 1$) would be distinct elements of N belonging to the same equivalence class, which is impossible.

Suppose now that $\mu\colon \mathcal{P}(\mathbb{R}) \to [0,\infty]$ satisfies (1), (2) and (3). By (1) and (2),

$$\mu(N) = \mu(N \cap [0, 1-r)) + \mu(N \cap [1-r, 1]) = \mu(N_r)$$

for any $r \in R$. Also, since R is countable and $[0,1)$ is the disjoint union of the N_r's,

$$\mu([0,1]) = \sum_{r \in R} \mu(N_r)$$

by (1) again. But $\mu([0,1]) = 1$ by (3), and since $\mu(N_r) = \mu(N)$, the sum on the right is either 0 (if $\mu(N) = 0$) or ∞ (if $\mu(N) > 0$). Hence no such μ can exist. $\square$

Theorem 25.2 (Banach–Tarski paradox)**.** *Let U and V be arbitrary bounded open sets in $\mathbb{R}^n$, $n \geq 3$. There exist $k \in \mathbb{N}$ and subsets $E_1, \ldots, E_k, F_1, \ldots, F_k$ of $\mathbb{R}^n$ such that*

1. *the E_j's are disjoint and their union is U;*
2. *the F_j's are disjoint and their union is V;*
3. *E_j is congruent to F_j for $j = 1, \ldots, k$.*

25.2 σ-algebras

Definition 25.3 (σ-algebra)**.** Let X be a non-empty set. We say a non-empty collection of subsets $\mathcal{A} \subset \mathcal{P}(X)$ is an ***algebra*** on X if

(i) if $E_1, \dots, E_n \in \mathcal{A}$, then $\bigcup_{i=1}^n E_i \in \mathcal{A}$; (closed under finite unions)

(ii) if $E \in \mathcal{A}$, then $E^c \in \mathcal{A}$. (closed under complements)

A ***σ-algebra*** is an algebra that is closed under *countable unions*.
If $\mathcal{A}$ is a σ-algebra on X, then we say $(X, \mathcal{A})$ is a ***measurable space***, and the members of $\mathcal{A}$ are called the *measurable sets* in X.

Notation. If the σ-algebra $\mathcal{A}$ is clear, we simply omit it and denote a measurable space as X.

Suppose $\mathcal{A}$ is an algebra of subsets of X. Then we immediately deduce the following properties.

- $X = E \cup E^c \in \mathcal{A}$ and $\emptyset = X^c \in \mathcal{A}$.
- If $E_1, \dots, E_n \in \mathcal{A}$, then by de Morgan's laws,

$$\bigcap_{i=1}^n E_i = \left(\bigcup_{i=1}^n {E_i}^c \right)^c \in \mathcal{A}$$

  so $\mathcal{A}$ is closed under finite intersections. If $\mathcal{A}$ is a σ-algebra, then $\mathcal{A}$ is closed under countable intersections.

- If $A, B \in \mathcal{A}$, since $A \setminus B = B^c \cap A$, then $A \setminus B \in \mathcal{A}$.

Example 25.4.

- If X is any set, $\mathcal{P}(X)$ is a σ-algebra.
- If X is any set, $\{\emptyset, X\}$ is a σ-algebra.
- If X is any set, and $E \subset X$, $\{\emptyset, E, E^c, X\}$ is a σ-algebra.
- If X is uncountable, then

$$\mathcal{A} = \{E \subset X \mid E \text{ or } E^c \text{ is countable}\}$$

  is a σ-algebra, called the *σ-algebra of countable or co-countable sets*.

Definition 25.5. Given $\mathcal{E} \subset \mathcal{P}(X)$, the σ-algebra *generated* by $\mathcal{E}$, denoted by $\mathcal{M}(\mathcal{E})$, is the smallest σ-algebra containing $\mathcal{E}$.

That is, $\mathcal{M}(\mathcal{E})$ is the intersection of all σ-algebras containing $\mathcal{E}$:

$$\mathcal{M}(\mathcal{E}) = \bigcap_{\substack{\mathcal{A} \supset \mathcal{E} \\ \mathcal{A} \text{ is } \sigma\text{-algebra}}} \mathcal{A}.$$

$\mathcal{M}(\mathcal{E})$ exists since $\mathcal{P}(X)$ is a σ-algebra (so the intersection is non-empty), and the intersection of any family of σ-algebras on X is itself a σ-algebra.

