1VGLA

Vectors, Geometry & Linear Algebra

2021 – 2022

LECTURE NOTES

Chris Parker

Preamble

1. The lecture notes cover material presented in Semester 2 of the 20 credit module 1VGLA.

2. The notes are supported by a comprehensive week by week collection of recordings.

3. The material presented in these notes can be supplemented by reading the recommended chapters of the text book by Adams and Essex [2], online resources available via links on the American Institute of Mathematics website (such as [4] and [5]) or other recommended text books. Recommendations to specific resources are given at the beginning of each chapter.

4. In these notes, at the beginning of each chapter, the “Learning Outcomes" for that chapter are included. The definition of “Learning Outcomes", given below, is taken from the *Quality Assurance Agency* [6].

Definition 0.1. ***(Learning Outcomes)*** *“Learning outcomes" are statements of what a learner is expected to know, understand and/or be able to demonstrate after completion of a process of learning.*

- *Each module will have a number of formally identified learning outcomes where each outcome is expected to relate to a fundamental aspect of the module. The learning outcome informs the curriculum content and the assessment strategy for the module.*
- *The learning outcomes must be assessable and formal assessment criteria provided which tell the learners how it will be determined that the outcomes have been met.*
- *Normally the learner should satisfy all specific outcomes before credit is awarded. Quality or standard of performance may be reflected by the use of a grading system.*

Contents

Chapter 1

Vectors

▸ **Learning Outcomes** ◂ After the completion of the lectures and support sessions associated with this chapter you should be able to:

- understand basic properties vectors in 2 and 3 dimensions;
- use vectors to define and work with lines and planes in 2 and 3 dimensions; and
- understand scalar products and vector products and their applications.

[2, p.564-594] contains an alternative presentation of material in this chapter that you may find helpful.

Definition 1.1. *A **vector** is a mathematical object (thing) that has both magnitude and direction.*

For example, displacement, velocity and acceleration are all vector quantities as they have both magnitude and direction. Temperature is an example of a **scalar** object which has only magnitude.

In two and three dimensions, vectors are often represented as in Figure 1.1 by an initial point and a final point of a line segment.

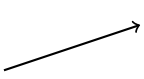


Figure 1.1: Here, the arrow's head indicates the direction of the vector and the length of the line represents the magnitude of the vector.

Suppose that P and Q are points (locations). By $\vec{PQ}$, we represent a vector in the direction P to Q with magnitude given by the distance between P and Q.

The **zero vector** is defined to be the vector which has magnitude 0. Since it has no length, it can't point in any particular direction (or it points in all directions at once). We choose to express this property as saying that it has no direction. We represent the zero vector by $\mathbf{0}$. We have

$$\mathbf{0} = \vec{PP}$$

for P any point.

Definition 1.2. *The set of vectors in three dimensions, denoted by E^3, is the set containing all such vectors described above (with $P, Q \in \mathbb{R}^3$) together with the zero vector. A similar definition applies for E^n, vectors in n-dimensions for $n \in \mathbb{N} \setminus \{1\}$.*

$$E^n = \{\vec{PQ} : P, Q \in \mathbb{R}^n\}.$$

Remember that a vector only has direction and magnitude and so $\vec{PQ} = \vec{RS}$ does not mean $P = R$ and $Q = S$. Rather, it only means that the direction from P to Q is the same as the direction from R to S and that P and Q are the same distance apart as R and S.

1.1 Notation and definitions

We will use lowercase bold letters e.g. $\mathbf{u}$, $\mathbf{a}$ and $\mathbf{v}$ to denote vectors (some other variations of this may appear). Sometimes, in typed text, underlined symbols is used.

Notation 1.3. *Suppose that $\mathbf{u}$ is a vector. Then $|\mathbf{u}|$ to denotes the magnitude of the vector $\mathbf{u}$.*

For all vectors $\mathbf{u}$, we have $|\mathbf{u}|$ is a non-negative real number.

Definition 1.4. *Let $\mathbf{u}$ and $\mathbf{v}$ be vectors. Then $\mathbf{u} = \mathbf{v}$ if and only if $\mathbf{u}$ and $\mathbf{v}$ have the same magnitude and direction.*

If $\mathbf{v}$ is a vector, $-\mathbf{v}$ is the unique vector which has magnitude the same as $\mathbf{v}$ and opposite direction. Hence if $\mathbf{v} = \vec{PQ}$, then $-\mathbf{v} = \vec{QP}$.

Definition 1.5. ***(Vector Addition)*** *Suppose that $\mathbf{u}$ and $\mathbf{v}$ are vectors. Choose point P, Q and R such that $\mathbf{u} = \vec{PQ}$ and $\mathbf{v} = \vec{QR}$. Then*

$$\mathbf{u} + \mathbf{v} = \vec{PR}.$$

With Definition 1.5, we obtain that vector addition is commutative and associative, that is for all vectors $\mathbf{u}, \mathbf{v}, \mathbf{v} \in E^n$,

$$(\mathbf{u} + \mathbf{v}) = (\mathbf{v} + \mathbf{u}), \tag{1.1}$$

$$(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w}). \tag{1.2}$$

Notation 1.6. ***(Vector Subtraction)*** *Vector subtraction is defined as follows:*

$$\mathbf{u} - \mathbf{v} = \mathbf{u} + (-\mathbf{v}).$$

Given a real number $\alpha \in \mathbb{R}$ and a vector $\mathbf{u}$ we can change the length of $\mathbf{u}$ by "scaling" it by α. For this reason, in this context, we call the elements of $\mathbb{R}$ **scalars**.

Definition 1.7. ***(Scalar Multiplication)*** *Given a vector* $\mathbf{v}$ *and a real number* α *(called a scalar), we define the* ***scalar multiple*** $\alpha\mathbf{v}$ *of* $\mathbf{v}$ *by:*

(1) If $\alpha > 0$*, then* $\alpha\mathbf{v}$ *has magnitude* $\alpha|\mathbf{v}|$ *and same direction as* $\mathbf{v}$*.*

(2) If $\alpha < 0$*, then* $\alpha\mathbf{v}$ *has magnitude* $|\alpha||\mathbf{v}|$ *and the same direction as* $-\mathbf{v}$*.*

(3) If $\alpha = 0$*, then* $\alpha\mathbf{v} = \mathbf{0}$*.*

If $\mathbf{u}$ and $\mathbf{v}$ are vectors and α, β are scalars, then it follows from Definition 1.7 that:

$$(\alpha + \beta)\mathbf{v} = \alpha\mathbf{v} + \beta\mathbf{v}, \tag{1.3}$$

$$\alpha(\mathbf{u} + \mathbf{v}) = \alpha\mathbf{u} + \alpha\mathbf{v}, \tag{1.4}$$

$$\alpha(\beta\mathbf{v}) = (\alpha\beta)\mathbf{v}. \tag{1.5}$$

Notice that in (1.3), the addition sign on the left-hand side represents addition of real numbers, whereas the addition sign on the right-hand side represents addition of vectors.

Definition 1.8. *A non-zero vector* $\mathbf{v}$ *is a* ***unit vector*** *if and only if*

$$|\mathbf{v}| = 1.$$

1.2 Parallel Vectors

Definition 1.9. *Non-zero vectors* $\mathbf{u}$ *and* $\mathbf{v}$ *in* E^n *are* ***parallel*** *(sometimes denoted* $\mathbf{u} \parallel \mathbf{v}$*) if and only if either*

1. *they have the same direction; or*
2. *they have the opposite direction.*

Thus non-zero the non-zero vectors that are parallel to $\mathbf{u}$ are

$$\{\alpha\mathbf{u} : \alpha \in \mathbb{R} \setminus \{0\}\}.$$

Three distinct points P, Q and R are **co-linear** if and only if $\vec{PQ}$ and $\vec{PR}$ are parallel, which is if and only if

$$\exists\ \alpha \in \mathbb{R} \setminus \{0\} \quad \text{such that} \quad \vec{PQ} = \alpha\vec{PR}.$$

Notice that, as $Q \neq R$, $\alpha \neq 1$.

1.3 Position Vectors

Definition 1.10. *Let* A *be a point in* $\mathbb{R}^n$ *and denote the origin of* $\mathbb{R}^n$ *by* O*. The position vector* $\mathbf{a}$ *of* A *(with respect to the origin* O*) is the vector* $\vec{OA}$*.*

If points A and B have position vectors $\mathbf{a}$ and $\mathbf{b}$ relative to an origin O, then:

$$\vec{AB} = \mathbf{b} - \mathbf{a}.$$

This follows since[1]

$$\begin{aligned}
\vec{OA} + \vec{AB} = \vec{OB} &\iff \vec{AO} + (\vec{OA} + \vec{AB}) = \vec{AO} + \vec{OB} && \text{(via Definition 1.5)} \\
&\iff (\vec{AO} + \vec{OA}) + \vec{AB} = \vec{AO} + \vec{OB} && \text{(via (1.2))} \\
&\iff \vec{AB} + (\vec{AO} + \vec{OA}) = \vec{OB} + \vec{AO} && \text{(via (1.1))} \\
&\iff \vec{AB} + (\vec{AO} - \vec{AO}) = \vec{OB} - \vec{OA} && \text{(via Definition 1.4)} \\
&\iff \vec{AB} + \mathbf{0} = \mathbf{b} - \mathbf{a} && \text{(via Definition 1.5 (2))} \\
&\iff \vec{AB} = \mathbf{b} - \mathbf{a} && \text{(via Definition 1.5 (3)).} \qquad (1.6)
\end{aligned}$$

[1] Typically we do not require this amount of detail when giving justification of a mathematical statement. We do so here for absolute clarity.

Example 1: Suppose A, B and C have distinct position vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ relative to O and that C is on the line segment joining A to B. Assume that the distance from A to C and the distance from C to B are in the ratio $\alpha : \beta$. Express $\mathbf{c}$ as a sum of scalar multiples of $\mathbf{a}$ and $\mathbf{b}$.

Answer: Since C divides AB internally, in the ratio $\alpha : \beta$, we have

$$\frac{|\vec{AC}|}{|\vec{CB}|} = \frac{\alpha}{\beta}$$

from which it follows that

$$\beta|\vec{AC}| = \alpha|\vec{CB}|. \tag{1.7}$$

As $\vec{AC}$ and $\vec{CB}$ have the same direction, it follows from (1.7) that

$$\begin{aligned} \beta\vec{AC} &= \alpha\vec{CB} \\ \beta(\mathbf{c}-\mathbf{a}) &= \alpha(\mathbf{b}-\mathbf{c}) \\ \alpha\mathbf{c}+\beta\mathbf{c} &= \alpha\mathbf{b}+\beta\mathbf{a} \\ (\alpha+\beta)\mathbf{c} &= \alpha\mathbf{b}+\beta\mathbf{a}. \end{aligned}$$

Therefore,

$$\mathbf{c} = \frac{\beta}{\alpha+\beta}\mathbf{a} + \frac{\alpha}{\alpha+\beta}\mathbf{b}.$$

SPECIAL CASE: If C is the midpoint of AB then $\alpha = \beta = 1$ and the position vector of the midpoint of AB is:

$$\frac{1}{2}(\mathbf{a}+\mathbf{b}).$$

1.4 Rectangular Cartesian Coordinate Systems

In a rectangular Cartesian[2] coordinate system in three dimensions, the reference framework consists of a fixed point called the origin (usually denoted by O) and three **mutually perpendicular** straight lines through O known as the coordinate axes. Very often the coordinate axes are labelled the x-axis, the y-axis and the z-axis. We usually think of the x- and y-axes as lying in a flat plane with the z-axis being vertical. Each point P in 3-dimensional space can then be identified by the means of an ordered triple (x_1, y_1, z_1) of real numbers (see Figure 1.2). The real number 'x_1' is the signed perpendicular distance of P from the plane containing the y- and z-axes (known as the yz-plane) and the real numbers y_1 and z_1 are the signed perpendicular distances of P from the zx-plane and the xy-plane respectively. In Figure 1.2 the xy-plane can be considered as the paper, and the z coordinate of the point P (namely z_1) is obtained by attaching a + or − to the perpendicular distance of P to the xy-plane according to whether or not the point P is above or below the paper, as indicated by the arrow on the z-axis.

[2]Named after René Descartes [10].

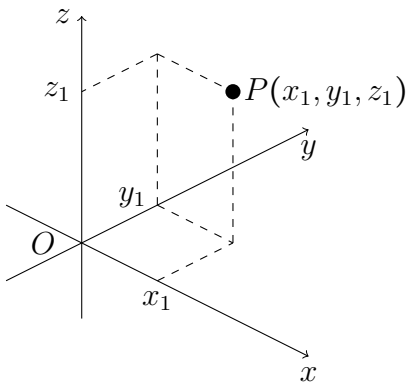


Figure 1.2: A representation of a 3-dimensional Cartesian system.

Similarly the arrows on the y- and z-axes indicate the signs that are to be attached to the perpendicular distances of P from the zx-plane and the xy-plane respectively to give the y and z coordinates of P. The rectangular Cartesian coordinate system $Oxyz$ illustrated in Figure 1.2 is an example of a **right-handed** Cartesian coordinate system. A rectangular coordinate system in three dimensions which is not right-handed is said to be **left-handed**.

We shall now show how vectors in 3-dimensional space as well as points in 3-dimensional space can be labelled by ordered triples of real numbers.

We shall arrange things in such a way that if (x_1, y_1, z_1) are the coordinates of a point P in a rectangular Cartesian coordinate system $Oxyz$, the position vector $\mathbf{p}$ (equal to $\vec{OP}$) of P relative to O is also labelled by the same ordered triple (x_1, y_1, z_1).

For ease of illustration, we shall consider a point $P = (x_1, y_1, z_1)$ whose coordinates are all positive.

Let $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ be three vectors whose directions are parallel to the x-axis, the y-axis and the z-axis respectively and in the direction of increasing x, increasing y, and increasing z respectively. Another way of saying this, is to say that $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are the position vectors relative to O of the points $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$ respectively. Note that the unit vectors $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ (in that order) form a right handed triad (of mutually perpendicular unit vectors).

Consider the rectangular block $OPQRSTUV$ as shown in Figure 1.3. Since $P = (x_1, y_1, z_1)$ with x_1, y_1 and z_1 all positive, we have $\vec{OR} = x_1\mathbf{i}$, $\vec{OS} = y_1\mathbf{j}$ and $\vec{OT} = z_1\mathbf{k}$.

Thus,

$$\begin{aligned} \mathbf{p} &= \vec{OP} \\ &= \vec{OR} + \vec{RP} \\ &= \vec{OR} + \vec{RU} + \vec{UP} \\ &= x_1\mathbf{i} + y_1\mathbf{j} + z_1\mathbf{k}. \end{aligned} \tag{1.8}$$

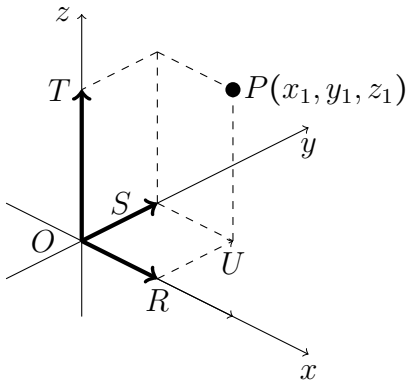


Figure 1.3: P in a 3-dimensional Cartesian system as a sum of three perpendicular vectors.

The numbers x_1, y_1 and z_1 in (1.8) are known as the components of $\mathbf{p}$ relative to the base $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$. It is usual to omit reference to the base $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$ and to simply write $\mathbf{p} = (x_1, y_1, z_1)$. We are therefore using $\mathbf{p} = (x_1, y_1, z_1)$ as notation for $\mathbf{p} = x_1\mathbf{i} + y_1\mathbf{j} + z_1\mathbf{k}$ i.e. we are representing the vector $\mathbf{p}$ by its components x_1, y_1 and z_1 relative to the base $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$. The components of the direction vector of the point P relative to O are the same as the coordinates of $\mathbf{p}$ in the rectangular Cartesian coordinate system $Oxyz$.

1.5 Uses of Components

Let $\mathbf{p} = (x_1, y_1, z_1)$ and $\mathbf{q} = (x_2, y_2, z_2)$ be vectors in 3-dimensional space relative to some base. Then:

(i) $\mathbf{p} + \mathbf{q} = (x_1 + x_2, y_1 + y_2, z_1 + z_2)$.

(ii) $\mathbf{p} - \mathbf{q} = (x_1 - x_2, y_1 - y_2, z_1 - z_2)$.

(iii) $\alpha\mathbf{p} = (\alpha x_1, \alpha y_1, \alpha z_1)$ for $\alpha \in \mathbb{R}$.

(iv) $|\mathbf{p}| = \sqrt{x_1^2 + y_1^2 + z_1^2}$. This is referred to as the Euclidean length of $\mathbf{p}$.

