Fundamental Tools

Jonathan Noel and Dominykas Norgilas, 2018

Based on original notes by:

Ysette Weiß-Pidstrigatch, 1999

Igor Kozin, 2000

David Wood, 2001

Peter Windridge, 2010

Matija Vidmar, 2014

Alex Tse, 2015

David Moxey, 2016

Nick Simm, 2017

September 12, 2018

Contents

Foreword

Fundamental Tools week is a preparatory course to refresh your knowledge of mathematics, and introduce you to the style of questions you are likely to encounter in courses containing mathematics at Warwick.

It is likely that these notes may seem more formal than other courses you have taken before. Do not let this concern you too much, as most of the course will involve calculation using the formal results laid out here, and indeed for the most part, we shall not prove a lot of theorems here. However, you are expected to know and understand the proofs that are presented here.

You should be aware that the material covered here is considered to be basic mathematical content which, hopefully, you will have encountered before. Importantly, if you have not covered a topic before, then these notes **should not** serve as the means to learn about it – the material presented here is a brief summary of these topics and is not meant to be exhaustive. Ideally, you should consult other textbooks which will provide far more detail.

In the first half of the course we will study basic linear algebra, followed by differentation of multi-dimensional functions and recipies for solving both ordinary and partial differential equations. In the second half we study basic probability theory.

Structure and Assessment

The course is held over a week, with four lectures on the mathematical content and another four on the probability component. As such, we will not be able to cover all of the material, so it is important you have read through these notes beforehand. After each lecture you will have the chance to attend a two-hour seminar and work through various questions on the main topics of the course.

There is an examination at the end of the week which broadly assumes the knowledge you will find here, along with the example sheets completed in seminars. As a general guide, if you understand most of the material here and have completed all of the questions from the assignment sheets then you will not have a problem passing the exam.

Partly, this course will also introduce you to the way that examinations are held at Warwick, and as such there are certain rules and procedures that you **must** abide by.

These rules must be taken very seriously. Failure to comply with them could result in you having to resit the exam.

You **must**:

- bring your university card and writing implements – a pen will suffice;
- enter the room silently.

You **must not**:

- bring a bag – no bags are allowed in exam halls;
- use a calculator.
- talk or communicate in any part of the examiniation **until scripts are collected and you have left the examination hall**;
- bring, or use at any point, a mobile (or cellular) phone.

You **can**:

- leave the examiniation early, provided it is after 30 minutes after the start of the examination and before 15 minutes before the end;
- bring bottled water.

Course webpage

There is a webpage for this course, which you can find at

`http://go.warwick.ac.uk/ma901`

It will contain a copy of these notes, as well as assignment sheets as they are set through the week. No solution sheets will be provided for the assignment sheets.

Finally, please note that these notes are new for the course in 2010/2011, and differ substantially from previous years. As such, they almost certainly contain errors and typos, although we believe that the mathematical content itself is correct. If you do find errors, then please contact us so that they may be corrected. You can always find an up-to-date copy of the notes on the course webpage.

Chapter 1

Linear Algebra

1.1 Vector Spaces

Fix a point O in space. The vectors whose starting points are located at O can be multiplied by numbers and added by the parallelogram rule.

Now forget about the three-dimensionality of physical space, replace real numbers by an arbitrary field and postulate the simplest properties of addition and multiplication as axioms. This gives the definition of a vector space.

Definition 1.1.1 (Vector space). A set V is said to be a **vector space** over a field $\mathbb{K}$ if it is equipped with a binary operation $V \times V \to V$ usually denoted as addition $(\mathbf{v}_1, \mathbf{v}_2) \mapsto \mathbf{v}_1 + \mathbf{v}_2$, and a binary operation $\mathbb{K} \times V \to V$ usually denoted as multiplication $(k, \mathbf{v}) \mapsto k \cdot \mathbf{v}$, which satisfy the following axioms:

1. (a) **Commutativity:** $\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{v}_2 + \mathbf{v}_1$;
   (b) **Associativity:** $(\mathbf{v}_1 + \mathbf{v}_2) + \mathbf{v}_3 = \mathbf{v}_1 + (\mathbf{v}_2 + \mathbf{v}_3)$;
   (c) **Zero element:** $\exists\, \mathbf{0}_V \in V$ such that $\mathbf{v} + \mathbf{0}_V = \mathbf{v}\ \forall\, \mathbf{v} \in V$;
   (d) **Additive inverse:** $\forall \mathbf{v} \in V\, \exists\, (-\mathbf{v}) \in V$ such that $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}_V$.
2. Multiplication of vectors by scalars (elements of the field $\mathbb{K}$) satisfies;
   (a) $1_K \cdot \mathbf{v} = \mathbf{v}$;
   (b) $\alpha \cdot (\beta \cdot \mathbf{v}) = (\alpha\beta) \cdot \mathbf{v}$ where $\alpha, \beta \in \mathbb{K}, \mathbf{v} \in V$.
3. Addition and multiplication satisfy the **distributivity laws**, i.e.
   (a) $\alpha \cdot (\mathbf{v}_1 + \mathbf{v}_2) = \alpha \cdot \mathbf{v}_1 + \alpha \cdot \mathbf{v}_2$ where $\alpha \in \mathbb{K}, \mathbf{v}_1, \mathbf{v}_2 \in V$;
   (b) $(\alpha + \beta) \cdot \mathbf{v} = \alpha \cdot \mathbf{v} + \beta \cdot \mathbf{v}$ where $\alpha, \beta \in \mathbb{K}, \mathbf{v} \in V$.

Remark 1.1.2.

1. Notice that the word *field* is left undefined. Whilst there is a general definition similar to the one above, for the purposes of this course we will only consider the fields $\mathbb{K} = \mathbb{R}$ and $\mathbb{K} = \mathbb{C}$ (real and complex numbers).

   In general, you can think of a field as being a set of scalars on which you can add, subtract, multiply and divide elements and have them remain in the same set.

Also note that fields have their own zero element, $0_{\mathbb{K}}$, which will usually be different from the zero element of the vector field. Since it is often clear which we are considering, we will simply write $\mathbf{0}_V = \mathbf{0}$ and $0_{\mathbb{K}} = 0$ throughout the rest of this chapter.

2. All other 'standard' identities can be derived from these axioms. For example, it is easy (although tedious) to show that $0 \cdot \mathbf{v} = \mathbf{0}$ for all $\mathbf{v} \in V$.

3. It is standard notation to drop the use of the $\cdot$ symbol unless it is unclear in the setting of the proof or definition. Throughout the rest of the chapter, it will be omitted.

Definition 1.1.3 (Cartesian product). Given two sets A and B, we define the **Cartesian product** as the set

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

In addition, we denote

$$A^n = \underbrace{A \times \cdots \times A}_{n \text{ times}} = \{(a_1, \ldots, a_n) \mid a_k \in A\}.$$

Example 1.1.4. Here are some trivial examples of vector spaces.

1. $V = \{\mathbf{0}\}$;
2. $V = \mathbb{K}^n = \{(k_1, \ldots, k_n) \mid k_i \in \mathbb{K}\}$.

Example 1.1.5. Let $\mathcal{P}_n[x]$ denote the set of polynomials in one variable x of degree $\leq n$ with real coefficients. That is,

$$\mathcal{P}_n[x] = \left\{ \sum_{k=0}^{n} \alpha_k x^k \,\middle|\, \alpha_k \in \mathbb{R} \right\}$$

You should check that this does indeed constitute a vector space!

1.2 Basis and Dimension

Now that we have the basic definition of a vector space, it is time to do something useful with them! In the following definitions, V is a general vector space over a field $\mathbb{K}$.

Definition 1.2.1 (Linear independence). The finite subset of vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\} \subset V$ are said to be **linearly independent** if for any $\alpha_1, \ldots \alpha_n \in \mathbb{K}$,

$$\sum_{k=1}^{n} \alpha_k \mathbf{e}_k = \mathbf{0} \Rightarrow \alpha_k = 0 \text{ for } 1 \leq k \leq n. \tag{1.1}$$

Similarly, the subset is said to be **linearly dependent** if it is not linearly independent.

Example 1.2.2. Let us consider the second vector space of example 1.1.4, with $\mathbb{K} = \mathbb{R}$ and $n = 3$. Then consider the subsets:

- { (1,0,0), (0,1,0), (0,0,1) }: Writing out (1.1) in full,

  $$\alpha_1(1, 0, 0) + \alpha_2(0, 1, 0) + \alpha_3(0, 0, 1) = (0, 0, 0)$$

  and we see immediately that $\alpha_1 = \alpha_2 = \alpha_3 = 0$. Hence this set is linearly independent.

- { (1,1,0), (1,0,0), (0,1,0) }: Since it is clear that $\mathbf{e}_1 = \mathbf{e}_2 + \mathbf{e}_3$, we see that this set must be linearly dependent.

Definition 1.2.3 (Span). The finite subset of vectors $\{\mathbf{e}_1, \dots, \mathbf{e}_n\} \subset V$ are said to **span** V if for any $\mathbf{v} \in V$ there exists $\alpha_1, \dots \alpha_n \in \mathbb{K}$ such that

$$\mathbf{v} = \sum_{k=1}^{n} \alpha_k \mathbf{e}_k$$

Definition 1.2.4 (Basis). The finite subset of vectors $\{\mathbf{e}_1, \dots, \mathbf{e}_n\} \subset V$ are said to form a **basis** of V if they are linearly independent and span V.

Definition 1.2.5 (Standard basis). If $V = \mathbb{K}^n$, then the set

$$\{(1_K, 0_K, \dots, 0_K), (0_K, 1_K, \dots, 0_K), \dots (0_K, 0_K, \dots 1_K)\}$$

is a basis of V and is called the **standard basis**.

Example 1.2.6. The standard basis defined above can be extended to some other spaces by a suitable mapping from the vector space to $\mathbb{K}^n$. For example, consider the polynomial vector space in example 1.1.5. Define a function $f : \mathcal{P}_n[x] \to \mathbb{R}^{n+1}$ by

$$f(\alpha_0 + \alpha_1 x + \dots + \alpha_n x^n) = (\alpha_0, \dots, \alpha_n).$$

This defines a bijection between $\mathcal{P}_n[x]$ and $\mathbb{R}^{n+1}$ with an inverse function

$$f^{-1}(\alpha_0, \dots, \alpha_n) = \sum_{k=0}^{n} \alpha_k x^k$$

Provided that f is also linear (see the next section), a basis for $\mathcal{P}_n[x]$ can be defined by

$$\{f^{-1}(1, 0, \dots, 0), f^{-1}(0, 1, \dots, 0), \dots, f^{-1}(0, 0, \dots, 1)\} = \{1, x, x^2, \dots x^n\}.$$

In general, a linear bijection between two vector spaces V and W is called an **isomorphism** and V and W are said to be **isomorphic**.

Whilst in this case the basis is somewhat obvious from the definition, this demonstrates a technique which is useful for determining a basis of more complicated spaces.

Remark 1.2.7. Remember that a basis is not, in general, unique. For example, the vector space $V = \mathbb{R}^3$ admits uncountably many sets which form a basis, since any set of the form

$$\{(a, 0, 0), (0, a, 0), (0, 0, a) \mid a \in \mathbb{R}\backslash\{0\}\}$$

is a basis of V.

Definition 1.2.8 (Dimension). If a basis of V contains n elements where $n \in \mathbb{N}$, then the space is said to be **finite dimensional**. We denote the **dimension** of V by $\dim V$ and write $\dim V = n$.

It is not necessarily the case that the basis set contains only finitely many elements. Many interesting and fundamental vector spaces (often containing functions) are so-called *infinite dimensional*. For example, one may consider the space of polynomials of indeterminate order, or alternatively the space of square-integrable functions. Whilst you should have some familiarity with these spaces, here we will only consider finite dimensional spaces.

Finally, we conclude with an interesting (and simple) theorem about isomorphisms.

Theorem 1.2.9. Every n-dimensional vector space V over a field $\mathbb{K}$ is isomorphic to $\mathbb{K}^n$.

Proof 1.2.10. Choose a basis $\{\mathbf{e}_i\}_{i=1}^n \subset V$. Then for all $\mathbf{v} \in V$, there exists a unique choice of co-efficients $\{a_i\}_{i=1}^n \subset \mathbb{K}$ such that $\mathbf{v} = \sum a_i \mathbf{e}_i$. So construct a map $f : V \to \mathbb{K}^n$ such that $\mathbf{v} \mapsto (a_1, a_2, \ldots, a_n) \in \mathbb{K}^n$. This is clearly linear and invertible, so V is isomorphic to $\mathbb{K}^n$. □

One key part of the previous proof does need clarification – can you spot (and prove) it?

1.3 Subspaces

Another important property of vector spaces is that, in some cases, we may obtain smaller sets within them which themselves form vector spaces.

Definition 1.3.1 (Subspace). Let V be a vector space over a field $\mathbb{K}$ and $W \subseteq V$. Then W is a **subspace** of V if it is itself a vector space over $\mathbb{K}$ defined using the same vector space operations as V.

In principle then, all one has to do to prove that a particular subset forms a subspace is to check all of the axioms of a vector space. However this is somewhat of a tedious process, and indeed we can form a small theorem which does a lot of the work for us.

Theorem 1.3.2. Let V be a vector space and $W \subset V$. Then W is a subspace of V if and only if the three conditions below hold:

- $\mathbf{0}_V \in W$;
- If $\mathbf{u}, \mathbf{v} \in W$ then $\mathbf{u} + \mathbf{v} \in W$;
- If $\mathbf{u} \in W$ and $\lambda \in \mathbb{K}$ then $\lambda\mathbf{u} \in W$.

Proof 1.3.3. See example sheet 1. □

Example 1.3.4. We will now consider a few examples of subspaces and show how theorem 1.3.2 can be applied to show various subsets do and do not form subspaces.

(i) Let $\mathbb{K} = \mathbb{R}$ and $V = \mathbb{R}^3$. Consider the subset

$$W_1 = \{(x, y, z) \mid x + 2y + 3z = 0\}.$$

Checking the conditions of theorem 1.3.2:

- $\mathbf{0}_V = (0, 0, 0) \in W_1$;
- Let $\mathbf{u}, \mathbf{v} \in W$. Then we can write $\mathbf{u} = (u_1, 2u_2, 3u_3)$ and $\mathbf{v} = (v_1, 2v_2, 3v_3)$, and hence $\mathbf{u} + \mathbf{v} = (u_1 + v_1, 2(u_2 + v_2), 3(u_3 + v_3)) \in W_1$.
- Let $\mathbf{u} = (u_1, 2u_2, 3u_3) \in W$ and $\lambda \in \mathbb{R}$. Then $\lambda\mathbf{u} = ((\lambda u_1), 2(\lambda u_2), 3(\lambda u_3)) \in W_1$.

Hence W_1 forms a subspace of V.

(ii) Using this theorem sometimes makes it easy to spot when subsets do not form subspaces. Consider a slight variation of the previous example, with

$$W_2 = \{(x, y, z) \mid x + 2y + 3z = 1\}.$$

This is clearly not a subspace of $\mathbb{R}^3$ since the zero element is not in W_2.

(iii) The theorem does not only limit us to 'simple' spaces such as the ones considered above. For example, let $\mathbb{K} = \mathbb{R}$, $V = \mathcal{P}_n[x]$ and consider the subset

$$W_3 = \left\{ \sum_{k=0}^{n} a_k x^k \mid a_k \in \mathbb{R},\ a_0 + a_1 + \cdots + a_n = 0 \right\};$$

that is, the set of polynomials such that the co-efficients sum to zero. Checking the conditions of theorem 1.3.2:

- $\mathbf{0}_V = \sum_{k=0}^{n} 0x^k \in W_3$.
- Let $\mathbf{u} = \sum_{k=0}^{n} a_k x^k$ and $\mathbf{v} = \sum_{k=0}^{n} b_k x^k$ be elements of W_3. Then $\mathbf{u}+\mathbf{v} = \sum_{k=0}^{n}(a_k + b_k)x^k$, and hence the co-efficients of $\mathbf{u}+\mathbf{v}$ satisfy $\sum_{k=0}^{n}(a_k+b_k) = \sum_{k=0}^{n} a_k + \sum_{k=0}^{n} b_k = 0 + 0 = 0$, hence $\mathbf{u} + \mathbf{v} \in W_3$.
- The third condition is clear.

Hence, W_3 is a subspace of V. (Can you find a basis for W_3?)

(iv) As a brief foray into the study of infinite dimensional spaces, consider the vector space of functions

$$V = \{f \mid f : \mathbb{R} \to \mathbb{R}\}.$$

Then the subset

$$W_4 = \{f : \mathbb{R} \to \mathbb{R} \mid f \text{ is continuous}\}$$

is a subspace of V, since

- the zero function is in W_4;
- the sum of two continuous functions is continuous;
- given a scalar $\lambda \in \mathbb{K}$, λf is continuous for any continuous function f.

To conclude this section, we note one of many intuitive properties of subspaces; that the dimension of a subspace must be less than or equal to that of its enclosing space.

Theorem 1.3.5. Let W be a subspace of a finite dimensional vector space V. Then $\dim W \leq \dim V$.

Proof 1.3.6. Let $\{\mathbf{w}_i\}_{i=1}^{n}$ be a basis of W so that $\dim W = n$. Then, by definition, these vectors are linearly independent, and also a subset of V. If they form a basis of V then $\dim V = n$. Otherwise, we can form a basis $\{\mathbf{v}_i\}_{i=1}^{m}$ such that $m > n$ with $\mathbf{v}_i = \mathbf{w}_i$ for $1 \leq i \leq n$, and so $\dim V > \dim W$. □

1.4 Linear maps and matrices

Definition 1.4.1 (Linear map). Let U be a n-dimensional vector space and V be a m-dimensional vector space. $T : U \to V$ is called a **linear map** if it satisfies

$$T(\alpha\mathbf{u} + \beta\mathbf{v}) = \alpha T(\mathbf{u}) + \beta T(\mathbf{v}), \text{ where } \mathbf{u}, \mathbf{v} \in U,\ \alpha, \beta \in \mathbb{K}.$$

The space of all such linear maps is denoted $\mathcal{L}(V_1, V_2)$.

Linear maps are extremely important and form one of the cornerstones of linear algebra. Indeed, the entire field of linear algebra originates from the investigation of systems of linear equations, and developing the techniques required to solve such systems. Additionally, the behaviour of many non-linear phenomena can be examined by considering an appropriate linearisation – particularly in the study of non-linear differential equations. Such equations are prevalent throughout applied mathematics, so it is essential that you have a good understanding of the fundamental properties.

By this stage, you should have also encountered matrices and how they behave under some elementary operations (e.g. addition, multiplication, inverses). You should also hopefully be aware that linear maps and matrices share a very close link and can be considered in some respects the same as each other. However, for completeness we provide the definition below.

Definition 1.4.2 (Matrix representation). Let $T : U \to V$ be a linear map. Given a basis $\{\mathbf{e}_i\}_{i=1}^n$ of U and $\{\mathbf{f}_i\}_{i=1}^m$ of V, there exists a unique set of coefficients $\alpha_{ij} \in \mathbb{K}$ such that

$$T(\mathbf{e}_j) = \sum_{i=1}^{m} \alpha_{ij}\mathbf{f}_i, \quad 1 \leq j \leq n.$$

Then the **matrix representation** of T is given by

$$\begin{bmatrix} \alpha_{11} & \dots & \alpha_{1n} \\ \vdots & \ddots & \vdots \\ \alpha_{m1} & \dots & \alpha_{mn} \end{bmatrix}$$

As a shorthand notation we write $A = (\alpha_{ij})$.

Remark 1.4.3. Remember that the choice of basis for both the domain *and* the range of the map is essential when finding the matrix representation. For example, the matrix representation of the identity map $\mathrm{id} : \mathbb{R}^3 \to \mathbb{R}^3, \mathbf{x} \mapsto \mathbf{x}$ under the standard basis is clearly

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

where the $n \times n$ diagonal matrix I_n is called the **identity matrix**. However, under the basis $\{(1,1,0),(0,1,0),(0,0,1)\}$ in the domain (but not the range), the matrix representation is

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \neq I_3$$

Perhaps the two most important definitions you need to remember about linear maps and matrices – asides from their definition – is that of the image and kernel. These sets describe the subset of the codomain that is covered by a linear map, and what elements of the domain map to the zero element of the codomain respectively.

Definition 1.4.4 (Kernel and image). Let $T : U \to V$ be a linear map. The **kernel** of T, written $\ker T$, is the set of vectors $\mathbf{u} \in U$ such that $T(\mathbf{u}) = \mathbf{0}_V$. The **image** of T, written $\mathrm{Im}\, T$ is the set of vectors $\mathbf{v} \in V$ such that $T(\mathbf{u}) = \mathbf{v}$ for some $\mathbf{u} \in U$. In set notation, these are written as

$$\ker T = \{\mathbf{u} \in U \mid T(\mathbf{u}) = \mathbf{0}_V\}, \quad \mathrm{Im}\, T = \{T(\mathbf{u}) \mid \mathbf{u} \in U\}.$$

Remark 1.4.5. Equivalently given A, a matrix representation of T, we may write

$$\ker A = \{\mathbf{u} \in U \mid A\mathbf{u} = \mathbf{0}_V\}, \quad \operatorname{Im} A = \{A\mathbf{u} \mid \mathbf{u} \in U\}.$$

where $A\mathbf{u}$ denotes matrix-vector multiplication. Consequently $\ker T = \ker A$ and $\operatorname{Null} T = \operatorname{Null} A$.

Theorem 1.4.6. For T as above, $\operatorname{Im} T$ is a subspace of V and $\ker T$ is a subspace of U.

Proof 1.4.7. See example sheet 1. □

Definition 1.4.8 (Rank and nullity). By the above theorem, we may define the **rank** of a linear map $T : U \to V$ (or a matrix representation) as the dimension of its image, and the **nullity** of the map as the dimension of its kernel. That is,

$$\operatorname{Rank} T = \dim(\operatorname{Im} T), \quad \operatorname{Null} T = \dim(\ker T).$$

Whilst these concepts may not at first seem connected, they are linked through the remarkable dimension formula.

Theorem 1.4.9 (Dimension formula). Let $T : U \to V$ be a linear map. Then $\dim U = \operatorname{Rank} T + \operatorname{Null} T$.

We shall not prove this result here. However, the dimension formula does give a very useful method of calculating the rank of the matrix knowing the nullity or vice-versa. Sometimes it is the case that a linear map is defined in a complex fashion, and calculating the dimension of one of these spaces is hard whereas the other is trivial or at least moderately straight-forward. The dimension formula gives us a way to quickly calculate the rank or nullity of one without knowing the other. It is also an extremely useful tool in the proof of many results throughout linear algebra.

In practice it is often quicker to calculate the rank and then the nullity of a matrix. We will demonstrate a method here which works quickly and easily using the matrix representation of the map; however, we first need some basic definitions.