The following observation is often useful:

Lemma 25.6. *If $\mathcal{E} \subset \mathcal{M}(\mathcal{F})$, then $\mathcal{M}(\mathcal{E}) \subset \mathcal{M}(\mathcal{F})$.*

Proof. $\mathcal{M}(\mathcal{F})$ is a σ-algebra containing $\mathcal{E}$; it therefore contains $\mathcal{M}(\mathcal{E})$. □

We have come to an important example of a σ-algebra.

Definition 25.7 (Borel σ-algebra)**.** If X is any metric space (or topological space), the ***Borel σ-algebra*** on X, denoted by $\mathcal{B}(X)$, is the σ-algebra generated by the family of open sets in X. Its members are called ***Borel sets***.

B_X thus includes open sets, closed sets, countable intersections of open sets, countable unions of closed sets, and so forth.

We introduce some standard terminology for the levels in this hierarchy.

- A countable intersection of open sets is called a G_δ set.
- A countable union of closed sets is called a F_σ set.
- A countable union of G_δ sets is called a $G_{\delta\sigma}$ set; a countable intersection of F_σ sets is called a $F_{\sigma\delta}$ set; and so forth.

(δ and σ stand for the German "Durchschnitt" and "Summe", that is, intersection and union.)

The Borel σ-algebra on $\mathbb{R}$ will play a fundamental role in what follows. It can be generated in a number of different ways:

Proposition 25.8. *$\mathcal{B}(\mathbb{R})$ is generated by each of the following:*

(i) the open intervals: $\mathcal{E}_1 = \{(a,b) \mid a < b\}$;

(ii) the closed intervals: $\mathcal{E}_2 = \{[a,b] \mid a < b\}$;

(iii) the half-open intervals: $\mathcal{E}_3 = \{(a,b] \mid a < b\}$ or $\mathcal{E}_4 = \{[a,b) \mid a < b\}$;

(iv) the open rays: $\mathcal{E}_5 = \{(a,\infty) \mid a \in \mathbb{R}\}$ or $\mathcal{E}_6 = \{(-\infty,a) \mid a \in \mathbb{R}\}$;

(v) the closed rays: $\mathcal{E}_7 = \{[a,\infty) \mid a \in \mathbb{R}\}$ or $\mathcal{E}_8 = \{(-\infty,a] \mid a \in \mathbb{R}\}$.

Proof. □

Suppose $(X, \mathcal{M}, \mu)$ is a measure space. We say $E \in \mathcal{M}$ is a *null set* if $\mu(E) = 0$.

If $\mu(E) = 0$ and $F \subset E$, then $\mu(F) = 0$ by monotonicity provided that $F \in \mathcal{M}$, but in general it need not be true that $F \in \mathcal{M}$. A measure whose domain includes all subsets of null sets is called ***complete***. The next result states that completeness can be achieved by enlarging the domain of μ.

Proposition 25.9. *Suppose* $(X, \mathcal{M}, \mu)$ *is a measure space. Let* $\mathcal{N} = \{N \in \mathcal{M} \mid \mu(N) = 0\}$ *and*

$$\overline{\mathcal{M}} = \{E \cup F \mid E \in \mathcal{M}, F \subset N \textit{ for some } N \in \mathcal{N}\}.$$

Then $\overline{\mathcal{M}}$ *is a* σ*-algebra, and there exist a unique extension* $\overline{\mu}$ *of* μ *to a complete measure on* $\overline{\mathcal{M}}$*.*

$\overline{\mu}$ is called the *completion* of μ; $\overline{\mathcal{M}}$ is called the *completion* of $\mathcal{M}$ with respect to μ.

Proof. We check that $\overline{\mathcal{M}}$ is a σ-algebra.

(i) Since $\mathcal{M}$ and $\mathcal{N}$ are closed under countable unions, so is $\overline{\mathcal{M}}$.

(ii) Let $E \cup F \in \overline{\mathcal{M}}$, where $E \in \mathcal{M}$ and $F \subset N \in \mathcal{N}$.

□

25.3 Measures

Definition 25.10 (Measure)**.** Let X be a set equipped with a σ-algebra $\mathcal{M}$. A ***measure*** on $\mathcal{M}$ is a function $\mu\colon \mathcal{M} \to [0, \infty]$ such that

(i) $\mu(\emptyset) = 0$;

(ii) if $E_1, E_2, \ldots$ is a sequence of disjoint sets in $\mathcal{M}$, then $\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i)$.