Note that for $x \in [0, \infty)$, $\sqrt{x}$ indicates the non-negative real number which when multiplied by itself, is equal to x.

Also, note that for E^n with $n \in \mathbb{N} \setminus \{1\}$, identities analogous to (i)-(iv) hold with the Euclidean length of $\vec{PQ}$ for initial point P and final point Q (where $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ and $\mathbf{q} = (q_1, q_2, \ldots, q_n)$), given by

$$|\vec{PQ}| = |\mathbf{q} - \mathbf{p}| = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

So, for example, in E^3 if $P = (x_1, y_1, z_1)$ and $Q = (x_2, y_2, z_2)$, the Euclidean length of $\vec{PQ}$ is given by:

$$|\vec{PQ}| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}.$$

Example 2: If $P = (2, 1, 3)$, $Q = (3, -2, 4)$ and $R = (0, 7, 5)$, then

$$\vec{PQ} = \vec{OQ} - \vec{OP} = \mathbf{q} - \mathbf{p} = (3, -2, 4) - (2, 1, 3) = (1, -3, 1) \text{ and } \vec{PR} = \vec{OR} - \vec{OP} = (-2, 6, 2).$$

Since $\vec{PR}$ is not a multiple of $\vec{PQ}$ it follows that P, R and Q are not co-linear.

Fact 1.11. *There are two unit-vectors parallel to any non-zero vector* $\mathbf{v}$ *in* E^n. *These two unit vectors are:*

$$\frac{1}{|\mathbf{v}|}\mathbf{v} \text{ and } -\frac{1}{|\mathbf{v}|}\mathbf{v}.$$

Example 3: Taking $\mathbf{v} = (1, -3, 4)$, then

$$|\mathbf{v}| = \sqrt{1^2 + 3^2 + 4^2} = \sqrt{26}.$$

Therefore Fact 1.11 implies there are exactly two unit vectors parallel to $\mathbf{v}$:

$$\left(\frac{1}{\sqrt{26}}, \frac{-3}{\sqrt{26}}, \frac{4}{\sqrt{26}}\right) \text{ and } \left(\frac{-1}{\sqrt{26}}, \frac{3}{\sqrt{26}}, \frac{-4}{\sqrt{26}}\right).$$

1.6 Vector Equation of a Line

Suppose that $\mathbf{q}$ is a non-zero vector and P is a point. We want to define a line L such that P is on the line L and if R is on the line L with $R \neq P$, then $\mathbf{q}$ and $\vec{PR}$ are parallel. This means that we want

$$\vec{PR} = \alpha\mathbf{q}$$

for some real number $\alpha \neq 0$. Now

$$\vec{PR} = \mathbf{r} - \mathbf{p}$$

hence

$$\mathbf{r} - \mathbf{p} = \alpha\mathbf{q}. \tag{1.9}$$

In (1.9), $\mathbf{r}$ and $\mathbf{p}$ are the position vectors of R and P respectively. It follows that

$$\mathbf{r} = \mathbf{p} + \alpha\mathbf{q}. \tag{1.10}$$

Equation (1.10) is the vector equation of the straight line that passes through the point P and which has the same direction as $\mathbf{q}$. Note that $\mathbf{q}$ is sometimes referred to as the **direction vector** of L and we will say that the line L is a **parallel** to $\mathbf{q}$.

Formally, the line L through the point P parallel to the vector $\mathbf{q}$ is defined as the set of points

$$\left\{R \in \mathbb{R}^3 : \mathbf{r} = \mathbf{p} + \alpha\mathbf{q} \text{ for } \alpha \in \mathbb{R}\right\}.$$

Example 4: Suppose that $P = (1, 2, 3)$ and $\mathbf{q} = (3, 2, 1)$. Then the line through P and parallel to $\mathbf{q}$ is

$$L = \{(1 + 3\alpha, 2 + 2\alpha, 3 + \alpha) : \alpha \in \mathbb{R}\}.$$

1.7 Scalar Product of Two Vectors

The scalar product of two vectors is also referred to as the 'dot product'.

Definition 1.12. *Given two vectors* $\mathbf{u}$ *and* $\mathbf{v}$ *in* E^3 *(or* E^2*), the* ***scalar product*** *is denoted by* $\mathbf{u} \cdot \mathbf{v}$ *and is defined as follows:*

(1) If $\mathbf{u} \neq 0$ *and* $\mathbf{v} \neq 0$*, then*

$$\mathbf{u} \cdot \mathbf{v} = |\mathbf{u}||\mathbf{v}| \cos(\theta_{uv})$$

where θ_{uv} *is the non-reflex angle between* $\mathbf{u}$ *and* $\mathbf{v}$*.*

(2) If $\mathbf{u} = \mathbf{0}$ *or* $\mathbf{v} = \mathbf{0}$*, then*

$$\mathbf{u} \cdot \mathbf{v} = 0.$$

In relation to Definition 1.12 we have the following comments:

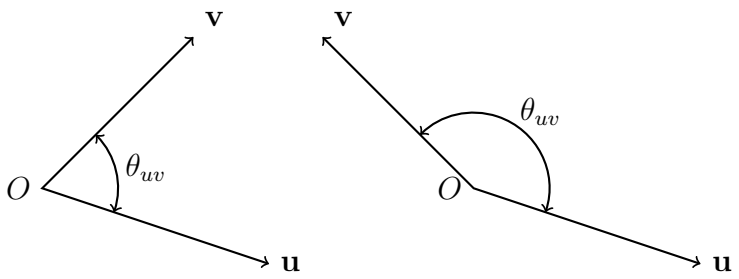


Figure 1.4: Illustration of the terms in the scalar product of two pairs of vectors in E^3.

1. If necessary, translate $\mathbf{u}$ or $\mathbf{v}$ to arrange the angle is as shown in Figure 1.4.

2. By **non-reflex** angle, we mean that $0 \leq \theta_{u,v} \leq \pi$.

3. If $\mathbf{u}$ and $\mathbf{v}$ point in the same direction, then $\theta_{u,v} = 0$ and so,

$$\mathbf{u} \cdot \mathbf{v} = |\mathbf{u}||\mathbf{v}| \cos(0) = |\mathbf{u}||\mathbf{v}|.$$

Similarly, if $\mathbf{u}$ and $\mathbf{v}$ point in **opposite directions**, then $\theta_{u,v} = \pi$ and so,

$$\mathbf{u} \cdot \mathbf{v} = |\mathbf{u}||\mathbf{v}| \cos(\pi) = -|\mathbf{u}||\mathbf{v}|.$$

4. If $\mathbf{u}$ and $\mathbf{v}$ are **perpendicular** (sometimes denoted $\mathbf{u} \perp \mathbf{v}$), then $\theta_{u,v} = \frac{\pi}{2}$, and so,

$$\mathbf{u} \cdot \mathbf{v} = |\mathbf{u}||\mathbf{v}| \cos\left(\tfrac{\pi}{2}\right) = 0.$$

Therefore, if $\mathbf{u} \cdot \mathbf{v} = 0$, then either

$$\mathbf{u} = \mathbf{0}, \ \mathbf{v} = \mathbf{0}, \text{ or } \mathbf{u} \text{ and } \mathbf{v} \text{ are perpendicular.} \tag{1.11}$$

5. Since $\mathbf{u} \cdot \mathbf{u} = |\mathbf{u}||\mathbf{u}| \cos(0) = (|\mathbf{u}|)^2$, we observe that

$$|\mathbf{u}| = \sqrt{\mathbf{u} \cdot \mathbf{u}}.$$

6. If $\mathbf{u}$ and $\mathbf{v}$ are vectors in E^2 (or E^3) and $\alpha \in (0, \infty)$ then,

$$\begin{aligned} \mathbf{u} \cdot (\alpha \mathbf{v}) &= |\mathbf{u}||\alpha \mathbf{v}| \cos(\theta_{u,\alpha v}) \\ &= \alpha |\mathbf{u}||\mathbf{v}| \cos(\theta_{u,v}) \\ &= \alpha(\mathbf{u} \cdot \mathbf{v}). \end{aligned}$$

More generally, for any $\alpha \in \mathbb{R}$ we have:

$$(\alpha \mathbf{v}) \cdot \mathbf{u} = \mathbf{v} \cdot (\alpha \mathbf{u}) = \alpha(\mathbf{v} \cdot \mathbf{u}).$$

7. If $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ are vectors in E^2 (or E^3), then the scalar product is **distributive** over addition of vectors, i.e.

$$\mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w} \quad \text{and} \quad (\mathbf{v} + \mathbf{w}) \cdot \mathbf{u} = \mathbf{v} \cdot \mathbf{u} + \mathbf{w} \cdot \mathbf{u}. \tag{1.12}$$

To see that (1.12) holds in E^2, see Figure 1.5.

Example 5: For vectors $\mathbf{i} = (1, 0, 0)$, $\mathbf{j} = (0, 1, 0)$ and $\mathbf{k} = (0, 0, 1)$, we have

$$\mathbf{i} \cdot \mathbf{i} = |\mathbf{i}||\mathbf{i}| \cos(0) = 1$$

$$\mathbf{j} \cdot \mathbf{j} = |\mathbf{j}||\mathbf{j}| \cos(0) = 1$$

$$\mathbf{k} \cdot \mathbf{k} = |\mathbf{k}||\mathbf{k}| \cos(0) = 1$$

$$\mathbf{i} \cdot \mathbf{j} = |\mathbf{i}||\mathbf{j}| \cos\left(\tfrac{\pi}{2}\right) = 0.$$

Similarly,

$$\mathbf{i} \cdot \mathbf{k} = \mathbf{j} \cdot \mathbf{k} = 0.$$

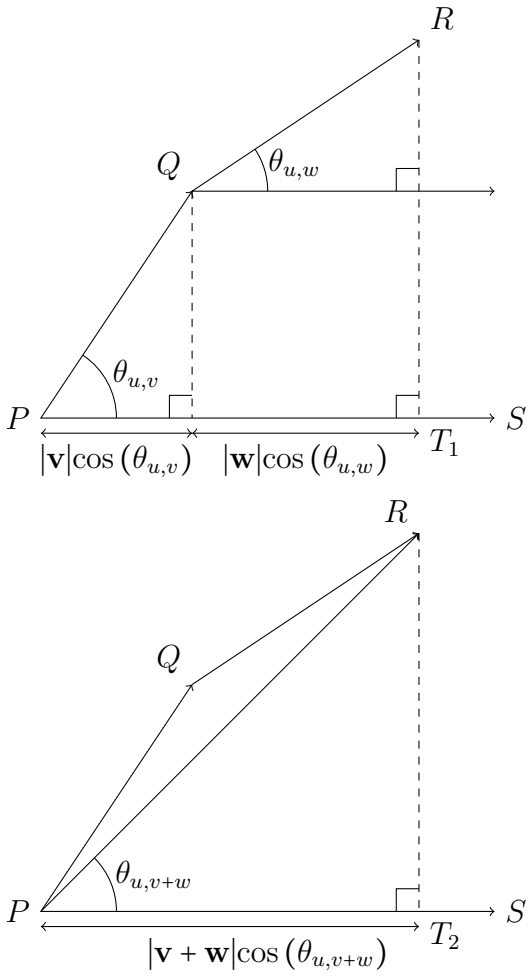


Figure 1.5: The distributive property in equations (1.12) in E^2 can be seen to hold from the diagrams above. Here $\mathbf{u} = \vec{PS}$, $\mathbf{v} = \vec{PQ}$ and $\mathbf{w} = \vec{QR}$. Observe that $|\vec{PT_1}|$ is given by $\frac{1}{|\mathbf{u}|}\mathbf{u}\cdot\mathbf{v} + \frac{1}{|\mathbf{u}|}\mathbf{u}\cdot\mathbf{w}$, $|\vec{PT_2}|$ is given by $\frac{1}{|\mathbf{u}|}\mathbf{u}\cdot(\mathbf{v}+\mathbf{w})$. As $|\vec{PT_1}| = |\vec{PT_2}|$, the distributive law follows. The result in E^3 follows similarly.

Proposition 1.13. *For any two vectors in E^3 given in Cartesian form,*

$$(x_1, y_1, z_1)\cdot(x_2, y_2, z_2) = x_1x_2 + y_1y_2 + z_1z_2.$$

Similarly for two vectors in E^2 given in Cartesian form,

$$(x_1, y_1)\cdot(x_2, y_2) = x_1x_2 + y_1y_2.$$

Proof: We provide a proof in E^3 with the proof in E^2 following similarly. Since for $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ in Example 5, we have for $\mathbf{u}, \mathbf{v} \in E^3$ that

$$\mathbf{u} = (x_1, y_1, z_1) = x_1\mathbf{i} + y_1\mathbf{j} + z_1\mathbf{k} \quad \text{and} \quad \mathbf{v} = (x_2, y_2, z_2) = x_2\mathbf{i} + y_2\mathbf{j} + z_2\mathbf{k}, \tag{1.13}$$

for scalars $x_i, y_i, z_i \in \mathbb{R}$ for $i = 1, 2$. Since the dot product distributes over addition (from 7. above) and is linear under scalar multiplication (from 6. above), it follows from (1.13) that

$$\begin{aligned}\mathbf{u}\cdot\mathbf{v} &= (x_1\mathbf{i} + y_1\mathbf{j} + z_1\mathbf{k})\cdot(x_2\mathbf{i} + y_2\mathbf{j} + z_2\mathbf{k})\\ &= x_1x_2\mathbf{i}\cdot\mathbf{i} + x_1y_2\mathbf{i}\cdot\mathbf{j} + x_1z_2\mathbf{i}\cdot\mathbf{k}\\ &\quad + y_1x_2\mathbf{j}\cdot\mathbf{i} + y_1y_2\mathbf{j}\cdot\mathbf{j} + y_1z_2\mathbf{j}\cdot\mathbf{k}\\ &\quad + z_1x_2\mathbf{k}\cdot\mathbf{i} + z_1y_2\mathbf{k}\cdot\mathbf{j} + z_1z_2\mathbf{k}\cdot\mathbf{k}\\ &= x_1x_2 + y_1y_2 + z_1z_2,\end{aligned}$$

via Example 5, as required.

Example 6: Find all $\lambda \in \mathbb{R}$ for which $\mathbf{u} = (7, 2, \lambda)$ is perpendicular to $\mathbf{v} = (1, -3, \lambda)$.

Answer: Via (1.11) $\mathbf{u}$ and $\mathbf{v}$ are non-zero, and perpendicular if and only if $\mathbf{u}\cdot\mathbf{v} = 0$. Observe that

$$\begin{aligned}\mathbf{u}\cdot\mathbf{v} &= (7, 2, \lambda)\cdot(1, -3, \lambda)\\ &= (7\mathbf{i} + 2\mathbf{j} + \lambda\mathbf{k})\cdot(1\mathbf{i} - 3\mathbf{j} + \lambda\mathbf{k})\\ &= 7\mathbf{i}\cdot\mathbf{i} - 6\mathbf{j}\cdot\mathbf{j} + \lambda^2\mathbf{k}\cdot\mathbf{k}\end{aligned}$$

where we note that all expressions involving $\mathbf{i}\cdot\mathbf{j}$ etc are equal to zero.

Hence $\mathbf{u} \perp \mathbf{v}$ if and only if

$$7 - 6 + \lambda^2 = 0 \quad (\lambda \in \mathbb{R}).$$

This quadratic equation has no real number solutions. Therefore, for every $\lambda \in \mathbb{R}$, $\mathbf{u}$ is not perpendicular to $\mathbf{v}$.

Using

$$\cos(\theta) = \frac{\mathbf{u}\cdot\mathbf{v}}{|\mathbf{u}||\mathbf{v}|}$$

we see that the scalar product encapsulates the idea of “angle" and the important idea of orthogonality.

Example 7: Find the (non-reflex) angle between $\mathbf{u} = (1, -2, 2)$ and $\mathbf{v} = (-2, 2, 1)$.

Answer: Here

$$\mathbf{u}\cdot\mathbf{v} = (1, -2, 2)\cdot(-2, 2, 1) = -2 - 4 + 2 = -4$$

and

$$|\mathbf{u}| = \sqrt{1^2 + (-2)^2 + 2^2} = 3,$$

which is the same as $|\mathbf{v}|$. Therefore,

$$\cos(\theta) = \frac{-4}{9},$$

and moreover, $\theta = \arccos\left(-\frac{4}{9}\right)$, as required.

We can check for the five key properties of the scalar product as an operation between two vectors:

1. The scalar product is not an internal operation, since it takes two vectors as input and outputs a real number, not a vector.

2. The scalar product is not associative. To clarify, for every $\mathbf{u}, \mathbf{v}, \mathbf{w} \in E^3$ the statement $(\mathbf{u} \cdot \mathbf{v}) \cdot \mathbf{w}$ is meaningless, since $\mathbf{u} \cdot \mathbf{v} \notin E^3$ and the scalar product requires both arguments to be (in this case) in E^3.