Definition 1.4.10 (Elementary operations). Let $A = (a_{ij})$ be a $m \times n$ matrix with rows $\mathbf{r}_i$ and columns $\mathbf{c}_j$. Then the **elementary row (or column) operations** on A are defined by:

1. Swapping a row (or column) of A, denoted by $\mathbf{r}_i \leftrightarrow \mathbf{r}_j$.
2. Multiplying a row (or column) of A by a scalar $\lambda \in \mathbb{K}$, denoted by $\mathbf{r}_i \to \lambda\mathbf{r}_i$.
3. Adding a multiple of a row (or column) of A to another row or column, denoted by $\mathbf{r}_i \to \mathbf{r}_i + \lambda\mathbf{r}_j$.

The following theorem then allows us to connect elementary row and column operations to calculating ranks of matrices.

Theorem 1.4.11. Let B be a $m \times n$ matrix obtained by use of elementary row and column operations on a matrix A. Then $\operatorname{Rank}(B) = \operatorname{Rank}(A)$.

Definition 1.4.12 (Row-echelon form). A matrix is said to be in **row-echelon form** if

(i) All non-zero rows are above any rows of all zeroes;

(ii) The leading coefficient (the first non-zero number from the left) of each row is 1 and always strictly to the right of the leading coefficient of the row above it.

Further, it is said to be in **reduced row-echelon form** if

(i) the matrix is in row-echelon form;

(ii) every leading coefficient is 1 and is the only non-zero entry in its column.

Theorem 1.4.13. Any $m \times n$ matrix $A = (\alpha_{ij})$ can be transformed into row-echelon or reduced row-echelon form through use of elementary row and column operations. Furthermore, this transformation is unique.

Proof 1.4.14. We shall not explicitly prove this theorem, but the following algorithm yields the unique row-echelon form of a matrix. At any particular stage we will consider a point (i, j) corresponding to the matrix element α_{ij}, called the *pivot point*. Set $(i, j) = (1, 1)$ and then proceed in the following fashion:

1. If $\alpha_{kj} = 0$ for all $k \geq i$ then move the pivot to $(i, j + 1)$ and repeat this step. If $j = n$ then stop.
2. If $\alpha_{ij} = 0$ but $\alpha_{kj} \neq 0$ for some $k > j$ then perform $\mathbf{r}_k \leftrightarrow \mathbf{r}_j$.
3. If $\alpha_{ij} \neq 1$ then perform $\mathbf{r}_i \to \alpha_{ij}^{-1}\mathbf{r}_i$.
4. If for any $k > i, \alpha_{kj} \neq 0$ then apply $\mathbf{r}_k \to \mathbf{r}_k - \alpha_{kj}\mathbf{r}_i$.
5. If $i = m$ or $j = n$ then stop. Otherwise, move the pivot to position $(i+1, j+1)$ and go back to step 1.

For reduced row-echelon form, replace step 4 by

4. If for any $k \neq i, \alpha_{kj} \neq 0$ then apply $\mathbf{r}_k \to \mathbf{r}_k - \alpha_{kj}\mathbf{r}_i$. □

Example 1.4.15. This algorithm is a little dense without an example. We demonstrate the algorithm on a simple 4×5 matrix in the table below.

Matrix	Pivot point	Step number	Operation
$\begin{bmatrix} 0 & 0 & 1 & 2 & 1 \\ 2 & 4 & 2 & -4 & 2 \\ 3 & 6 & 3 & -6 & 3 \\ 1 & 2 & 3 & 3 & 3 \end{bmatrix}$	$(1, 1)$	2	$\mathbf{r}_1 \leftrightarrow \mathbf{r}_2$
$\begin{bmatrix} 2 & 4 & 2 & -4 & 2 \\ 0 & 0 & 1 & 2 & 1 \\ 3 & 6 & 3 & -6 & 3 \\ 1 & 2 & 3 & 3 & 3 \end{bmatrix}$	$(1, 1)$	3	$\mathbf{r}_1 \to \mathbf{r}_1/2$
$\begin{bmatrix} 1 & 2 & 1 & -2 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 3 & 6 & 3 & -6 & 3 \\ 1 & 2 & 3 & 3 & 3 \end{bmatrix}$	$(1, 1)$	3	$\mathbf{r}_3 \to \mathbf{r}_3 - 3\mathbf{r}_1$ $\mathbf{r}_4 \to \mathbf{r}_4 - \mathbf{r}_1$

$\begin{bmatrix} 1 & 2 & 1 & -2 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 2 & 5 \end{bmatrix}$	$(1,1) \to (2,2) \to (3,3)$	5, 1
$\begin{bmatrix} 1 & 2 & 1 & -2 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 2 & 5 \end{bmatrix}$	$(2,3)$	4 $\quad \mathbf{r}_4 \to \mathbf{r}_4 - 2\mathbf{r}_2$
$\begin{bmatrix} 1 & 2 & 1 & -2 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$	$(2,3) \to (3,4)$	5, 2 $\quad \mathbf{r}_3 \leftrightarrow \mathbf{r}_4$
$\begin{bmatrix} 1 & 2 & 1 & -2 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$(3,4) \to (4,5) \to$ stop	5, 1

The reduced row-echelon form can be found in a very similar fashion with resultant matrix

$$\begin{bmatrix} 1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

So, how does the algorithm help us obtain the rank? We know by theorem 1.4.13 that row operations do not change the rank of a matrix, but we need one final theorem.

Theorem 1.4.16. The rank of a matrix is equal to the number of non-zero rows of its row-echelon or reduced row-echelon form.

Hence in the example above, $\text{Rank}(A) = 3$.

Finally, we introduce some special forms of matrices that you should be aware of. Usually when a matrix has a special form (or can be transformed into a special form through elementary row and column operations) then calculations can be made significantly easier.

Definition 1.4.17 (Triangular matrix). Let $A = (\alpha_{ij})$ be a $n \times n$ matrix. A is said to be **lower triangular** if $\alpha_{ij} = 0$ for all $i > j$. Similarly, A is **upper triangular** if $\alpha_{ij} = 0$ for all $i < j$.

The two matrices below represent upper- and lower-triangular matrices in their general form.

$$U = \begin{bmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ 0 & u_{22} & u_{23} & \dots & u_{2n} \\ 0 & 0 & u_{33} & \dots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & u_{nn} \end{bmatrix} \qquad L = \begin{bmatrix} l_{11} & 0 & 0 & \dots & 0 \\ l_{21} & l_{22} & 0 & \dots & 0 \\ l_{31} & l_{32} & l_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \dots & l_{nn} \end{bmatrix}$$

1.5 The inverse of a matrix

Matrices are commonly used to represent systems of linear equations. For example,

$$\begin{aligned} x_1 + 4x_2 + \ \ x_3 &= 3 \\ x_1 + 2x_2 - 2x_3 &= 6 \\ 2x_1 + 3x_2 + 3x_3 &= 1 \end{aligned} \quad \Longleftrightarrow \quad \begin{bmatrix} 1 & 4 & 1 \\ 1 & 2 & -2 \\ 2 & 3 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \\ 1 \end{bmatrix}.$$

If we can find an *inverse* of the above matrix, we can easily find a solution to the system.

Definition 1.5.1. A $n \times n$ matrix A is **invertible** (or non-singular) if there exists a matrix B such that $AB = BA = I_n$. B is called the **inverse** of A and is written $A^{-1} = B$. If A is not invertible then it is called **singular**.

1.5.1 Determinants

There are a number of methods to determine whether a matrix is invertible. One such method is through the determinant of a matrix.

Definition 1.5.2 (Determinant). Let $A = (a_{ij})$ be a $n \times n$ matrix. Then formally, we define the **determinant** of A by the summation

$$\det(A) = \sum_{\phi \in S_n} \operatorname{sgn}(\phi) a_{1\phi(1)} a_{2\phi(2)} \dots a_{n\phi(n)},$$

where

$$S_n = \{\sigma : \{1, \dots, n\} \to \{1, \dots, n\} \mid \sigma \text{ is a bijection}\}, \qquad \operatorname{sgn}(\phi) = \begin{cases} 1, & \phi \text{ is even}, \\ 0, & \phi \text{ is odd}. \end{cases}$$

Whilst this is a very formal definintion, you will have almost certainly seen determinants for the cases $n = 2$ and $n = 3$:

$$n = 2: \qquad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

$$n = 3: \qquad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

$$= a_{11}a_{22}a_{33} - a_{11}a_{23}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{32}a_{32} - a_{13}a_{22}a_{31}$$

The precise definition of S_n (called the permutation group) and odd/even permutations is beyond the scope of this course, although you can look them up in nearly every textbook on basic linear algebra. However, this formal definition can be manipulated to yield some fascinating results and properties which we outline in the following theorem.

Definition 1.5.3 (Transpose). Let $A = (\alpha_{ij})$ be an $m \times n$ matrix. Then the **transpose** of A, denoted $A^\top$, is an $n \times m$ matrix (β_{ij}) such that $\beta_{ij} = \alpha_{ji}$.

Theorem 1.5.4 (Properties of the determinant). The determinant of an $n \times n$ matrix A satisfies the following properties:

(i) $\det(I_n) = 1$.

(ii) $\det(A^\top) = \det(A)$.

(iii) If B is a matrix obtained by swapping two rows of A then $\det(B) = -\det(A)$.

(iv) If B is a matrix obtained by multiplying a row (or column) of A by $\lambda \in \mathbb{K}$ then $\det(B) = \lambda \det(A)$.

(v) If B is a matrix obtained by adding a multiple of a row (or column) of A to another row (or column) of A then $\det(B) = \det(A)$.

(vi) If B is an $n \times n$ matrix then $\det(AB) = \det(BA)$.

(vii) If $A = (\alpha_{ij})$ is an $n \times n$ upper-triangular or lower-triangular matrix then $\det(A) = \alpha_{11} \cdot \alpha_{22} \cdots \alpha_{nn}$.

Example 1.5.5. There are many ways to calculate the determinant of a matrix, particularly if the matrix has a special structure such as triangularity. However for a general matrix, there are common techniques which are easily utilised. We will outline two methods here.

(i) The first is known as *Cramer's rule*, and is the most natural method following nearly directly from the definition of the determinant. We pick a row or column of the matrix, and then 'expand' about this row or column. In general, if A_{ij} is the $(n-1) \times (n-1)$ matrix obtained by removing the i-th row and j-th column of $A = (a_{ij})$, then expanding about the k-th row or column gives

$$\begin{aligned}\det(A) &= \sum_{i=1}^{n} (-1)^{i+k} a_{ik} \det(A_{ik}) \\ &= \sum_{j=1}^{n} (-1)^{k+j} a_{kj} \det(A_{kj}).\end{aligned}$$

You have seen this already when calculating a 3×3 determinant as the summation of three 2×2 determinants. As a larger example, consider the matrix

$$A = \begin{bmatrix} 1 & 0 & 1 & 2 \\ -1 & 3 & 2 & 3 \\ 2 & 5 & 1 & -3 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Expanding about the last row, one obtains

$$\det(A) = (-1)^6 \cdot 1 \cdot \begin{vmatrix} 1 & 1 & 2 \\ -1 & 2 & 3 \\ 2 & 1 & -3 \end{vmatrix} + (-1)^7 \cdot 1 \cdot \begin{vmatrix} 1 & 0 & 2 \\ -1 & 3 & 3 \\ 2 & 5 & -3 \end{vmatrix} = 30.$$

You are ill advised to use this method on large matrices, since in general at the n-th stage, this algorithm will require evaluating n subdeterminants of size $(n-1) \times (n-1)$, meaning that this algorithm has $\mathcal{O}(n!)$ complexity! This being said, it is very useful for evaluating determinants of sparse matrices, since a careful choice of row or column containing many zeros to expand about may yield a quick answer.

(ii) Theorem 1.5.4 tells us the effect of elementary row operations on the determinant. By repeatedly applying swapping and subtraction operations it is easy to put the matrix in

upper-triangular form. For example,

$$\begin{vmatrix} 0 & 1 & 1 & 2 \\ 1 & 2 & 1 & 1 \\ 2 & 1 & 3 & 1 \\ 1 & 2 & 4 & 2 \end{vmatrix} \xrightarrow{\mathbf{r}_1 \leftrightarrow \mathbf{r}_2} - \begin{vmatrix} 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 2 \\ 2 & 1 & 3 & 1 \\ 1 & 2 & 4 & 2 \end{vmatrix} \xrightarrow[\mathbf{r}_4 \to \mathbf{r}_4 - \mathbf{r}_1]{\mathbf{r}_3 \to \mathbf{r}_3 - 2\mathbf{r}_2} - \begin{vmatrix} 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & -3 & 1 & -1 \\ 0 & 0 & 3 & 1 \end{vmatrix} \xrightarrow{\mathbf{r}_3 \to \mathbf{r}_3 + 3\mathbf{r}_2}$$

$$- \begin{vmatrix} 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 4 & 5 \\ 0 & 0 & 3 & 1 \end{vmatrix} \xrightarrow{\mathbf{r}_4 \to \mathbf{r}_4 - \frac{3}{4}\mathbf{r}_3} - \begin{vmatrix} 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 4 & 5 \\ 0 & 0 & 0 & -\frac{11}{4} \end{vmatrix}$$

and hence the determinant is 11. This is a much faster approach for larger matrices.

As a last word on singularity, here is an *extremely useful and important* theorem which ties together all of our knowledge so far.

Theorem 1.5.6. The following statements are equivalent:

(i) A is a singular $n \times n$ matrix;

(ii) $\det A = 0$;

(iii) $\operatorname{Rank} A < n$;

(iv) There exists $\mathbf{v} \in \ker A$ such that $\mathbf{v} \neq \mathbf{0}$.

1.5.2 Calculating the inverse

We have, in fact, already seen one of the most efficient methods for calculating the inverse of a matrix. If we begin with an invertible $n \times n$ matrix A and compute its reduced row-echelon form, we obtain the identity matrix I_n. Curiously, if we start from the identity matrix, and apply the same operations we used to compute the reduced row-echelon form of A, then we obtain A^{-1}.

Whilst this might seem a little counter-intuitive, this technique works because any elementary row operation can be viewed as a matrix multiplication. Each elementary row operation has an equivalent matrix form so that the resulting product performs the desired operation. For example to swap rows 2 and 3 of a 3×3 matrix A we can perform a multiplication by a matrix U where

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \\ 4 & 5 & 6 \end{bmatrix}.$$

Such matrices are called **elementary matrices**. If a matrix is invertible, then there exist a sequence of elementary matrices $E_1, \ldots, E_k$ such that

$$(E_1 E_2 \cdots E_k) A = I_n,$$

and so post-multiplying both sides of this equation by A^{-1} we find $(E_1 E_2 \cdots E_k) I_n = A^{-1}$. Finally then, we present an example to illustrate this method and demonstrate how to find the inverse of a 3×3 matrix.

Matrix	Operation	Identity
$\begin{bmatrix}1&2&3\\1&1&0\\3&2&1\end{bmatrix}$	$\mathbf{r}_2 \to \mathbf{r}_2 - \mathbf{r}_1$ $\mathbf{r}_3 \to \mathbf{r}_3 - 3\mathbf{r}_1$	$\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$
$\begin{bmatrix}1&2&3\\0&-1&-3\\0&-4&-8\end{bmatrix}$	$\mathbf{r}_2 \to -\mathbf{r}_2$	$\begin{bmatrix}1&0&0\\-1&1&0\\-3&0&1\end{bmatrix}$
$\begin{bmatrix}1&2&3\\0&1&3\\0&-4&-8\end{bmatrix}$	$\mathbf{r}_1 \to \mathbf{r}_1 - 2\mathbf{r}_2$ $\mathbf{r}_3 \to \mathbf{r}_3 + 4\mathbf{r}_2$	$\begin{bmatrix}1&0&0\\1&-1&0\\-3&0&1\end{bmatrix}$
$\begin{bmatrix}1&0&-3\\0&1&3\\0&0&4\end{bmatrix}$	$\mathbf{r}_3 \to \frac{1}{4}\mathbf{r}_3$	$\begin{bmatrix}-1&2&0\\1&-1&0\\1&-4&1\end{bmatrix}$
$\begin{bmatrix}1&0&-3\\0&1&3\\0&0&1\end{bmatrix}$	$\mathbf{r}_1 \to \mathbf{r}_1 + 3\mathbf{r}_3$ $\mathbf{r}_2 \to \mathbf{r}_2 - 3\mathbf{r}_3$	$\begin{bmatrix}-1&2&0\\1&-1&0\\\frac{1}{4}&-1&\frac{1}{4}\end{bmatrix}$
$\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$	**done**	$\begin{bmatrix}-\frac{1}{4}&-1&\frac{3}{4}\\\frac{1}{4}&2&-\frac{3}{4}\\\frac{1}{4}&-1&\frac{1}{4}\end{bmatrix}$

Therefore we have that

$$\begin{bmatrix}1&2&3\\1&1&0\\3&2&1\end{bmatrix}^{-1} = \begin{bmatrix}-\frac{1}{4}&-1&\frac{3}{4}\\\frac{1}{4}&2&-\frac{3}{4}\\\frac{1}{4}&-1&\frac{1}{4}\end{bmatrix}.$$

1.6 Eigenvalues and eigenvectors

Calculating the eigenvalues and eigenvectors of a matrix are particularly important in many areas of applied mathematics. We will make heavy use of them in the following chapter, and we therefore give their definition, a brief outline of some important results and some examples of how to calculate them here.

Definition 1.6.1 (Eigenvalues and eigenvectors). Let V be an n-dimensional vector space and $T : V \to V$ a linear transformation of V with a matrix representation A. Then $\lambda \in \mathbb{K}$ is an **eigenvalue** of A if there exists a non-zero vector $\mathbf{v} \in V$ such that

$$T(\mathbf{v}) = \lambda\mathbf{v} \Leftrightarrow A\mathbf{v} = \lambda\mathbf{v}.$$

The vector $\mathbf{v}$ is called the **eigenvector** corresponding to λ.

Definition 1.6.2 (Characteristic polynomial). For A as defined above, the **characteristic polynomial** of degree n is defined by

$$f_A(z) = \det(A - zI_n)$$

where I_n is the identity matrix.

Theorem 1.6.3. λ is an eigenvalue of A if and only if $f_A(\lambda) = 0$.

Proof 1.6.4. We first prove the forward direction. Let λ be an eigenvalue of A. Then there exists $\mathbf{v} \in V$ with $\mathbf{v} \neq \mathbf{0}$ such that $A\mathbf{v} = \lambda\mathbf{v}$, and so $(A - \lambda I_n)\mathbf{v} = \mathbf{0}$. By definition, this means that $\mathbf{v} \in \ker(A - \lambda I_n)$ and $\text{Null}(A - \lambda I_n) > 0$. Applying the dimension formula we see that

$$\text{Rank}(A - \lambda I_n) = n - \text{Null}(A - \lambda I_n) < n,$$

hence any n vectors belonging to the image of $A - \lambda I_n$ are linearly dependent. It follows that $f_A(z) = \det(A - \lambda I_n) = 0$.

Suppose that $f_A(\lambda) = 0$. Then by theorem 1.5.6, the transformation $(A - \lambda I_n)$ is singular. Moreover, by the same theorem, there exists $\mathbf{v} \in \ker(A - \lambda I_n)$ with $\mathbf{v} \neq 0$ such that $(A - \lambda I_n)\mathbf{v} = 0$. This implies $A\mathbf{v} = \lambda\mathbf{v}$, so λ is an eigenvalue of A. □

Definition 1.6.5 (Eigenspace). Let λ be an eigenvalue of A. Then the set

$$\{\mathbf{v} \mid \mathbf{v} \text{ is an eigenvector of } A\} \cup \{\mathbf{0}\} = \ker(A - \lambda I_n)$$

is called the *eigenspace* of λ.

Example 1.6.6. In this example we will calculate the eigenvalues and eigenvectors of real-valued 2×2 matrices, and demonstrate the three outcomes which are possible. These correspond to all of the possible roots of a polynomial $f(z)$: we can either obtain two real values, a single repeated value or a complex conjugate pair.

(i) Consider the matrix

$$A = \begin{bmatrix} -4 & 2 \\ 3 & -5 \end{bmatrix}.$$

Theorem 1.5.6 gives us an ideal way to find all of the eigenvalues of A. We first calculate $f_A(z)$ by expanding the determinant;

$$f_A(z) = \begin{vmatrix} -4 - z & 2 \\ 3 & -5 - z \end{vmatrix} = (4 + z)(5 + z) - 6 = z^2 + 9z + 14 = (z + 7)(z + 2).$$

Hence we see that the eigenvalues of A are $\lambda_1 = -2$ and $\lambda_2 = -7$.

To find the eigenvectors of A, we simply use the definition. Since the eigenvector $\mathbf{v}$ correspoding to λ_1 will satisfy $(A + 2I_2)\mathbf{v} = \mathbf{0}$, writing $\mathbf{v} = (v_1, v_2)^\top$ we see that

$$\begin{bmatrix} -2 & 2 \\ 3 & -3 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

This immediately shows that $-2v_1 + 2v_2 = 0$, and hence $v_1 = v_2$. Also we see that $3v_1 - 3v_2 = 0$, which also shows $v_1 = v_2$. We therefore see that the eigenvector for λ_1 is

$$v = \begin{bmatrix} v_1 \\ v_1 \end{bmatrix} = v_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

This highlights an important point: **eigenvectors are not generally unique**. This is clear because the eigenspace of an eigenvalue is (for real matrices) at least dimension 1, and thus is spanned by a single vector. However, the subspace will in general have more than one element. Usually we omit the constant and write $\mathbf{v} = (1, 1)^\top$.

For λ_2, in a similar fashion we find that $3v_1 + 2v_2 = 0$ and hence $v_2 = -\frac{3}{2}v_1$. So $\mathbf{v} = (3, -2)^\top$.

(ii) Consider the matrix

$$A = \begin{bmatrix} 7 & -1 \\ 4 & 3 \end{bmatrix} \Rightarrow f_A(z) = z^2 - 10z + 25 = (z-5)^2.$$

In this example, we only obtain one **repeated** eigenvalue – in this case, the eigenvalue is said to have **multiplicity of 2**. Typically to denote this we write $\lambda_{1,2} = 5$. Following the usual procedure we obtain one eigenvector, $\mathbf{v} = (1,2)^\top$.

Note that in some cases, you may be able to find more than one eigenvector for a repeated eigenvalue. For example, the identity matrix has a single eigenvalue $\lambda = 1$, and any non-zero vector in $\mathbb{R}^2$ is an eigenvector.

(iii) Consider the matrix

$$A = \begin{bmatrix} 1 & -2 \\ 1 & 3 \end{bmatrix} \Rightarrow f_A(z) = z^2 - 4z + 5.$$

In this case the characteristic polynomial has no real roots. Instead we find that

$$\lambda_{1,2} = \tfrac{1}{2}(4 \pm \sqrt{-4}) = 2 \pm i,$$

so that the eigenvalues of the matrix are not real. This is perfectly acceptable, and we can continue as usual to find the eigenvectors $(2, -1-i)^\top$ and $(2, -1+i)^\top$.