If μ is a measure on $(X, \mathcal{M})$, we call $(X, \mathcal{M}, \mu)$ a ***measure space***.

Property (ii) is called *countable additivity*. It implies *finite additivity*:

(ii') if $E_1, \ldots, E_n$ are disjoint sets in $\mathcal{M}$, then $\mu(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} \mu(E_i)$,

because one can take $E_i = \emptyset$ for $i > n$. A function μ that satisfies (i) and (ii') but not necessarily (ii) is called a *finitely additive measure*.

We introduce some standard terminology concerning the "size" of μ.

- If $\mu(X) < \infty$ (which implies that $\mu(E) < \infty$ for all $E \in \mathcal{M}$), μ is called *finite*.
- If $X = \bigcup_{i=1}^{\infty} E_i$ where $E_i \in \mathcal{M}$ and $\mu(E_i) < \infty$ for all i, μ is called *σ-finite*.
- More generally, if $E = \bigcup_{i=1}^{\infty} E_i$ where $E_i \in \mathcal{M}$ and $\mu(E_i) < \infty$ for all i, the set E is said to be *σ-finite* for μ.
- If for each $E \in \mathcal{M}$ with $\mu(E) = \infty$ there exists $F \in \mathcal{M}$ with $F \subset E$ and $0 < \mu(F) < \infty$, μ is called *semifinite*.

Example 25.11.

- For any $E \subset X$, where X is any set, define $\mu(E) = \infty$ if E is an infinite set, and let $\mu(E)$ be the number of points in E if E is finite. This μ is called the *counting measure* on X.
- Fix $x_0 \in X$. For any $E \subset X$, let
$$\mu(E) = \begin{cases} 1 & (x_0 \in E) \\ 0 & (x_0 \notin E) \end{cases}$$
This μ is called the *unit mass* concentrated at x_0.
- A *probability measure* on Ω is a measure $\mathbb{P}$ such that $\mathbb{P}(\Omega) = 1$.
- On $\mathcal{B}(\mathbb{R})$, define a measure $\mu\left((a, b)\right) = b - a$ for any $a, b \in \mathbb{R}$, $a < b$. This is called the *Lebesgue measure* on $\mathbb{R}$.

Lemma 25.12 (Basic properties of measures)**.** *Suppose $(X, \mathcal{M}, \mu)$ is a measure space.*

(i) If $E, F \in \mathcal{M}$ and $E \subset F$, then $\mu(E) \le \mu(F)$. *(monotonicity)*

(ii) If $E_1, E_2, \dots \in \mathcal{M}$, then $\mu\left(\bigcup_{n=1}^{\infty} E_i\right) \le \sum_{n=1}^{\infty} \mu(E_n)$. *(subadditivity)*

(iii) If $E_1, E_2, \dots \in \mathcal{M}$ and $E_1 \subset E_2 \subset \cdots$, then *(continuity from below)*

$$\mu\left(\bigcup_{n=1}^{\infty} E_n\right) = \lim_{n\to\infty} \mu(E_n).$$

(iv) If $E_1, E_2, \dots \in \mathcal{M}$, $E_1 \supset E_2 \supset \cdots$, and $\mu(E_1) < \infty$, then *(continuity from above)*

$$\mu\left(\bigcap_{n=1}^{\infty} E_n\right) = \lim_{n\to\infty} \mu(E_n).$$

Proof.

(i) If $E \subset F$, note that $F = E \cup (F \setminus E)$. Then

$$\mu(F) = \mu(E) + \mu(F \setminus E) \ge \mu(E).$$

(ii) Let $F_1 = E_1$ and $F_n = E_n \setminus \left(\bigcup_{i=1}^{n-1} E_i\right)$ for $n > 1$. Then the F_n's are disjoint and $\bigcup_{i=1}^{n} F_i = \bigcup_{i=1}^{n} E_i$ for all n. Hence by (i),

$$\begin{aligned}
\mu\left(\bigcup_{n=1}^{\infty} E_n\right) &= \mu\left(\bigcup_{n=1}^{\infty} F_n\right) \\
&= \sum_{n=1}^{\infty} \mu(F_n) && \text{[by countable additivity]} \\
&\le \sum_{n=1}^{\infty} \mu(E_n) && \text{[by monotonicity, since } F_n \subset E_n\text{]}
\end{aligned}$$