3. The scalar product has no identity, since the output of the product $\mathbf{a} \cdot \mathbf{b}$ can never be a vector, so definitively not $\mathbf{a}$.

4. Via 3., the scalar product cannot have an inverse.

5. The scalar product is commutative since $\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}$.

Example 8: Express the vector $\mathbf{w} = (2, 3, -1)$ in the form $\mathbf{w} = \mathbf{u} + \mathbf{v}$ where $\mathbf{u}$ is parallel to $(1, 1, 1)$ and $\mathbf{v}$ is perpendicular to $\mathbf{u}$.

Answer: Here we require $\mathbf{w} = \alpha(1, 1, 1) + \mathbf{v}$ where $\mathbf{v}$ is perpendicular to $\mathbf{u}$ i.e.

$$\mathbf{w} = (2, 3, -1) = \alpha(1, 1, 1) + \mathbf{v}. \tag{1.14}$$

Now

$$\mathbf{v} = (2, 3, -1) - \alpha(1, 1, 1) = (2 - \alpha, 3 - \alpha, -1 - \alpha) \tag{1.15}$$

and since $\mathbf{u}, \mathbf{v} \neq \mathbf{0}$ and $\mathbf{u} \cdot \mathbf{v} = 0$, we have

$$\begin{aligned} (1, 1, 1) \cdot (2 - \alpha, 3 - \alpha, -1 - \alpha) &= 0 \\ 2 - \alpha + 3 - \alpha - 1 - \alpha &= 0 \\ 4 - 3\alpha &= 0 \\ \alpha &= \frac{4}{3}. \end{aligned} \tag{1.16}$$

Substituting (1.16) and (1.15) into (1.14) gives

$$\mathbf{w} = \frac{4}{3}(1, 1, 1) + \frac{1}{3}(2, 5, -7),$$

as required.

Definition 1.14. *Given two vectors* $\mathbf{u}$ *and* $\mathbf{w}$*, the* ***projection of*** $\mathbf{w}$ ***onto*** $\mathbf{u}$ *is the vector parallel to* $\mathbf{u}$ *given by*

$$\text{proj}_{\mathbf{u}}(\mathbf{w}) = \left(\mathbf{w} \cdot \frac{\mathbf{u}}{|\mathbf{u}|}\right) \frac{\mathbf{u}}{|\mathbf{u}|}.$$

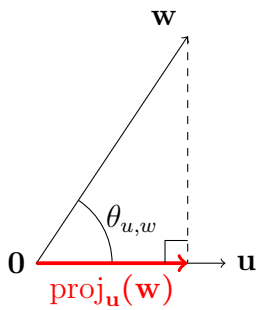


Figure 1.6: Geometric representation of the projection of $\mathbf{w}$ onto $\mathbf{u}$.

The projection onto a unit vector $\mathbf{e}$ simplifies to

$$\text{proj}_{\mathbf{e}}(\mathbf{w}) = (\mathbf{w} \cdot \mathbf{e})\ \mathbf{e}.$$

Example 9: Express the vector $\mathbf{w} = (2, 3, -1)$ as $\mathbf{w} = \mathbf{u} + \mathbf{v}$ where $\mathbf{u}$ is parallel to $(1, 1, 1)$ and $\mathbf{v}$ is perpendicular to $\mathbf{u}$.

Answer: We know that $\mathbf{u}$ is the projection of $\mathbf{w}$ onto $\mathbf{t} = (1, 1, 1)$. So we calculate

$$\mathbf{u} = \text{proj}_{\mathbf{t}}(\mathbf{w}) = \left(\frac{\mathbf{w} \cdot \mathbf{t}}{|\mathbf{t}|}\right)\frac{\mathbf{t}}{|\mathbf{t}|} = \left(\frac{(2, 3, -1) \cdot (1, 1, 1)}{\sqrt{3}}\right)\frac{1}{\sqrt{3}}(1, 1, 1) = \left(\frac{4}{3}, \frac{4}{3}, \frac{4}{3}\right).$$

Then $\mathbf{v}$ is given by

$$\mathbf{v} = \mathbf{w} - \mathbf{u} = (2, 3, -1) - \left(\frac{4}{3}, \frac{4}{3}, \frac{4}{3}\right) = \left(\frac{2}{3}, \frac{5}{3}, -\frac{7}{3}\right).$$

By construction, $\mathbf{v}$ is perpendicular to $\mathbf{t}$, and can be checked as follows:

$$\left(\frac{2}{3}, \frac{5}{3}, -\frac{7}{3}\right) \cdot (1, 1, 1) = \frac{2 + 5 - 7}{3} = 0.$$

Since $\mathbf{u}$ is parallel to $\mathbf{t}$ is follows that $\mathbf{v}$ is perpendicular to $\mathbf{u}$, as required.

1.8 General Right Handed Systems

Let $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ be three non co-planar vectors. Let $\theta_{u,v}$ be the non-reflex angle from $\mathbf{u}$ to $\mathbf{v}$.

Then $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ form a **right handed triad** or a right handed system. This system, as depicted in Figure 1.7 is sometimes referred to as “perturbed".

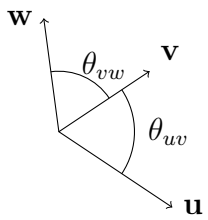


Figure 1.7: A perturbed right handed system.

1.9 Vector Product of Two Vectors

Definition 1.15. *The* ***vector product*** *of two vectors* $\mathbf{u}$ *and* $\mathbf{v}$ *in* E^3 *is denoted by* $\mathbf{u} \times \mathbf{v}$. *The definition is split into three cases:*

(1) If $\mathbf{u}, \mathbf{v} \neq \mathbf{0}$ *and* $\mathbf{u}$ *is not parallel to* $\mathbf{v}$. *Let* $\vec{OU}$ *and* $\vec{OV}$ *represent* $\mathbf{u}$ *and* $\mathbf{v}$ *respectively and let* θ *be the non reflex angle between* $\mathbf{u}$ *and* $\mathbf{v}$. *Then* $\mathbf{u} \times \mathbf{v}$ *is a vector with:*

(a) ***magnitude*** *given by,*

$$|\mathbf{u} \times \mathbf{v}| = |\mathbf{u}||\mathbf{v}| \sin\theta.$$

(b) ***direction perpendicular*** *to both* $\mathbf{u}$ *and* $\mathbf{v}$ *such that* $\mathbf{u}$, $\mathbf{v}$ *and* $\mathbf{u} \times \mathbf{v}$ *are a right handed triad.*

(2) If $\mathbf{u}$ *and* $\mathbf{v}$ *are non-zero parallel vectors, then*

$$\mathbf{u} \times \mathbf{v} = \mathbf{0}.$$

(3) If $\mathbf{u} = \mathbf{0}$ *or* $\mathbf{v} = \mathbf{0}$, *then*

$$\mathbf{u} \times \mathbf{v} = \mathbf{0}$$

and in particular,

$$\mathbf{0} \times \mathbf{0} = \mathbf{0}.$$

In relation to Definition 1.15 we have the following comments:

- Note that

$$\mathbf{u} \times \mathbf{v} \neq |\mathbf{u}||\mathbf{v}| \sin\theta.$$

  Geometrically, $\mathbf{u} \times \mathbf{v}$, defines a vector that is perpendicular to $\mathbf{u}$ and $\mathbf{v}$ with direction determined via the right hand rule which has length defined by the area $|\mathbf{u}||\mathbf{v}| \sin\theta$ of the parallelogram defined by $\mathbf{u} \times \mathbf{v}$, depicted in Figure 1.8.

- Also, importantly, it can be shown that the vector product is distributive over addition of vectors (see the end of the section) i.e. for any $\mathbf{u}, \mathbf{v}, \mathbf{w} \in E^3$, it follows

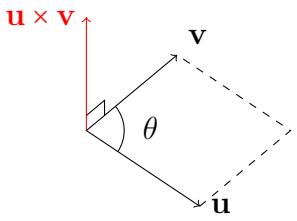


Figure 1.8: Depiction of Definition 1.15 (1). Note that the area of the parallelogram with sides $\mathbf{u}$ and $\mathbf{v}$ is $|\mathbf{u}||\mathbf{v}|\sin(\theta)$.

that

$$(\mathbf{u}+\mathbf{v})\times\mathbf{w} = (\mathbf{u}\times\mathbf{w}) + (\mathbf{v}\times\mathbf{w}).$$

- The vector product (as a result of the right hand rule) is anti-symmetric, i.e. for $\mathbf{u}, \mathbf{v} \in E^3$, it follows that

$$\mathbf{u}\times\mathbf{v} = -\mathbf{v}\times\mathbf{u}.$$

Example 10: Observe that since $\mathbf{i}$ is parallel to itself

$$\mathbf{i}\times\mathbf{i} = \mathbf{0}.$$

Now observe that since $\mathbf{i} \perp \mathbf{j} \perp \mathbf{k} \perp \mathbf{i}$ we have

$$\mathbf{i}\times\mathbf{j} = \mathbf{k},$$

$$\mathbf{j}\times\mathbf{k} = \mathbf{i},$$

$$\mathbf{k}\times\mathbf{i} = \mathbf{j}.$$

We can check for the five key properties of the vector product as an operation between two vectors:

1. The vector product is an internal operation, since it takes two vectors as input and outputs a vector.

2. The vector product is not associative. For example the vector triple product

$$\mathbf{i}\times(\mathbf{i}\times\mathbf{j}) = -\mathbf{j} \text{ and } (\mathbf{i}\times\mathbf{i})\times\mathbf{j} = \mathbf{0}.$$

   Hence to write $\mathbf{i}\times\mathbf{j}\times\mathbf{j}$ is meaningless (it is not clear which operation to perform first).

3. The vector product has no identity since the output of the product $\mathbf{u}\times\mathbf{v}$ is perpendicular to both $\mathbf{u}$ and $\mathbf{v}$, so can never be $\mathbf{u}$.

4. Since the vector product has no identity, the vector product cannot have an inverse.

5. The vector product is not commutative since $\mathbf{u}\times\mathbf{v} = -\mathbf{v}\times\mathbf{u}$. This follows since the non-reflex angle from $\mathbf{u}$ to $\mathbf{v}$ is the reverse of that of $\mathbf{v}$ to $\mathbf{u}$ i.e. $\mathbf{u}\times\mathbf{v}$ and $\mathbf{v}\times\mathbf{u}$ are in opposite directions with the same magnitude.

Example 11: Observe that

$$\mathbf{j}\times\mathbf{i} = -\mathbf{i}\times\mathbf{j} = -\mathbf{k},$$
$$\mathbf{k}\times\mathbf{j} = -\mathbf{i},$$
$$\mathbf{i}\times\mathbf{k} = -\mathbf{j} = -\mathbf{k}\times\mathbf{i}.$$

Proposition 1.16. ***(Vector product in Cartesian form)*** *Let*

$$\mathbf{u} = u_1\mathbf{i} + u_2\mathbf{j} + u_3\mathbf{k} \text{ and } \mathbf{v} = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}. \tag{1.17}$$

Then

$$\mathbf{u}\times\mathbf{v} = (u_2v_3 - u_3v_2)\mathbf{i} + (u_3v_1 - u_1v_3)\mathbf{j} + (u_1v_2 - u_2v_1)\mathbf{k}.$$

Proof: Since the vector product is distributive over vector addition, we have,

$$\begin{aligned}\mathbf{u}\times\mathbf{v} =& (u_1\mathbf{i} + u_2\mathbf{j} + u_3\mathbf{k})\times(v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k})\\ =& u_1v_1(\mathbf{i}\times\mathbf{i}) + u_1v_2(\mathbf{i}\times\mathbf{j}) + u_1v_3(\mathbf{i}\times\mathbf{k})\\ &+ u_2v_1(\mathbf{j}\times\mathbf{i}) + u_2v_2(\mathbf{j}\times\mathbf{j}) + u_2v_3(\mathbf{j}\times\mathbf{k})\\ &+ u_3v_1(\mathbf{k}\times\mathbf{i}) + u_3v_2(\mathbf{k}\times\mathbf{j}) + u_3v_3(\mathbf{k}\times\mathbf{k}).\end{aligned}$$

So, via Examples 10 and 11 it follows that

$$\mathbf{u}\times\mathbf{v} = (u_2v_3 - u_3v_2)\mathbf{i} + (u_3v_1 - u_1v_3)\mathbf{j} + (u_1v_2 - u_2v_1)\mathbf{k},$$

as required.

Another way to calculate the vector product of two vectors $\mathbf{u}$ and $\mathbf{v}$ is by using a method that is analogous to calculating the determinant of a 3×3 matrix (we will see this in Chapter 6).

Let $\mathbf{u} = (u_1, u_2, u_3)$ and $\mathbf{v} = (v_1, v_2, v_3)$ denote the vectors in (1.17). Then [3]

$$\mathbf{u}\times\mathbf{v} = \begin{vmatrix}\mathbf{i} & \mathbf{j} & \mathbf{k}\\ u_1 & u_2 & u_3\\ v_1 & v_2 & v_3\end{vmatrix}$$

[3] Formulas for determinants of 2×2 and 3×3 matrices are derived in Example 78.

$$\begin{aligned} &= \mathbf{i}\begin{vmatrix} u_2 & u_3 \\ v_2 & v_3 \end{vmatrix} - \mathbf{j}\begin{vmatrix} u_1 & u_3 \\ v_1 & v_3 \end{vmatrix} + \mathbf{k}\begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix} \\ &= (u_2v_3 - u_3v_2)\mathbf{i} + (u_3v_1 - u_1v_3)\mathbf{j} + (u_1v_2 - u_2v_1)\mathbf{k}. \end{aligned}$$

Example 12: Calculate $\mathbf{u} \times \mathbf{v}$ where $\mathbf{u} = (2, 1, -2)$ and $\mathbf{v} = (-4, 3, 1)$. Hence determine a unit vector perpendicular to $\mathbf{u}$ and $\mathbf{v}$.

Answer:

$$\begin{aligned} \mathbf{u} \times \mathbf{v} &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 2 & 1 & -2 \\ -4 & 3 & 1 \end{vmatrix} \\ &= \mathbf{i}\begin{vmatrix} 1 & -2 \\ 3 & 1 \end{vmatrix} - \mathbf{j}\begin{vmatrix} 2 & -2 \\ -4 & 1 \end{vmatrix} + \mathbf{k}\begin{vmatrix} 2 & 1 \\ -4 & 3 \end{vmatrix} \\ &= \mathbf{i}(1+6) - \mathbf{j}(2-8) + \mathbf{k}(6+4) \\ &= 7\mathbf{i} + 6\mathbf{j} + 10\mathbf{k}. \end{aligned}$$

Therefore $\mathbf{u} \times \mathbf{v} = 7\mathbf{i} + 6\mathbf{j} + 10\mathbf{k} = (7, 6, 10)$. Since $\mathbf{u} \times \mathbf{v}$ is perpendicular to $\mathbf{u}$ and $\mathbf{v}$ we set $\mathbf{w} = \mathbf{u} \times \mathbf{v}$ and calculate

$$|\mathbf{w}|^2 = 7^2 + 6^2 + 10^2 = 185.$$

Therefore a unit vector perpendicular to $\mathbf{u}$ and $\mathbf{v}$ is given by

$$\frac{7}{\sqrt{185}}\mathbf{i} + \frac{6}{\sqrt{185}}\mathbf{j} + \frac{10}{\sqrt{185}}\mathbf{k} = \frac{1}{\sqrt{185}}(7, 6, 10).$$

1.10 Equations of a Plane

1.10.1 Scalar Equation of a plane

The important vectors in relation to a scalar representation of a plane are those which are perpendicular to it. We call a vector which is perpendicular to a given plane a **normal vector** to the plane.

A non-zero vector $\mathbf{n}$ is a normal vector to a plane Π if every directed line segment representing $\mathbf{n}$ is perpendicular to Π. If $\mathbf{n}$ is a normal vector, then $\alpha\mathbf{n}$ will be a normal vector for $\alpha \in \mathbb{R} \setminus \{0\}$.

Suppose P is a fixed point in the plane Π and that R is another point in Π which have position vectors $\mathbf{p}$ and $\mathbf{r}$ respectively with respect to some origin O.

Suppose $\mathbf{n}$ is a normal vector to the plane Π and is represented by $\vec{PS}$. The vector $\vec{PR}$ is $\mathbf{r} - \mathbf{p}$ and is perpendicular to $\mathbf{n}$. Hence,

$$\mathbf{n} \cdot (\mathbf{r} - \mathbf{p}) = 0$$

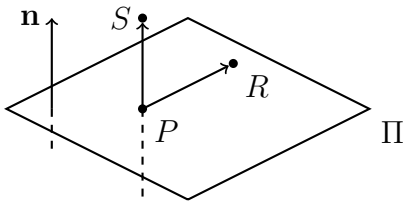


Figure 1.9: Depiction of the points P, R and S, and normal vector $\mathbf{n}$ to the plane Π.

i.e.

$$\mathbf{n} \cdot \mathbf{r} = \mathbf{n} \cdot \mathbf{p}. \tag{1.18}$$

Thus, a scalar equation for a plane Π which is perpendicular to the vector $\mathbf{n}$ is of the form

$$\mathbf{n} \cdot \mathbf{r} = d \tag{1.19}$$

for some constant d (i.e. every point $R \in \Pi$ with position vector $\mathbf{r}$ satisfies (1.19)).