Finally, we finish the section with two small results which allow you to calculate the eigenvalues of triangular matrices very easily.

Theorem 1.6.7. Suppose that $A = (\alpha_{ij})$ is a $n \times n$ upper- or lower-triangular matrix. Then for all $1 \leq i \leq n$, α_{ii} is an eigenvalue of A.

Corollary 1.6.8. Let A be an $n \times n$ matrix. Then $\det(A) = \lambda_1 \lambda_2 \cdots \lambda_n$, where λ_i is an eigenvalue of A.

1.7 Diagonalising a matrix

Definition 1.7.1. An $n \times n$ matrix is said to be **diagonalisable** if there exists a matrix P such that $P^{-1}AP$ is a diagonal matrix.

We will quickly set up the framework to allow us an easy way to find whether or not a matrix is diagonalisable, and how to find the matrix P. As we shall see, the diagonalisation of a matrix is intricately linked to the eigenvector decomposition of the matrix. We start by proving that the eigenvectors of a linear map form a basis for the domain.

Lemma 1.7.2. Let $T : V \to V$ be a linear map with matrix representation A. Then A is diagonal with respect to some basis of V if and only if V has a basis consisting of eigenvectors of T.

Proof 1.7.3. Suppose that $A = (\alpha_{ij})$ is diagonal with respect to the basis $\{\mathbf{e}_i\}_{i=1}^n$. This implies that $A\mathbf{e}_i = \alpha_{ii}\mathbf{e}_i$, and so $\mathbf{e}_i$ is an eigenvector of A.

Now suppose that $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is a basis of V consisting of eigenvectors. Then, by definition, $T(\mathbf{e}_i) = \lambda \mathbf{e}_i$ and so the matrix is diagonal. $\square$

Theorem 1.7.4. A matrix A has n linearly independent eigenvectors if and only if A is diagonalisable.

Example 1.7.5. Consider the matrix from example 1.6.6(i). Then the matrix P is given by

$$P = \begin{bmatrix} 1 & -2 \\ 1 & 3 \end{bmatrix} \Rightarrow P^{-1}AP = \frac{1}{5}\begin{bmatrix} 3 & 2 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} -4 & 2 \\ 3 & -5 \end{bmatrix}\begin{bmatrix} 1 & -2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} -2 & 0 \\ 0 & -7 \end{bmatrix}$$

Notice that the diagonal matrix we obtain contains the eigenvalues along the diagonal.

Whilst the above theorem works well for those matrices which have linearly independent eigenvectors, we are so far not able to say anything for those which do not. The matrix of example 1.6.6(ii) only has a single eigenvector for its repeated eigenvalue. It is clear from the contrapositive of the above theorem that this matrix is not diagonalisable, but how close can we get to a diagonal matrix?

The final piece of the puzzle is to be able to find vectors which are 'nearly' eigenvectors, in the sense that we can use them to 'almost' diagonalise a matrix.

Definition 1.7.6 (Generalised eigenvector). Given a matrix A with eigenvalue λ with multiplicity $k \geq 1$, then a non-zero vector $\mathbf{v} \in V$ is said to be a **generalised eigenvector** if

$$(A - \lambda I_n)^k \mathbf{v} = \mathbf{0}.$$

It can be proven that if λ is an eigenvalue of multiplicity k, then $\text{Null}(A - \lambda I_n)^k = k$ which guarantees we can always find a basis consisting of eigenvectors and generalised eigenvectors.

Calculating a generalised eigenvector is relatively straightforward after proving some properties of the null-space of the powers of a matrix which we will omit here. We construct a sequence of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ such that

$$(A - \lambda I_n)\mathbf{x}_i = \mathbf{x}_{i-1},$$

where we define $\mathbf{x}_0 = \mathbf{0}$ so that $\mathbf{x}_1$ is the regular eigenvector of the system. Each resulting $\mathbf{x}_i$ for $i > 1$ is a generalised eigenvector. Using this method, for a eigenvalue of multiplicity k we therefore find $k - 1$ linearly independent generalised eigenvectors which span the eigenspace of λ. Coupled with the other eigenvectors of the system, this gives us a basis for the entire space.

Example 1.7.7. Let us consider the matrix in example 1.6.6(ii). We found that the characteristic polynomial had a root $\lambda_{1,2} = 5$ of multiplicity 2 with an associated eigenvector $\mathbf{x}_1 = (1, 2)^\top$.

To find the generalised eigenvector, we solve the equation

$$(A - 5I_2)\mathbf{x}_2 = \mathbf{x}_1,$$

so that

$$(A - 5I_2)^2\mathbf{x}_2 = 0_{2\times 2}$$

where the **zero matrix** $0_{n\times m}$ is the $n \times m$ matrix of zeroes. Hence we obtain the equations

$$\begin{bmatrix} 2 & -1 \\ 4 & -2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Notice that just like regular eigenvectors, the generalised eigenvector is not uniquely determined by this equation. We see that $2x - y = 1$, and so $y = 2x - 1$ giving $\mathbf{x}_2 = (x, 2x - 1)^\top$. At this point we can pick any value of x to give us a solution. Picking $x = 0$ gives the generalised eigenvector $\mathbf{x}_2 = (0, -1)^\top$.

Importantly, generalised eigenvectors allow us to 'nearly' diagonalise a matrix. In the example above, consider the matrix P formed by regular eigenvector $\mathbf{x}_1$ and the generalised eigenvector $\mathbf{x}_2$. The obvious choice for P then is

$$P = \begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix}.$$

Since $\mathbf{x}_1$ and $\mathbf{x}_2$ are linearly independent by construction, P is invertible, and so

$$P^{-1}AP = \begin{bmatrix} 5 & 1 \\ 0 & 5 \end{bmatrix}$$

The resultant matrix gives us a special form of A which has the (repeated) eigenvalue along the diagonal and an entry of 1 above the diagonal. This remarkable fact can be generalised to any $n \times n$ matrix, as the next theorem describes.

Theorem 1.7.8. Let A be an $n \times n$ matrix with eigenvalues $\lambda_1, \ldots \lambda_n$ which are not necessarily distinct. Then there exists an invertible matrix P such that

$$P^{-1}AP = \begin{bmatrix} \lambda_1 & a_1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2 & a_2 & \cdots & 0 & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_{n-1} & a_{n-1} \\ 0 & 0 & 0 & \cdots & 0 & \lambda_n \end{bmatrix}$$

where each a_i is either 0 or 1. $P^{-1}AP$ is said to be the **Jordan canonical form** or **Jordan normal form** of the matrix A.

Finding the Jordan canonical form of a general $n \times n$ matrix is not covered in this course - at most we will consider a 3×3 matrix. However the same approach outlined here will work for larger matrices, at the expense of a greater degree of difficulting in determining the eigenvalues and eigenvectors.

To conclude this material on eigenvalues and eigenvectors, we state a remarkable theorem connecting the characteristic polynomial to the matrix from which it is derived.

Theorem 1.7.9 (Cayley-Hamilton). Let A be a $n \times n$ matrix with characteristic equation $f_A(z)$. Then $f_A(A) = 0_{n \times n}$.

The proof of the Cayley-Hamilton theorem falls beyond the scope of this short course. However, do not be tempted to use the false proof of

$$f_A(A) = \det(A - AI_n) = \det(0_{n \times n}) = 0_{n \times n}$$

since the first equality does not make sense. As an example, if the characteristic equation for a 2×2 matrix A is $f_A(z) = z^2 + 2z + 5$, then the Cayley-Hamilton theorem states that $A^2 + 2A + 5I_2 = 0$.

Example 1.7.10. The Cayley-Hamilton theorem can be applied in a number of interesting ways. Here, we show two such examples.

(i) In this first example, we shall find a non-diagonal matrix A such that $A^2 + 5A + 6I = 0$.

One obvious approach is to assume that the matrix is a 2×2 matrix of arbitrary form, evaluate the left hand side and find a solution to the system of equations. However this takes quite a lot of time! Instead, we note that, by the Cayley-Hamilton theorem, we need only find a matrix which has characteristic polynomial

$$f_A(z) = z^2 + 5z + 6.$$

Clearly $f_A(z) = (z+3)(z+2)$, and so if we can find a matrix with eigenvalues $\lambda_1 = -3$, $\lambda_2 = -2$ then the problem is solved. An obvious choice is the matrix

$$A = \begin{bmatrix} -3 & 1 \\ 0 & -2 \end{bmatrix}$$

which is upper-triangular and so has the appropriate eigenvalues. To verify, manually evaluating the expression we see that

$$A^2 + 5A + 6I = \begin{bmatrix} 9 & -5 \\ 0 & 4 \end{bmatrix} + \begin{bmatrix} -15 & 5 \\ 0 & -10 \end{bmatrix} + \begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix} = 0_{n \times n}.$$

(ii) We will show that the matrix

$$A = \begin{bmatrix} 0 & 1 & -1 \\ 1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

satisfies the horrific equality $A^6 = 4A^4 + 4A^3 - 7A^2 - 4A + 4I_3$. Whilst this is possible through direct calculation, working out the sixth power of a 3×3 matrix is clearly extremely time consuming! We will show an alternative method using the Cayley-Hamilton theorem.

First, we find the characteristic equation. This is somewhat more tricky for 3×3 matrices, but the methodology is essentially the same.

$$\begin{aligned} f_A(z) = \begin{bmatrix} -z & 1 & -1 \\ 1 & 1-z & 0 \\ -1 & 0 & 1-z \end{bmatrix} &= -z(1-z)^2 - (1-z) - (1-z) \\ &= -z^3 + 2z^2 + z - 2 \end{aligned}$$

If we were to calculate the eigenvalues at this point then the process is a little convoluted. The standard method is to guess a root of the polynomial and factor it from there.

However in this case we simply apply the Cayley-Hamilton theorem which implies that

$$-A^3 + 2A^2 + A - 2I_3 = 0_{3\times 3} \Rightarrow A^3 = 2A^2 + A - 2I_3.$$

This immediately implies that

$$\begin{aligned} A^6 = (2A^2 + A - 2I_3)^2 &= 2A^2(2A^2 + A - 2I_3) + A(2A^2 + A - 2I_3) - 2I_3(2A^2 + A - 2I_3) \\ &= 4A^4 + 4A^3 - 7A^2 - 4A + 4I_3. \end{aligned}$$

1.8 Dual spaces

Let V be a finite dimensional real vector space with basis $\{\mathbf{e}_i\}_{i=1}^n$. The maps $f_j : V \to \mathbb{R}$ defined by

$$f_j : \mathbf{v} = \sum_{i=1}^{n} a_i \mathbf{e}_i \mapsto a_j \quad \text{for } j = 1, \ldots, n$$

give you the coordinates of any vector $\mathbf{e}$ in the basis of $\mathbf{e}_i$. This is an example of a **linear functional**.

Definition 1.8.1 (Linear functional). A **linear functional** on a vector space V over $\mathbb{K}$ is a map $f : V \to \mathbb{K}$ satisfying

$$f(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha f(\mathbf{x}) + \beta f(\mathbf{y}), \qquad \alpha, \beta \in \mathbb{K},\ \mathbf{x}, \mathbf{y} \in V.$$

Example 1.8.2. Let $V = \mathcal{P}_n[x]$, and take $x \in [0, 1]$. Then

1. $f_j\left(\sum_{i=0}^n \alpha_i x^i\right) = \alpha_j$ for $j = 0, \ldots, n$;
2. $f\left(\sum_{i=0}^n \alpha_i x^i\right) = \sum_{i=1}^n \alpha_i$;
3. $f\left(\sum_{i=0}^n \alpha_i x^i\right) = \sum \alpha_i x_0^i$ for some $x_0 \in [0, 1]$;

are all examples of linear functionals over V.

Definition 1.8.3 (Dual space). Denote the set of all linear functionals on V by V^*. V^* is called the **dual space** to V.

Whilst we will not study them in detail here, dual spaces are typically used in the study of infinite dimensional spaces. They are particularly important in measure theory (and hence probability theory), and also in the general study of functional analysis. They can be quite tricky to study in infinite dimensional cases. But like a lot of other things, dual spaces are a lot easier to deal with in a finite dimensional setting.

Dual spaces are useful because we may construct isomorphisms which allow us to transfer results from one space easily to results about its dual space. Quite often the proof in one may be far easier, or more succinct, than the corresponding proof in the other. We state a few theorems here.

Lemma 1.8.4. Let V be a finite-dimensional vector space. Then V^* is finite dimensional and $\dim V^* = \dim V$.

Proof 1.8.5. This follows from one of the simple examples given above. Take $\{\mathbf{e}_i\}_{i=1}^n$ as a basis of V. Then define $\mathbf{e}^i : V \to \mathbb{K}$ defined by

$$\mathbf{e}^i(a_1\mathbf{e}_1 + \cdots + \cdots a_n\mathbf{e}_n) = a_i$$

It is clear that this forms a basis for V^*. In fact, this is known as the **dual basis** of $\{\mathbf{e}_i\}_{i=1}^n$. Since there are precisely n elements in the basis and dual basis, we immediately see that $\dim V^* = \dim V$. □

This proof leads to a very nice result, and one that is fundamental to the study of dual vector spaces.

Theorem 1.8.6. If V is a finite dimensional vector space then V is isomorphic to V^{**}.

Proof 1.8.7. This follows since every vector $\mathbf{v}$ in V can be thought of as a linear functional on V^*. Indeed, given $\mathbf{v}^*$ in V^*, define $\mathbf{v}(\mathbf{v}^*)$ to be $\mathbf{v}^*(\mathbf{v})$. We can rephrase this as follows: there is a map $g : V \to V^{**}$ which maps $\mathbf{v}$ in V to $g(\mathbf{v})$ in V^{**}, where $g(\mathbf{v})$ is the linear functional on V^* defined by the formula $g(\mathbf{v})(\mathbf{v}^*) = \mathbf{v}^*(\mathbf{v})$.

It is clear that g is linear, and so if we can show that $\ker g = \{0\}$ then it follows by theorem 1.5.6 that the image of g is V^{**}, and hence g is an isomorphism.

Let $\mathbf{v}$ be a non-zero vector in V. Then $\mathbf{v} = a_1\mathbf{e}_1 + \cdots + a_n\mathbf{e}_n$ and at least one a_i is non-zero. Then $\mathbf{e}^i(\mathbf{v}) = a_i$ and so $\mathbf{e}^i$ is non-zero. Hence (by taking the contrapositive), $\ker g = \{0\}$ and g is an isomorphism. □

1.9 Inner product (scalar product)

The definition of a vector space gives the description of lines, planes, parallel lines and subspaces.

In order to define the length of a vector and the angles between vectors (in particular to define orthogonality) we need to introduce an additional structure on the vector space - the **inner product**, otherwise known as the **scalar product**.

Definition 1.9.1 (Inner product). Let V be a finite dimensional vector space. Let $h : V \times V \to \mathbb{R}$ be

- **bilinear:** $h(\lambda x_1 + y_1, \mu x_2 + y_2) = \lambda\mu h(x_1, x_2) + \lambda h(x_1, y_2) + \mu h(y_1, x_2) + h(y_1, y_2)$;
- **symmetric:** $h(x, y) = h(y, x)$ for $x, y \in V$;
- **positive definite:** $h(x, x) > 0\ \forall x \neq 0$ and $h(x, x) = 0$ if and only if $x = 0$.

Then h defines an **inner product** on V, and V is called a **Euclidean space**. Usually we do not denote the inner product as a function, but instead by the symbol $\langle u, v\rangle = h(u, v)$.

An important property of the inner product is that it allows us to define whether two vector or orthogonal to one another.

Definition 1.9.2. Suppose that V is a vector space and $W = \{\mathbf{f}_i\}_{i=1}^n \subset V$. Then W is said to be an **orthogonal set** under an inner product if $\langle \mathbf{f}_i, \mathbf{f}_j\rangle = 0$ for $i \neq j$.

Further, W is an **orthonormal set** if

$$\langle \mathbf{f}_i, \mathbf{f}_j\rangle = \delta_{i,j}$$

where

$$\delta_{i,j} = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$$

is the **Kronecker delta**.

Theorem 1.9.3 (Gram-Schmidt). Suppose that V is a vector space and $W = \{\mathbf{f}_i\}_{i=1}^n \subset V$ are linearly independent. Then, given an inner product $\langle \cdot, \cdot\rangle$ on V, there exists vectors $\{\mathbf{g}_i\}_{i=1}^n$ which are orthonormal and a linear combination of the vectors in W.

Proof 1.9.4. We prove this theorem by an algorithm for constructing $\mathbf{g}_i$ inductively. Note that since W contains linearly indepdent vectors, by definition they are all non-zero.

- Set $\mathbf{g}_1 = \dfrac{\mathbf{f}_1}{\langle \mathbf{f}_1, \mathbf{f}_1 \rangle}$.
- For $2 \leq i \leq n$, define $\mathbf{x}_i = \mathbf{f}_i - \langle \mathbf{f}_i, \mathbf{g}_{i-1} \rangle \mathbf{g}_{i-1}$.
- Finally, set $\mathbf{g}_i = \dfrac{\mathbf{x}_i}{\langle \mathbf{x}_i, \mathbf{x}_i \rangle}$.

It is easy to show that this process generates a set of vectors with the desired properties. $\square$

Chapter 2

Differential Equations

Differential equations form a fundamental part of any applied mathematics as they are so adept at describing real-world phenomena. Therefore it is vital that you have a good understanding of basic ordinary and partial differential equations.

The goal of this chapter is to equip you with the knowledge and techniques required to solve various simple ordinary and partial differential equations. Whilst there is a large amount of material outlining quantitative methods for analysis of differential equations, we will **not** cover that here.

Firstly we will introduce a variety of definitions and theorems from the study of multi-variate calculus. Despite the fact that most of the differential equations you will work with are partial differential equations (PDEs) we will then continue with an overview of ODEs. The main reason for this is the fact that an important tool of PDEs is to reduce the problem to one or more system of ODEs. Finally, we will solve the one-dimensional wave equation.

2.1 Differentiation

The definitions and theorems stated in this chapter are taken from a summary of a second year course in differentiation by J. Rawnsley. They form the basic knowledge in the theory of multivariate functions. The main aim of this section is to form the knowledge you will need in order to find local extrema of functions both with and without constraints.

2.1.1 Norms

Norms are common to a lot of areas of mathematics and have a close relationship with the inner products we defined in the last chapter. They define the concept of magnitude of a vector.

Definition 2.1.1 (Norm). Let V be a vector space over a field $\mathbb{K}$ containing complex numbers. A **norm** is a function $f : V \to \mathbb{R}$ which, for $\mathbf{u}, \mathbf{v} \in V$ and $a \in \mathbb{K}$ satisfies

(i) $p(a\mathbf{v}) = |a|\, p(\mathbf{v})$;

(ii) $p(\mathbf{u} + \mathbf{v}) \leq p(\mathbf{u}) + p(\mathbf{v})$;

(iii) $p(\mathbf{u}) = 0$ if and only if $\mathbf{u} = \mathbf{0}$.

We typically write $\|\mathbf{u}\| = p(\mathbf{u})$.

Example 2.1.2. There are many important norms you should be aware of; however three of the most common norms are defined below.

(i) Of course, the standard example of a norm is the Euclidean norm. Let $V = \mathbb{R}^n$ and $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Then define

$$\|\mathbf{x}\| = \sqrt{\sum_{k=1}^{n} x_k^2}.$$

(ii) In fact, the Euclidean norm is a special case of a more generalised set of norms. Let $p \geq 1$, and define the norm

$$\|\mathbf{x}\|_p = \left(\sum_{k=1}^{n} |x_k|^p \right)^{1/p}$$

This is called the ℓ^p **norm**. We can also define the 'limiting' norm

$$\|\mathbf{x}\|_\infty = \max_{1 \leq k \leq n} |x_k|$$

which is called the ℓ^∞ **norm**.

(iii) We can also define norms on function spaces. For a set $\Omega \subset \mathbb{R}$, let

$$V(\Omega) = \{f : \Omega \to \mathbb{R} \mid f \text{ is integrable}\}.$$

Then for any $p \geq 1$ we may define the norm

$$\|f\|_p = \left(\int_\Omega |f(x)|^p \, \mathrm{d}x \right)^{1/p}$$

is called the L^p **norm**, and is the continuous analogue of the discrete version we saw above.

Definition 2.1.3 (Open ball). Given a norm $\|\cdot\|$ on a vector space V we define an **open ball** of radius ε around a point $\mathbf{x}_0 \in V$ as the set

$$B_\varepsilon(\mathbf{x}_0) = \{x \in V \mid \|\mathbf{x} - \mathbf{x}_0\| < \varepsilon\}.$$

2.1.2 Limit and continuity

Definition 2.1.4 (Limit). Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a map and $\mathbf{x}_0 \in \mathbf{x}^n$. $A \in \mathbb{R}^m$ is called a **limit** of f at $\mathbf{x}_0$ if for any $\varepsilon > 0$ there exists $\delta > 0$ such that for any $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x} - \mathbf{x}_0\| < \delta$ one has $\|f(\mathbf{x}) - A\| < \varepsilon$. In this case one writes

$$\lim_{\mathbf{x} \to \mathbf{x}_0} f(\mathbf{x}) = A.$$

Definition 2.1.5 (Continuous). f is called **continuous** at $\mathbf{x}_0$ if $\lim_{\mathbf{x} \to \mathbf{x}_0} f(\mathbf{x}) = f(\mathbf{x}_0)$. f is called **continuous** if it is continuous at every point of its domain of definition.

The typical analogy of a continous function is one which, if you were to draw it, would not require you to lift your pen off of the paper.

2.1.3 Differentiability

Definition 2.1.6 (Differentiable). $f : \mathbb{R}^n \to \mathbb{R}^m$ is **differentiable** at $\mathbf{a} \in \mathbb{R}^n$ if there is a linear map $L : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$\lim_{\mathbf{x}\to\mathbf{a}} \frac{\|f(\mathbf{x}) - f(\mathbf{a}) - L(\mathbf{x}-\mathbf{a})\|}{\|\mathbf{x}-\mathbf{a}\|} = 0.$$

L is unique and is called the derivative $d_{\mathbf{a}}f$ of f at $\mathbf{a}$. Differentiable maps are continuous.

Theorem 2.1.7. The following statements define several important properties of differentiable functions.

(i) **Chain rule**: If $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\mathbf{a}$ and $g : \mathbb{R}^m \to \mathbb{R}^p$ is differentiable at $\mathbf{b} = f(\mathbf{a})$ then the composition $f \circ g$ is differentiable at $\mathbf{a}$ and

$$d_{\mathbf{a}}(g \circ f) = d_{\mathbf{b}}g \circ d_{\mathbf{a}}f.$$

(ii) If L is a linear map, then L is differentiable and $d_{\mathbf{a}}L = L$.