(iii) Suppose $E_1 \subset E_2 \subset \cdots$. Then we have

$$\begin{aligned}
\mu\left(\bigcup_{n=1}^{\infty} E_n\right) &= \mu\left(\bigcup_{n=1}^{\infty} E_n \setminus E_{n-1}\right) && \text{[setting } E_0 = \emptyset\text{]} \\
&= \sum_{n=1}^{\infty} \mu(E_n \setminus E_{n-1}) && \text{[by countable additivity]} \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} \mu(E_i \setminus E_{i-1}) \\
&= \lim_{n\to\infty} \mu(E_n) && \text{[by finite additivity]}
\end{aligned}$$

(iv) Let $F_n = E_1 \setminus E_n$; then $F_1 \subset F_2 \subset \cdots$, $\mu(E_1) = \mu(F_n) + \mu(E_n)$, and $\bigcup_{n=1}^{\infty} F_n = E_1 \setminus$

$(\bigcap_{n=1}^{\infty} E_n)$. By (iii),

$$\begin{aligned}
\mu(E_1) &= \mu\left(\bigcap_{n=1}^{\infty} E_n\right) + \lim_{n\to\infty} \mu(F_n) \\
&= \mu\left(\bigcap_{n=1}^{\infty} E_n\right) + \lim_{n\to\infty} (\mu(E_1) - \mu(E_n)) \\
&= \mu\left(\bigcap_{n=1}^{\infty} E_n\right) + \mu(E_1) - \lim_{n\to\infty} \mu(E_n)
\end{aligned}$$

Since $\mu(E_1) < \infty$, we may subtract it from both sides to yield the desired result.

□

Lemma 25.13 (Inclusion–exclusion formula)**.** *Suppose $(X, \mathcal{M}, \mu)$ is a measure space. If $E, F \in \mathcal{M}$, then*

$$\mu(E) + \mu(F) = \mu(E \cup F) + \mu(E \cap F).$$

Proof. We have

$$\begin{aligned}
\mu(E) + \mu(F) &= \mu\left((E \setminus F) \cup (E \cap F)\right) + \mu\left((F \setminus E) \cup (E \cap F)\right) \\
&= \mu(E \setminus F) + \mu(E \cap F) + \mu(F \setminus E) + \mu(E \cap F) \\
&= \mu(E \cup F) + \mu(E \cap F).
\end{aligned}$$

□

25.4 Outer Measures

Recall the procedure used in calculus to define the area of a bounded region $E \subset \mathbb{R}^2$: one draws a grid of rectangles in the plane and approximates the area of E from above and below by the sum of areas of rectangles. The limits of these approximations as the grid is taken finer and finer give the "inner area" and "outer area" of E; if they are equal, their common value is the "area" of E.

Definition 25.14 (Outer measure)**.** An ***outer measure*** on a non-empty set X is a function $\mu^* : \mathcal{P}(X) \to [0, \infty]$ such that

(i) $\mu^*(\emptyset) = 0$;

(ii) if $A \subset B$, then $\mu^*(A) \le \mu^*(B)$; (monotonicity)

(iii) $\mu^* \left(\bigcup_{i=1}^{\infty} A_i\right) \le \sum_{i=1}^{\infty} \mu^*(A_i)$. (subadditivity)

The most common way to obtain outer measures is to start with a family $\mathcal{E}$ of "elementary sets" on which a notion of measure is defined and then to approximate arbitrary sets "from the outside" by countable unions of members of $\mathcal{E}$.

Proposition 25.15. *Let $\mathcal{E} \subset \mathcal{P}(X)$ and $\rho \colon \mathcal{E} \to [0, \infty]$ be such that $\emptyset \in \mathcal{E}$, $X \in \mathcal{E}$, and $\rho(\emptyset) = 0$. For any $A \subset X$, define*

$$\mu^*(A) = \inf \left\{ \sum_{n=1}^{\infty} \mu(E_n) \;\middle|\; E_n \in \mathcal{E},\ A \subset \bigcup_{n=1}^{\infty} E_n \right\}.$$

Then μ^* is an outer measure.

Theorem 25.16 (Carathéodory's theorem)**.**

Our first applications of Carathéodory's theorem will be in the context of extending measures from algebras to σ-algebras.