So if we write $\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$, $\mathbf{n} = a\mathbf{i} + b\mathbf{j} + c\mathbf{k}$ and $\mathbf{p} = p_1\mathbf{i} + p_2\mathbf{j} + p_3\mathbf{k}$, then via (1.19),

$$(a\mathbf{i} + b\mathbf{j} + c\mathbf{k}) \cdot (x\mathbf{i} + y\mathbf{j} + z\mathbf{k}) = (a\mathbf{i} + b\mathbf{j} + c\mathbf{k}) \cdot (p_1\mathbf{i} + p_2\mathbf{j} + p_3\mathbf{k})$$

and hence

$$ax + by + cz = d,$$

with $d = p_1a + p_2b + p_3c$.

In relation to the scalar equation for a plane given in (1.18) and (1.19) we have the following comments:

(i) Since $\mathbf{n} \neq 0$, not all of a, b and c can be zero.

(ii) If $d = 0$, then the plane intersects the origin.

(iii) In 2-dimensional geometry, the equations $x = d$, $y = d$ and $ax + by = d$ all represent lines. However, in three dimensions, $ax + by = d$ represents a plane as well as $x = d$ and $y = d$.

(iv) If U, V and W are three non co-linear points in 3-dimensional space, there is exactly one plane containing all three of them.

Example 13: Determine the scalar equation of a plane, denoted by Π, that contains the points $U = (1, -2, 6)$, $V = (-2, -3, 4)$ and $W = (4, -1, 7)$.

Answer: Let $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ be position vectors of U, V and W. Since, $\vec{VU} = \mathbf{u} - \mathbf{v} = (3, 1, 2)$ and $\vec{VW} = \mathbf{w} - \mathbf{v} = (6, 2, 3)$, it follows that a normal vector to any plane that contains U, V and W is given by $\mathbf{n}$ with,

$$\mathbf{n} = \vec{VU} \times \vec{VW} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 3 & 1 & 2 \\ 6 & 2 & 3 \end{vmatrix}$$

$$\begin{aligned} &= (3-4)\mathbf{i} - (9-12)\mathbf{j} + (6-6)\mathbf{k} \\ &= -\mathbf{i} + 3\mathbf{j}. \end{aligned}$$

Thus, let $\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ be a point on Π. Since $U = (1, -2, 6)$ is on the plane Π, let $\mathbf{u} = \mathbf{i} - 2\mathbf{j} + 6\mathbf{k}$ be the position vector of U. Finally, via (1.18), the plane Π is determined by $\mathbf{n} \cdot \mathbf{r} = \mathbf{n} \cdot \mathbf{u}$, i.e.

$$-x + 3y = -1 - 6 \iff -x + 3y = -7. \tag{1.20}$$

So

$$\Pi = \{(x, y, z) \in \mathbb{R}^3 \mid -x + 3y + 7 = 0\}.$$

Note that to check the validity of the answer (1.20), substitution of $\mathbf{r}$ equal to U, V or W into (1.20) should satisfy the equation, since these 3 points should be contained within the plane. Additionally, one should also check that the equation in (1.20) is that of a plane (which it is).

1.10.2 Vector Equation of a plane*

The important vectors in relation to a vector representation of a plane are those which are parallel to it.

We can represent a plane in 3-dimensional space using a vector representation. By choosing a vector $\mathbf{p}$ on the plane and 2 non-colinear vectors parallel to the plane, $\mathbf{v}_1$ and $\mathbf{v}_2$, we can represent any other point $\mathbf{r}$ on the plane as

$$\mathbf{r} = \mathbf{p} + \lambda_1 \mathbf{v}_1 + \lambda_2 \mathbf{v}_2$$

for some $\lambda_1, \lambda_2 \in \mathbb{R}$.

Returning to Example 13 we have

$$\begin{aligned} \Pi &= \{R : \vec{OR} = \vec{OU} + \lambda_1 \vec{VU} + \lambda_2 \vec{VW} \text{ with } \lambda_1, \lambda_2 \in \mathbb{R}\} \\ &= \{(1, -2, 6) + \lambda_1(3, 1, 2) + \lambda_2(6, 2, 3) : \lambda_1, \lambda_2 \in \mathbb{R}\} \\ &= \{(1 + 3\lambda_1 + 6\lambda_2, -2 + \lambda_1 + 2\lambda_2, 6 + 2\lambda_1 + 3\lambda_2) : \lambda_1, \lambda_2 \in \mathbb{R}\}. \end{aligned}$$

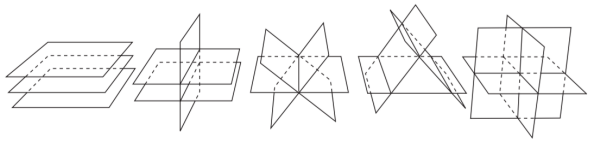
1.11 Intersection of Planes

Consider three distinct planes Π_1, Π_2 and Π_3. Then one of the following statements is true:

(i) The planes do not intersect. For example, consider the planes $x = 0$, $x = 1$ and $x = 2$. Clearly these 3 planes do not intersect (and are parallel).

(ii) The planes intersect at one point. For example, consider the planes $x = 0$, $y = 0$ and $z = 0$. Clearly these planes intersect at one point, $(0, 0, 0)$.

(iii) The planes intersect on a straight line. For example, consider the planes $z = 0$, $z = y$ and $z = -y$. The intersection of these 3 planes is the set $\{(x, 0, 0) : \ x \in \mathbb{R}\}$, alternatively known as the x-axis.



More-or-less all configurations of three planes.

1.12 Lines in Three Dimensions

Let $\mathbf{v}$ be a direction vector defined by a line L (this it not unique since although the direction is fixed up to sign of the vector, the length is not) and the position vector $\mathbf{p}$ correspond to a point P on L. Let R (with position vector $\mathbf{r}$) be any other point on the line L. Then,

$$\vec{PR} = \mathbf{r} - \mathbf{p} = \alpha\mathbf{v}$$

for some $\alpha \in \mathbb{R}$. Therefore a **vector equation** of L is:

$$\mathbf{r} = \mathbf{p} + \alpha\mathbf{v} \quad \alpha \in \mathbb{R}. \tag{1.21}$$

If $\mathbf{r} = (x, y, z)$, $\mathbf{p} = (p_1, p_2, p_3)$ and $\mathbf{v} = (v_1, v_2, v_3)$, then (1.21) is expressed as,

$$(x, y, z) = (p_1 + \alpha v_1, p_2 + \alpha v_2, p_3 + \alpha v_3). \tag{1.22}$$

Equation (1.22) allows ps to define the **parametric** equations for a line L in 3-dimensions:

$$x = p_1 + \alpha v_1, \ y = p_2 + \alpha v_2 \text{ and } z = p_3 + \alpha v_3,$$

with parameter $\alpha \in \mathbb{R}$. So we have:

$$\alpha = \frac{x - p_1}{v_1}, \quad \alpha = \frac{y - p_2}{v_2}, \quad \alpha = \frac{z - p_3}{v_3}$$

which implies

$$\frac{x - p_1}{v_1} = \frac{y - p_2}{v_2} = \frac{z - p_3}{v_3}. \tag{1.23}$$

(1.24) gives the equations for a line in **standard form** whenever v_1, v_2 and v_3 are non-zero.

In the case that say $v_2 = 0$, we may still write

$$\frac{x - p_1}{v_1} = \frac{z - p_3}{v_3} \text{ and } y = p_2. \tag{1.24}$$

Example 14: Find the vector equation, parametric form and standard form for the line L through the points $A = (3, 2, 0)$ and $B = (5, -2, 3)$.

Answer: Since $\mathbf{v} = \vec{AB} = (2, -4, 3)$ and for a point on L, we take A, it follows that $\mathbf{u} = (3, 2, 0)$. Then a **vector equation** of L is $\mathbf{r} = \mathbf{u} + \alpha\mathbf{v}$ i.e.

$$\mathbf{r} = (3, 2, 0) + \alpha(2, -4, 3)$$

where $\alpha \in \mathbb{R}$. Setting $\mathbf{r} = (x, y, z)$ implies

$$(x, y, z) = (3, 2, 0) + \alpha(2, -4, 3)$$

which can be written in **parametric form**:

$$x = 3 + 2\alpha, \ y = 2 - 4\alpha \text{ and } z = 3\alpha.$$

By eliminating $\alpha \in \mathbb{R}$, we have

$$\alpha = \frac{x - 3}{2} = \frac{2 - y}{4} = \frac{z}{3},$$

so in **standard form** L is given by,

$$\frac{x - 3}{2} = \frac{2 - y}{4} = \frac{z}{3}.$$

Example 15: Find where the line from the previous example intersects the plane Π with equation

$$x - 4y + 3z = 4.$$

Answer: We can use the parametric form for L, i.e.

$$x = 3 + 2\alpha, \ y = 2 - 4\alpha \text{ and } z = 3\alpha$$

for $\alpha \in \mathbb{R}$. The point on L intersects the plane Π if any only if,

$$\begin{aligned} (3 + 2\alpha) - 4(2 - 4\alpha) + 3(3\alpha) &= 4 \\ 3 + 2\alpha - 8 + 16\alpha + 9\alpha &= 4 \\ 27\alpha &= 9 \\ \alpha &= \frac{1}{3} \end{aligned}$$

i.e. the point of intersection between L and Π occurs (when $\alpha = \frac{1}{3}$) at

$$\left(\frac{11}{3}, \frac{2}{3}, 1\right).$$

Note, when considering problems which include various lines defined in parametric form, it is advisable to use a different parameter for each line i.e. σ, τ etc.

1.13 Perpendicular Distance From Point to Plane

Let a plane Π be given by

$$\mathbf{r} \cdot \mathbf{n} = d. \tag{1.25}$$

Let a point P have position vector $\mathbf{p}$ and consider a vector perpendicular to Π that connects Π to a point P' on the plane Π (P' has position vector $\mathbf{p}'$, as depicted in Figure 1.10). We wish to find $|\mathbf{p} - \mathbf{p}'|$ or alternatively, the distance of P to the the plane Π.

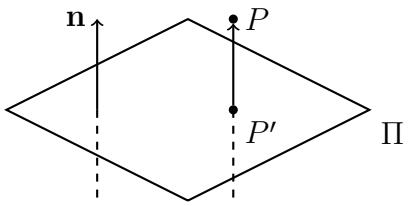


Figure 1.10: Depiction of the points P and P' in relation to the distance of P from the plane Π.

Since $\mathbf{n}$ is parallel to $\mathbf{p} - \mathbf{p}'$, for some $\alpha \in \mathbb{R}$

$$\mathbf{p} - \mathbf{p}' = \alpha \mathbf{n}. \tag{1.26}$$

Taking the scalar product with $\mathbf{n}$, via (1.26) we get

$$\mathbf{p} \cdot \mathbf{n} - \mathbf{p}' \cdot \mathbf{n} = \alpha |\mathbf{n}|^2,$$

and moreover, from (1.25) and $P' \in \Pi$, we conclude that

$$\mathbf{p} \cdot \mathbf{n} - d = \alpha |\mathbf{n}|^2.$$

Hence,

$$\alpha = \frac{\mathbf{p} \cdot \mathbf{n} - d}{|\mathbf{n}|^2}.$$

Then perpendicular distance from P to Π is

$$|\mathbf{p} - \mathbf{p}'| = \frac{|\mathbf{p} \cdot \mathbf{n} - d|}{|\mathbf{n}|^2} \, |\mathbf{n}| = \frac{|\mathbf{p} \cdot \mathbf{n} - d|}{|\mathbf{n}|}. \tag{1.27}$$

Perpendicular distance between two parallel planes: If in addition, P lies in a parallel plane (to Π), given by

$$\Pi' : \ \mathbf{r} \cdot \mathbf{n} = d',$$

then $\mathbf{p} \cdot \mathbf{n} = d'$ and hence the perpendicular distance between two parallel planes $\mathbf{n} \cdot \mathbf{v} = d$ and $\mathbf{n} \cdot \mathbf{v} = d'$ is

$$\frac{|d - d'|}{|\mathbf{n}|}.$$

Two lines in $\mathbb{R}^2$ or $\mathbb{R}^3$: Consider 2 distinct lines L_1 and L_2. Then at least one of the following statements is true:

1. L_1 and L_2 are parallel ($L_1 \parallel L_2$) and do not intersect.
2. L_1 and L_2 intersect.
3. L_1 and L_2 are not parallel and do not intersect (sometimes referred to as skew lines).

Angle between lines: Consider the lines defined by vector equations

$$L : \mathbf{r} = \mathbf{u} + \alpha\mathbf{v}$$

$$L' : \mathbf{r} = \mathbf{u}' + \beta\mathbf{v}'.$$

Then the acute angle θ between L and L' is given by

$$\cos\theta = \left| \frac{\mathbf{v} \cdot \mathbf{v}'}{|\mathbf{v}||\mathbf{v}'|} \right|.$$

1.14 Scalar Triple Product

Definition 1.17. *The scalar triple product of three vectors* $\mathbf{u}$, $\mathbf{v}$ *and* $\mathbf{w}$ *is*

$$\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}).$$

Note that the scalar triple product can be interpreted as the signed volume of a parallelepiped (a cuboid with every face a parallelogram), i.e.

$$|\mathbf{u}||\mathbf{v}||\mathbf{w}||\sin(\theta_{v,w})|\cos(\theta_{u,v\times w}) = \mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}).$$

Also note that the order of vectors defining the parallelepiped matters, due to the right hand rule used to define the vector product.

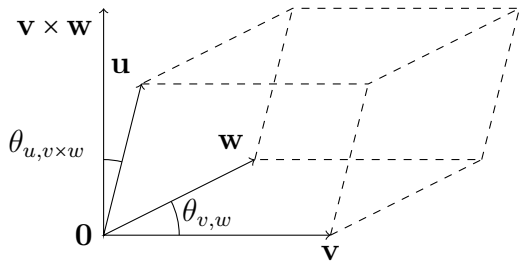


Figure 1.11: A parallelepiped, as described in Definition 1.17.

If we switch the labels for the vectors $\mathbf{v}$ and $\mathbf{w}$ in Figure 1.11 (see Figure 1.12), it follows that $\mathbf{v} \times \mathbf{w}$ is oriented in the opposite direction, which changes the sign of the scalar triple product.

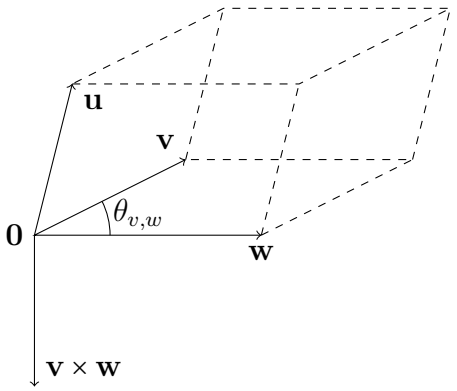


Figure 1.12: If we re-label **v** and **w** in Figure 1.11. Since the parallelogram defined by adjacent edges **v** and **w** has the same area it follows that the magnitude of $\mathbf{v} \times \mathbf{w}$ does not change. However, due to the right hand rule, the direction of $\mathbf{v} \times \mathbf{w}$ changes.

To calculate the volume of a parallelepiped with adjacent edges **u**, **v** and **w** one can simply compute the absolute value of $\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w})$. Note that if **u**, **v** and **v** have coordinates (u_1, u_2, u_3), (v_1, v_2, v_3) and (w_1, w_2, w_3) with respect to **i**, **j** and **k**, then it follows that[4]

$$\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) = (u_1, u_2, u_3) \cdot \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} = \begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix}.$$

Example 16: Calculate the volume of the parallelepiped with adjacent edges $\vec{OU}$, $\vec{OV}$ and $\vec{OW}$ with

$$\vec{OU} = \mathbf{u} = (3, 0, 0), \quad \vec{OV} = \mathbf{v} = (2, 1, 1) \text{ and } \vec{OW} = \mathbf{w} = (0, 0, -2).$$

Answer: The volume of the parallelepiped is given by the absolute value of

$$\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}) = (3, 0, 0) \cdot \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 2 & 1 & 1 \\ 0 & 0 & -2 \end{vmatrix} = \begin{vmatrix} 3 & 0 & 0 \\ 2 & 1 & 1 \\ 0 & 0 & -2 \end{vmatrix} = 3.1.(-2) = -6,$$

and hence the volume of the parallelepiped is 6.

1.15 An explanation of the distributive rule for the vector product

This section contains material which is additional to the course.