(iii) **Leibniz rule**: If $f, g : \mathbb{R}^n \to \mathbb{R}$ (note image is $\mathbb{R}$) are differentiable at $\mathbf{a}$ then so is the product fg and

$$d_{\mathbf{a}}(fg) = f(\mathbf{a})d_{\mathbf{a}}g + g(\mathbf{a})d_{\mathbf{a}}f.$$

(iv) **Quotient Rule**: If $f, g : \mathbb{R}^n \to \mathbb{R}$ are differentiable at $\mathbf{a}$ and $g(\mathbf{a}) \neq 0$ then so is the quotient f/g and

$$d_{\mathbf{a}}(f/g) = \frac{g(\mathbf{a})d_{\mathbf{a}}f - f(\mathbf{a})d_{\mathbf{a}}g}{g(\mathbf{a})^2}.$$

(v) $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\mathbf{a}$ if and only if all the components f_i, $i = 1, \ldots, m$ are differentiable at $\mathbf{a}$.

2.1.4 Partial derivatives

Definition 2.1.8 (Partial derivative). Suppose that $f : \mathbb{R}^n \to \mathbb{R}$. Then the i-th **partial derivative** of f is defined as

$$\frac{\partial f}{\partial x_i}(a_1, \ldots, a_n) = \lim_{t\to 0} \frac{f(a_1, \ldots, a_{i-1}, a_i + t, a_{i+1}, \ldots, a_n) - f(a_1, \ldots, a_n)}{t},$$

where the limit exists.

Lemma 2.1.9. If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable then all of its partial derivatives exist and

$$\frac{\partial f}{\partial x_i}(a_1, \ldots, a_n) = d_{\mathbf{a}}f(\mathbf{e}_i)$$

where $\mathbf{e}_1, \ldots, \mathbf{e}_n$ is the standard basis of $\mathbb{R}^n$. In addition, if the partial derivatives of f are continuous then f is differentiable.

Definition 2.1.10 (Jacobian). The **Jacobian** of $f : \mathbb{R}^n \to \mathbb{R}^m$ is the $m \times n$ matrix of partial derivatives

$$J_{\mathbf{a}} f = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{a}) & \dots & \frac{\partial f_1}{\partial x_n}(\mathbf{a}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{a}) & \dots & \frac{\partial f_m}{\partial x_n}(\mathbf{a}) \end{bmatrix}$$

and is the matrix of $d_{\mathbf{a}} f$ in the standard basis when f is differentiable at $\mathbf{a}$. The chain rule translates as matrix multiplication

$$J_{\mathbf{a}}(g \circ f) = J_{f(\mathbf{a})} g \, J_{\mathbf{a}} f.$$

In addition, the special case of $m = 1$ allows the Jacobian to be written as a vector in $\mathbb{R}^n$ called the **gradient**:

$$\nabla f(\mathbf{a}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{a}), \dots, \frac{\partial f}{\partial x_n}(\mathbf{a}) \right).$$

2.1.5 Higher derivatives

Definition 2.1.11 (Higher derivative). If $d_{\mathbf{x}} f$ exists for all $\mathbf{x} \in U$, we may view the derivative as a map $df : U \to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ and say f is twice differentiable at $\mathbf{a}$ if df is differentiable at $\mathbf{a}$. Set $d^2_{\mathbf{a}} = d_{\mathbf{a}}(df)$ and $d^2_{\mathbf{a}}(\mathbf{v}, \mathbf{w}) = (d^2_{\mathbf{a}}(\mathbf{v}))(\mathbf{w})$.

In the special case of functions $f : \mathbb{R}^n \to \mathbb{R}$, second order derivatives are denoted by

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \partial^2_{x_i x_j} f.$$

Definition 2.1.12 (Hessian). Let $f : \mathbb{R}^n \to \mathbb{R}$. Suppose that the partial derivatives of f have partial derivatives. Then the matrix of second partial derivatives is called the **Hessian** matrix $H_{\mathbf{a}} f = (\alpha_{ij})$, where

$$\alpha_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{a}).$$

Theorem 2.1.13. Let $f : \mathbb{R}^n \to \mathbb{R}$, and suppose that f has second-order partial derivatives which are continuous. Then

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{a}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{a}).$$

Corollary 2.1.14. Let f be as in the previous theorem. Then $H_{\mathbf{a}} f$ is symmetric.

This corollary can actually be quite a time-saver, since we can simply calculate a single second-order mixed derivative and see if it is continuous. If it is, then we only need to calculate the upper- or lower-triangular part of the Hessian.

Finally, we conclude this section with a brief application of higher derivatives.

Definition 2.1.15 (C^k class). Let f be as in the previous definition. If U is an open set in $\mathbb{R}^n$ and $f : U \to \mathbb{R}^m$ a map then f is of class C^k (or $f \in C^k(U, \mathbb{R}^m)$) if f has k-fold partial derivatives at each point of U and the k-th order partial derivatives are continuous.

Theorem 2.1.16 (Taylor's theorem). If f is in C^{k+1} on U, $\mathbf{a} \in U$ and $B_\varepsilon(\mathbf{a}) \subset U$ then f has a Taylor expansion for $\|\mathbf{h}\| < \varepsilon$:

$$f(\mathbf{a}+\mathbf{h}) = f(\mathbf{a}) + \sum_i h_i \frac{\partial f}{\partial x_i}(\mathbf{a}) + \cdots + \frac{1}{k!} \sum_{i_1 \ldots i_k} h_{i_1} \ldots h_{i_k} \frac{\partial^k f}{\partial x_{i_1} \ldots \partial x_{i_k}}(\mathbf{a}) + R_k(\mathbf{a}, \mathbf{h})$$

where $\|R_k(\mathbf{a}, \mathbf{h})\| / \|\mathbf{h}\|^k \to 0$ as $\mathbf{h} \to 0$. $R_k(\mathbf{a}, \mathbf{h})$ is called the **remainder function**.

2.1.6 Maxima and minima

Having defined the concept of a derivative in multiple dimensions, we now turn our attention to utilising these techniques to locate minima and maxima of some differentiable functions.

Definition 2.1.17 (Local extrema). A point $\mathbf{a}$ is a local **maximum** (or **minimum**) of $f : U \to \mathbb{R}$ if there exists $\varepsilon > 0$ such that for all $x \in B_\varepsilon(\mathbf{a})$, $f(\mathbf{x}) \leq f(\mathbf{a})$ (or $f(\mathbf{x}) \geq f(\mathbf{a})$).

Definition 2.1.18 (Critical point). A point $\mathbf{a}$ is a **critical point** of f if $d_\mathbf{a} f = 0$.

Theorem 2.1.19. Every local maximum or minimum is a critical point.

Remark 2.1.20. It is extremely important to remember that the converse of the previous theorem is **not** true. For example, the one-dimensional function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^3$ has a critical point at $x = 0$ which is neither a local maximum or minimum. However, we can use some properties of the derivative in order to attempt to classify the critical points that we find.

Definition 2.1.21 (Definite matrix). Let A be an $n \times n$ matrix with eigenvalues λ_k. A is:

- **Positive definite** if $\lambda_i > 0$ for all $1 \leq i \leq n$;
- **Negative definite** if $\lambda_i < 0$ for all $1 \leq i \leq n$;

Theorem 2.1.22. Let $f : \mathbb{R}^n \to \mathbb{R}$ and suppose $\mathbf{a} \in \mathbb{R}^n$ is a critical point of f with Hessian $H_\mathbf{a} f$. Then:

- if $H_\mathbf{a} f$ is positive definite then $\mathbf{a}$ is a local minimum;
- if $H_\mathbf{a} f$ is negative definite then $\mathbf{a}$ is a local maximum;
- if the eigenvalues of $H_\mathbf{a} f$ are all non-zero then $\mathbf{a}$ is a **saddle point**;
- otherwise, classification is not possible.

Example 2.1.23. We will now study some simple examples demonstrating the approach you should take when attempting to find local extrema. The approach is quite straightforward. We first find all critical points of f, and then apply theorem 2.1.22 to determine what the critical points are.

Whilst in general it will not always be the case that the points can be classified, in this course we will not consider functions where this is the case.

(i) First consider a map $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(\mathbf{x}) = f(x, y) = x^2 + y^2$. This is a nice initial example, since it defines the surface obtained by rotating the function $g(x) = x^2$ around the z-axis. It is therefore obvious that there should be a local minimum at $(x, y) = (0, 0)$ – in fact it is easy to prove that this will be a global minimum.

We first calculate the derivative. In this case, we do not need to calculate $d_{\mathbf{x}}f$ explicitly since the partial derivatives are continuous and so the gradient will suffice. This gives

$$\nabla f(\mathbf{x}) = (\partial_x f(\mathbf{x}), \partial_y f(\mathbf{x})) = (2x, 2y).$$

Setting this equal to zero yields only the trivial solution $(x, y) = (0, 0)$. The Hessian of f for any point $\mathbf{x} \in \mathbb{R}^2$ is given by

$$H_{\mathbf{x}}f = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

yielding a single eigenvalue $\lambda = 2$ of multiplicity 2. Since this is positive, $(0, 0)$ is therefore a local minimum of f.

(ii) Now consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ with $f(x, y) = 3xy - x^3 - y^3$. Calculating the partial derivatives, we see that

$$\nabla f(\mathbf{x}) = (3y - 3x^2, 3x - 3y^2).$$

Setting this equal to zero, we see that $y = x^2$ and so $x - x^4 = x(1 - x^3) = 0$. Hence the critical points are $(x, y) = (0, 0)$ and $(1, 1)$. Finally, the Hessian is given by

$$H_{\mathbf{x}}f = \begin{bmatrix} -6x & 3 \\ 3 & -6y \end{bmatrix} \Rightarrow H_{(0,0)}f = \begin{bmatrix} 0 & 3 \\ 3 & 0 \end{bmatrix}, \; H_{(1,1)}f = \begin{bmatrix} -6 & 3 \\ 3 & -6 \end{bmatrix}.$$

For the first matrix we therefore obtain eigenvalues $\lambda_{1,2} = \pm 3$ and hence $(0, 0)$ is a saddle point. In the second case, we obtain the characteristic equation $(6 + z)^2 - 9 = 0$, giving solutions $6 + z = \pm 3 \Rightarrow \lambda_{1,2} = -6 \pm 3 < 0$ and so $(1, 1)$ is a local maximum.

2.1.7 Constrained maxima and minima

Theorem 2.1.24. Let $U \subset \mathbb{R}^n$ and assume we have two functions $f, g : U \to \mathbb{R}$ that are C^1 (continuous first partial derivatives). Suppose that f has a local maximum (or minimum) $f(\mathbf{x}_0)$ at $\mathbf{x}_0 \in U$ when $\mathbf{x}$ is subject to the constraint $g(\mathbf{x}) = 0$. If $\nabla g(\mathbf{x}_0) \neq 0$ then there exists a real number λ, the **Lagrange multiplier**, such that

$$\nabla f(\mathbf{x}_0) = \lambda \nabla g(\mathbf{x}_0).$$

Example 2.1.25. We consider two examples to demonstrate how the previous theorem can be applied to find constrained maxima and minima.

(i) First, we will find the extrema of the function $f : \mathbb{R}^2 \to \mathbb{R}$ with $f(x, y) = xy$ assuming that (x, y) is restricted to the ellipse $4x^2 + y^2 = 4$.

In this example the constraint is $g(x, y) = 4x^2 + y^2 - 4 = 0$. Setting $\nabla f(x, y) = \lambda \nabla g(x, y)$, we obtain

$$(y, x) = \lambda(8x, 2y)$$

which when combined with the original condition give us the equations

$$\begin{cases} y = 8\lambda x, \\ x = 2\lambda y, \\ 0 = 4x^2 + y^2 - 4. \end{cases}$$

Now, for example, we can substitute to obtain

$$x = 2\lambda y = 2\lambda(8x\lambda) = 16x\lambda^2,$$

and so $x(1-16\lambda^2)=0$. Hence $x=0$ or $\lambda=\pm\frac{1}{4}$.

If $x=0$ then the constraint gives us that $y=\pm 2$. Since $y\neq 0$, the second equation implies $\lambda=0$; however if this is true then the first equation gives $y=0$ which is a contradiction, therefore $(0,\pm 2)$ cannot be a critical point.

If $\lambda=\pm\frac{1}{4}$, then $y=8x\lambda=\pm 2x$. Again using the constraint, in either case we see that $8x^2=4$ and so $x=\pm\frac{1}{\sqrt{2}}$ with corresponding y values of $y=\pm\sqrt{2}$. This then yields four extrema,

$$(x,y)=(0,\pm 2),(\tfrac{1}{\sqrt{2}},\pm\sqrt{2}),(-\tfrac{1}{\sqrt{2}},\pm\sqrt{2}).$$

We therefore find that

$$f(0,\pm 2)=0,\quad f(\tfrac{1}{\sqrt{2}},\pm\sqrt{2})=\pm 1,\quad f(-\tfrac{1}{\sqrt{2}},\pm\sqrt{2})=\mp 1.$$

It is clear then that $(\frac{1}{\sqrt{2}},\sqrt{2})$ and $(-\frac{1}{\sqrt{2}},-\sqrt{2})$ maximise f under this constraint, whereas $(-\frac{1}{\sqrt{2}},\sqrt{2})$ and $(\frac{1}{\sqrt{2}},-\sqrt{2})$ minimise f.

(ii) For a more applied example, let us use Lagrange multipliers to calculate the minimum distance between the point $(x,y)=(0,1)$ and the hyperbola $x^2-2y^2=1$.

The square of the distance from the point $(0,1)$ to any point (x,y) in the Cartesian plane is given by the function $f(x,y)=x^2+(y-1)^2$. We minimize this with respect to the function $g(x,y)=x^2-2y^2$ under the condition $g(x,y)=1$. Proceeding in the standard fashion then, we solve the equations

$$\begin{cases}\nabla f(x,y)=\lambda\nabla g(x,y)\\ g(x,y)=1\end{cases}\Rightarrow\begin{cases}2x=2\lambda x\\ 2y-2=-4\lambda y\\ x^2-2y^2=1\end{cases}$$

From the first equation we see that $\lambda=1$ since by equation 3, $x\neq 0$. Using this in the second equation this gives $6y=2\Rightarrow y=\frac{1}{3}$. Substituting this into the third equation gives $x^2=\frac{11}{9}$. So $f(x,y)=\frac{11}{9}+\frac{4}{9}=\frac{15}{9}=\frac{5}{3}$. Hence, the minimum distance between $(0,1)$ and the hyperbola is $\sqrt{\frac{5}{3}}$.

2.2 Ordinary differential equations

The equation

$$\mathcal{F}(t,y(t),y'(t),\ldots,y^{(n)}(t))=0 \tag{2.1}$$

where $y^{(k)}=\frac{d^k y}{dt^k}$ is an **ordinary differential equation** (**ODE**) of the n-th order. The unknown real valued function y depends on a single independent variable and only ordinary derivatives appear in the differential equation.

Definition 2.2.1. The ordinary differential equation (2.1) is said to be

(i) **linear** if it can be written in the form

$$a_n(t)\frac{\mathrm{d}^n y}{\mathrm{d}t^n}+a_{n-1}(t)\frac{\mathrm{d}^{n-1}y}{\mathrm{d}t^{n-1}}+\cdots+a_1(t)\frac{\mathrm{d}y}{\mathrm{d}t}+a_0(t)y=f(t)$$

(ii) **homogeneous** if it is linear and $f(t)=0$;

(iii) **inhomogeneous** if it is linear and $f(t) \neq 0$;

(iv) **autonomous** if it may be written in the form $\mathcal{F}(y(t), y'(t), \ldots, y^{(n)}(t)) = 0$.

As stated at the beginning of the chapter, we will not pause to reflect on the quantitative aspects of these differential equations, but simply outline a number of methods which can be used to solve a variety of the most commonly seen equations.

One important point is that the solutions of an ODE will only be unique up to a constant unless additional conditions are imposed. Usually, this is something in the form

$$y(t_0) = y_0,\ t_0, y_0 \in \mathbb{R}$$

in which case it is known as an **initial condition**. A differential equation and an initial condition together form a **initial value problem**. Usually the independent variable will be time, t when the initial value is at time $t = 0$.

Also note that the value of the function at the initial time is sometimes insufficient to uniquely determine a solution; we may also require knowledge of the value of the gradient(s) of a function as well for higher order equations.

2.2.1 First order equations

2.2.1.1 Separable equations

Consider the equation

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

Then, integrating both sides of the equation yields

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

and so we get an implicit equation for y. Depending on the equation then, we may be able to obtain an explicit solution $y(x)$.

Example 2.2.2. Here are two simple examples which use the above technique.

(i) We first consider the equation

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -9x^2y^2.$$

Re-arranging this in the form above, we find that

$$\int \frac{\mathrm{d}y}{y^2} = -9\int x^2\,\mathrm{d}x \Rightarrow -\frac{1}{y} = -3x^3 + c$$

where $c \in \mathbb{R}$ is a constant of integration. Therefore we obtain the solution

$$y = -\frac{1}{3x^3 + c}$$

(ii) Let us solve the equation

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\sin x}{y\cos y}$$

Rewriting this in the form above, we see that

$$\int y\cos y\,\mathrm{d}y = \int \sin x\,\mathrm{d}x \Rightarrow y\sin y + \cos y = -\cos x + c.$$

In this case, we are unable to invert the implicit equation, and so this is as simple a solution as we can obtain.

2.2.1.2 Integrating factors

Now consider the linear first order differential equation,

$$y' + p(x)y = g(x). \tag{2.2}$$

We cannot use the above technique since this equation is not separable. However, through the use of a clever multiplication, we can solve this equation. First, we define the function

$$\mu(x) = \exp\left(\int p(x)\,\mathrm{d}x\right) \Rightarrow \mu'(x) = p(x)\mu(x).$$

$\mu(x)$ is called the **integrating factor**. Multiplying equation (2.2) by $\mu(x)$ yields

$$\frac{\mathrm{d}}{\mathrm{d}x}[y(x)\mu(x)] = \mu(x)g(x) \Rightarrow y = \frac{1}{\mu(x)}\int \mu(x)g(x)\,\mathrm{d}x + c,$$

Example 2.2.3. Consider the equation

$$y' = e^x - 2y.$$

Using the integrating factor

$$\mu(x) = \exp\left(\int 2\,\mathrm{d}x\right) = e^{2x},$$

we get the family of solutions

$$y = \frac{1}{3}e^x + ce^{-2x}, \quad c \in \mathbb{R}.$$

2.2.2 Second-order equations

Consider the differential equation given by

$$A\frac{\mathrm{d}^2y}{\mathrm{d}t^2} + B\frac{\mathrm{d}y}{\mathrm{d}t} + Cy = f(t). \tag{2.3}$$

This is a general second-order linear inhomogeneous differential equation with constant coefficients, and is commonly found in many simple physics problems.

First let us tackle the inhomogeneous problem; that is, the case where $f(t) = 0$. The first step is to assume that y can be written as $y(t) = e^{\lambda t}$. Substituting this into (2.3) we see that

$$A\lambda^2 y(t) + B\lambda y(t) + Cy(t) = 0. \tag{2.4}$$

Since $y(t) \neq 0$ for any t, we may divide this equation by $y(t)$ to get

$$A\lambda^2 + B\lambda + C = 0. \tag{2.5}$$

This quadratic equation in λ is called the **auxiliary equation**. Depending on the values of the constants, it is clear that there are three cases that we must consider.

1. **Distinct real roots**. If $B^2 - 4AC > 0$, then we obtain two distinct real roots λ_1 and λ_2 of (2.5). This indicates that the functions $y_1(t) = c_1 e^{\lambda_1 t}$ and $y_2(t) = c_2 e^{\lambda_2 t}$ are both solutions to (2.4), where c_1 and c_2 are real constants.

   Indeed, since the equation is linear, (2.4) will have a **general solution**

   $$y(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}, \quad c_1, c_2 \in \mathbb{R}.$$

   This indicates that in order to determine a unique solution we will need additional restrictions on the equation (such as an initial condition).

2. **Repeated real root**. If $B^2 - 4AC = 0$ then we obtain one real root λ of (2.5). This would indicate that the only solution is one of the form $c_1 e^{\lambda t}$. Whilst this is indeed a solution, it is not the *only* solution since the dimension of the vector space of solutions to the equation (2.4) should be two. So we need to find another linearly independent function which is also a solution.

   The procedure is actually relatively simple. Let us assume that $y_2(t) = v(t)y_1(t)$ is a solution of (2.4), where $v(t)$ is a function that needs to be determied. By substituting y_2 into (2.4) it is easy to prove that $v''(t) = 0$. Combined with the fact that y_2 must be linearly independent to y_1, we may take $v(t) = c_2 t$ where c_2 is an arbitrary constant. Hence, the general solution is given by

   $$y(t) = c_1 e^{\lambda t} + c_2 t e^{\lambda t}$$

3. **Distinct complex roots**. If $B^2 - 4AC < 0$ then we obtain a pair of complex roots $\lambda_{1,2} = a \pm ib$ where $i = \sqrt{-1}$ and $a, b \in \mathbb{R}$. Now notice that

   $$e^{\lambda_{1,2} t} = e^{(a \pm ib)t} = e^{at} e^{\pm ibt} = e^{at}(\cos(\pm bt) + i \sin(\pm bt)) = e^{at}(\cos bt \pm i \sin bt).$$

   A little manipulation will yield a solution not involving i:

   $$y(t) = e^{at} \left[c_1 \cos bt + c_2 \sin bt \right], \quad c_1, c_2 \in \mathbb{R}.$$

Now consider the general inhomogeneous case where $f(t) \neq 0$. This is, in fact, remarkably straight-forward for certain cases of $f(t)$. Let us assume that the homogeneous problem has general solution $y_h(t)$. If we can 'guess' a solution $y_p(t)$, called the **particular solution**, then by linearity of the differential operator, $y(t) = y_h(t) + y_p(t)$ is a solution of (2.3).

Of course, the 'guess' you make needs to depend on $f(t)$. You can use the following table as a guide to the functions you might choose.

$f(t)$	$y_p(t)$
at^n, $n \geq 0$	$b_n t^n + b_{n-1} t^{n-1} + \cdots + b_0$
ae^{kt}	be^{kt}
$a\cos(kt)$ or $a\sin(kt)$	$b_1 \cos(kt) + b_2 \sin(kt)$
$ae^{k_1 t}\cos(k_2 t)$ or $ae^{k_1}\sin(k_2 t)$	$e^{k_1 t}[b_1 \cos(k_2 t) + b_2 \sin(k_2 t)]$

Be aware though that y_p must be linearly independent from y_h. For example, if $y_h(t) = e^t$ and $f(t) = e^t$ then you should try $y_p(t) = cte^t$ where $c \in \mathbb{R}$ and re-arrange to obtain c.

Example 2.2.4. To conclude this section on second-order equations, we solve three equations showing each of the individual cases above.