Definition 25.17 (Premeasure)**.** Suppose $\mathcal{A} \subset \mathcal{P}(X)$ is an algebra. We say $\mu_0 \colon \mathcal{A} \to [0, \infty]$ is a ***premeasure*** if

(i) $\mu_0(\emptyset) = 0$;

(ii)

Proposition 25.18. *If μ_0 is a premeasure on $\mathcal{A}$, and μ^* is defined by (1.12), then*

(i) $\mu^*|_{\mathcal{A}} = \mu_0$;

(ii) every set in $\mathcal{A}$ is μ^* measurable.

Theorem 25.19.

25.5 Borel Measures on the Real Line

We are now in a position to construct a definitive theory for measuring subsets of $\mathbb{R}$ based on the idea that the measure of an interval is its length. We begin with a more general construction that yields a large family of measures on $\mathbb{R}$ whose domain is the Borel σ-algebra $\mathcal{B}(\mathbb{R})$; such measures are called *Borel measures* on $\mathbb{R}$.

Chapter 26

Integration

26.1 Measurable Functions

Definition 26.1 (Measurable function)**.** Let X be a measurable space, Y be a topological space. We say $f\colon X \to Y$ is ***measurable*** if $f^{-1}(U)$ is a measurable set in X for every open set U in Y.

That is, f is measurable if the pre-image of every open set is measurable.

The next result states that continuous functions of measurable functions are measurable.

Lemma 26.2. *Let X be a measurable space, Y and Z be topological spaces. If $f\colon X \to Y$ is measurable and $g\colon Y \to Z$ is continuous, then $h = g \circ f$ is measurable.*

Proof. Let U be open in Z. By continuity of g, we have that $g^{-1}(U)$ is open in Y. Note that

$$h^{-1}(U) = f^{-1}\left(g^{-1}(U)\right).$$

If f is measurable, it follows that $h^{-1}(U)$ is measurable, so h is continuous. $\square$

The next result states that the tuple of measurable functions is measurable.

Lemma 26.3. *Suppse X is a measurable space, Y is a topological space. Let $u, v\colon X \to \mathbb{R}$ and $\Phi\colon \mathbb{R}^2 \to Y$. Define*

$$h(x) = \Phi\left(u(x), v(x)\right).$$

If u and v are measurable and Φ is continuous, then $h\colon X \to Y$ is measurable.

Proof. Define $f\colon X \to \mathbb{R}^2$ by $f(x) = (u(x), v(x))$. Then $h = \Phi \circ f$. By 26.2, it suffices to show that f is measurable.

Idea. We want to show that any open set in $\mathbb{R}^2$ has a measurable pre-image. Hint: every open set in $\mathbb{R}^2$ is a countable union of sets of the form $I_1 \times I_2$, where I_1 and I_2 are open intervals in $\mathbb{R}$.

If $R \subset \mathbb{R}^2$ is any open rectangle with sides parallel to the axes, then $R = I_1 \times I_2$ for two open intervals

$I_1, I_2 \subset \mathbb{R}$. Then

$$\begin{aligned} f^{-1}(R) &= f^{-1}(I_1 \times I_2) \\ &= (u,v)^{-1}(I_1 \times I_2) \\ &= u^{-1}(I_1) \cap v^{-1}(I_2) \end{aligned}$$

Let $x \in f^{-1}(R)$ so that $f(x) \in R$. Then $u(x) \in I_1$ and $v(x) \in I_2$. Since u is measurable, $u^{-1}(I_1) \in \mathcal{M}$; since v is measurable, $v^{-1}(I_2) \in \mathcal{M}$. Since $\mathcal{M}$ is a σ-algebra, it is closed under intersections, so $u^{-1}(I_1) \cap v^{-1}(I_2) \in \mathcal{M}$. Thus $f^{-1}(R)$ is measurable, for any rectangle R.

Every open set U in $\mathbb{R}^2$ is a countable union of such rectangles R_i. Thus

$$f^{-1}(U) = f^{-1}\left(\bigcup_{i=1}^{\infty} R_i\right) = \bigcup_{i=1}^{\infty} f^{-1}(R_i).$$

Since each term in the union is in $\mathcal{M}$, and since $\mathcal{M}$ is closed under countable unions, we have that $f^{-1}(U) \in \mathcal{M}$. Hence f is measurable. □

We now prove some consequences of the above results, concerning the measurability of functions that we shall frequently encounter.

Proposition 26.4. *Let X be a measurable space.*

(i) Let $f\colon X \to \mathbb{C}$ with $f = u + iv$, where u and v are real measurable functions on X. Then f is complex measurable.

(ii) If $f = u + iv$ is a complex measurable function on X, then u, v, and $|f|$ are real measurable functions on X.