[4] At this stage we have not covered determinants (see Chapter 6). Please review this section after we have covered Chapter 6, and this calculation will be clearer.

To show that for any $\mathbf{u}, \mathbf{v}, \mathbf{w} \in E^3$, it follows that

$$(\mathbf{u} + \mathbf{v}) \times \mathbf{w} = (\mathbf{u} \times \mathbf{w}) + (\mathbf{v} \times \mathbf{w}),$$

we give the following justification.

Let $\mathbf{v}$ be a non-zero vector and Π be the plane perpendicular to $\mathbf{v}$ that intersects $\mathbf{0}$. Moreover, let $\mathbf{u}^*$ be the projection of $\mathbf{u}$ onto Π.

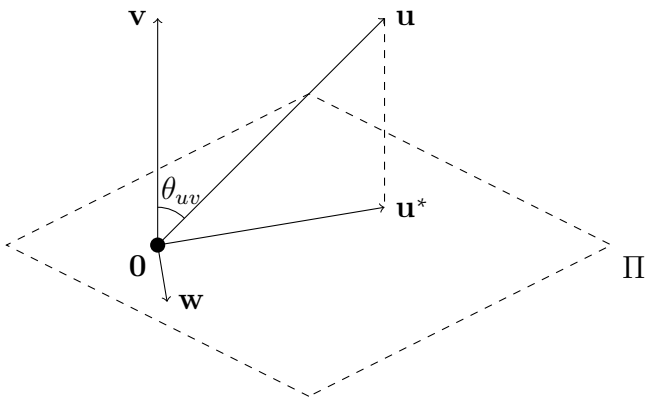


Figure 1.13: Illustration of $\mathbf{u}$, $\mathbf{v}$, $\mathbf{u}^*$ and Π. Here $\mathbf{w} \parallel (\mathbf{u} \times \mathbf{v})$ and $\mathbf{w} \parallel (\mathbf{u}^* \times \mathbf{v})$.

Lemma 1.18. $\mathbf{u} \times \mathbf{v} = \mathbf{u}^* \times \mathbf{v}$.

Proof: Let θ_{uv} be the angle between $\mathbf{u}$ and $\mathbf{v}$. Then $|\mathbf{u}^*| = |\mathbf{u}| \sin(\theta_{uv})$. Since $\mathbf{u}^* \perp \mathbf{v}$ we have,

$$\begin{aligned} |\mathbf{u} \times \mathbf{v}| &= |\mathbf{u}||\mathbf{v}| \sin(\theta_{uv}) \\ &= |\mathbf{v}||\mathbf{u}^*| \\ &= |\mathbf{v}||\mathbf{u}^*| \sin(\pi/2) \\ &= |\mathbf{u}^* \times \mathbf{v}|. \end{aligned} \tag{1.28}$$

Since $\mathbf{u}$, $\mathbf{u}^*$ and $\mathbf{v}$ are coplanar (see Figure 1.13), it follows from the right hand rule that $\mathbf{u}^* \times \mathbf{v}$ and $\mathbf{u} \times \mathbf{v}$ point in the same direction. Therefore, via (1.28), we conclude that

$$\mathbf{u} \times \mathbf{v} = \mathbf{u}^* \times \mathbf{v},$$

as required.

Lemma 1.19. $(\mathbf{u}_1 + \mathbf{u}_2)^* = \mathbf{u}_1^* + \mathbf{u}_2^*$.

Proof: Since

$$\mathbf{u}_i - \frac{\mathbf{v}}{|\mathbf{v}|}|\mathbf{u}_i|\cos(\theta_{vu_i}) = \mathbf{u}_i^*$$

for $i = 1, 2$, it follows from the distributivity over addition of the scalar product (depicted in Figure 1.5) that

$$\begin{aligned}
(\mathbf{u}_1 + \mathbf{u}_2)^* &= (\mathbf{u}_1 + \mathbf{u}_2) - \left(\frac{\mathbf{v}}{|\mathbf{v}|}\right)|\mathbf{u}_1 + \mathbf{u}_2|\cos(\theta_{u_{12}v}) \\
&= (\mathbf{u}_1 + \mathbf{u}_2) - \left(\frac{\mathbf{v}}{|\mathbf{v}|}\right)\left(\frac{\mathbf{v}}{|\mathbf{v}|}\cdot(\mathbf{u}_1 + \mathbf{u}_2)\right) \\
&= (\mathbf{u}_1 + \mathbf{u}_2) - \left(\frac{\mathbf{v}}{|\mathbf{v}|}\right)\left(\frac{\mathbf{v}}{|\mathbf{v}|}\cdot\mathbf{u}_1 + \frac{\mathbf{v}}{|\mathbf{v}|}\cdot\mathbf{u}_2\right) \\
&= (\mathbf{u}_1 + \mathbf{u}_2) - \left(\frac{\mathbf{v}}{|\mathbf{v}|}\right)(|\mathbf{u}_1|\cos(\theta_{vu_1}) + |\mathbf{u}_2|\cos(\theta_{vu_2})) \\
&= \left(\mathbf{u}_1 - \left(\frac{\mathbf{v}}{|\mathbf{v}|}\right)|\mathbf{u}_1|\cos(\theta_{vu_1})\right) + \left(\mathbf{u}_2 - \left(\frac{\mathbf{v}}{|\mathbf{v}|}\right)|\mathbf{u}_2|\cos(\theta_{vu_2})\right) \\
&= \mathbf{u}_1^* + \mathbf{u}_2^*,
\end{aligned}$$

as required.

Lemma 1.20. $(\mathbf{u}_1 + \mathbf{u}_2) \times \mathbf{v} = (\mathbf{u}_1 \times \mathbf{v}) + (\mathbf{u}_2 \times \mathbf{v})$.

Proof: Since $\mathbf{u} \times \mathbf{v} = \mathbf{u}^* \times \mathbf{v}$, via Lemma 1.18, it follows that $\mathbf{u} \times \mathbf{v}$ can be determined by rotating $\mathbf{u}^*$ by $\pi/2$ radians in the appropriate direction and multiplying the resulting vector by $|\mathbf{v}|$ (see Figure 1.14).

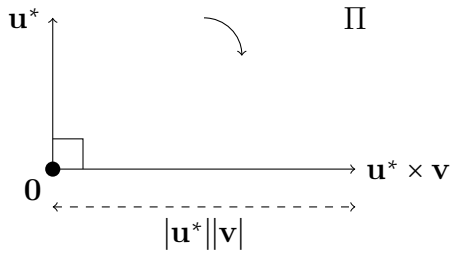


Figure 1.14: Picture of Π with $\mathbf{v}$ pointing straight out of the page towards you. Here, $\mathbf{u}^* \times \mathbf{v}$ is determined by rotating Π clockwise and rescaling. Recall $\mathbf{u}^* \perp \mathbf{v}$.

Therefore, from vector addition in E^2 (see Figure 1.15) it follows that

$$(\mathbf{u}_1^* + \mathbf{u}_2^*) \times \mathbf{v} = (\mathbf{u}_1^* \times \mathbf{v}) + (\mathbf{u}_2^* \times \mathbf{v}). \tag{1.29}$$

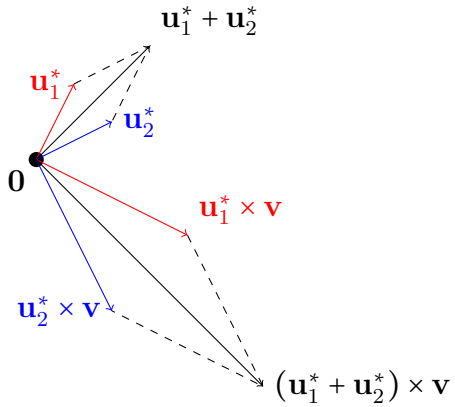


Figure 1.15: Depiction of (1.29).

Finally, we have,

$$\begin{aligned}
(\mathbf{u}_1 + \mathbf{u}_2) \times \mathbf{v} &= (\mathbf{u}_1 + \mathbf{u}_2)^* \times \mathbf{v} && \text{(via Lemma 1.18)} \\
&= (\mathbf{u}_1^* + \mathbf{u}_2^*) \times \mathbf{v} && \text{(via Lemma 1.19)} \\
&= (\mathbf{u}_1^* \times \mathbf{v}) + (\mathbf{u}_2^* \times \mathbf{v}) && \text{(via (1.29))} \\
&= (\mathbf{u}_1 \times \mathbf{v}) + (\mathbf{u}_2 \times \mathbf{v}) && \text{(via Lemma 1.18),}
\end{aligned}$$

as required.

Note that $\mathbf{v} \times (\mathbf{u}_1 + \mathbf{u}_2) = (\mathbf{v} \times \mathbf{u}_1) + (\mathbf{v} \times \mathbf{u}_2)$ follows from Lemma 1.20 and anti-symmetry of the vector product.

Chapter 2

Groups and Fields

The main purpose of this chapter is to introduce some of the basic algebraic objects and concepts which underlie modern mathematics.

▸ **Learning Outcomes** ◂ After the completion of the lectures and support sessions associated with this chapter you should be able to:

- check that an operation is an internal binary operation;
- determine whether a set with a binary operation is a group; and
- determine if a set with two binary operations is a field.

2.1 Binary operation

Consider a set V, which contains elements that can be combined by some "operation".

Definition 2.1. *A **binary operation** on a set V to a set C is a function*

$$B : V \times V \to C.$$

For $x_1, x_2 \in V$, we often denote the operation $B(x_1, x_2)$ by $x_1 * x_2$ or by some other appropriate symbol.

Examples 17:

- $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$: '$+$' and '$\times$' are binary operations on all of these sets, but '$\div$' is generally not (unless we remove 0 from the sets which contain it).

- E^3: '+', '$\cdot$', '$\times$'. All of these operations are binary operations on pairs of vectors in E^3 the first and the last have $C = E^3$ and the middle one has $C = \mathbb{R}$. ⌟

Definition 2.2. *An **internal binary operation** on a set V is a binary operation for which $C = V$, or equivalently,*

$$x_1 * x_2 \in V \quad \forall x_1, x_2 \in V.$$

*If a binary operation is internal, then the set V is said to be **closed** under the binary operation.*

Examples 18:

- $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$: '+' and '$\times$' are internal binary operations for each of these sets, but not '$\div$' (since it is not necessarily a binary operation and for $\mathbb{N}$ ... it is clearly not an internal binary operation).
- E^3: '+' and '$\times$' are internal binary operations on E^3, but '$\cdot$' is not, since $\mathbf{v}_1 \cdot \mathbf{v}_2 \in \mathbb{R} \notin E^3$ for all $\mathbf{v}_1, \mathbf{v}_2 \in E^3$. ⌟

Definition 2.3. *A binary operation on a set V is **commutative** if and only if for all $x_1, x_2 \in V$,*

$$x_1 * x_2 = x_2 * x_1.$$

Definition 2.4. *A binary operation on a set V is **associative** if and only if for all $x_1, x_2, x_3 \in V$,*

$$(x_1 * x_2) * x_3 = x_1 * (x_2 * x_3).$$

Definition 2.5. *An element $e \in V$ is called an **identity** for a binary operation on a set V if and only if for all $x_1 \in V$,*

$$e * x_1 = x_1 * e = x_1.$$

Theorem 2.6. *If a binary operation on a set V has an identity, then the identity is unique.*

Proof: Assume there are two elements e_1 and e_2 in V which are both identities. Then, $e_1 * e_2 = e_2$ since e_1 is an identity. However, $e_1 * e_2 = e_1$ since e_2 is an identity. Hence, $e_1 = e_2$, as required.

Definition 2.7. *Consider a binary operation on a set V which has identity $e \in V$. An element $a \in V$ has an **inverse** if there exists an element $b \in V$ such that*

$$a * b = b * a = e.$$

Example Problem 19: Check whether the following sets with binary operations satisfy the properties in Definitions 2.2 (closed), 2.3 (commutative), 2.4 (associative), 2.5 (identity element) and 2.7 (inverse elements).

- $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$: '$+$' and '$\times$'.
- E^3: '$+$', '$\times$'. ⌟

2.2 Groups and Fields

Special names exist for sets with a binary operation that satisfy a given number of properties. The first structure we define below is that of a **group** which has been developed into an important branch of modern mathematics called *Group Theory*.

Definition 2.8. *A set V endowed with an internal binary operation $*$ is called a **group** with respect to the binary operation $*$ if and only if the following hold:*

1. *$*$ is associative;*
2. *$*$ has an identity V; and*
3. *Every element of V have an inverse element.*

Example Problem 20: Decide which of the following sets with binary operations are groups and give a justification if they are not:

- $(\mathbb{Z}, +)$; $(\mathbb{Q}, +)$; $(\mathbb{R}, +)$.
- $(\mathbb{Q} \setminus \{0\}, \times)$; $(\mathbb{R} \setminus \{0\}, \times)$.
- $(E^3, +)$.

Definition 2.9. *A group with binary operation $*$ is called **abelian** if and only if $*$ is commutative.*

These groups are named after Neils Henrik Abel, a famous Norwegian mathematician after which, the Abel prize[1] is named. This prize, is one of the most prestigious in mathematics.

It is possible to define more than one binary operation on a set. If one does, the combination of properties of each binary operation gives rise to other structures. The most important one is a **field**:

Definition 2.10. *A set F with binary operations $+$ and $\times$ is called a* ***field*** $(F, +, \times)$ *if and only if:*

1. $(F, +)$ *is an abelian group, with identity denoted by* 0*;*
2. $(F \setminus \{0\}, \times)$ *is an abelian group, with identity denoted by* 1*;*
3. *for all* $x \in F$, $0 \times x = x \times 0 = 0$*; and*
4. ***distributivity*** *of* $\times$ *with respect to* $+$*: for all* $x_1, x_2, x_3 \in F$*, we have*

$$x_1 \times (x_2 + x_3) = (x_1 \times x_2) + (x_1 \times x_3).$$

Example 21: The following two sets with binary operations defining addition and multiplication are fields: $(\mathbb{Q}, +, \times)$, $(\mathbb{R}, +, \times)$.

We will often work with the field $(\mathbb{R}, +, \times)$ of real numbers. Let's just check it is a field.

1. $(\mathbb{R}, +)$ is an abelian group:
   (a) $\mathbb{R}$ is closed under $+$ since $\forall\ a, b \in \mathbb{R},\ a + b \in \mathbb{R}$;
   (b) $+$ is associative since $\forall\ a, b, c \in \mathbb{R},\ (a + b) + c = a + (b + c)$;
   (c) $+$ has an identity since $\forall\ a \in \mathbb{R}, 0 \in \mathbb{R}$ and $0 + a = a + 0 = a$;
   (d) existence of inverses. $\forall\ a \in \mathbb{R}, \exists\ b \in \mathbb{R}$ such that $a + b = b + a = 0$;
   (e) $+$ is commutative since $\forall\ a, b \in \mathbb{R}$, $a + b = b + a$.
2. $(\mathbb{R} \setminus \{0\}, \times)$ is an abelian group:
   (a) $(\mathbb{R} \setminus \{0\})$ is closed under $\times$ since $\forall\ a, b \in \mathbb{R} \setminus \{0\}$, $a \times b \in \mathbb{R} \setminus \{0\}$;
   (b) $\times$ is associative since $\forall\ a, b, c \in \mathbb{R} \setminus \{0\}$, $(a \times b) \times c = a \times (b \times c)$;
   (c) $\times$ has an identity since $\forall\ a \in \mathbb{R} \setminus \{0\}$, $1 \in \mathbb{R} \setminus \{0\}$ and $1 \times a = a \times 1 = a$;
   (d) existence of inverses. $\forall\ a \in \mathbb{R} \setminus \{0\}$, $\exists\ b \in \mathbb{R} \setminus \{0\}$ such that $a \times b = b \times a = 1$;
   (e) $\times$ is commutative since $\forall\ a, b \in \mathbb{R} \setminus \{0\}$, $a \times b = b \times a$.
3. $\forall\ a \in \mathbb{R}$, $0 \times a = a \times 0 = 0$; and

[1] For more details, see [12].

4. Distributivity: $\forall\ a, b, c \in \mathbb{R},\ a \times (b + c) = (a \times b) + (a \times c)$. ⌟

There are many different types of fields. You will see how to make some next year if you take 2AC.

Example 22: Suppose that $\mathbb{F}_2 = \{0, 1\}$. Define $+$ to be the internal binary operation which makes $(\mathbb{F}_2, +)$ into an abelian group with identity 0 and $1 + 1 = 0$. Define an internal binary operation $\times$ on $\mathbb{F}_2$ so that $(\mathbb{F}_2 \setminus \{0\}, \times)$ is a group with identity 1. Then $(\mathbb{F}_2, +, \times)$ is a field with 2-elements.