(i) Let us solve the inhomogeneous problem

$$\frac{\mathrm{d}^2y}{\mathrm{d}t^2} + \frac{\mathrm{d}y}{\mathrm{d}t} - 6y = 4e^{2t} \tag{2.6}$$

We first solve the homogeneous problem $y_h''(t)+y_h'(t)-6y_h(t) = 0$. Substituting $y_h(t) = e^{\lambda t}$ gives the auxiliary equation

$$\lambda^2 + \lambda - 6 = 0 \Rightarrow (\lambda + 3)(\lambda - 2) = 0 \Rightarrow \lambda = -3, 2.$$

Therefore the general solution is given by $y_h(t) = Ae^{-3t} + Be^{2t}$ for some $A, B \in \mathbb{R}$.

Now, using our table above, we search for a particular solution. The table would indicate a choice of something along the lines of $y_p(t) = Ce^{2t}$; however, this would not be linearly independent with the function y_h. Instead then, we choose $y_p(t) = Cte^{2t}$. Substituting into (2.6) yields

$$y_p'(t) = Ce^{2t} + 2Cte^{2t},$$
$$y_p''(t) = 4Ce^{2t} + 4Cte^{2t}.$$

Equating coefficients of e^{2t}, we see $4C + C = 4 \Rightarrow C = \frac{4}{5}$. So the general solution is $y(t) = Ae^{-3t} + Be^{2t} + \frac{4}{5}te^{2t}$.

(ii) Consider the homogeneous problem

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + 6\frac{\mathrm{d}x}{\mathrm{d}t} + 9x = 0.$$

subject to the initial condition that $x(0) = 2$ and $\dot{x}(1) = 2e^{-3}$. This ODE has auxiliary equation $\lambda^2 + 6\lambda + 9 = 0$ so that $(\lambda + 3)^2 = 0$ and hence $\lambda = -3$. In this case then, we will have a general solution

$$x(t) = Ae^{-3t} + Bte^{-3t}.$$

Since $x(0) = 2$, substituting this into the solution immediately yields $A = 2$. The solution has derivative $\dot{x}(t) = -6e^{-3t} + Be^{-3t} - 3Bte^{-3t}$. Substituting the second boundary condition and dividing by e^{-3} implies that $2 = -6 - 2B$, and hence $B = -4$. Hence,

$$x(t) = 2e^{-3t} - 4te^{-3t}.$$

(iii) Consider the inhomogeneous problem

$$\frac{\mathrm{d}^2y}{\mathrm{d}t^2} + 6\frac{\mathrm{d}y}{\mathrm{d}t} + 13y = 169t.$$

Let us first consider the homogeneous problem $y_h(t)$. This has auxiliary equation $\lambda^2 + 6\lambda + 13 = 0$, so $\lambda = \frac{1}{2}(-6 \pm \sqrt{-16}) = -3 \pm 2i$. Hence, the general solution is

$$y_h(t) = e^{-3t}[A\cos(2t) + B\sin(2t)].$$

Guessing a particular solution $y_p(t) = Ct + D$ for $C, D \in \mathbb{R}$, we see that

$$6C + 13(Ct + D) = 169t.$$

Equating co-efficients of t on both sides, we see that $13C = 169$, so $C = 13$. Similarly, equating co-efficients of t^0, $6C + 13D = 0$ so $D = -6$. Hence the general solution is

$$y(t) = y_h(t) + y_p(t) = e^{-3t}[A\cos(2t) + B\sin(2t)] + 2t - 6.$$

2.2.3 Systems of ODEs

A common occurance in many applied fields is the coupling of two or more quantities by differential equations. Such systems can be quite complex to solve, and in general are in the form

$$\begin{cases} x_1 = F_1(t, x_1, \ldots, x_n), \\ \quad \vdots \\ x_n = F_n(t, x_1, \ldots, x_n). \end{cases}$$

Obviously, in general this has a lot of complexity and is difficult to solve. Here then, we will look at a much simpler case. Suppose that A is an $n \times n$ matrix over $\mathbb{R}$. Let use consider the system of linear ODEs given by

$$\dot{\mathbf{x}} = A\mathbf{x} \tag{2.7}$$

If $n = 1$ so that the system is one-dimensional, this simply boils down to the equation

$$\dot{x} = \lambda x \Rightarrow x(t) = ce^{\lambda t},\ c \in \mathbb{R}$$

In the multi-dimensional case then, it seems like we should be able to take a very similar approach. The key to this is to extend the exponential function to a function on the space of matrices.

Definition 2.2.5 (Matrix exponential). Let A be an $n \times n$ matrix over $\mathbb{C}$. Then the **matrix exponential** $e : \mathcal{L}(\mathbb{C}^n, \mathbb{C}^n) \to \mathcal{L}(\mathbb{C}^n, \mathbb{C}^n)$ is defined by the infinite summation

$$e^A = \sum_{n=0}^{\infty} \frac{A^n}{n!}.$$

Theorem 2.2.6. The following statements are true:

(i) The matrix exponential is defined for any $n \times n$ matrix A.

(ii) If $t \in \mathbb{R}$ then $\dfrac{\mathrm{d}}{\mathrm{d}t}(e^{At}) = A \cdot e^{At}$.

(iii) Given matrices P and D such that $A = PDP^{-1}$, $e^A = Pe^D P^{-1}$.

(iv) If $D = \operatorname{diag}\{d_1, \ldots, d_n\}$ is a diagonal matrix then $e^D = \operatorname{diag}\{e^{d_1}, \ldots, e^{d_n}\}$.

(v) If A, B are $n \times n$ matrices which commute (i.e. $AB = BA$), then $e^{A+B} = e^A e^B$.

Proof 2.2.7.

(i) Suppose $M \in \mathbb{R}$ is such that $|A_{ij}| < M$ for all $i, j \leq n$. Then $|A_{ij}^2| < nM^2$, and hence, in general, $|A_{ij}^k| < n^k M^{k+1}$. Since

$$\sum_{k=0}^{\infty} \frac{n^k M^{k+1}}{k!} = M \sum_{k=0}^{\infty} \frac{(nM)^k}{k!} = Me^{nM} < \infty$$

the largest entry of the matrix e^A is finite, and hence e^A is well-defined.

(ii) Differentiating, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}(e^{At}) = \frac{\mathrm{d}}{\mathrm{d}t} \sum_{k=0}^{\infty} \frac{(At)^k}{k!} = \sum_{k=0}^{\infty} \frac{\mathrm{d}}{\mathrm{d}t} \left[\frac{(At)^k}{k!} \right] = \sum_{k=0}^{\infty} \frac{A^{k+1} t^k}{k!} = A \sum_{k=0}^{\infty} \frac{(At)^k}{k!} = Ae^{At}$$

as required.

(iii) First note that for any $k \geq 0$, we have that

$$\begin{aligned}(PDP^{-1})^k &= \underbrace{(PDP^{-1})(PDP^{-1})\cdots(PDP^{-1})}_{k \text{ times}} \\ &= P\underbrace{(DP^{-1}P)(DP^{-1}P)\cdots(DP^{-1}P)}_{k-1 \text{ times}}DP^{-1} \\ &= PD^{k-1}DP^{-1} \\ &= PD^kP^{-1},\end{aligned}$$

and so

$$e^A = e^{PDP^{-1}} = \sum_{k=0}^{n} \frac{PD^kP^{-1}}{k!} = P\left(\sum_{k=0}^{n}\frac{D^k}{k!}\right)P^{-1} = Pe^DP^{-1}.$$

(iv) For any $k \geq 0$ we have that $D^k = \text{diag}\{d_1^k, \ldots, d_n^k\}$. Hence,

$$e^D = \sum_{k=0}^{\infty}\frac{\text{diag}\{d_1^k,\ldots,d_n^k\}}{k!} = \text{diag}\left\{\sum_{k=0}^{\infty}\frac{d_1^k}{k!},\ldots,\sum_{k=0}^{\infty}\frac{d_n^k}{k!}\right\} = \text{diag}\{e^{d_1},\ldots,e^{d_n}\}.$$

(v) The proof of this property is beyond the scope of this course. □

By part (ii) of the above theorem then, we see that the general solution of the linear system of equation (2.7) is given by

$$\mathbf{x}(t) = e^{At}\mathbf{c}.$$

where $\mathbf{c} \in \mathbb{R}^n$ is a column vector of length n.

Calculating the matrix exponential can be somewhat of a complex task. However, parts (iii) and (iv) give us a clue on how we can do this if A is diagonalizable.

Example 2.2.8. Let us consider the case where $n = 2$ so that we have the system of equations

$$\begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\end{bmatrix} = \begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}\begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix}.$$

There are three cases to consider.

(i) Suppose that $A = (a_{ij})$ has real, distinct eigenvalues λ_1 and λ_2. Then by theorem 1.7.4, the matrix $P = [\mathbf{v}_1\ \mathbf{v}_2]$ of eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ diagonalises A. Applying part (iii) of theorem 2.2.6 then,

$$e^{At} = e^{P(Dt)P^{-1}} = Pe^{Dt}P^{-1} = [\mathbf{v}_1\ \mathbf{v}_2]\begin{bmatrix}e^{\lambda_1 t} & 0\\ 0 & e^{\lambda_2 t}\end{bmatrix}[\mathbf{v}_1\ \mathbf{v}_2]^{-1} = [e^{\lambda_1 t}\mathbf{v}_1\ e^{\lambda_2 t}\mathbf{v}_2][\mathbf{v}_1\ \mathbf{v}_2]^{-1}$$

where the third equality follows from part (iv) of theorem 2.2.6. Finally, if we define $(\tilde{c}_1, \tilde{c}_2)^\top = \tilde{\mathbf{c}} = P^{-1}\mathbf{c}$, we see that the solution is given by

$$\begin{aligned}\mathbf{x}(t) = e^{At}\mathbf{c} &= [e^{\lambda_1 t}\mathbf{v}_1\ e^{\lambda_2 t}\mathbf{v}_2]\tilde{\mathbf{c}} \\ &= \tilde{c}_1 e^{\lambda_1 t}\mathbf{v}_1 + \tilde{c}_2 e^{\lambda_2 t}\mathbf{v}_2.\end{aligned}$$

(ii) Suppose that A has complex eigenvalues $\lambda_{1,2} = a \pm ib$. This is actually not too different from case (i), since we are guaranteed that the eigenvectors will be distinct and so A will be diagonalisable. Indeed, we will obtain a solution

$$\mathbf{x}(t) = Ce^{(a+ib)t}[\mathbf{v}_1 + i\mathbf{v}_2] + \overline{C}e^{(a-ib)t}[\mathbf{v}_1 - i\mathbf{v}_2],$$

which, with a little effort, we can show is actually real and can be written in the form

$$\mathbf{x}(t) = e^{at}[(A\cos(bt) + B\sin(bt))\mathbf{v}_1 + (B\cos(bt) - A\sin(bt))\mathbf{v}_2].$$

(iii) Suppose that A has a real, repeated eigenvalue λ. Then we must construct the matrix P from the solitary eigenvector and a generalised eigenvector. The resultant matrix D can either be diagonal, in which case we can find the solution from part (i), or it is of the form

$$D = P^{-1}AP = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix} = \underbrace{\begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}}_{D_1} + \underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}}_{D_2}.$$

Notice that D_1 and D_2 commute, so using property (iv) of theorem 2.2.6 we get

$$e^{Dt} = e^{(D_1+D_2)t} = e^{D_1 t}e^{D_2 t} = \begin{bmatrix} e^{\lambda t} & 0 \\ 0 & e^{\lambda t} \end{bmatrix} \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}$$

since $D_2^2 = 0_{2\times 2}$. From this, we obtain the general solution

$$\mathbf{x}(t) = e^{At}\mathbf{c} = Pe^{Dt}\tilde{\mathbf{c}} = [\tilde{c}_1 e^{\lambda t} + \tilde{c}_2 t e^{\lambda t}]\mathbf{v}_1 + \tilde{c}_1 e^{\lambda t}\mathbf{v}_2.$$

Essentially then, one only needs to calculate the eigenvalues and eigenvectors of the system, and then use one of the above formulae to find the general solution of the system of equations.

In higher dimensions the general approach is the same; we find the Jordan canonical form J of A, and use some clever techniques to calculate the exponential of J. However, such discussion is beyond the scope of this course.

2.3 Partial differential equations

A **partial differential equation** of order m is an equation

$$\mathcal{F}\left(x_1, x_2, \ldots, x_n, \frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_n}, \frac{\partial^2 u}{\partial x_1 \partial x_2}, \ldots, \frac{\partial^m u}{\partial x_1^{h_1} \cdots \partial x_n^{h_n}}\right) = 0,$$

where $h_1 + \cdots + h_n = m$ and the unknown function $u : \Omega \to \mathbb{R}$ is a multivariate function on the domain $\Omega \subset \mathbb{R}^n$.

Partial differential equations (PDEs) are notoriously difficult to solve. Even simple linear equations can prove tricky at times to find an explicit solution for. To this end, we will consider a simple example using a technique that you can use to solve a variety of

One of the most famous partial differential equations in financial mathematics is the Black-Scholes equation. With a little effort and the use of elementary transformations this equation can be transformed to the following equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \tag{2.8}$$

This equation is known as the **heat** or **diffusion equation**, and like many other PDEs is obtained through the analysis of a physical problem. In this case, the physical setting itself is rather simple. Let us consider a one-dimensional rod of length L. Now consider the function $u : [0, L] \times [0, \infty) \to \mathbb{R}$ which describes the temperature of the rod at given point x in space and t in time. It can be shown that, under certain simplifications, u will evolve according to the above partial differential equation.

There is a wide range of literature devoted to the solution of the heat equation with analytical or numerical methods. Analogous to our discussion of ODEs we will study a method to solve the heat equation.

In a similar sense to an ordinary differential equation, we must provide additional information in order to obtain a unique solution. This is true for most types of PDEs, in fact, and so they are given special names:

- **Initial condition**: a function $f(x)$ which describes the initial distribution of the temperature at time $t = 0$;
- **Boundary conditions**: At each end of the domain $\Omega = [0, L]$, we must impose additional restrictions on u. That is, we must specify $u(x, t)$ for $x \in \partial\Omega = \{0, L\}$.

In this case, the initial condition is given by

$$u(x, 0) = f(x), \quad 0 \leq x \leq L, \tag{2.9}$$

where $f(x)$ is a given function describing the initial temperature distribution in the bar.

In addition, we will consider two types of boundary conditions.

(i) Assume that the ends of the bar are held at a fixed temperatures. It can be shown that this problem can be reduced to the following boundary condition,

$$u(0, t) = 0 = u(L, t) = 0 \quad \forall t > 0. \tag{2.10}$$

This is known as a **homogeneous Dirichlet** boundary condition – homogeneous since we insist that $u(x, t) = 0$.

(ii) We could also assume that the flux of heat through the boundary remains constant; i.e. heat is not allowed to escape through the boundary. This immediately implies that the gradient of u in the spatial direction must be zero; or, in other words,

$$\frac{\partial u}{\partial x}(t,0) = \frac{\partial u}{\partial x}(t,L) = 0 \quad \forall t > 0. \tag{2.11}$$

This is known as a homogenenous **Neumann** boundary condition.

The problem described by (2.8), (2.9) and one of (2.10) or (2.11) is an initial value problem in the time variable t. With respect to the space variable x, the problem is a boundary value problem.

2.3.1 Separation of variables

Regardless of the boundary conditions, we need an approach to get us started in the solution. One (of many) approaches, and possibly the simplest, is the method of separation of variables. In this, we assume that the function is separable so that

$$u(x,t) = X(x)T(t)$$

for $X : \Omega \to \mathbb{R}$ and $T : [0,\infty) \to \mathbb{R}$. In this case we have that

$$\frac{\partial u}{\partial t} = X(x)\frac{\mathrm{d}T}{\mathrm{d}t}(t), \quad \frac{\partial^2 u}{\partial x^2} = \frac{\mathrm{d}^2 X}{\mathrm{d}x^2}(x)T(t),$$

and substituting this into (2.8) gives

$$\frac{X''(x)}{X(x)} = \frac{\dot{T}(t)}{T(t)}.$$

Importantly, this implies that

$$\frac{X''(x)}{X(x)} = \frac{\dot{T}(t)}{T(t)} = \sigma.$$

for some $\sigma \in \mathbb{R}$. (Can you prove this?) Hence, we have reduced our PDE problem to the decoupled ODEs

$$X''(x) = \sigma X(x), \tag{2.12}$$

$$\dot{T}(t) = \sigma T(t). \tag{2.13}$$

We can now use standard ODE theory to progress from here. Let us first consider solutions to $X(x)$. There are three cases to investigate which depend on the value of σ.

(i) If $\sigma > 0$ then define $\lambda^2 = \sigma$. Then,

$$X(x) = A\cosh(\lambda x) + B\sinh(\lambda x), \quad A, B \in \mathbb{R},$$

where

$$\cosh(x) = \frac{e^x + e^{-x}}{2}, \quad \sinh(x) = \frac{e^x - e^{-x}}{2}$$

are the hyperbolic sine and cosine functions.

(ii) If $\sigma = 0$ then we obtain the linear solution

$$X(x) = Ax + B, \quad A, B \in \mathbb{R}.$$

(iii) If $\sigma < 0$ then define $\lambda^2 = -\sigma$. Then,

$$X(x) = A\cos(\lambda x) + B\sin(\lambda x), \quad A, B \in \mathbb{R}.$$

For $T(t)$, we again have two possible outcomes depending on the value of σ.

(i) If $\sigma \neq 0$ then define $\lambda^2 = \sigma$ if $\sigma > 0$; otherwise, set $\lambda^2 = -\sigma$. Then

$$T(t) = e^{\pm\lambda^2 t}.$$

(ii) If $\sigma = 0$ then we obtain the constant solution

$$T(t) = A, \quad A \in \mathbb{R}.$$

In order to progress further than this, we need to consider the two types of boundary conditions. First we discuss the slightly simpler Dirichlet boundary conditions, before moving on to Neumann boundary conditions.

2.3.2 Dirichlet boundary conditions

The idea here is to apply equation (2.10) to the solutions of the ODEs we found for $X(x)$. So, in turn, we must consider three separate cases.

(i) If $\sigma = 0$ then $X(x) = Ax + B$. Equation (2.10) implies that $X(0) = X(L) = 0$. We therefore see that $A = B = 0$, and so the only solution is the trivial one, $u(x,t) = 0\ \forall x, t$. Whilst this is a valid solution, it is rather boring (and obvious). So the standard convention at this point is to consider all other cases and ensure that it is not the *only* solution.

(ii) If $\sigma > 0$ then $X(x) = A\cosh(\lambda x) + B\sinh(\lambda x)$ for $\lambda^2 = \sigma$. Using the fact that $\cosh(0) = 1$ and $\sinh(0) = 0$, we see that $X(0) = 0 \Rightarrow A = 0$. Further, $X(L) = 0 \Rightarrow 0 = B\sinh(\lambda L)$. So either $\sinh(\lambda L) = 0$ in which case $\lambda = 0$, or $B = 0$. In either case, we again obtain the trivial solution.

(iii) If $\sigma < 0$ then $X(x) = A\cos(\lambda x) + B\sin(\lambda x)$ for $\lambda^2 = -\sigma$. Now, $X(0) = 0 \Rightarrow A = 0$. However, $X(L) = 0$ implies that $B\sin(\lambda L) = 0$. If $B = 0$ then this implies that $X(x) = 0$ and hence the only solution is the trivial one. Otherwise,

$$\sin(\lambda L) = 0 \Rightarrow \lambda = \frac{n\pi}{L} \text{ for } n \in \mathbb{N}.$$

So, we obtain infinitely many solutions of the form

$$X_n(x) = B_n \sin(\lambda_n x) \text{ where } \sigma_n = -\lambda_n^2 = \frac{n^2\pi^2}{L^2}, n \geq 0.$$

These functions and constants actually have special names that you have already seen. Consider that we can re-write equation (2.12) in the form

$$\mathbb{L}X = \sigma X$$

where $\mathbb{L} = \mathrm{d}/\mathrm{d}x^2$ is the second order differential operator, and X belongs to the vector space $C^\infty([0, L])$. This bears a remarkable – indeed, identical – resemblence to the definition of eigenvalues and eigenvectors we found for matrices earlier in these notes. This is further backed

up by the fact that, just as when we found eigenvectors for the matrices, the eigenfunctions here are not unique and are defined up to a constant; in this case, $B \in \mathbb{R}$.

In case (iii) above then, we have found the **eigenvalues** (σ_n) and their correspoding **eigenvectors** (X_n). However, since $C^\infty([0, L])$ is a space of functions (and as such is infinite dimensional), we usually refer to the X_n as **eigenfunctions**.

Armed with all of this knowledge, we see that all non-trivial solutions must be of the form

$$u_n(x,t) = X_n(x)T_n(t) = B_n e^{-\lambda_n^2 t}\sin(\lambda_n x), \quad n \geq 1.$$

We can, however, go further. Since the heat equation is linear, the **principle of superposition** states that we can combine all of these solutions so that

$$u(x,t) = \sum_{n=0}^{\infty} B_n e^{-\lambda_n^2 t}\sin(\lambda_n x)$$

is the general form of our solution. The question now is, given the initial condition (2.9), how do we determine the constants B_n? To answer this question, we must take a brief excursion into the realms of functional analysis.

2.3.3 Fourier series

Lemma 2.3.1. Let $L > 0$. Then the set of functions $\{\cos(\lambda_n x)|\ n \in \mathbb{N}\} \cup \{\sin(\lambda_n x)|\ n \in \mathbb{N}\}$ is orthogonal in the vector space $C^\infty([-L, L])$ under the inner product

$$\langle u, v\rangle = \int_{-L}^{L} u(x)v(x)\,\mathrm{d}x.$$

Proof 2.3.2. To prove this we use the following common trigonometric formulae:

$$\begin{aligned}
\cos(x)\cos(y) &= \tfrac{1}{2}[\cos(x-y) + \cos(x+y)] \\
\sin(x)\sin(y) &= \tfrac{1}{2}[\cos(x-y) - \cos(x+y)] \\
\cos(x)\cos(y) &= \tfrac{1}{2}[\sin(x+y) + \sin(x-y)]
\end{aligned}$$

which can be derived from the standard double angle formulae for sin and cos. This immediately yields the three results

$$\begin{aligned}
\int_{-L}^{L} \cos(\lambda_m x)\cos(\lambda_n x)\,\mathrm{d}x &= \begin{cases} 0 & m \neq n, \\ L & m = n; \end{cases} \\
\int_{-L}^{L} \cos(\lambda_m x)\sin(\lambda_n x)\,\mathrm{d}x &= 0 \text{ for all } m, n; \\
\int_{-L}^{L} \sin(\lambda_m x)\sin(\lambda_n x)\,\mathrm{d}x &= \begin{cases} 0 & m \neq n, \\ L & m = n, \end{cases}
\end{aligned}$$

and hence proves the lemma. □

In fact, in certain circumstances, the set above is not only orthogonal but provides a basis. This is a very important result which has many applications, and forms the theory of Fourier analysis.