(iii) If f and g are complex measurable functions on X, then so are $f + g$ and fg.

(iv) If E is a measurable set in X and the characteristic function *of E is defined as*

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

then χ_E is a measurable function.

(v) If f is a complex measurable function on X, there is a complex measurable function α on X such that $|\alpha| = 1$ and $f = \alpha|f|$.

Proof.

(i) This follows from 26.3, by taking $\Phi(z) = z$.

(ii) This follows from 26.2, by taking $g(z) = \mathrm{Re}(z)$, $g(z) = \mathrm{Im}(z)$, and $g(z) = |z|$ respectively.

(iii) For real f and g, this follows from 26.3, by taking $\Phi(s,t) = s + t$ and $\Phi(s,t) = st$ respectively.

The complex case then follows from (i) and (ii).

(iv) Let U be open in $\mathbb{R}$. Then

$$\chi_E^{-1}(U) = \begin{cases} X & (0, 1 \in U) \\ E & (1 \in U,\ 0 \notin U) \\ E^c & (1 \notin U,\ 0 \in U) \\ \emptyset & (\text{otherwise}) \end{cases}$$

all those sets are measurable, since E is measurable.

(v) Let $E = \{x \mid f(x) = 0\}$. Let $Y = \mathbb{C} \setminus \{0\}$, and define $\phi(z) = \dfrac{z}{|z|}$ for $z \in Y$. Let

$$\alpha(x) = \phi\left(f(x) + \chi_E(x)\right) \quad (x \in X).$$

If $x \in E$, $\alpha(x) = 1$; if $x \notin E$, $\alpha(x) = \dfrac{f(x)}{|f(x)|}$. This shows that $|\alpha| = 1$.

Since ϕ is continuous on Y, and E is measurable (since $|f|$ is real measurable and $E^c = |f|^{-1}\left((0, \infty)\right)$), the measurability of α follows from (iii), (iv), and 26.2.

□

Definition 26.5. Suppose X is a measurable space with Borel σ-algebra $\mathcal{B}$, and Y is a topological space. If $f\colon X \to Y$ is such that $f^{-1}(U) \in \mathcal{B}$ for all $U \in \mathcal{T}$, we say f is ***Borel measurable***.

In particular, continuous functions are Borel measurable.

Proposition 26.6. *Suppose $\mathcal{M}$ is a σ-algebra in X, and Y is a topological space. Let $f\colon X \to Y$.*

(i) If Ω is the collection of all sets $E \subset Y$ such that $f^{-1}(E) \in \mathcal{M}$, then Ω is a σ-algebra in Y.

(ii) If f is measurable and E is a Borel set in Y, then $f^{-1}(E) \subset \mathcal{M}$.

(iii) If $Y = [-\infty, \infty]$ and $f^{-1}\left((\alpha, \infty]\right) \in \mathcal{M}$ for every real α, then f is measurable.

(iv) If f is measurable, if Z is a topological space, if $g\colon Y \to Z$ is a Borel mapping, and if $h = g \circ f$, then $h\colon X \to Z$ is measurable.

Suppose (f_n) is a sequence of extended-real functions on a set X. Then define

$$\left(\sup_n f_n\right)(x) := \sup_n \left(f_n(x)\right),$$
$$\left(\limsup_{n\to\infty} f_n\right)(x) := \limsup_{n\to\infty} \left(f_n(x)\right).$$

If

$$f(x) = \lim_{n\to\infty} f_n(x) \quad (x \in X),$$

then we call f the *pointwise limit* of (f_n).

Proposition 26.7. *If $f_n : X \to [-\infty, \infty]$ is measurable, and*

$$g = \sup_{n \geq 1} f_n, \qquad h = \limsup_{n \to \infty} f_n,$$

then g and h are measurable.

Proof. Note that

$$g^{-1}((\alpha, \infty]) = \bigcup_{n=1}^{\infty} f_n^{-1}((\alpha, \infty]).$$

Hence by Theorem 1.12(c), g is measurable.

The same result holds with inf in place of sup, and since

$$h = \inf_{k \geq 1} \left(\sup_{n \geq k} f_n \right),$$

it follows that h is measurable. □

Corollary 26.8.