Example 23: Suppose that Ω is a non-empty set. Let $\mathrm{Sym}(\Omega)$ be the set of all bijections from Ω to Ω. Define a binary operation on $\mathrm{Sym}(\Omega)$ by composition of functions. Explain why $\mathrm{Sym}(\Omega)$ is a group. We call $\mathrm{Sym}(\Omega)$ the **symmetric group** on Ω.

Chapter 3

Complex Numbers

▸ **Learning Outcomes** ◂ After the completion of the lectures and support sessions associated with this chapter you should be able to:

- perform basic arithmetic operations (addition, subtraction, multiplication and division) on complex numbers;
- define and calculate the complex conjugate of a complex number and be able to cite the basic properties involving the complex conjugate;
- plot complex numbers on an Argand diagram;
- define and calculate the modulus-argument form of a complex number and perform basic arithmetic operations on complex numbers using the modulus-argument form;
- cite, prove and use De Moivre's theorem;
- calculate the n-th roots of complex numbers;
- cite and apply properties related to polynomials with real coefficients;
- solve quadratic equations, both with real and complex coefficients; and
- solve basic inequalities in $\mathbb{C}$.

[13, Ch.1] and [8, Ch.1] contain alternative presentations of material in this chapter that you may find helpful.

3.1 Introduction

You should be familiar with the following sets of numbers:

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}. \tag{3.1}$$

All these sets are *closed* under addition and multiplication (+ and × are internal operations on V). We mostly suppress × and use juxtaposition ir $\cdot$ whenever we can without confusion. So for any set V as in (3.1) we have:

(i) If $a \in V$ and $b \in V$ then $a + b \in V$.

(ii) If $a \in V$ and $b \in V$ then $ab \in V$.

One can imagine that the need for expanding a set of numbers is motivated by the desire to solve problems which cannot be solved in a particular set of numbers:

- In $\mathbb{N}$, the equation $7 + x = 3$ has no solution. In $\mathbb{Z}$ this equation can be solved.
- In $\mathbb{Z}$, the equation $7x = 3$ has no solution. In $\mathbb{Q}$ this equation can be solved.
- In $\mathbb{Q}$, the equation $x^2 = 2$ has no solution. A set in which this equation can be solved is $\mathbb{R}$ which includes $\sqrt{2}$.

To get all of the real numbers, we also need to add limits so that numbers such as π can be defined.

Complex numbers are the extension of $\mathbb{R}$ which is required to solve general polynomial equations, for instance $x^2+1 = 0$. This equation has no real solutions, and as a consequence, we also have cases where the quadratic equation $ax^2 + bx + c = 0$ cannot be solved (for $x \in \mathbb{R}$). However every polynomial equation can be solved for values in $\mathbb{C}$.

3.2 $\mathbb{C}$ - the field of complex numbers,

Following the historic process of expanding the number set to allow the solution of given problems, we will define a new type of number which solves the equation $x^2 = -1$:

Definition 3.1. *The imaginary number ' i ' is defined such that* $i^2 = -1$.

We can then obtain the set of all purely imaginary numbers as the product of i with all real numbers, e.g., $2i, -3.7i, \sqrt{3}i, \ldots$. The set of all purely imaginary numbers can be written as $\mathbb{R}i$.

In the set $\mathbb{R} \cup \mathbb{R}i$, all equations of the form $x^2 = a$, with $a \in \mathbb{R}$ can be solved. But how do we add a real number and a purely imaginary number? To allow for the definition of this addition, we need to expand the set further to the set of complex numbers:

Definition 3.2. *The set of* ***complex numbers****, denoted* $\mathbb{C}$*, is defined as*

$$\mathbb{C} = \{a + bi : a \in \mathbb{R} \text{ and } b \in \mathbb{R}\}.$$

It follows from Definition 3.2 that we can view $\mathbb{R} \subset \mathbb{C}$ and $\mathbb{R}i \subset \mathbb{C}$.

Example 24:

- $3 + 2i$, $7 - 7i$, $-3 + i$ and $-5 - 2i$ are in $\mathbb{C}$ but not in $\mathbb{R}$ or $\mathbb{R}i$.
- 3, -8 and $\sqrt{5}$ are in $\mathbb{R} \subset \mathbb{C}$ ($b = 0$ in Definition 3.2).
- i, $-5i$, $\sqrt{7}i$ and $2i/3$ are in $\mathbb{R}i \subset \mathbb{C}$.

Normally write a single label for a complex variable, e.g. $z = a + bi$. We say that $a + bi$ is the algebraic form of a complex number.

Definition 3.3. *The* ***real part*** *of a complex number* $z = a + bi$ *is given by* $\mathrm{Re}(z) = a$.

Definition 3.4. *The* ***imaginary part*** *of a complex number* $z = a + bi$ *is given by* $\mathrm{Im}(z) = b$.

Note that both the real and the imaginary part of a complex number are real numbers. Specifically, $\mathrm{Re}(a + bi) = a \in \mathbb{R}$ and $\mathrm{Im}(a + bi) = b \in \mathbb{R}$.

Example 25: Simplify the following expressions:

- $\mathrm{Re}(2 + 3i) + \mathrm{Im}(2 + 3i)i$.
- $\mathrm{Re}(-3)\mathrm{Im}(-3)$.
- $\mathrm{Re}(5i) + \mathrm{Im}(5i)$.

3.2.1 Definitions

We now need to redefine various operations on the set of complex numbers.

Definition 3.5. *Complex numbers $a + bi$ and $c + di$ are **equal** if and only if $a = c$ and $b = d$.*

Definition 3.6. *The sum of two complex numbers $a + bi$ and $c + di$ is the complex number $(a + bi) + (c + di) = (a + c) + (b + d)i$.*

Addition is an internal binary operation on $\mathbb{C}$.

Example 26: Calculate the following:

- $(5 + 3i) + (3 + 2i)$.
- $(-3 + 2i) + (2 - 4i)$.
- $3i + 5i$.

So we see that this definition is compatible with the use of the addition symbol in the general complex number notation $a + bi$. Also, addition is commutative the complex number 0 (which is shorthand for $0 + 0i$) is an identity with respect to addition, i.e.

$$0 + (a + bi) = a + bi = (a + bi) + 0.$$

We note that the negative of the complex number $a + bi$ is $-a - bi$.

Example 27: Simplify the following expressions:

- $(5 + 3i) - (3 + 2i)$.
- $(-3 + 2i) - (2 - 4i)$.
- $2 - (7 - 3i)$. ⌟

Definition 3.7. *The product of two complex numbers $a + bi$ and $c + di$ is the complex number $(a + bi) \times (c + di) = (ac - bd) + (ad + bc)i$.*

Multiplication is an internal binary operation on ℂ. This definition is compatible with the usual rules of associativity and distributivity concerning addition and multiplication. This can be illustrated as

$$\begin{aligned}(a+bi)\cdot(c+di) &= ((a+bi)\cdot c)+((a+bi)\cdot di)\\ &= (a\cdot c)+(bi\cdot c)+(a\cdot di)+(bi\cdot di)\\ &= ac+(bc)i+(ad)i+(bd)\cdot(i^2)\\ &= ac+(ad+bc)i+(bd)\cdot(-1)\\ &= (ac-bd)+(ad+bc)i.\end{aligned}$$

Example 28: Calculate the following:

- $(-2+i)(3-2i)$.
- $3(2+i)$.
- $(2i)(5i)$.
- $(3+2i)(3-2i)$. ⌟

The complex number 1 (which is shorthand for $1+0i$) is the identity for multiplication, i.e.

$$1\cdot(a+bi) = a+bi = (a+bi)\cdot 1.$$

Now suppose that $a+bi \in C$ we would like to see that if $a+bi \neq 0$, then $a+bi$ has an inverse. Notice that, as $a+bi\neq 0$, $a^2+b^2>0$. Let $z=\frac{a}{a^2+b^2}-\frac{c}{a^2+b^2}i$. Then

$$(a+bi)z = (a+bi)(\frac{a}{a^2+b^2}-\frac{c}{a^2+b^2}i) = \frac{1}{a^2+b^2}(a+bi)(a-bi) = \frac{a^2+b^2}{a^2+b^2} = 1.$$

Hence every non-zero complex number has a multiplicative inverse.

Example 28 indicates there is a special relationship between a complex number $a+bi$ and the complex number $a-bi$. To explore this further, we need to be able to refer to the two real numbers that appear in a complex number separately. Thus we have:

Definition 3.8. *The **complex conjugate** $\overline{z}$ of a complex number $z=a+bi$ is given by $\overline{z}=a-bi$.*

The complex conjugate of a complex number has the same real part, i.e. $\mathrm{Re}(\overline{z}) = \mathrm{Re}(z)$ and opposite imaginary part, i.e. $\mathrm{Im}(\overline{z}) = -\mathrm{Im}(z)$. An alternative notation for the complex conjugate of z is z^*.

Example 29: Simplify the following expressions:

- $\overline{(2+3i)}$.
- $\overline{5}$.
- $\overline{-7i}$.
- $(a+bi)(a-bi)$. ⌟

Lemma 3.9. *For $z \in \mathbb{C} \setminus \{0\}$ a non-zero complex number, the inverse of z is given by*

$$z^{-1} = \frac{1}{z\overline{z}}\overline{z}.$$

Lemma 3.10. *The quotient of two complex numbers $\alpha = a+bi$ and $\beta = c+di$, with $\beta \neq 0$, is the complex number*

$$\alpha\beta^{-1} = \frac{\alpha}{\beta} = \frac{1}{\beta\overline{\beta}}\alpha\overline{\beta}. \tag{3.2}$$

Note that the denominator in the right-hand side of (3.2) is a non-zero real number. In detail, (3.2) states that

$$\begin{aligned} \frac{a+bi}{c+di} &= \frac{\alpha}{\beta} = \frac{\alpha\cdot\overline{\beta}}{\beta\cdot\overline{\beta}} \\ &= \frac{(a+bi)\cdot(c-di)}{(c+di)\cdot(c-di)} = \frac{(ac+bd)+(bc-ad)i}{c^2+d^2} \\ &= \left(\frac{ac+bd}{c^2+d^2}\right) + \left(\frac{bc-ad}{c^2+d^2}\right)\, i. \end{aligned} \tag{3.3}$$

You should remember Lemma 3.10 as a method.

Example 30:

$$\frac{3+5i}{1-2i} = \frac{(3+5i)(1+2i)}{(1-2i)(1+2i)} = \left(\frac{3-10}{1+4}\right) + \left(\frac{5+6}{1+4}\right)i = \frac{-7}{5} + \frac{11}{5}i.$$

One can also show that addition and multiplication are *associative*, i.e.

$$(\alpha+\beta)+\gamma = \alpha+(\beta+\gamma) = \alpha+\beta+\gamma,$$

$$(\alpha\cdot\beta)\cdot\gamma = \alpha\cdot(\beta\cdot\gamma) = \alpha\cdot\beta\cdot\gamma.$$

Hence for $z_1, z_2, \ldots, z_n \in \mathbb{C}$, the expressions

$$\sum_{i=1}^{n} z_i \text{ and } \prod_{i=1}^{n} z_i$$

make sense. In particular, $(\mathbb{C}, +)$ and $(\mathbb{C} \setminus \{0\}, \times)$ are abelian groups and, as we can also check that the distributive law holds we have

Theorem 3.11. *We have $(\mathbb{C}, +, \times)$ is a field.*

Example 31:

1. Calculate
$$\frac{(2+3i)\cdot(3-4i)}{1-5i}.$$
2. Find all the solutions of the equation
$$z\overline{z} - 2i\overline{z} - 7 + 4i = 0.$$
3. By writing $\alpha = a + bi$, with $a, b \in \mathbb{R}$, find all the complex numbers α such that $\alpha^2 = 3 - 4i$. ⌟

It is useful to take note of the following properties of $\overline{z}$:

Lemma 3.12. *We have the following properties of complex conjugation:*

1. $\overline{(\overline{z})} = z$ *for any complex number* $z \in \mathbb{C}$.
2. *If* $z_1, z_2, \ldots, z_n \in \mathbb{C}$, *then*
$$\overline{(z_1 + z_2 + \ldots + z_n)} = \overline{z_1} + \overline{z_2} + \ldots + \overline{z_n}.$$
3. *If* $z_1, z_2, \ldots, z_n \in \mathbb{C}$, *then*
$$\overline{(z_1 \cdot z_2 \cdot \ldots \cdot z_n)} = \overline{z_1} \cdot \overline{z_2} \cdot \ldots \cdot \overline{z_n}.$$
4. *The complex number z is a* real number *if and only if* $\overline{z} = z$.
5. *The complex number z is* purely imaginary *if and only if* $\overline{z} = -z$.

3.3 The Argand diagram

The set of all real numbers $\mathbb{R}$ can be visualised by a number line with “no gaps". The set of complex numbers consists of elements, $a + bi$, which in general depend on two real numbers, a and b. This requires a two-dimensional graphical representation.

In the *Argand diagram* (also called the *complex plane*, see Figure 3.1), a complex number $z = a+bi$ is represented by the point with coordinates (a, b). The Argand diagram is named after Jean-Robert Argand. (1768–1822) In other words, the point with x-coordinate equal to a and y-coordinate equal to b.

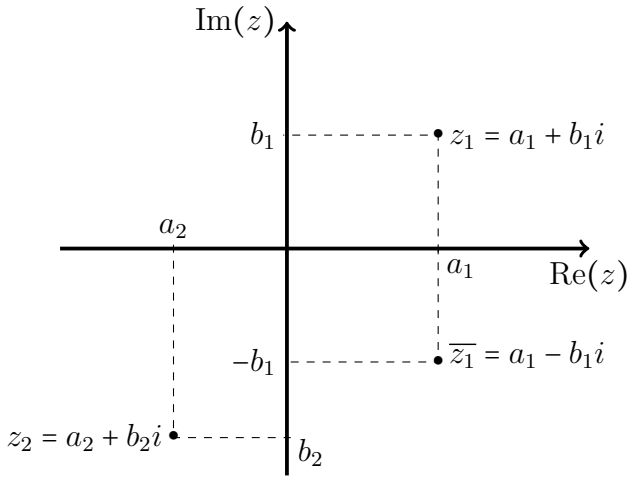


Figure 3.1: An illustration of the Argand diagram which can be used to graphically represent points in $z_1, \overline{z_1}, z_2 \in \mathbb{C}$. Here $z_1 = a_1 + b_1 i$ and $z_2 = a_2 + b_2 i$ with $a_1, b_1, a_2, b_2 \in \mathbb{R}$.

One should note the following:

1. The x-axis in the Argand diagram represents the real number line. Hence, it is called the *real axis*.

2. The y-axis in the Argand diagram represents the set of purely imaginary numbers, i.e, $\mathbb{R}i$. Hence it is called the *imaginary axis*.

Since one often uses z to denote complex numbers, you might also find that the Argand diagram is referred to as the *z-plane*.

Example 32: Plot the following complex numbers on the Argand diagram:

$$3 + i, 5,\ -2i,\ 4 - 2i,\ -2 + 3i,\ -1 - 2i.$$

3.4 Modulus-argument form (polar form)

We in addition to the standard Cartesian co-ordinates, every point in the plane can also be identified by *polar coordinates*(r, θ) rather than *Cartesian coordinates* (x, y). This can be applied to the Argand diagram as well, with the zero on the real axis acting as the origin O and the positive real axis as the initial ray.

We set $x = r\cos(\theta)$ and $y = r\sin(\theta)$, so that[1]

$$z = x + yi = r\cos(\theta) + r\sin(\theta)i = r\left(\cos(\theta) + i\sin(\theta)\right),$$

where by basic trigonometry

$$r = \sqrt{x^2 + y^2} \text{ and } \tan(\theta) = \frac{y}{x}.$$

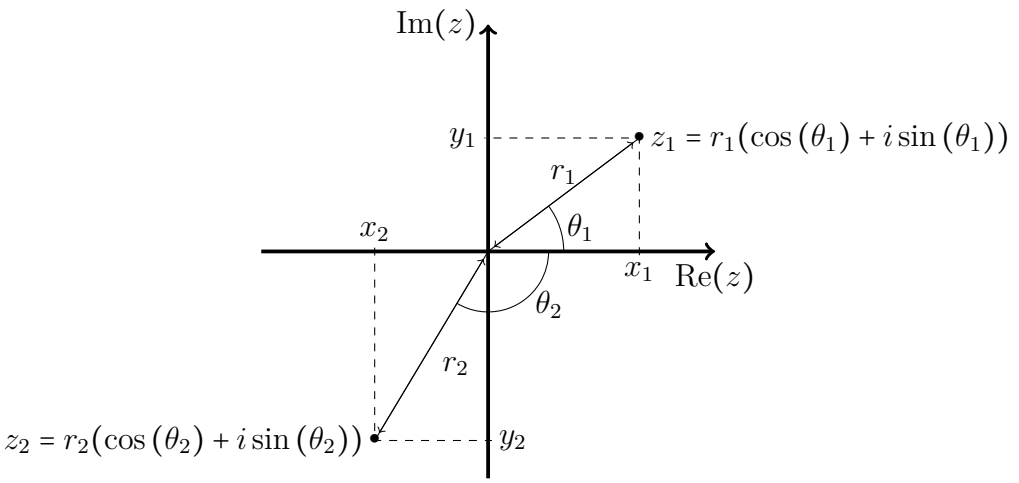


Figure 3.2: An illustration of the modulus and argument of $z_1, z_2 \in \mathbb{C}$. Here $z_1 = x_1 + y_1 i$ and $z_2 = x_2 + y_2 i$ with $x_1, y_1, x_2, y_2 \in \mathbb{R}$, $|z_1| = r_1$, $|z_2| = r_2$ and $\theta_1 \in \arg(z_1)$ and $\theta_2 \in \arg(z_2)$.