Theorem 2.3.3 (Fourier series). Let $f \in L^2([-L, L])$, where

$$L^2([-L, L]) = \left\{ g : [-L, L] \to \mathbb{R} \;\middle|\; \int_{-L}^{L} |g(x)|^2 \, \mathrm{d}x < \infty \right\}.$$

Then there exists $a_n, b_n \in \mathbb{R}$ such that

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos(\lambda_n x) + b_n \sin(\lambda_n x)]. \tag{2.14}$$

In other words, the set of functions in the previous lemma constitute an orthogonal basis in the space $L^2([-L, L])$. This series expansion is called the **Fourier series** of f.

The proof of this theorem, whilst reasonably simple, is far beyond the scope of this course (see, for example, the fourth year course MA433 Fourier Analysis for a rigorous proof).

However, the method of calculating the co-efficients a_n and b_n uses the orthogonality relations we proved in lemma 2.3.1. Namely, if we multiply equation (2.14) by either $\sin(\lambda_m x)$ or $\cos(\lambda_m x)$ and integrate both sides, we obtain

$$b_n = \frac{1}{L} \int_{-L}^{L} f(x) \sin(\lambda_n x) \, \mathrm{d}x, \quad a_n = \frac{1}{L} \int_{-L}^{L} f(x) \cos(\lambda_n x) \, \mathrm{d}x$$

Example 2.3.4. Let us calculate the Fourier expansions of some common functions.

(i) Of course, any function of the form $\cos(\lambda_k x)$ or $\sin(\lambda_k x)$ has a trivial expansion. In the former case, we choose $a_n = \delta_{nk}$; in the latter, $b_n = \delta_{nk}$.

(ii) Consider $f(x) = x$. Then,

$$\begin{aligned} a_0 &= \frac{1}{L} \int_{-L}^{L} x \, \mathrm{d}x = 0, \\ a_n &= \frac{1}{L} \int_{-L}^{L} x \cos(\lambda_n x) \, \mathrm{d}x = \frac{1}{L\lambda_n} \left(\Big[x \cos(\lambda_n x) \Big]_{-L}^{L} - \int_{-L}^{L} \sin(\lambda_n x) \, \mathrm{d}x \right) = 0, \end{aligned}$$

and

$$\begin{aligned} b_n = \frac{1}{L} \int_{-L}^{L} x \sin(\lambda_n x) \, \mathrm{d}x &= -\frac{1}{L\lambda_n} \left(\Big[x \cos(\lambda_n x) \Big]_{-L}^{L} - \int_{-L}^{L} \cos(\lambda_n x) \, \mathrm{d}x \right) \\ &= -\frac{1}{L\lambda_n} \left(2L \cos(\lambda_n L) - \left[\frac{1}{\lambda_n} \sin(\lambda_n x) \right]_{-L}^{L} \right) \\ &= -\frac{1}{L\lambda_n} \left(2L(-1)^n + \frac{2}{\lambda_n} \sin(n\pi) \right) \\ &= 2\frac{(-1)^{n+1}}{\lambda_n}. \end{aligned}$$

Hence,

$$x = \sum_{n=1}^{\infty} 2\frac{(-1)^{n+1}}{\lambda_n} \sin(\lambda_n x).$$

2.3.4 Dirichlet boundary conditions (continued)

Recall that we obtained a general solution

$$u(x,t) = \sum_{n=0}^{\infty} b_n e^{-\lambda_n^2 t} \sin(\lambda_n x)$$

to the heat equation with homogeneous Dirichlet boundary conditions. Substituting $t = 0$ and using the initial condition from equation (2.9) we obtain

$$u(x,0) = f(x) = \sum_{n=1}^{\infty} b_n \sin(\lambda_n x)$$

This is extremely close to the expansion in terms of sines and cosines that we saw in equation (2.14). However, whilst the previous section *nearly* gives us the tools we need to determine b_n, we wish to find Fourier representations of functions in the domain $[0, L]$, not $[-L, L]$. Thankfully we can use some simple rules to get us out of trouble.

Definition 2.3.5 (Even and odd functions). Let $f : \mathbb{R} \to \mathbb{R}$. Then f is an **even** function if $f(x) = f(-x)$ and an **odd** function if $f(x) = -f(-x)$.

These functions obey have some very useful properties. Indeed, the terms 'even' and 'odd' result from the properties we obtain when multiplying even and odd functions, which bear a striking resemblance to those we obtain when multiplying even and odd natural numbers.

Lemma 2.3.6. Suppose $f, g : \mathbb{R} \to \mathbb{R}$. Then

- If f is odd and g is odd then fg is even.
- If f is even and g is odd then fg is odd.
- If f is even and g is even then fg is even.

Lemma 2.3.7. Let $L > 0$ and $f : [-L, L] \to \mathbb{R}$ be an integrable function. Then

- f is even $\Rightarrow \displaystyle\int_{-L}^{L} f(x)\,\mathrm{d}x = 2\int_{0}^{L} f(x)\,\mathrm{d}x.$
- f is odd $\Rightarrow \displaystyle\int_{-L}^{L} f(x)\,\mathrm{d}x = 0.$

Proof 2.3.8 (of Lemma 2.3.6 and 2.3.7). See example sheet 4. □

Corollary 2.3.9. Suppose that $f \in L^2([0, L])$. Then f has a Fourier series representation of the form

$$f(x) = \sum_{n=1}^{\infty} b_n \sin(\lambda_n x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(\lambda_n x)$$

where

$$a_n = \frac{2}{L}\int_0^L f(x)\cos(\lambda_n x)\,\mathrm{d}x, \quad b_n = \frac{2}{L}\int_0^L f(x)\sin(\lambda_n x)\,\mathrm{d}x.$$

This corollary is quite nice; it says that for functions which lie on the domain $[0, L]$, we can choose a representation consisting entirely of cosines (called a **cosine series**) or a sines (called a **sine series**). The proof is also quite straight-forward by our previous lemmas.

Proof 2.3.10. Let us define the functions $f_{\text{even}}, f_{\text{odd}} : [-L, L] \to \mathbb{R}$ by

$$f_{\text{even}}(x) = \begin{cases} f(x), & x \geq 0, \\ f(-x), & x < 0, \end{cases} \qquad f_{\text{odd}}(x) = \begin{cases} f(x), & x \geq 0, \\ -f(-x), & x < 0. \end{cases}$$

These are called the **even** and **odd extensions** of f respectively. It is immediately clear from the definition that f_{even} is even and f_{odd} is odd.

By theorem 2.3.3, both f_{odd} and f_{even} have a Fourier expansion. First, consider the expansion of f_{odd}. Since cos is even and f_{odd} is odd it follows that their product is odd by lemma 2.3.6, and hence

$$a_n = \frac{2}{L} \int_{-L}^{L} f(x) \cos(\lambda_n x) \, \mathrm{d}x = 0$$

by lemma 2.3.7. Also, the product of sin and f_{odd} is odd and by lemma 2.3.7 the desired equation for b_n is found.

By the same argument, the even extension of f has the desired a_n and b_n. □

At last, we have proved that the heat equation (2.8) on the domain $\Omega = [0, L]$ coupled with the boundary conditions (2.10) and initial condition (2.9) has a unique solution

$$u(x,t) = \sum_{n=1}^{\infty} b_n e^{-\lambda_n^2 t} \sin(\lambda_n x)$$

where

$$\lambda_n = \frac{n\pi}{L}, \quad b_n = \frac{2}{L} \int_0^L f(x) \sin(\lambda_n x) \, \mathrm{d}x.$$

Example 2.3.11. We will consider two specific examples of initial conditions.

1. Consider the domain with $L = 1$ and initial condition $f(x) = \sin(k\pi x)$ for $k \geq 1$. Then, by the orthogonality relationship of lemma 2.3.1, we see that $b_n = \delta_{kn}$, so that the solution is
$$u(x,t) = e^{-k^2\pi^2 t} \sin(k\pi x).$$

2. Now with $L = 1$ consider $f(x) = x(1-x)$. Then,
$$\begin{aligned} b_n = 2\int_0^1 x(1-x)\sin(n\pi x)\,\mathrm{d}x &= -\frac{2}{n\pi}\left[x(1-x)\sin(n\pi x)\right]_0^1 + \frac{2}{n\pi}\int_0^1 (1-2x)\cos(n\pi x)\,\mathrm{d}x \\ &= \frac{2}{n\pi}\left(\left[(1-2x)\frac{\sin(n\pi x)}{n\pi}\right]_0^1 + \frac{2}{n\pi}\int_0^1 \sin(n\pi x)\,\mathrm{d}x\right) \\ &= \frac{4}{n^3\pi^3}\left[\cos(n\pi x)\right]_0^1 \\ &= \frac{4}{n^3\pi^3}[\cos(n\pi) - 1] = \frac{4}{n^3\pi^3}[(-1)^n - 1]. \end{aligned}$$

   Hence we obtain the solution
$$u(x,t) = \frac{4}{\pi^3} \sum_{n=1}^{\infty} \frac{(-1)^n - 1}{n^3} e^{-n^2\pi^2 t} \sin(n\pi x).$$

2.3.5 Neumann boundary conditions

Now that we have found the groundwork in the previous sections to solve homogeneous Dirichlet boundary conditions, dealing with the Neumann boundary condition in equation (2.11) is much easier. Recall from section 2.3.1 that we have three cases to consider when finding the eigenfunctions and eigenvalues of equation (2.12).

First note that if $u(x,t) = X(x)T(t)$ then $\partial_x u(x,t) = X'(x)T(t)$, and so if $X'(0)T(t) = X'(L)T(t) = 0$ for every t then $X'(0) = X'(L) = 0$.

(i) If $\sigma = 0$ then $X(x) = Ax + B$. Since $X'(x) = A$ the boundary conditions imply $A = 0$. The other boundary condition $X'(L) = 0$ places no restriction on B, and therefore we have an eigenfunction $X(x) = 1$ with eigenvalue $\sigma = 0$.

(ii) If $\sigma > 0$ then $X(x) = A\cosh(\lambda x) + B\sinh(\lambda x)$ with $\lambda^2 = \sigma$, so that $X'(x) = A\lambda\sinh(\lambda x) + B\lambda\cosh(\lambda x)$. Just as in the Dirichlet case, this implies the only solution is the trivial one.

(iii) If $\sigma < 0$ then $X(x) = A\cos(\lambda x) + B\sin(\lambda x)$ and $X'(x) = -A\lambda\sin x + B\lambda\cos x$. Applying the boundary conditions we see that $B = 0$ and also that $A\lambda\sin(\lambda L) = 0$. If $A = 0$ or $\lambda = 0$ then we obtain the trivial solution. Otherwise,

$$\sin(\lambda L) = 0 \Rightarrow \lambda = \lambda_n = \frac{n\pi}{L}.$$

So we obtain the eigenfunction $X_n(x) = \cos(\lambda_n x)$ with associated eigenvalues $\sigma_n = -\lambda_n^2$.

From all of these cases then, we see that the solution must be of the form

$$u(x,t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n e^{-\lambda_n^2 t}\cos(\lambda_n x),$$

where by corollary 2.3.9

$$a_n = \frac{2}{L}\int_0^L f(x)\cos(\lambda_n x)\,\mathrm{d}x$$

and $f(x)$ is the initial condition defined in equation (2.9).

Example 2.3.12. Consider the intial condition $f(x) = x(1-x)$ where $L = 1$. Then we see that

$$a_0 = \frac{2}{1}\int_0^1 x(1-x)\,\mathrm{d}x = 2\left[\frac{x^2}{2} - \frac{x^3}{3}\right] = \frac{1}{3},$$

and

$$\begin{aligned}
a_n &= 2\int_0^1 x(1-x)\cos(n\pi x)\,\mathrm{d}x \\
&= 2\left(\left[\frac{1}{n\pi}x(1-x)\sin(n\pi x)\right]_0^1 - \frac{1}{n\pi}\int_0^1 (1-2x)\sin(n\pi x)\,\mathrm{d}x\right) \\
&= \frac{2}{n\pi}\left(\left[\frac{1}{n\pi}(1-2x)\cos(n\pi x)\right]_0^1 + \frac{2}{n\pi}\int_0^1 \cos(n\pi x)\,\mathrm{d}x\right) \\
&= -\frac{2}{n^2\pi^2}[\cos(n\pi) + 1] \\
&= \begin{cases} -\dfrac{4}{n^2\pi^2}, & n \text{ even}, \\ 0, & n \text{ odd}. \end{cases}
\end{aligned}$$

So, finally,

$$u(x,t) = \frac{1}{6} - \frac{1}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}e^{-4n^2\pi^2 t}\cos(2n\pi x).$$

Notice that since

$$\left|\sum_{n=1}^{\infty} e^{-4n^2\pi^2 t}\cos(2n\pi x)\right| \leq e^{-4\pi^2 t}\sum_{n=1}^{\infty}\frac{1}{n^2} \xrightarrow{t\to\infty} 0,$$

this solution has the interesting property that

$$\lim_{t\to\infty} u(x,t) = \frac{1}{6}.$$

Chapter 3

Probability theory

3.1 Introduction

In this chapter we summarise the basic notions of probability theory. If you are not familiar with the material here, work hard at the relevant exercises and ask lots of questions. Don't be afraid of "exposing" your lack of understanding! Becoming comfortable with the lingo and discussing problems is an important part of learning.

3.1.1 Some books worth looking at

There are many excellent books about probability theory. You can find them in the library under class mark QA.273 (that's on the second floor). The following are good for getting to grips with the subject.

- Ross, *A First Course in Probability*;
- Pitman, *Probability*;
- Grimmett and Stirzaker, *Probability and Random Processes*;
- Chung, *A Course in Probability Theory*;
- Durrett, *The essentials of probability theory*;
- Feller, *An Introduction to Probability Theory and Its Applications*, Volume I;
- Klenke, *Probability theory*, A comprehensive course.

3.2 Probability space

We begin with basic terms. The **sample space** Ω is the set of all possible outcomes of some hypothetical experiment. Informally, an **event** A is a subset of Ω. Then $\emptyset$ is the impossible event, which never happens, and Ω is the certain event, which always happens.

Example 3.2.1. The outcome of throwing a die is the number on its top when it comes to rest, so a suitable sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. The event that "the die shows an even number" is $\{2, 4, 6\} \subset \Omega$.

Example 3.2.2. If we roll two dice, then $\Omega = \{1,2,3,4,5,6\}^2 = \{(i,j) : 1 \leq i,j \leq 6\}$ is a suitable sample space and the event "the sum is 8" is $\{(2,6),(3,5),(4,4),(5,3),(6,2)\}$.

In many cases, we cannot observe every event. For example, if two dice are rolled ($\Omega = \{1,2,3,4,5,6\}^2$) but the second is hidden from us, we can only observe the events $\{i\} \times \{1,2,\ldots,6\}$, $i = 1,2,\ldots,6$. Moreover, there are deep mathematical reasons why a probability cannot always be ascribed "consistently" to every subset of Ω (e.g. there does not exist a translationally invariant measure m on $2^{\mathbb{R}}$ that has $m(I) > 0$ for bounded non-degenerate intervals I). Subsets of the sample space that can be observed (resp. for which a probability measure can be (or is) introduced) are called **events**. The collection of events is denoted by $\mathcal{F}$ ("curly F"). It is assumed that $\mathcal{F}$ is a **σ-algebra**, i.e. a collection of subsets of Ω that satisfy,

1. $\Omega \in \mathcal{F}$,
2. if $A \in \mathcal{F}$ then $A^c = \Omega \backslash A \in \mathcal{F}$,
3. if $A_i \in \mathcal{F}$, $i \geq 1$, is a countable sequence, then $\cup_i A_i \in \mathcal{F}$.

Recall that $\emptyset$ is the empty set, A^c is the complement of A in Ω and **countable** means finite or in bijection with the natural numbers $\mathbb{N} := \{1,2,3,\ldots\}$.

Since $\cap_i A_i = (\cup_i A_i^c)^c$, it follows that a σ-field is closed under countable intersections. It contains the empty set. The above are our first set of axioms. They are quite natural, for example (2) is that "if we can say when an event does occur, it is also possible to say that it doesn't occur".

Example 3.2.3. Let $\Omega = \{\xi_1,\xi_2,\xi_3,\xi_4\}$ consist of 4 outcomes. The classes

$$\mathcal{F}_1 = \{\emptyset, \{\xi_1,\xi_2\}, \{\xi_3,\xi_4\}, \{\xi_1,\xi_2,\xi_3,\xi_4\}\},$$
$$\mathcal{F}_2 = 2^\Omega = \text{the set of all subsets of } \Omega,$$

define a σ-field on Ω, whereas the class

$$\mathcal{F} = \{\emptyset, \{\xi_1\}, \{\xi_2\}, \{\xi_1,\xi_2,\xi_3,\xi_4\}\}$$

is not a σ-field.

Note that the intersection of any family of σ-algebras on Ω is a σ-algebra on Ω in turn. This allows to introduce (here for a set Ω, 2^Ω = set of all subsets of Ω):

Definition 3.2.4 (Generated σ-algebra). Let $\mathcal{A} \subset 2^\Omega$. The σ-algebra generated by $\mathcal{A}$ on Ω, denoted $\sigma_\Omega(\mathcal{A})$ (or $\sigma(\mathcal{A})$ for short, if there can be no confusion), is defined to be the smallest (with respect to inclusion) σ-algebra on Ω containing $\mathcal{A}$, i.e.:

$$\sigma(\mathcal{A}) = \cap\{\mathcal{F} \in 2^{2^\Omega} : \mathcal{A} \subset \mathcal{F}, \mathcal{F} \text{ is a } \sigma\text{-algebra on } \Omega\}.$$

Note that (i) notation-wise, for a non-empty set $\mathcal{C}$, $\cap\mathcal{C} = \bigcap_{c\in\mathcal{C}} c$ and (ii) 2^Ω is always a σ-algebra on Ω.

Example 3.2.5 (Borel σ-algebra). Let $\Omega = \mathbb{R}$ be the real line. Then the Borel σ-algebra $\mathcal{B}$ on Ω is the smallest σ-algebra containing every open interval (a,b), $a < b$. Most subsets of $\mathbb{R}$ that you will encounter are Borel (see exercises), however you are warned that not all are.

Definition 3.2.6 (Probability measure). Recall A and B are **disjoint** if $A \cap B = \emptyset$. A **probability measure** $\mathbb{P}$ is a way of assigning numbers to measurable events. It is a mapping, defined on $\mathcal{F}$, and assumed to satisfy the properties:

1. for any event $A \in \mathcal{F}$, $0 \leq \mathbb{P}(A) \leq 1$;
2. $\mathbb{P}(\Omega) = 1$;
3. for a countable sequence of pairwise disjoint events $A_i \in \mathcal{F}$, $\mathbb{P}\left(\cup A_i\right) = \sum_i \mathbb{P}(A_i)$.

These assumptions are motivated by the frequentist interpretation of probability, which says that if we repeat an experiment a large number of times then the fraction of times the event A occurs will be close to $\mathbb{P}(A)$. Some simple consequences of the definition are,

- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$;
- $\mathbb{P}(\emptyset) = 0$;
- if $A \subset B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$;
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$;
- $\mathbb{P}(\cup_i A_i) \leq \sum_i \mathbb{P}(A_i)$.

Definition 3.2.7 (Probability space). A **probability space** is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where $\mathcal{F}$ is a set of events, and $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a probability measure.

Formally, then, we model a random experiment via a probability space.

Example 3.2.8 (Discrete probability spaces). Let Ω be a countable set and $\mathcal{F} = 2^\Omega =$ set of all subset of Ω. Let

$$\mathbb{P}(A) = \sum_{\omega \in A} p(\omega) \text{ where } p(\omega) \geq 0 \text{ and } \sum_{\omega \in \Omega} p(\omega) = 1.$$

A little thought reveals that this is the most general probability measure on this space.

A large part of classical probability theory deals with finite sample spaces where each outcome has equal probability. Then, for any $A \subset \Omega$, $\mathbb{P}(A) = |A|/|\Omega|$, where $|A|$ is the number of elements in the set A. This is typically assumed when drawing cards, rolling dice etc.

Example 3.2.9. Roll two dice and let A be "the sum is 8" $= \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}$. Then $\mathbb{P}(A) = 5/36$ since it contains 5 of the 36 equally possible outcomes.

Example 3.2.10 ((0, 1), Borel, Lebesgue – the big three). Let $\Omega = \mathbb{R}$ be the real line, $\mathcal{B}$ be the Borel sets and $\lambda =$ **Lebesgue measure** (the measure that assigns lengths to intervals, $\lambda([a, b]) = b - a$ for all $a < b$). To get a probability space let $\Omega = (0, 1), \mathcal{F} = \{A \cap (0, 1) \mid A \in \mathcal{B}\}$ and $\mathbb{P}(B) = \lambda(B)$ for $B \in \mathcal{F}$. This is a canonical space on which we can build the absolutely continuous random variables introduced below.

Example 3.2.11 (Product spaces). If $(\Omega_i, \mathcal{F}_i, P_i), i = 1, \ldots, n$ are probability spaces, we can let $\Omega = \Omega_1 \times \cdots \times \Omega_n = \{(\omega_1, \ldots, \omega_n) : \omega_i \in \Omega_i, 1 \leq i \leq n\}$; $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_n =$ the smallest σ-algebra containing the collection of sets $\{A_1 \times \cdots \times A_n : A_i \in \mathcal{F}_i, 1 \leq i \leq n\}$; finally $\mathbb{P} = \mathbb{P}_1 \times \cdots \times \mathbb{P}_n =$ the measure on $\mathcal{F}$ that has,

$$\mathbb{P}(A_1 \times \cdots \times A_n) = \mathbb{P}_1(A_1) \cdot \mathbb{P}_2(A_2) \cdots \mathbb{P}_n(A_n).$$

Concrete examples of product spaces are;

(i) **Roll** $n = 2$ **dice**. $\Omega = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} =$ all subsets of Ω, $\mathbb{P}(A) = |A|/36$.

(ii) **Unit cube**. If $\Omega_i = (0,1)$, $\mathcal{F}_i$ = Borel sets, and $\mathbb{P}_i$ = Lebesgue measure, then the product space defined above is the unit cube $\Omega = (0,1)^n$, $\mathcal{F}$ = the Borel subsets of Ω, and $\mathbb{P}$ is n-dimensional Lebesgue measure restricted to $\mathcal{F}$ (i.e. $\mathbb{P}((a_1,b_1)\times\ldots\times(a_n,b_n)) = (a_1-b_1)\times\ldots\times(a_n-b_n)$ for $a_i < b_i$).

3.3 Combinatorics

In the finite (discrete) probability space with equally likely outcomes, $\mathbb{P}(A)$ is the fraction of outcomes that lie in A. Thus, to compute probabilities we have to be able to count the number of outcomes. To do this, the following results are useful and fall under the umbrella of the branch of mathematics known as combinatorics. The featured list is, however, not exhaustive and you will become familiar with some extra concepts through the exercise sheets.

3.3.1 Three basic principles

We begin with the **multiplication rule**. This is fundamental to most counting methods. Suppose that m experiments are performed one after another, and that, no matter what the outcomes of experiments $1,\ldots,k-1$ are, experiment k has n_k possible outcomes. Then the total number of outcomes is $n_1\cdot n_2\cdots n_m$.