(i) *The limit of every pointwise convergent sequence of complex measurable functions is measurable.*

(ii) *If f and g are measurable (with range in $[-\infty, \infty]$), then so are $\max\{f, g\}$ and $\min\{f, g\}$. In particular, this is true of the functions*

$$f^+ = \max\{f, 0\} \quad \text{and} \quad f^- = -\min\{f, 0\}.$$

Proof. Let (f_n) be pointwise convergent, and

$$f(x) = \lim_{n \to \infty} f_n(x) \quad (x \in X).$$

□

We call f^+ and f^- the *positive and negative parts* of f.

We have $|f| = f^+ + f^-$ and $f = f^+ - f^-$, a standard representation of f as a difference of two non-negative functions, with a certain minimality property:

Lemma 26.9. *If $f = g - h$, $g \geq 0$ and $h \geq 0$, then $f^+ \leq g$ and $f^- \leq h$.*

Proof. $f \leq g$ and $0 \leq g$ clearly implies $\max\{f, 0\} \leq g$. □

26.2 Integration

26.2.1 Simple Functions

Definition 26.10. Let X be a measurable space. We say $s : X \to \mathbb{C}$ is a ***simple function*** if its range consists of only finitely many points.

Among these are the *non-negative simple functions*, whose range is a finite subset of $[0, \infty)$. Note that we explicitly exclude ∞ from the values of a simple function.

If $\alpha_1, \ldots, \alpha_n$ are the distinct values of a simple function s, and if we set $A_i = \{x \mid s(x) = \alpha_i\}$, then clearly

$$s = \sum_{i=1}^{n} \alpha_i \chi_{A_i}$$

where χ_{A_i} is the characteristic function of A_i.

It is also clear that s is measurable if and only if each of the sets A_i is measurable.

Proposition 26.11. *Let $f\colon X \to [0, \infty]$ be measurable. There exist simple measurable functions s_n on X such that*

(i) $0 \leq s_1 \leq s_2 \leq \cdots \leq f$*;*

(ii) $\lim_{n\to\infty} s_n(x) = f(x)$*, for every* $x \in X$*.*

26.2.2 Integration of Positive Functions

Let $\mathcal{M}$ be a σ-algebra in a set X, and let $\mu : \mathcal{M} \to [0, \infty]$ be a measure.

Definition 26.12 (Lebesgue integral)**.** If $s : X \to [0, \infty)$ is a measurable simple function, of the form

$$s = \sum_{i=1}^{n} \alpha_i \chi_{A_i},$$

where $\alpha_1, \ldots, \alpha_n$ are the distinct values of s, and if $E \in \mathcal{M}$, we define

$$\int_E s \, \mathrm{d}\mu := \sum_{i=1}^{n} \alpha_i \mu(A_i \cap E).$$

If $f\colon X \to [0, \infty]$ is measurable, and $E \in \mathcal{M}$, we define the ***Lebesgue integral*** of f over E as

$$\int_E f \, \mathrm{d}\mu := \sup \int_E s \, \mathrm{d}\mu \,. \tag{26.1}$$

Lemma 26.13 (Properties of Lebesgue integral)**.**

We now come to the interesting part of the theory. One of its most remarkable features is the ease with which it handles limit operations.

Theorem 26.14 (Lebesgue's monotone convergence theorem). *Let (f_n) be a sequence of measurable functions on X, and suppose that*

(i) $0 \le f_1(x) \le f_2(x) \le \cdots \le \infty$ *for every* $x \in X$,

(ii) $f_n(x) \to f(x)$ *pointwise for every* $x \in X$.

Then f is measurable, and

$$\lim_{n\to\infty} \int_X f_n \, \mathrm{d}\mu = \int_X f \, \mathrm{d}\mu \,. \tag{26.2}$$

Proposition 26.15. *If $f_n : X \to [0, \infty]$ are measurable, and*

$$f(x) = \sum_{n=1}^{\infty} f_n(x) \quad (x \in X),$$

then

$$\int_X f \, \mathrm{d}\mu = \sum_{n=1}^{\infty} \int_X f_n \, \mathrm{d}\mu \,. \tag{26.3}$$

If we let μ be the counting measure on a countable set, Theorem 1.27 is a statement about double series of nonnegative real numbers (which can of course be proved by more elementary means):

Corollary 26.16. *If $a_{ij} \ge 0$ for $i, j = 1, 2, \ldots$, then*

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} a_{ij} .$$

Theorem 26.17 (Fatau's lemma). *Let $f_n : X \to [0, \infty]$ be a sequence of measurable functions. Then*

$$\int_X \left(\liminf_{n\to\infty} f_n \right) \mathrm{d}\mu \le \liminf_{n\to\infty} \int_X f_n \, \mathrm{d}\mu \,. \tag{26.4}$$

Proposition 26.18 (Change of variables). *Suppose $f \colon X \to [0, \infty]$ is measurable, and*

$$\phi(E) = \int_E f \, \mathrm{d}\mu \quad (E \in \mathcal{M}.)$$

Then ϕ is a measure on $\mathcal{M}$, and

$$\int_X g \, \mathrm{d}\phi = \int_X g f \, \mathrm{d}\mu \tag{26.5}$$

for every measurable $g \colon X \to [0, \infty]$.