Notice that complex numbers of the form $\cos(\theta) + i\sin(\theta)$ lie on the unit circle in the Argand diagram.

When applied to the representation of complex numbers, we use the following terminology:

Definition 3.13. *The **modulus** of a complex number $z = x+yi$, denoted $|z|$, is given by*

$$|z| = \sqrt{x^2 + y^2}.$$

[1] Do not confuse the complex number i with the vector $\mathbf{i}$ here.

It is clear that $|z| \geq 0$ for all $z \in \mathbb{C}$ and $|z| = 0$ if and only if $z = 0$. The modulus is sometimes referred to as the *absolute value* of z. Also note that

$$|z| = \sqrt{x^2 + y^2} = \sqrt{z \cdot \overline{z}} \text{ for } z = x + yi \in \mathbb{C}.$$

Definition 3.14. *The values of $\theta \in \mathbb{R}$ for which*

$$x = |z|\cos(\theta) \text{ and } y = |z|\sin(\theta),$$

*are called **arguments** of the complex number $z = x + yi$, where x and y are not both equal to 0. The set of all arguments of z is denoted by $\arg(z)$.*

Complex numbers have infinitely many 'arguments'. If $\theta \in \arg(z)$, then

$$\arg(z) = \{\theta + k2\pi : k \in \mathbb{Z}\}.$$

This multiplicity of possibilities can be made unique as follows:

Definition 3.15. *The **principal value** of $\arg(z)$, denoted $\text{Arg}(z)$, is the member of of $\arg(z)$ such that $\text{Arg}(z) \in (-\pi, \pi]$.*

This means that

$$\{\text{Arg}(z)\} = \arg(z) \cap (-\pi, \pi].$$

Every complex number $z \in \mathbb{C} \setminus \{0\}$, has a unique principal value of its argument. In practice, to determine the principal value of a complex number, one typically starts by solving

$$\tan(\theta) = \frac{y}{x}, \tag{3.4}$$

which yields two possible values of θ in the interval $(-\pi, \pi]$, a value of π apart if $x \neq 0$ (if $x = 0$ and $y \neq 0$, then θ is either $\frac{\pi}{2}$ or $-\frac{\pi}{2}$). We have learned to restrict the tangent function $\tan : (-\frac{\pi}{2}, \frac{\pi}{2}) \to \mathbb{R}$ in order to define its inverse, so the solution we obtain analytically should be in this interval. The value of θ obtained from solving (3.4) needs to be checked against the position in the Argand diagram, and, where necessary, discarded and replaced by the value $\theta + \pi$ or $\theta - \pi$. It is good practice to plot the complex number in the Argand diagram to check whether the calculated value of θ is the principal value of the argument or not.

Example 33: Calculate the modulus, $|z|$, and principal value of the argument, $\text{Arg}(z)$ for:

- $z = 2 + i$.

- $z = 2 - i$.
- $z = -2 + i$.
- $z = -2 - i$. ⌟

From Definition 3.15, it follows that $z_1, z_2 \in \mathbb{C} \setminus \{0\}$ satisfy $z_1 = z_2$ if and only if $|z_1| = |z_2|$ and $\text{Arg}(z_1) = \text{Arg}(z_2)$.

3.5 Product and Quotient using modulus-argument form

Proposition 3.16. *Given two non-zero complex numbers* $z_1 = r_1(\cos(\theta_1) + i\sin(\theta_1))$ *and* $z_2 = r_2(\cos(\theta_2) + i\sin(\theta_2))$, *i.e.* $r_1 = |z_1|$, $r_2 = |z_2|$, θ_1 *is a value in* $\arg(z_1)$ *and* θ_2 *is a value in* $\arg(z_2)$. *Then,*

(i) $|z_1 \cdot z_2| = |z_1||z_2|$,

(ii) $\theta_1 + \theta_2$ *is a value in* $\arg(z_1 \cdot z_2)$.

Proof: We use the definition and trigonometric formulae to calculate

$$\begin{aligned} z_1 \cdot z_2 &= r_1 \cdot r_2 \cdot (\cos(\theta_1) + i\sin(\theta_1)) \cdot (\cos(\theta_2) + i\sin(\theta_2)) \\ &= r_1 r_2 \left[(\cos(\theta_1)\cos(\theta_2) - \sin(\theta_1)\sin(\theta_2)) + i(\cos(\theta_1)\sin(\theta_2) + \sin(\theta_1)\cos(\theta_2))\right] \\ &= r_1 r_2 (\cos(\theta_1 + \theta_2) + i\sin(\theta_1 + \theta_2)). \end{aligned}$$

So $|z_1 \cdot z_2| = r_1 r_2 = |z_1||z_2|$ and a value of the argument is given by $\theta_1 + \theta_2$, as required.

Note that even when θ_1 and θ_2 are the principal values of the argument of z_1 and z_2 respectively, $\theta_1 + \theta_2$ is not necessarily the principal value of the argument of $z_1 \cdot z_2$, since $\theta_1 + \theta_2$ is not necessarily in the interval $(-\pi, \pi]$.

Example 34: For

$$z_1 = 2\left(\cos\left(\frac{\pi}{3}\right) + i\sin\left(\frac{\pi}{3}\right)\right) \text{ and } z_2 = 3\left(\cos\left(\frac{3\pi}{4}\right) + i\sin\left(\frac{3\pi}{4}\right)\right)$$

calculate $|z_1 \cdot z_2|$ and $\text{Arg}(z_1 \cdot z_2)$.

Proposition 3.17. *Given two non-zero complex numbers* $z_1 = r_1(\cos(\theta_1) + i\sin(\theta_1))$ *and* $z_2 = r_2(\cos(\theta_2) + i\sin(\theta_2)) \neq 0$. *Then,*

(i) $\frac{1}{z_2} = z_2^{-1} = \frac{1}{r_2}(\cos(\theta_2) - i\sin(\theta_2)) = \frac{1}{r_2}(\cos(-\theta_2) + i\sin(-\theta_2))$;

(ii) $\left|\frac{z_1}{z_2}\right| = \frac{|z_1|}{|z_2|}$,

(iii) $\theta_1 - \theta_2$ *is a value of* $\arg\left(\frac{z_1}{z_2}\right)$.

Proof: Part (i) follows from Lemma 3.9. To understand part (ii) and (iii), we use Proposition 3.16 and calculate

$$\begin{aligned} z_1 z_2^{-1} &= r_1(\cos(\theta_1) + i\sin(\theta_1))\frac{1}{r_2}(\cos(-\theta_2) + i\sin(-\theta_2)) \\ &= \frac{r_1}{r_2}(\cos(\theta_1 - \theta_2) + i\sin(\theta_1 - \theta_2)). \end{aligned}$$

Therefore,

$$\left|\frac{z_1}{z_2}\right| = |z_1 z_2^{-1}| = \frac{r_1}{r_2} \text{ and } \theta_1 - \theta_2 \in \arg\left(\tfrac{z_1}{z_2}\right),$$

as required.

Similarly when θ_1 and θ_2 are the principal values of the argument of z_1 and z_2 respectively, $\theta_1 - \theta_2$ is not necessarily the principal value of the argument of $z_1 z_2^{-1}$.

Example 35: For,

$$z_1 = 2\left(\cos\left(\frac{\pi}{3}\right) + i\sin\left(\frac{\pi}{3}\right)\right) \text{ and } z_2 = 3\left(\cos\left(-\frac{3\pi}{4}\right) + i\sin\left(-\frac{3\pi}{4}\right)\right),$$

calculate $|z_1/z_2|$ and $\mathrm{Arg}(z_1/z_2)$.

Corollary 3.18. *Given non-zero complex numbers of the form* $z_j = r_j(\cos(\theta_j) + i\sin(\theta_j))$ *for* $j = 1, \ldots, n$ *then*

$$z_1 \cdot z_2 \cdot \ldots \cdot z_n = \left(\prod_{j=1}^{n} r_j\right)(\cos(\theta_1 + \theta_2 + \ldots + \theta_n) + i\sin(\theta_1 + \theta_2 + \ldots + \theta_n)).$$

Proof: Use mathematical induction and Proposition 3.16 to establish the result.

In particular, the product of two or more complex numbers with modulus one also has modulus one. The special case of taking the n-th power of a complex number leads to the formulation of De Moivre's Theorem.

Example 36: Let $S = \{z \in \mathbb{C} : |z| = 1\}$. Show that $(S, \times)$ is a group.

3.6 de Moivre's Theorem

Theorem 3.19. ***(de Moivre)*** *If $n \in \mathbb{N}$ and $\theta \in \mathbb{R}$, then*

$$(\cos(\theta) + i\sin(\theta))^n = \cos(n\theta) + i\sin(n\theta).$$

Proof: Apply Corollary 3.18.

This was discovered by Abraham de Moivre (1667, 1754).

Example 37: Via de Moivre's Theorem calculate

$$1^n, i^2, i^3, (-1)^4 \text{ and } (-i)^3.$$

Corollary 3.20. *If $n \in \mathbb{N}$ and $\theta \in \mathbb{R}$, then*

$$(\cos(\theta) + i\sin(\theta))^{-n} = \cos(n\theta) - i\sin(n\theta) = \cos(-n\theta)i + \sin(-n\theta).$$

Proof: Use Proposition 3.17 and de Moivre's Theorem.

3.7 Euler's formula

If we consider the modulus-argument notation for a complex number with modulus equal to one, together with the MacLaurin series for cos and sin, we discover a peculiar formula. Since the following series are absolutely convergent for all $\theta \in \mathbb{R}$ (see [15]),

$$\begin{aligned} \cos(\theta) &= 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \cdots \\ \sin(\theta) &= \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \cdots, \end{aligned}$$

one finds that

$$\begin{aligned} \cos(\theta) + i\sin(\theta) &= \left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \cdots\right) + i\left(\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \cdots\right), \\ &= 1 + (i\theta) + \frac{(i\theta)^2}{2!} + \frac{(i\theta)^3}{3!} + \frac{(i\theta)^4}{4!} + \frac{(i\theta)^5}{5!} + \frac{(i\theta)^6}{6!} + \frac{(i\theta)^7}{7!} + \ldots \\ &= e^{i\theta}, \end{aligned} \tag{3.5}$$

where we recall that the Maclaurin series for e^x is given by (at least for $x \in \mathbb{R}$)[2]

$$e^x = \frac{1}{0!} + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots.$$

So it appears that we can use $e^{i\theta}$ as a short notation for $\cos(\theta) + i\sin(\theta)$. But does this use of the exponential function satisfy useful properties we associate with e^x? A crucial property of the exponential function is the index law, i.e. for $x, y \in \mathbb{R}$,

$$e^x \cdot e^y = e^{x+y}. \tag{3.6}$$

Now, observe that via Corollary 3.18 and (3.5), we have

$$\begin{aligned} e^{i\theta_1} \cdot e^{i\theta_2} &= (\cos(\theta_1) + i\sin(\theta_1)) \cdot (\cos(\theta_2) + i\sin(\theta_2)) \\ &= \cos(\theta_1 + \theta_2) + i\sin(\theta_1 + \theta_2) \\ &= e^{i(\theta_1+\theta_2)} \\ &= e^{i\theta_1 + i\theta_2}. \end{aligned} \tag{3.7}$$

So we can indeed use $e^{i\theta}$ in the same way as we use the exponential function with respect to the index law in (3.6). The formula

$$e^{i\theta} = \cos(\theta) + i\sin(\theta),$$

is known as **Euler's formula**[3]. Using Euler's formula, de Moivre's Theorem can be written as

$$\left(e^{i\theta}\right)^n = e^{in\theta} \text{ for all } \theta \in \mathbb{R} \text{ and } n \in \mathbb{N}.$$

Similarly,

$$\left(e^{-i\theta}\right)^n = e^{-in\theta} = \left(e^{i\theta}\right)^{-n} \text{ for all } \theta \in \mathbb{R} \text{ and } n \in \mathbb{N}.$$

In particular, we have the famous expression

$$e^{i\pi} = -1.$$

Hence, we can now write a complex number in three different forms:

$$z = a + bi = r\left(\cos(\theta) + i\sin(\theta)\right) = re^{i\theta},$$

with $r^2 = a^2 + b^2$ and $\theta = \arg(z)$. These forms for a complex number z are respectively referred to as: the *algebraic form*, the *modulus-argument form*, and *exponential form.*

When doing calculations with complex numbers, one should choose the most suitable form/notation to increase efficiency and simplification of associated operations.

Example 38: Write the following complex numbers in modulus-argument form and in exponential form:

[2]Laurent series allows one to expand smooth complex valued functions in a similar way to Taylor/Maclaurin expansions for real functions of 1 variable. For details see [13, p.195]

[3]Named after the Swiss mathematician Leonhard Euler [11].

1. $z = \dfrac{\sqrt{2}}{2} + \dfrac{\sqrt{2}}{2}i.$

2. $z = \dfrac{\sqrt{3}}{2} - \dfrac{1}{2}i.$ ⌟

3.8 De Moivre's Theorem and trigonometric formulae

If we combine de Moivre's Theorem with the Binomial Theorem in the form

$$\begin{aligned}(a+b)^n &= a^n + na^{n-1}b + \frac{n(n-1)}{2!}a^{n-2}b^2 + \cdots + \frac{n(n-1)\ldots 2}{(n-1)!}ab^{n-1} + \frac{n(n-1)\ldots 2.1}{n!}b^n \\ &= \binom{n}{0}a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \cdots + \binom{n}{n-1}ab^{n-1} + \binom{n}{n}b^n \\ &= \sum_{k=0}^{n}\binom{n}{k}a^{n-k}b^k\end{aligned}$$

then we can construct trigonometric identities which include $\cos(n\theta)$ and $\sin(n\theta)$ terms. For example, De Moivre's Theorem with $n = 2$ states that

$$(\cos(\theta) + i\sin(\theta))^2 = \cos(2\theta) + i\sin(2\theta). \tag{3.8}$$

The Binomial Theorem, which can also be applied to complex quantities, yields

$$(\cos(\theta) + i\sin(\theta))^2 = (\cos^2(\theta) - \sin^2(\theta)) + 2\cos(\theta)\sin(\theta)i. \tag{3.9}$$

Hence, by equating the real and imaginary parts (3.8) and (3.9) we obtain the familiar identities

$$\begin{aligned}\cos(2\theta) &= \cos^2(\theta) - \sin^2(\theta), \\ \sin(2\theta) &= 2\cos(\theta)\sin(\theta).\end{aligned}$$

Example 39: Establish that $\cos(3\theta) = 4\cos^3(\theta) - 3\cos(\theta)$.

3.9 Euler's formula and trigonometric formulae

Euler's formula allows us to deduce various trigonometric formulae that are more difficult to deduce otherwise. Consider Euler's formula for θ and $-\theta$:

$$e^{i\theta} = \cos(\theta) + i\sin(\theta), \tag{3.10}$$

$$\begin{aligned} e^{-i\theta} &= \cos(\theta) + i\sin(-\theta) \\ &= \cos(\theta) - i\sin(\theta). \end{aligned} \tag{3.11}$$

Equations (3.10) and (3.11) yield the relation between exponential and trigonometric functions,

$$\begin{aligned} e^{i\theta} + e^{-i\theta} &= (\cos(\theta) + i\sin(\theta)) + (\cos(\theta) - i\sin(\theta)) \\ &= 2\cos(\theta) \end{aligned}$$

which allows us to write $\cos(\theta)$ as

$$\cos(\theta) = \frac{1}{2}\left(e^{i\theta} + e^{-i\theta}\right) \quad \forall\theta \in \mathbb{R}.$$

Similarly,

$$\begin{aligned} e^{i\theta} - e^{-i\theta} &= (\cos(\theta) + i\sin(\theta)) - (\cos(\theta) - i\sin(\theta)) \\ &= 2i\sin(\theta) \end{aligned}$$

so that

$$\sin(\theta) = \frac{1}{2i}\left(e^{i\theta} - e^{-i\theta}\right) \quad \forall\theta \in \mathbb{R}.$$

Similarly, one can find formulae for $\cos(n\theta)$ and $\sin(n\theta)$. For any $n \in \mathbb{N}$, via de Moivre's Theorem,

$$\begin{aligned} e^{in\theta} &= \cos(n\theta) + i\sin(n\theta), \\ e^{-in\theta} &= \cos(n\theta) - i\sin(n\theta), \end{aligned}$$

and hence,

$$e^{in\theta} + e^{-in\theta} = 2\cos(n\theta) \Longrightarrow \cos(n\theta) = \frac{1}{2}\left(e^{in\theta} + e^{-in\theta}\right) \quad \forall\theta \in \mathbb{R}. \tag{3.12}$$

Moreover,

$$e^{in\theta} - e^{-in\theta} = 2i\sin(n\theta) \Longrightarrow \sin(n\theta) = \frac{1}{2i}\left(e^{in\theta} - e^{-in\theta}\right) \quad \forall\theta \in \mathbb{R}. \tag{3.13}$$

Equations (3.12) and (3.13) allow us to rewrite powers of $\cos(\theta)$ and $\sin(\theta)$ in terms of *linear combinations* of $\cos(n\theta)$ and $\sin(n\theta)$ for $n \in \mathbb{N}$.