In set-theoretic terms, the multiplication rule expresses the fact that the number of elements in a cartesian product of n sets A_1, ..., A_n is the product of the number of elements in each of these respective sets. Symbolically:

$$|A_1\times\cdots\times A_n| = |A_1|\cdots|A_n|.$$

Here $|A|$ denotes, as always, the number of elements in a given set A.

The second is the **rule of sum**. Suppose n groups are given, the i-th group containing m_i individuals ($i=1,\ldots,n$). Then the total number of all individuals in all the groups is $m_1+\cdots+m_n$. In set-theoretic terms, this is saying that the number of elements in a disjoint union is the sum of the number of elements in each of the respective sets (constituting said union). Symbolically:

$$|A_1\cup\cdots\cup A_n| = |A_1|+\cdots+|A_n|,$$

where $A_1,\ldots,A_n$ are pairwise disjoint sets.

Finally there is the **bijection rule**, which simply says that given two sets between which there is a one-to-one and onto correspondence (i.e. a bijection), then these two sets contain the same number of elements:

$$\exists f: A\to B,\ \text{bijection}\ \Rightarrow\ |A| = |B|.$$

These rules are often used tacitly when counting the number of objects, or ways in which something may transpire, and it is important to be aware of them, and how they are applied.

3.3.2 Variations and combinations

The first consequence of the multiplication rule is that if we have n people then the number of ways we can pick k of them to stand in line in order (an ordered k-tuple) is

$$P_{n,k} = n\cdot(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}.$$

These are known as **variations without repetition**. Here $n! := n \cdot (n-1) \cdots 2 \cdot 1$, and is called n **factorial**. In the special case when $k = n$, we speak of **permutations**.

Next, if we have n people then the number of ways we can pick a subgroup of k individuals from them is

$$C_{n,k} = \frac{n \cdot (n-1) \cdots (n-k+1)}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k},$$

These are known as **combinations**. Here $0! = 1$, so $\binom{n}{0} = \binom{n}{n} = 1$. We read $\binom{n}{k}$ as "n choose k".

The **binomial theorem** says,

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

and so the $\binom{n}{k}$ are also referred to as the **binomial coefficients**.

This can be generalised to **multinomial coefficients**. The number of ways we can divide n objects into m piles of sizes $n_1, \ldots, n_m$ is,

$$\frac{n!}{n_1! n_2! \cdots n_m!}.$$

3.3.3 Stirling numbers of the second kind and partitions

A **partition** of a set A is a collection of non-empty pairwise disjoint sets, whose union is A. The number of different partitions of a set of n elements into k disjoint non-empty sets is denoted $S(n,k)$. These are the so-called Stirling numbers of the second kind. Clearly $S(n,0) = 0$ and $S(n,n) = S(n,1) = 1$ for every natural $n \in \mathbb{N}$. Additionally we define $S(n,0) = 0$ for every $n \in \mathbb{N}$, and $S(0,0) = 1$. Then for every pair of natural numbers $k \leq n$ (see exercises):

$$S(n,k) = S(n-1,k-1) + kS(n-1,k).$$

In a similar problem, we consider the number of ways in which a natural number n can be written as the sum of k natural numbers $m_1 \leq m_2 \leq \cdots \leq m_k$ in nondecreasing order. This number is denoted by $p_k(n)$. Clearly $p_k(n) = 0$ for $k > n$. Additionally we define $p_0(0) = 1$ and $p_0(n) = 0$ for $n > 0$. Then for every pair of natural numbers n, k:

$$p_k(n) = p_{k-1}(n-1) + p_k(n-k).$$

3.3.4 Urn problems

Suppose we have an urn with M red balls and N black balls and we draw out n balls without replacement. Then the probability we get r red balls and hence $n-r$ black balls is

$$\frac{\binom{M}{r}\binom{N}{n-r}}{\binom{M+N}{n}},$$

where we have used the convention that $\binom{m}{j} = 0$ if $j < 0$ or $j > m$. The last answer generalizes in a straightforward way to situations with more than two colors.

3.3.5 Repeated experiments

The problems in this section concern situations akin to drawing with replacement from an urn with K balls numbered from 1 up to K. That is, we pick a ball, note its number and then return it to the urn. Since all K outcomes are possible on each trial the total number of outcomes is K^k. Two important special cases are:

The birthday problem. If we draw k balls with replacement, then the probability all the numbers are different is,

$$\frac{P_{K,k}}{K^k} = \frac{K}{K} \cdot \frac{(K-1)}{K} \cdots \frac{(K-k+1)}{K}.$$

Flipping coins. This corresponds to $K = 2$, where heads $= 1$ and tails $= 2$. The probability of exactly j heads in k trials is

$$\binom{k}{j} 2^{-k}.$$

The first factor gives the number of ways of picking j tosses on which heads occurs; the second, the probability of each of the 2^k outcomes.

3.4 Independence and conditional probability

3.4.1 Independence

Definition 3.4.1 (Independence). Two events A and B are **independent** intuitively if the occurrence of A has no effect on the probability of occurrence of B. Formally we can write this as

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B).$$

A finite or infinite sequence of events $A_1, \ldots, A_n, \ldots$ is said to be (a sequence of) **pairwise independent** (events) if

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\,\mathbb{P}(A_j) \text{ whenever } i \neq j,$$

and said to be (a sequence of) **independent** (events) if for any $i_1 < i_2 < \cdots < i_k$,

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k}).$$

Pairwise independent events are not necessarily independent!

As you will see in the exercises, the definition of independence can play tricks on you.

Example 3.4.2. Pick a card from a deck of 52 and let $A =$ "the card is an ace", $B =$ "the card is a spade." $\mathbb{P}(A) = 1/13$, $\mathbb{P}(B) = 1/4$, $\mathbb{P}(A \cap B) = 1/52$. A and B are independent. If, however, $A =$ "the first card is an ace" and $B =$ "the second card is an ace" then A and B are not independent.

Example 3.4.3. For any event A, $A \cap A = A$, so A is independent of itself, if and only if $\mathbb{P}(A) = 0$ or 1.

The concept of repeated independent trials crops up a lot. You should understand how the following distributions arise.

Binomial distribution — $\text{Bin}(n, p)$. Suppose that n trials (experiments) occur, and the outcome of each trial is 'success' with probability p or 'failure' with probability $1 - p$. The probability of k successes is

$$\binom{n}{k} p^k (1-p)^{n-k} \ (k = 0, \ldots, n).$$

The special case when $n = 1$ is that of a Bernoulli random variable — notation: $\text{Ber}(p)$.

Multinomial distribution — $\text{mult}(n; p_1, \ldots, p_k)$. If instead of just success and failure there are k possible outcomes with probabilities $p_i, 1 \leq i \leq k$, then the probability of getting exactly n_i outcomes of type i in $n = n_1 + \ldots + n_k$ independent trials is,

$$\frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}.$$

Geometric distribution — $\text{geom}(p)$. Suppose again that independent trials, each with success probability p are performed. The probability that the first success occurs on the k-th trial is

$$(1-p)^{k-1} p \text{ for } k = 1, 2, \ldots.$$

3.4.2 Conditional probability

Definition 3.4.4 (Conditional probability). Suppose that $A \in \mathcal{F}$ has positive probability of occurring. The **conditional probability** of $B \in \mathcal{F}$ occurring given that A has occurred is

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}. \tag{3.1}$$

With A fixed, let $\mathcal{F}_A = \mathcal{F} \cap 2^A$. Then $\mathbb{P}_A : \mathcal{F}_A \to [0, 1]$ defined by $\mathbb{P}_A(B) = \mathbb{P}(B|A)$ is a probability measure on $\mathcal{F}_A$.

Multiplying each side of (3.1) by $\mathbb{P}(A)$ gives the **multiplication rule**

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B|A). \tag{3.2}$$

Note that the probability of A and B happening $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B|A) = \mathbb{P}(B)\,\mathbb{P}(A|B)$.

3.4.3 Two stage experiments

Let the events $B_1, \ldots, B_n, \ldots$ constitute a finite or countably infinite **partition** of Ω; that is, a sequence of pairwise disjoint non-empty sets with union Ω. Then for an event A, assuming $\mathbb{P}(B_i) > 0$ for each i:

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap B_i) = \sum_i \mathbb{P}(B_i)\,\mathbb{P}(A|B_i).$$

This formula is sometimes called the law of total probability.

3.4.4 Bayes' formula

Bayes' formula allows to compute $\mathbb{P}(B_1|A)$ when $\mathbb{P}(B_i) > 0$ and $\mathbb{P}(A|B_i)$ are given (where B_1, B_2,...,B_n, ... continue to enjoy the property of being a partition of Ω; A being any event with $\mathbb{P}(A) > 0$) and proceeds in the following three steps:

$$\begin{aligned} \mathbb{P}(B_1|A) &= \frac{\mathbb{P}(B_1 \cap A)}{\mathbb{P}(A)} && \text{definition of conditional probability,} \\ \mathbb{P}(B_i \cap A) &= \mathbb{P}(B_i)\ \mathbb{P}(A|B_i) && \text{multiplication rule,} \\ \mathbb{P}(A) &= \sum_i \mathbb{P}(B_i \cap A) && \text{law of total probability.} \end{aligned}$$

Combining the three steps gives the formula:

$$\mathbb{P}(B_1|A) = \frac{\mathbb{P}(B_1)\ \mathbb{P}(A|B_1)}{\sum_i \mathbb{P}(B_i)\ \mathbb{P}(A|B_i)}$$

To show an application of Bayes' formula, let us consider a simple example.

Example 3.4.5. The alpha fetal protein test is meant to detect spina bifida in unborn babies, a condition that affects 1 out of 1000 children who are born. Let B be the event that the baby has spina bifida and B^c be the event that it does not. The literature on the test indicates that 5% of the time a healthy baby will cause a positive reaction. We will assume that the test is positive 100% of the time when spina bifida is present. Your doctor has just told you that your alpha fetal protein test was positive. What is the probability that your baby has spina bifida?

Let $A =$ 'a positive reaction'. We want to calculate $\mathbb{P}(B|A)$. By the definition of conditional probability (3.1),

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}.$$

To evaluate the numerator we use the multiplication rule (3.2),

$$\mathbb{P}(B \cap A) = \mathbb{P}(B)\,\mathbb{P}(A|B) = 0.001 \cdot 1 = 0.001.$$

Similarly,

$$\mathbb{P}(B^c \cap A) = \mathbb{P}(B^c)\,\mathbb{P}(A|B^c) = 0.999 \cdot 0.05 \approx 0.05.$$

Now $\mathbb{P}(A) = \mathbb{P}(B \cap A) + \mathbb{P}(B^c \cap A)$ so,

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} \approx \frac{0.001}{0.001 + 0.050} = \frac{1}{51}.$$

Thus the probability of spina bifida given the positive reaction is only about 2%. This situation comes about because it is much easier to have a positive reaction by having a healthy baby and then having a positive reaction, which has probability 0.05, than by having a baby with spina bifida, which has probability 0.001.

We would like to point out that while the conditional probability of spina bifida given a positive reaction is only 2%, this does not mean that the test is worthless. To introduce some terminology from Bayesian statistics, the **prior probability** (i.e. before the test) of spina bifida is 0.1%, whereas the **posterior probability** (i.e. after the test results are known) is about 2%. That is, the probability is now 20 times larger. The positive reaction is thus a warning that more accurate (and more expensive) tests should be done to see if the baby has spina bifida.

3.5 Real-valued random elements (a.k.a. random variables)

Definition 3.5.1 (Random variable). A real valued function $X : \Omega \to \mathbb{R}$ is said to be a **random variable** (abbrev. RV) on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ if for every Borel set $B \in \mathcal{B}$ we have

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}.$$

In other words, X is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ if it is $\mathcal{F}$-measurable. It is sufficient to check that, for all real t, $X^{-1}((-\infty, t]) \in \mathcal{F}$. Random variables are typically denoted by upper case letters.

Example 3.5.2. If $\mathcal{F}$ is discrete, i.e. $\mathcal{F}$ is the set of all subsets of Ω, then any function,

$$X : \Omega \to \mathbb{R},$$

is a random variable.

Example 3.5.3. The **indicator function** of a set $A \in \mathcal{F}$,

$$\mathbb{1}_A(\omega) = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \in \Omega \backslash A, \end{cases}$$

is always a random variable.

3.5.1 Law and distribution function

Any random variable X induces a probability measure μ_X on $\mathbb{R}$, called its **law**, defined by

$$\mu_X(B) = \mathbb{P}(X \in B)$$

for Borel sets B. Note that $\mathbb{P}(X \in B)$ is short-hand for

$$\mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\}).$$

Similarly, $\mathbb{P}(X > 0)$ really means

$$\mathbb{P}(\{\omega : X(\omega) > 0\})$$

and so forth. We write $X \sim \mu$, indicating X has the law μ.

If two random variables X and Y have the same law we say they are equal in law (or in distribution). Note that this need not entail $X(\omega) = Y(\omega)$ for every ω, or even that X and Y be defined on the same probability space!

X is called **discrete** if its law μ_X only puts mass on a countable subset of $\mathbb{R}$, that is, there is a countable set $\{x_1, x_2, \ldots\}$ such that

$$\mu_X(\{x_1, x_2, \ldots\}) = \sum_i \mathbb{P}(X = x_i) = 1.$$

If in addition $\mu_X(\{x_i\}) > 0$ for each i, we call x_i the **atoms** of μ_X and the mapping

$$(x_i \mapsto p_X(x_i) := \mathbb{P}(X = x_i) = \mu_X(\{x_i\}))$$

is called the **probability mass function** (pmf) of X (this function completely characterises the law). The domain of p_X, denoted S_X, is then called the **support** of X.

Example 3.5.4. Suppose that $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = 2^\Omega$ and $\mathbb{P}$ is uniform so that $\mathbb{P}(A) = |A|/6$; i.e. we have a probability space corresponding to throwing a die. Then $X(\omega) = \omega$ is a random variable (exercise) which is discrete, and

$$\mu_X(B) = |\{1 \leq i \leq 6 : i \in B\}|/6,$$

so $\mu_X(\{1, 2, 3, 4, 5, 6\}) = 1$, and $p_X(i) = 1/6$ for $1 \leq i \leq 6$.

The law of a random variable X is specified uniquely by its **distribution function** $F = F_X$,

$$F(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}),$$

sometimes called the **cumulative distribution function** (CDF).

Any distribution function F has the following properties:

(i) F is non-decreasing;

(ii) $\lim_{x\to\infty} F(x) = 1$;

(iii) $\lim_{x\to-\infty} F(x) = 0$;

(iv) F is right continuous: $\lim_{y\downarrow x} F(y) = F(x)$;

(v) If $F(x-) = \lim_{y\uparrow x} F(y)$ then, $F(x-) = \mathbb{P}(X < x)$;

(vi) $\mathbb{P}(X = x) = F(x) - F(x-)$.

Conversely if a function F satisfies properties (i-iv), then it is the distribution function of some random variable; i.e. there is a unique probability measure μ_F on $(\mathbb{R}, \mathcal{B})$ that has $\mu_F((a, b]) = F(b) - F(a)$.

It is common practice, e.g. in statistics, to introduce random variables by specifying their distribution only, and not to define the underlying probability space. For example, one might say "suppose X is a real valued random variable with cdf F", and then refer to the probability of some event involving X, say $\mathbb{P}(X > 0)$, without ever having defined $\mathbb{P}$ or the sample space. In this situation, it is understood that there *exists* a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ containing a random variable with the given distribution – we just don't bother to specify it.

Example 3.5.5 (Poisson distribution). Suppose that X is a random variable taking values in the non-negative integers, with mass function

$$\mathbb{P}(X = k) = \exp(-\lambda)\frac{\lambda^k}{k!},\ k = 0, 1, 2, 3, \ldots$$

for some parameter $\lambda > 0$. Then, X is called a Pois(λ) random variable.

For a discrete random variable, the distribution function only increases by jump discontinuities, and is constant between those jumps. The jumps occur at the discrete values taken by the random variable. It is helpful to sketch distribution functions of some common discrete random variables – you will see they can be somewhat messy and it is easier to work directly with the mass function.

X is called **absolutely continuous** if there exists a non-negative function f on $\mathbb{R}$ such that

$$F(x) = \int_{-\infty}^{x} f(x)\,\mathrm{d}x,\ x \in \mathbb{R}.$$

The function f is called the **probability density function** (pdf) of X, and by the Fundamental Theorem of Calculus, $f(x) = F'(x)$ for Lebesgue-almost every $x \in \mathbb{R}$ (indeed every $x \in \mathbb{R}$, if f is continuous). Since $F(x) \to 1$ as $x \to \infty$, we must have that $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$. For small $\mathrm{d}x > 0$ we have

$$\mathbb{P}(X \in (x, x+\mathrm{d}x)) = \int_x^{x+\mathrm{d}x} f(x)\,\mathrm{d}x \approx f(x)\,\mathrm{d}x.$$

Sometimes $\mathrm{d}x$ is thought of both as an infinitesimal interval and number and the above is written

$$\mathbb{P}(X \in \mathrm{d}x) = f(x)\,\mathrm{d}x.$$

This can be useful for intuition, but be careful – $\mathbb{P}(X = x) = 0$ for any x!

Example 3.5.6 (Unif$([0,1])$-random variable). Suppose that $\Omega = (0,1)$, $\mathcal{F}$ is the Borel σ-algebra of $(0,1)$ and $\mathbb{P}$ is the Lebesgue measure. Then the variable $X : \Omega \to (0,1)$ defined by $X(\omega) = \omega$ for $\omega \in (0,1)$ has the **uniform distribution**. The distribution function of X is $F(x) = \mathbb{P}(X \le x) = \mathbb{P}((0,x)) = x$; the density is $f(x) = 1$ (for $0 < x < 1$).

In general a random variable having the density $\mathbb{1}_{[a,b]}/(b-a)$ is the uniform distribution on the interval $[a,b]$ (for $a < b$) — notation: Unif$([a,b])$.

Example 3.5.7 (Exponential distribution — Exp(λ)). The random variable with density $f(x) = \lambda \exp(-\lambda x)$, $x > 0$ is called the **exponential distribution** of rate $\lambda > 0$.

Most of the random variables you encounter will either be discrete or absolutely continuous, but it is important to realise that there exist random variables which are neither absolutely continuous nor discrete.

3.5.2 Functions of random variables

Suppose X is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and $r : \mathbb{R} \to \mathbb{R}$ is some function. Define $Y(\omega) = r(X(\omega)) = (r \circ X)(\omega)$. We can ask, when is Y a random variable in turn? In general, r being *measurable* (i.e. $r^{-1}(B) = \{x \in \mathbb{R} : r(x) \in B\} \in \mathcal{B}$ for any Borel $B \in \mathcal{B}$) is sufficient to guarantee that Y is also a random variable.

Now suppose that X is absolutely continuous, $r : \mathbb{R} \to \mathbb{R}$ is a differentiable bijection with a strictly positive derivative. This means that r has a differentiable inverse, say $s : \mathbb{R} \to \mathbb{R}$. What is the distribution of Y? Well,

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(r(X) \le y) = \mathbb{P}(X \le s(y)) = F_X(s(y))$$

$$= \int_{-\infty}^{s(y)} f_X(u)du = \int_{-\infty}^{y} f_X(s(z))s'(z)dz.$$

Thus Y is absolutely continuous and has density

$$f_Y(y) = F_Y'(y) = f_X(s(y))s'(y).$$

More generally:

Theorem 3.5.8 (Density transformation formula). Let $X : \Omega \to \mathbb{R}$ be an absolutely continuous random variable, with pdf f_X. Let A be an open subset (resp. open interval) of $\mathbb{R}$, X taking values in A, $g : A \to \mathbb{R}$ differentiable with $g' > 0$ positive everywhere. Then g is an

increasing bijection onto its open (resp. what is then an open interval) image $g(A)$, $Y = g \circ X$ is absolutely continuous, and has pdf given by:

$$f_Y = \mathbb{1}_{g(A)} \frac{f_X \circ g^{-1}}{g' \circ g^{-1}}$$

(note that $(g^{-1})' = 1/g' \circ g^{-1}$).

Two more useful facts are,

1. If X is a continuous random variable, i.e. $\mathbb{P}(X = x) = 0$ for all real x, with distribution function F, then $F(X) = F \circ X$ is uniform on $(0, 1)$. Conversely, if X is not continuous, then $F \circ X$ does not have the uniform distribution on $(0, 1)$.
2. If U is uniform on $(0, 1)$ and $F^{-1}(y) = \min\{x : F(x) \geq y\}$ then $F^{-1}(U)$ has distribution function F. So, provided F^{-1} can be calculated, it is very easy to simulate X on a computer – simulate uniform $U \sim \text{Unif}([0, 1])$.

3.5.3 Joint distributions of several random variables

Suppose that a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ has $n > 0$ random variables $X_1, \ldots, X_n$ defined on it. Then, for Borel $B \in \mathcal{B}(\mathbb{R}^n)$,

$$X^{-1}(B) = \{\omega \in \Omega : (X_1(\omega), \ldots, X_n(\omega)) \in B\} \in \mathcal{F}.$$

$X = (X_1, \ldots, X_n)$ is called a random vector. For it, we can define the **joint law** $\mu_{X_1,\ldots,X_n}$ by

$$\mu_{X_1,\ldots,X_n}(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}((X_1, \ldots, X_n) \in B).$$

The joint distribution function is given by:

$$F_X(x) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n),\ x = (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Example 3.5.9. Suppose that X and Y are two random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. The probability that (X, Y) lies in the rectangle $(a_1, a_2) \times (b_1, b_2)$ is given by

$$\mathbb{P}(a_1 < X \leq b_1, a_2 < Y \leq b_2) = F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_2, a_2),$$

which is the two dimensional analogue of $\mathbb{P}(a < X \leq b) = F(b) - F(a)$.

If the X_i are all discrete, then X is discrete and the joint distribution is given by the **joint probability mass function**, $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$, where x_i ranges (at most) over the values taken by X_i.

The definition for the X_i to be jointly absolutely continuous is that there exists a non-negative function $f_X : \mathbb{R}^n \to \mathbb{R}$ such that

$$F_X(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_X(u_1, \ldots, u_n)\, \mathrm{d}u_1 \cdots \mathrm{d}u_n.$$

This implies that

$$\mathbb{P}(X \in A) = \int_A f_X(u_1, \ldots, u_n)\, \mathrm{d}u_1 \cdots \mathrm{d}u_n \tag{3.3}$$

for any Borel set $A \in \mathcal{B}(\mathbb{R}^n)$. The function f_X is, not surprisingly, called the **joint density**.

Note that the X_i being individually absolutely continuous (i.e. having a density) does NOT imply that they are jointly absolutely continuous (see exercises).