26.2.3 Integration of Complex Functions

As before, let μ be a measure on an arbitrary measurable space X.

Definition 26.19 (Class of Lebesgue integrable functions)**.** We define $L^1(\mu)$ to be the collection of complex measurable functions on X, for which

$$\int_X |f| \,\mathrm{d}\mu < \infty.$$

[Note that the measurability of f implies that of $|f|$, as we saw in Proposition 1.9(b); hence the above integral is defined.]
The members of $L^1(\mu)$ are called ***Lebesgue integrable functions*** (with respect to μ) or *summable functions*.

If $f = u + iv$, where u and v are real measurable functions on X, and if $f \in L^1(\mu)$, define

$$\int_E f \,\mathrm{d}\mu := \left(\int_E u^+ \,\mathrm{d}\mu - \int_E u^- \,\mathrm{d}\mu\right) + i\left(\int_E v^+ \,\mathrm{d}\mu - \int_E v^- \,\mathrm{d}\mu\right)$$

for every measurable set E.

Occasionally it is desirable to define the integral of a measurable function f with range in $[-\infty, \infty]$ to be

$$\int_E f \,\mathrm{d}\mu := \int_E f^+ \,\mathrm{d}\mu - \int_E f^- \,\mathrm{d}\mu$$

provided that at least one of the integrals on the RHS is finite. The LHS is then a number in $[-\infty, \infty]$.

Lemma 26.20 (Linearity)**.** *Suppose $f, g \in L^1(\mu)$, and $\alpha, \beta \in \mathbb{C}$. Then $\alpha f + \beta g \in L^1(\mu)$, and*

$$\int_X (\alpha f + \beta g) \,\mathrm{d}\mu = \alpha \int_X f \,\mathrm{d}\mu + \beta \int_X g \,\mathrm{d}\mu \,.$$

Lemma 26.21 (Triangle inequality)**.** *If $f \in L^1(\mu)$, then*

$$\left|\int_X f \,\mathrm{d}\mu\right| \leq \int_X |f| \,\mathrm{d}\mu \,.$$

We conclude this section with another important convergence theorem. This is like the monotone convergence theorem, but we are going to remove the increasing and non-negative measurable condition, and add in something else.

Theorem 26.22 (Lebesgue's dominated convergence theorem)**.** *Suppose (f_n) is a sequence of complex measurable functions on X such that*

$$f(x) = \lim_{n\to\infty} f_n(x) \quad (x \in X).$$

If there exists $g \in L^1(\mu)$ such that

$$|f_n(x)| \leq g(x) \quad (n = 1, 2, \ldots, \; x \in X),$$

then $f \in L^1(\mu)$, *and*

$$\lim_{n\to\infty} \int_X f_n \, \mathrm{d}\mu = \int_X f \, \mathrm{d}\mu \, . \tag{26.6}$$

Bibliography

[Apo57] T. M. Apostol. *Mathematical Analysis*. Addison-Wesley, 1957.

[Axl24] S. Axler. *Linear Algebra Done Right*. Springer, 2024.

[DF04] D. S. Dummit and R. M. Foote. *Abstract Algebra*. John Wiley & Sons, 2004.

[Fol99] G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, Inc., 1999.

[HS65] E. Hewitt and K. Stromberg. *Real and Abstract Analysis*. Springer-Verlag, 1965.

[Lan05] S. Lang. *Undergraduate Algebra*. Springer, 2005.

[Mun18] J. R. Munkres. *Topology*. Pearson Education Limited, 2018.

[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.

[Spi65] M. Spivak. *Calculus on Manifolds: A Modern Approach to Classical Theorems of Advanced Calculus*. Harper Collins Publishers, 1965.

[SS03] E. M. Stein and R. Shakarchi. *Complex Analysis*. Princeton University Press, 2003.

Index