Example 3.5.10. Let

$$f(x, y) = \begin{cases} e^{-y}, & 0 < x < y < \infty, \\ 0, & \text{otherwise.} \end{cases}$$

To check that f is a density function, we observe that,

$$\int_0^\infty \int_0^y e^{-y}\, \mathrm{d}x\, \mathrm{d}y = \int_0^\infty y e^{-y}\, \mathrm{d}y,$$

and integrating by parts with $g(y) = y, h'(y) = e^{-y}$ (so $g'(y) = 1, h(y) = -e^{-y}$),

$$\int_0^\infty y e^{-y}\, \mathrm{d}y = \Big[-y e^{-y} \Big]_0^\infty + \int_0^\infty e^{-y}\, \mathrm{d}y = 0 + \Big[-e^{-y} \Big]_0^\infty = 1.$$

To illustrate the use of (3.3) we will now compute $\mathbb{P}(X \leq 1)$, which can be written as $\mathbb{P}((X, Y) \in A)$ where $A = \{(x, y) : x \leq 1\}$. The formula in (3.3) tells us that we find $\mathbb{P}((X, Y) \in A)$

by integrating the joint density over A. However, the joint density is only positive on $B = \{(x, y) \mid 0 < x < y < \infty\}$ so we only need to integrate over $A \cap B = \{(x, y) : 0 < x \leq 1, x < y\}$, and doing this we find

$$\mathbb{P}(X \leq 1) = \int_0^1 \int_x^\infty e^{-y} \, \mathrm{d}x \, \mathrm{d}y.$$

To evaluate the double integral we begin by observing that,

$$\int_x^\infty e^{-y} \, \mathrm{d}y = \Big[-e^{-y} \Big]_x^\infty = 0 - (-e^{-x}) = e^{-x},$$

so $\mathbb{P}(X < 1) = \displaystyle\int_0^1 e^{-x} \, \mathrm{d}x = \Big[-e^{-x} \Big]_0^1 = 1 - e^{-1}$.

3.6 Marginal distributions, independence, conditional distributions

Suppose we are given the joint distribution of several random variables, $X_1, \ldots, X_n$ say. The distribution of a single variable, X_1 say, induced by the joint distribution is called the **marginal distribution**. It is recovered by "integrating out" all of the other variables.

For example, suppose X and Y are two discrete variables, or are jointly absolutely continuous. Then, to recover the distribution of X from the joint distribution of X and Y, we use

$$\mathbb{P}(X = x) = \sum_{y \in S_y} \mathbb{P}(X = x, Y = y),$$

in the discrete case (where the summation ranges over the countable set of values taken by Y), or

$$f_X(x) = \int_{\mathbb{R}} f(x, y) \, \mathrm{d}y, \ x \in \mathbb{R}$$

in the absolutely continuous case. The density f_X is called the **marginal density**.

Example 3.6.1. Let

$$f(x, y) = \begin{cases} e^{-y}, & 0 < x < y < \infty, \\ 0, & \text{otherwise.} \end{cases}$$

In this case,

$$f_X(x) = \int_x^\infty e^{-y} \, \mathrm{d}y = \Big[-e^{-y} \Big]_x^\infty = e^{-x}, \ x > 0,$$

since we only have $f(x, y) > 0$ when $y > x > 0$. Similarly,

$$f_Y(y) = \int_0^y e^{-y} \, \mathrm{d}x = y e^{-y}, \ y > 0.$$

Here and in similar instances below, the formula for the absolutely continuous case is obtained from the discrete case by writing density functions in place of the corresponding probabilities and replacing the sum by an integral.

Two random variables X and Y are said to be **independent** if

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \, \mathbb{P}(Y \in B)$$

for any Borel $A, B \in \mathcal{B}$. In the discrete case it is equivalent to

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\,\mathbb{P}(Y = y), \quad \forall x, y.$$

In the absolutely continuous case this is equivalent to

$$f(x, y) = f_X(x) f_Y(y) \quad \forall x, y.$$

To be completely precise, we should say in the above for Lebesgue almost every (x, y), rather than all (x, y), since a density function can always be changed on a set of Lebesgue measure zero. At the cost of rigour, and to the benefit of simplicity, we shall tend to omit these qualifications in the sequel.

A finite sequence $X_1, \ldots, X_n$ of random variables is said to be (a sequence of) independent (random variables) if:

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n),$$

for all Borel sets $A_1, \ldots, A_n$, i.e. if $\mu_{(X_1,\ldots,X_n)} = \mu_{X_1} \times \cdots \times \mu_{X_n}$.

Similarly, an infinite sequence of random variables $X_1, X_2, \ldots$ is said to be (a sequence of) independent (random variables), if any finite subsequence is so.

Independence is a very important notion in probability theory allowing for much greater generality than what we have considered above. For example, one can speak of two (finite or infinite) sequences $X_1, X_2, \ldots, X_n, \ldots$ and $Y_1, Y_2, \ldots, Y_m, \ldots$ of random variables being independent (if $\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n, Y_1 \in B_1, \ldots, Y_n \in B_n) = \mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n)\,\mathbb{P}(Y_1 \in B_1, \ldots, Y_n \in B_n)$ for any choice of Borel sets $A_1, \ldots, A_n, B_1, \ldots, B_n$, $n \geq 1$) and so on and so forth.

3.6.1 Conditional distributions

For discrete variables X and Y, the **conditional distribution** of X *given* $Y = y \in \mathbb{R}$ is defined, for $x \in \mathbb{R}$ by

$$\mathbb{P}(X = x | Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{\mathbb{P}(X = x, Y = y)}{\sum_u \mathbb{P}(X = u, Y = y)},$$

which is defined whenever $\mathbb{P}(Y = y) > 0$. When X and Y are jointly absolutely continuous, the conditional distribution of X given $Y = y$ is defined via the conditional density;

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{f(x, y)}{\int f(u, y)\,\mathrm{d}u},$$

provided $f_Y(y)$ is positive.

If X and Y are independent, we have

$$\mathbb{P}(X = x | Y = y) = \mathbb{P}(X = x)$$

and

$$f_{X|Y}(x|y) = f_X(x)$$

in the discrete and absolutely continuous cases respectively.

3.7 Sums of independent random variables

Suppose that X and Y are independent and absolutely continuous, with densities f_X and f_Y respectively. Then the sum $X+Y$ is a random variable too. What is its distribution? The cumulative distribution function of the sum is obtained as follows.

$$\begin{aligned} F_{X+Y}(a) &= \iint_{x+y\le a} f_X(x) f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{\infty}^{a-y} f_X(x)\,\mathrm{d}x f_Y(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} F_X(a-y) f_Y(y)\,\mathrm{d}y, \end{aligned}$$

where F_X is the distribution of X. Now differentiate this expression in a (through the integral!) to recover the density[1]

$$f_{X+Y}(a) = F'_{X+Y}(a) = \int_{-\infty}^{\infty} \frac{\partial}{\partial a} F_X(a-y) f_Y(y)\,\mathrm{d}y.$$

That is,

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(a-y) f_Y(y)\,\mathrm{d}y.$$

This is called the **convolution** of f_X and f_Y and is written $f_X * f_Y(a)$. The analogue for discrete variables is

$$\mathbb{P}(X+Y=z) = \sum_y \mathbb{P}(X=z-y)\,\mathbb{P}(Y=y). \tag{3.4}$$

Example 3.7.1. If $X \sim \mathrm{Bin}(n,p)$ and $Y \sim \mathrm{Bin}(m,p)$ are independent then $X+Y \sim \mathrm{Bin}(n+m,p)$.

The easiest way to see this is to note that if X is the number of successes in the first n trials and Y is the number of successes in the next m trials, then $X+Y$ is the number of successes in $n+m$ trials.

To get the conclusion by computation we use (3.4), note that $\mathbb{P}(X=j)=0$ when $j<0, \mathbb{P}(Y=k-j)=0$ when $j>k$, and plug in the definition of the binomial distribution to get

$$\begin{aligned} \mathbb{P}(X+Y=k) &= \sum_{j=0}^{k} \mathbb{P}(X=j)\,\mathbb{P}(Y=k-j), \\ &= \sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j} \binom{m}{k-j} p^{k-j} (1-p)^{m-(k-j)}, \\ &= p^k (1-p)^{n+m-k} \sum_{j=0}^{k} \binom{n}{j}\binom{m}{k-j}, \\ &= p^k (1-p)^{n+m-k} \binom{n+m}{k}. \end{aligned}$$

[1] In order to maintain some level of rigor, note that differentiation under the integral sign is less than a completely benign operation, needing explicit, non-trivial justification. In this case we can actually avoid it by effecting in $F_{X+Y}(a) = \iint_{x+y\le a} f_X(x) f_Y(y)\,\mathrm{d}x\,\mathrm{d}y$ the change of variables $x = x' - y'$ and $y = y'$. This allows to conclude directly that $X+Y$ is absolutely continuous with the density given below.

3.8 Expected value and moment generating functions

Definition 3.8.1 (Expected value). The **expected value** of X is defined to be

$$\mathbb{E}[X] = \sum_{x \in S} x \, \mathbb{P}(X = x),$$

for a discrete variable with support S, and

$$\mathbb{E}[X] = \int x f(x) \, \mathrm{d}x,$$

for an absolutely continuous variable with density f, whenever these are well-defined, e.g. if $\sum_{x \in S} |x| \, \mathbb{P}(X = x) < \infty$ and $\int |x| f(x) dx < \infty$, respectively.

Remarks:

(i) If X is, say, discrete, with support S, then we need in the above, in fact, only demand $\min\{\sum_{x \in S, x > 0} x \, \mathbb{P}(X = x), \sum_{x \in S, x < 0} (-x) \, \mathbb{P}(X = x)\} < \infty$ (using the convention $\sum_{\emptyset} = 0$), in which case $\mathbb{E}[X] = \sum_{x \in S, x > 0} x \, \mathbb{P}(X = x) - \sum_{x \in S, x < 0} (-x) \, \mathbb{P}(X = x) \in [-\infty, +\infty]$ (ultimately, all we want to avoid is having to make sense of "$\infty - \infty$"). Similarly for the absolutely continuous case.

(ii) More generally, $\mathbb{E}[X] = \int X d \, \mathbb{P} = \int x d\mu_X(x)$, ...

Expectation has the following frequentist interpretation. If $X_1, X_2, \ldots,$ are independent variables with the same distribution as X, then $(X_1 + \cdots + X_n)/n$ converges to $\mathbb{E}[X]$ with probability one ("almost surely").

Example 3.8.2 (Bernoulli distribution). Suppose X is 1 with probability p and 0 with probability $1 - p$. Then

$$\mathbb{E}[X] = p \cdot 1 + (1 - p) \cdot 0 = p.$$

Some important properties of expected values are:

(i) $\mathbb{E}[aX + b] = a \, \mathbb{E}[X] + b$ for $a, b \in \mathbb{R}$;

(ii) If $X \geq Y$ then $\mathbb{E}[X] \geq \mathbb{E}[Y]$;

(iii) If $X_1, \ldots, X_n$ are independent, then:

$$\mathbb{E}[X_1 \cdots X_n] = \mathbb{E}[X_1] \cdots \mathbb{E}[X_n],$$

but the equality may hold for random variables that are not independent.

Random variables X and Y that have $\mathbb{E}[XY] = \mathbb{E}[X] \, \mathbb{E}[Y]$ (these expectations needing to be well-defined and finite) are said to be **uncorrelated**. Independent variables are uncorrelated but the converse need not hold.

Theorem 3.8.3. If $X = (X_1, \ldots, X_d)$ is jointly absolutely continuous (resp. discrete) $\mathbb{R}^d$-valued with joint pdf (resp. pmf) f_X (resp. p_X and support S), then for any $r : \mathbb{R}^d \to \mathbb{R}$ (Borel measurable):

$$\mathbb{E}[r \circ X] = \int r(x) f_X(x) dx \text{ (resp. } = \sum_{x \in S} r(x) p_X(x)),$$

whenever these are well-defined.

Definition 3.8.4 (Variance). The **variance** of X is defined as (provided $\mathbb{E}[|X|]$ is finite):

$$\operatorname{var}[X] = \mathbb{E}\{(X - EX)^2\} = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

The variance satisfies (for $a, b \in \mathbb{R}$)

$$\operatorname{var}[aX + b] = a^2 \operatorname{var}[X],$$

and if $X_1, \ldots, X_n$ are pairwise uncorrelated then

$$\operatorname{var}[X_1 + \cdots + X_n] = \operatorname{var}[X_1] + \cdots + \operatorname{var}[X_n].$$

Example 3.8.5. Suppose X has density $f(x) = rx^{r-1}$ when $0 < x < 1$, and 0 otherwise. Here, $r > 0$ (when $r = 1$, X is uniform on $(0, 1)$). So, for $k > -r$

$$\mathbb{E}[X^k] = \int_0^1 x^k r x^{r-1} \, \mathrm{d}x = \left[\frac{r x^{k+r}}{k+r}\right]_0^1 = \frac{r}{k+r}.$$

Example 3.8.6. Suppose that X has an exponential density with parameter λ. Then,

$$\mathbb{E}[e^{tX}] = \int_0^\infty e^{tx} \lambda e^{-\lambda x} \, \mathrm{d}x = \lambda \int_0^\infty e^{-(\lambda - t)x} \, \mathrm{d}x = \frac{\lambda}{\lambda - t},$$

for $t < \lambda$. $\mathbb{E}[e^{tX}] = \infty$ for $t \geq \lambda$.

Definition 3.8.7 (Moments; moment generating function). $\mathbb{E}[X^k]$ is called the k-th **moment** of X; $\phi_X(t) = \mathbb{E}[e^{tX}]$, defined for all $t \in \mathbb{R}$ for which this expectation is finite, is called the **moment generating function** and can be used to compute the moments of X.

Suppose ϕ_X is defined on some open interval $(-t_0, t_0)$, $t_0 > 0$ (i.e. the domain of ϕ_X contains a neighborhood of zero), and let $\phi^{(n)}$ denote the n-th derivative of ϕ. Then,

$$\phi^{(n)}(0) = \mathbb{E}[X^n].$$

As an application of the last formula we can compute the moments of the exponential distribution. To begin, we note that from Example 3.8.6, $\phi_X(t) = \lambda/(\lambda - t)$, so,

$$\begin{aligned}
\phi_X'(t) &= \frac{\lambda}{(\lambda - t)^2}, & \mathbb{E}[X] &= \phi_X'(0) = \frac{1}{\lambda}, \\
\phi_X''(t) &= \frac{2\lambda}{(\lambda - t)^3}, & \mathbb{E}[X^2] &= \phi_X''(0) = \frac{2}{\lambda^2}, \\
\phi_X'''(t) &= \frac{3 \cdot 2\lambda}{(\lambda - t)^4}, & \mathbb{E}[X^3] &= \phi_X'''(0) = \frac{3!}{\lambda^3}.
\end{aligned}$$

From the first three we guess that in general,

$$\phi_X^{(n)}(t) = \frac{n!\lambda}{(\lambda - t)^{n+1}}, \quad \mathbb{E}[X^n] = \phi_X^{(n)}(0) = \frac{n!}{\lambda^n},$$

which one could prove rigorously by induction.

The moment generating function is defined for any random variable but if X is a nonnegative integer-valued random variable, it is often more convenient to look at its **probability generating function** defined by

$$\gamma_X(z) = \mathbb{E}[z^X] = \sum_{k=0}^\infty z^k \, \mathbb{P}(X = k),$$

for all $z \in \mathbb{R}$ for which this sum converges absolutely. Here we use the convention that $z^0 = 1$ even when $z = 0$. Note that if we set $z = e^t$ in the generating function we get the moment generating function,

$$\gamma_X(e^t) = \mathbb{E}[e^{tX}] = \phi_X(t).$$

Example 3.8.8 (Poisson distribution). If $\mathbb{P}(X = k) = e^{-\lambda}\lambda^k/k!$ for $k = 0, 1, 2, \ldots$ then the generating function is given by

$$\gamma_X(z) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!}z^k = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda z)^k}{k!} = e^{-\lambda}e^{\lambda z} = e^{\lambda(z-1)}.$$

Setting $z = e^t$ we get the moment generating function $\phi_X(t) = e^{\lambda(e^t-1)}$.

One of the reasons for the interest in the generating function is that it can also be used to compute moments. Let X be non-negative integer-valued random variable, let $\gamma_X(z) = \mathbb{E}[z^X]$ be its generating function, and let $\gamma_X^{(n)}$ be the n-th derivative of γ_X. Then,

$$\gamma_X^{(n)}(1) = \mathbb{E}[X(X-1)\cdots(X-n+1)],$$

provided γ_X is defined on some open ball containing 1. The quantity on the right-hand is sometimes called the n-th **factorial moment**.

Example 3.8.9. In the case of the Poisson distribution, $\gamma(z) = e^{\lambda(z-1)}$ so

$$\gamma'(z) = \lambda e^{\lambda(z-1)} \Rightarrow \gamma'(1) = \lambda.$$

Continuing to differentiate, we see that

$$\gamma''(z) = \lambda^2 e^{\lambda(z-1)} \Rightarrow \gamma''(1) = \lambda^2.$$

It should be clear from the first two formulae that, in general,

$$\gamma^{(n)}(z) = \lambda^n e^{\lambda(z-1)} \Rightarrow \gamma^{(n)}(1) = \lambda^n,$$

and so $\mathbb{E}[X(X-1)\cdots(X-n+1)] = \lambda^n$.

Moment generating function are important because they characterise distributions.

Theorem 3.8.10. Suppose X and Y have moment generating functions, $\phi_X(t)$ and $\phi_Y(t)$ which are defined (so finite) and equal for $t \in (-t_0, t_0)$ for some $t_0 > 0$. Then X and Y are equal in distribution.

3.8.1 A first look at conditional expectation

Definition 3.8.11 (Conditional mean). The **conditional mean** of Y given $X = x$ is just the mean of the conditional distribution,

$$\mathbb{E}[Y|X = x] = \sum_y y\,\mathbb{P}(Y = y|X = x),$$

$$\mathbb{E}[Y|X = x] = \int y f_{Y|X}(y|x)\,\mathrm{d}y.$$

provided $\mathbb{P}(X = x) > 0$ or $f_X(x) > 0$ are positive in the discrete and absolutely continuous cases, respectively. In the absolutely continuous case this definition is not completely unambiguous, but we do not go into such detail here. As an aside, however, note that f_X is only determined up to a set of Lebesgue measure zero.

If $h(x) = \mathbb{E}[Y|X = x]$ then $h(X)$ is a random variable usually denoted by $\mathbb{E}[Y|X]$. It tells us the conditional mean of Y for the value of X we have observed. This quantity has the following properties:

(i) $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$;

(ii) $\mathbb{E}[Y + Z|X] = \mathbb{E}[Y|X] + \mathbb{E}[Z|X]$;

(iii) $\mathbb{E}[h(X)Y|X] = h(X)\,\mathbb{E}[Y|X]$.

You will become very intimate with conditional expectations when studying the martingale approach to pricing derivatives!

3.9 Limit theorems for sums of independent random variables

Suppose $X_1, X_2, \ldots$ are independent and identically distributed random variables (lingo: we abbreviate this to **iid**). As mentioned above, the sample mean, $S_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ "should" converge to the common expectation $\mathbb{E}[X_i]$ of the X_i. In what sense is this true?

3.9.1 Weak law of large numbers

The first result is known as the **weak law of large numbers**.

Theorem 3.9.1 (Weak law of large numbers). Suppose $X_1, X_2, \ldots$ are iid random variables which have finite expectation $\mathbb{E}[X_i] = \mu$, and let

$$S_n = \frac{X_1 + \cdots + X_n}{n}.$$

Then $S_n \to \mu$ *in probability* as $n \to \infty$. That is, for any $\varepsilon > 0$,

$$\mathbb{P}(|S_n - \mu| > \varepsilon) \to 0, \text{ as } n \to \infty.$$

Proof 3.9.2. We can prove this easily when the variance

$$\sigma^2 = \mathrm{var}[X_i] = \mathbb{E}[(X_i - \mu)^2]$$

of the X_i is finite. The proof begins with **Markov's inequality**; for any non-negative random variable X with finite expectation,

$$\mathbb{P}(X \geq c) = \mathbb{E}[\mathbb{1}_{\{X \geq c\}}] \leq \frac{1}{c}\,\mathbb{E}[X],$$

where the indicator $\mathbb{1}_{\{X \geq c\}}$ is equal to 1 if $X \geq c$ and zero otherwise. Apply this inequality to the random variable $X = (Y - \mu)^2$, where Y is any random variable with finite variance to get

$$\mathbb{P}(|Y - \mu| \geq c) = \mathbb{P}((Y - \mu)^2 \geq c^2) \leq \frac{1}{c^2}\,\mathbb{E}[(Y - \mu)^2] = \frac{1}{c^2}\,\mathrm{var}[Y].$$

This new inequality is called **Chebyshev's inequality**. We will apply this with $Y = S_n$. Using linearity of the expectation operator, $\mathbb{E}[S_n] = \mu$, and as the X_i are all independent,

$$\mathrm{var}[S_n] = \frac{1}{n^2}\left(\mathrm{var}[X_1] + \ldots + \mathrm{var}[X_n]\right) = \frac{\sigma^2}{n},$$

(recall that $\mathrm{var}[aY] = a^2\,\mathrm{var}[Y]$ for constant a). Thus, Chebyshev's inequality gives

$$\mathbb{P}(|S_n - \mu| \geq \varepsilon) \leq \frac{1}{n\varepsilon^2}\sigma^2.$$

Of course, the right hand side converges to zero as $n \to \infty$ which concludes the proof. □

3.9.2 Strong law of large numbers

The weak law gives *convergence in probability* of the sample mean. A stronger form of convergence for the sample mean is *almost sure convergence*.

Theorem 3.9.3 (Strong law of large numbers). Suppose $X_1, X_2, \ldots$ are iid with finite mean $\mu = \mathbb{E}[X_i]$. Then, almost surely (with probability 1), the sample mean S_n converges to μ,

$$\mathbb{P}(S_n \to \mu \text{ as } n \to \infty) = 1.$$

In fact Etemadi shows it is sufficient in the above to have pairwise independence of the random variables $X_1, X_2, \ldots$, which is a weaker assumption than demanding their independence.

3.9.3 The central limit theorem

A most beautiful result in probability theory describes the situation when we 'zoom' in on the fluctuations of the sample mean around μ. Incredibly, we find they are normally distributed! This result is known as the **central limit theorem** and is stated as follows.

Theorem 3.9.4 (Central limit theorem). Suppose $X_1, X_2, \ldots$ are independent and identically distributed with $\mathbb{E}[X_i] = \mu$ and $\text{var}(X_i) = \sigma^2 \in (0, \infty)$. Then, the random variable

$$\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{S_n - \mu}{\sigma/\sqrt{n}}$$

converges *in distribution* to the standard normal as $n \to \infty$. That is, as $n \to \infty$,

$$\mathbb{P}\left(\frac{S_n - \mu}{\sigma/\sqrt{n}} \leq x\right) \to \mathbb{P}(Z \leq x),$$

where Z has density

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2).$$

A typical application of the central limit theorem is to approximate either the binomial or Poisson distribution by a normal distribution (see exercises).

Index