Example 40:

$$\begin{aligned} \cos^3(\theta) &= \left(\frac{1}{2}\left(e^{i\theta} + e^{-i\theta}\right)\right)^3 \\ &= \frac{1}{8}\left(e^{i3\theta} + 3e^{i2\theta}e^{-i\theta} + 3e^{i\theta}e^{-i2\theta} + e^{-i3\theta}\right) \end{aligned}$$

$$\begin{aligned} &= \frac{1}{8}\left(\left(e^{i3\theta}+e^{-i3\theta}\right)+3\left(e^{i\theta}+e^{-i\theta}\right)\right) \\ &= \frac{1}{8}\left(2\cos(3\theta)+3\cdot 2\cos(\theta)\right) \\ &= \frac{1}{4}\cos(3\theta)+\frac{3}{4}\cos(\theta). \end{aligned}$$

Example 41: Write $\sin^3(\theta)$ in terms of cosines or sines of multiples of θ:

3.10 n-th roots of complex numbers

The n-th root of a complex number is defined similarly to that of a real number.

Definition 3.21. *Suppose that $n \geq 1$ is a natural number. A complex number w is an n-th root of a non-zero complex number z, denoted by $w = z^{1/n}$ if $w^n = z$. We say that w is a n-th root of unity if w is a n-th root of 1.*

For example, $i^2 = -1$ and $(-i)^2 = -1$, so i and $-i$ are square roots (or 2-th roots) of -1. Since

$$1^4 = i^4 = (-1)^4 = (-i)^4 = 1,$$

$1, -1, i$ and $-i$ are all 4-th roots of 1.

To find the n-th root in general, we make use of the exponential notation of a complex number. We express z as $z = re^{i\theta}$ and determine the n-th roots w in their exponential form $w = \rho e^{i\phi}$. Remember that by definition, $r > 0$ and $\rho > 0$ provided that $z \neq 0$ (n-th roots of 0 are easy to find ... they consist of ... just 0). According to Definition 3.21, w is then an n-th root of z if and only if $z = w^n$ or equivalently

$$re^{i\theta} = \left(\rho e^{i\phi}\right)^n = \rho^n e^{in\phi},$$

by the Euler's version of de Moivre's Theorem.

The modulus of a complex number is unique, hence we have that $r = \rho^n$ or $\rho = r^{1/n}$, i.e. the positive, real n-th root of r. Also, both θ and $n\phi$ are values of the argument of z, or equivalently

$$n\phi = \theta + 2k\pi, \tag{3.14}$$

for some $k \in \mathbb{Z}$. Equation (3.14) can be solved for the unknown ϕ as,

$$\phi = \frac{\theta}{n} + k\frac{2\pi}{n}. \tag{3.15}$$

This yields n distinct values for ϕ in the interval $(-\pi, \pi]$, and hence n distinct n-th roots of z, given by

$$= r^{1/n} e^{i(\frac{\theta + k2\pi}{n})}$$

with $k = 0, 1, 2, \ldots n-1$ (or any other consecutive sequence of n integer numbers).

Note that:

- The n-th roots of $z = re^{i\theta} \neq 0$ have the same modulus, hence they are all on a circle in the Argand diagram with the origin at the centre and radius $r^{1/n}$.
- Unlike with the real n-th roots of real numbers, every $z \in \mathbb{C} \setminus \{0\}$ has exactly n distinct n-th roots.
- When plotted on the Argand diagram, these n n-th roots are equally spaced around the circle. The arguments of two neighbouring roots differ by $2\pi/n$.
- As a consequence, the n-th roots of z are the vertices of a regular polygon of n sides inscribed in the circle with centre at the origin and radius $r^{1/n}$.

Example 42: Find all the fifth roots of $z = 16\sqrt{3} + 16i$.

3.11 Polynomials

The discussion following (3.15) implies that a polynomial equation of the form $z^n = a$ with $a \in \mathbb{C} \setminus \{0\}$ has exactly n distinct solutions. Can the same be said about solutions of general polynomial equations?

Definition 3.22. *1. A **polynomial** of degree n in the complex variable z is an expression of the form*

$$p(z) = a_n z^n + a_{n-1} z^{n-1} + \ldots + a_1 z + a_0 = \sum_{i=0}^{n} a_i z^i,$$

where $a_i \in \mathbb{C}$ for $i = 0, 1, \ldots, n$ and $a_n \neq 0$.

*2. For $p(z)$ a degree n polynomial, a **polynomial equation** is an equation of the form*

$$p(z) = 0.$$

*3. For $p(z)$ a degree n polynomial, a complex number α is called a **zero** (or a **root**) of the polynomial $p(z)$ if $p(\alpha) = 0$. That is, α is a solution to the polynomial equation $p(z) = 0$.*

Example 43: The expression

$$5z^4 - 3z^2 + 1$$

is a polynomial of degree 4 with coefficients $a_4 = 5$, $a_3 = 0$, $a_2 = -3$, $a_1 = 0$ and $a_0 = 1$.

The equation

$$5z^4 - 3z^2 + 1 = 0$$

is a polynomial equation of degree 4

Theorem 3.23. *The complex number α is a root of the polynomial equation $p(z) = 0$, if and only if $z - \alpha$ is a factor of the polynomial $p(z)$, i.e.*

$$p(z) = (z - \alpha)q(z),$$

where $q(z)$ is a polynomial of degree $n - 1$.

Proof: If $z - \alpha$ is a factor of $p(z)$, i.e. $p(z) = (z - \alpha)q(z)$, with $q(z)$ a polynomial of degree $n - 1$, then it is clear that $p(\alpha) = (\alpha - \alpha)q(\alpha) = 0$. Hence, α is a root of $p(z) = 0$. This proves one part of the theorem.

We still need to prove that if α is a root of $p(z)$ then $z - \alpha$ is a factor of $p(z)$. Consider calculating $p(z)/(z - \alpha)$ by polynomial division. Then $p(z)$ can be written as

$$p(z) = (z - \alpha)q(z) + R, \tag{3.16}$$

where $q(z)$ is a polynomial of degree $n - 1$ and the remainder term, denoted by R, is a polynomial of degree less then the degree of $z - \alpha$, i.e., R is a constant. If α is a root of $p(z) = 0$ then $p(\alpha) = 0$ and, consequently via (3.16)

$$(\alpha - \alpha)q(\alpha) + R = 0 \iff R = 0.$$

Therefore, $p(z) = (z - \alpha)q(z)$ or equivalently, $z - \alpha$ is a factor of $p(z)$, as required.

3.12 Quadratic equations

We can now take a fresh look at the familiar quadratic equation, $az^2 + bz + c$ with $a, b, c \in \mathbb{C}$ $(a \neq 0)$. Zeros of this quadratic equation can be found by completing the square:

$$\begin{aligned} az^2 + bz + c = 0 &\iff a\left(z^2 + \frac{b}{a}z + \frac{c}{a}\right) = 0 \\ &\iff a\left(z^2 + \frac{b}{a}z + \left(\frac{b}{2a}\right)^2 - \left(\frac{b}{2a}\right)^2 + \frac{c}{a}\right) = 0 \end{aligned}$$

$$\begin{aligned} &\iff a\left(\left(z+\frac{b}{2a}\right)^2-\frac{b^2}{4a^2}+\frac{c}{a}\right)=0 \\ &\iff a\left(z+\frac{b}{2a}\right)^2=\frac{b^2}{4a}-c \\ &\iff a\left(z+\frac{b}{2a}\right)^2=\frac{b^2-4ac}{4a}. \end{aligned} \tag{3.17}$$

So we calculate

$$\begin{aligned} \left(z+\frac{b}{2a}\right)^2=\frac{b^2-4ac}{4a^2} &\iff z+\frac{b}{2a}=\left(\frac{b^2-4ac}{4a^2}\right)^{1/2} \\ &\iff z=-\frac{b}{2a}+\left(\frac{b^2-4ac}{4a^2}\right)^{1/2}. \end{aligned} \tag{3.18}$$

Theorem 3.24. *Suppose that $p(z)=az^2+bz+c$ with $a,b,c\in\mathbb{C}$ and $a\neq 0$. Let $w\in\mathbb{C}$ be such that*

$$w^2=\frac{b^2-4ac}{4a^2}.$$

Then counting multiplicities $p(z)$ has two roots

$$z_{1,2}=-\frac{b}{2a}\pm w$$

and we can factor

$$p(z)=az^2+bz+c=a(z-z_1)(z-z_2).$$

3.13 The Fundamental Theorem of Algebra

What we have seen with polynomials of degree 2 with complex number coefficients generalizes to polynomials of arbitrary degree.

Theorem 3.25. ***(Fundamental Theorem of Algebra)*** *Suppose that $p(z)$ is a polynomial of degree n with $n\geq 1$. Then the polynomial equation $p(z)=0$ has at least one root in $\mathbb{C}$.*

Corollary 3.26. *Suppose that $p(z)$ is a polynomial of degree n with $n\geq 1$. Then $p(z)$ is a product of polynomials of degree 1 which have complex factors.*

There are various proofs of Theorem 3.25 (see, for instance, [13, p.153-154]) all of which require methods outside the scope of this course, and most of which, require results concerning analysis of functions of a complex variable. In a more sophisticated language the fundamental theorem states that the field of complex numbers in algebraically closed (there are no missing roots of polynomial equations).

3.14 Polynomials with real coefficients

Suppose that

$$p(z) = a_n z^n + a_{n-1} z^{n-1} + \ldots + a_1 z + a_0 = \sum_{i=0}^{n} a_i z^i,$$

has degree n and $a_i \in \mathbb{R}$ for $i = 0, 1, \ldots, n$. Then we say that the polynomial $p(z)$ has **real coefficients**.

In the case of quadratic equations with real coefficients $p(z) = az^2 + bz + c$, $a, b, c \in \mathbb{R}$, $a \neq 0$, Theorem 3.24 gives roots

$$z_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

We define

$$D = b^2 - 4ac$$

and call D the **discriminant** . There are three possibilities, $D > 0$, $D = 0$ and $D < 0$.

In the case that $D > 0$, $p(z)$ can be factorised into two linear factors with real coefficients,

$$az^2 + bz + c = a(z - z_1)(z - z_2).$$

When $D = 0$, there is 1 (repeated) zero,

$$z_1 = -\frac{b}{2a},$$

and the quadratic can be factorised as

$$az^2 + bz + c = a(z - z_1)^2 = a(z - z_1)(z - z_1). \tag{3.19}$$

When $D < 0$, we can define a real variable ν as $b^2 - 4ac = -\nu^2$, and we have

$$z_{1,2} = \frac{-b \pm i\nu}{2a}$$

and in particular we observe the famous fact that the roots are complex conjugates of each other. That is $\overline{z_1} = z_2$.

The quadratic polynomial cannot be factorised in two real linear factors, and so we define it to be a real irreducible quadratic. One can write the quadratic using **complex** linear factors as

$$az^2 + bz + c = a(z - z_1)(z - z_2).$$

Hence,

Theorem 3.27. *A quadratic polynomial with real coefficients has either two distinct real zeros (positive discriminant), two identical real zeros (discriminant zero) or a pair of complex conjugate non-real complex zeros (negative discriminant).*

So when looking for zeros in the set of complex numbers, $\mathbb{C}$, the quadratic (or polynomial of degree 2) **always has 2 zeros** (that are not necessarily distinct).

The fact that complex conjugation permutes the roots of a quadratic equation (fixing every real root and swapping the complex roots) leads to the property is given in the following:

Theorem 3.28. *Suppose that $p(z)$ is a degree n polynomial with real coefficients. Then $\alpha \in \mathbb{C}$ is a zero of $p(z)$ if and only if the complex conjugate $\overline{\alpha}$ is a zero of $p(z)$. In particular, non-real zeros of $p(z)$ occur in conjugate pairs.*

Proof: We have seen the following identities:

$$\overline{(z_1 + z_2 + \ldots + z_n)} = \overline{z_1} + \overline{z_2} + \ldots + \overline{z_n}, \tag{3.20}$$

$$\overline{(z_1 \cdot z_2 \cdot \ldots \cdot z_n)} = \overline{z_1} \cdot \overline{z_2} \cdot \ \ldots \ \cdot \overline{z_n}. \tag{3.21}$$

It follows from (3.21) that $\overline{z^n} = \overline{z}^n$ and $\overline{az^n} = a\overline{z}^n$ for $a \in \mathbb{R}$ and $n \in \mathbb{N}$. Hence,

$$\begin{aligned} \overline{(p(z))} &= \overline{(a_n z^n + a_{n-1} z^{n-1} + \ldots + a_1 z + a_0)} \\ &= a_n (\overline{z})^n + a_{n-1} (\overline{z})^{n-1} + \ldots + a_1 \overline{z} + a_0 \quad \text{(via (3.20) and (3.21))} \\ &= p(\overline{z}). \end{aligned} \tag{3.22}$$

Therefore, (3.22) yields

$$p(z) = 0 \iff \overline{(p(z))} = \overline{0} \iff p(\overline{z}) = 0.$$

Hence

$$p(a + bi) = 0 \iff p(a - bi) = 0,$$

as required.

Please note that the above result does **not** hold in general for polynomial equations of degree n with *complex coefficients*.

Corollary 3.29. *A polynomial $p(z)$ with real coefficients of odd degree n has at least one real zero.*

Proof:[4] Suppose that $p(z)$ has only complex roots. Via Theorem 3.28, every root occurs in a complex conjugate pair i.e. $n = 2m$ for some $m \in \mathbb{N}$. This implies n is even which is a contradiction. We conclude that $p(z)$ has at least one real root, as required.

Corollary 3.30. *Suppose that $p(z)$ is a polynomial of degree n with real coefficients. The complex number $\alpha = a + bi$ is a non-real root of the polynomial equation $p(z) = 0$ if and only if the quadratic polynomial*

$$z^2 - 2az + a^2 + b^2$$

is a factor of $p(z)$.

Proof: Let us write $\alpha = a + ib$. Then $p(z) = (z - \alpha)q(z)$ by Theorem 3.23. Now Theorem 3.28 implies

$$0 = p(\overline{\alpha}) = (\overline{\alpha} - \alpha)q(\overline{\alpha})$$

so, as $\overline{\alpha} \neq \alpha$ because α is not real, we must have $q(\overline{\alpha}) = 0$. Hence Theorem 3.23 yields $q(z) = (z - \overline{\alpha})r(z)$. Hence

$$p(z) = (z - \alpha)(z - \overline{\alpha})r(z) = (z^2 - 2az + a^2 + b^2)r(z)$$

and this proves the claim.

The quadratic polynomial $z^2 - 2az + a^2 + b^2$ cannot be expressed as a product of two real linear factors and is therefore called a **real irreducible quadratic**.

Example 44: Given that $2 + 7i$ is a root of the polynomial equation $p(z) = 0$ where

$$p(z) = z^4 - 4z^3 + 55z^2 - 8z + 106,$$

find the remaining roots.

We are now a number of properties of polynomials of degree n with real coefficients. We will state them here without proof. The first one follows from the fundamental theorem of algebra. It demonstrates the deficiency of the real numbers.

Theorem 3.31. *A polynomial $p(z)$ with real coefficients can always be expressed as a product of real linear and real irreducible quadratic factors.*

[4] Consider for yourself how Corollary 3.29 can be established using properties of real valued functions (polynomials) $p : \mathbb{R} \to \mathbb{R}$ considered in [15].

Theorem 3.32. *A polynomial $p(z)$ of degree n with real coefficients has exactly n zeros, some of which may be repeated, and some of which may be complex conjugate pairs (the complex root and its complex conjugate).*

Note that zeros whose corresponding factor in the factorisation of the polynomial appears with a power μ (see, for example, (3.19) with $\mu = 2$) are counted as μ zeros. We say that such a zero has algebraic multiplicity μ. Specifically, if $p(z)$ is a polynomial of degree $n \geq 1$, then $p(z)$ has r distinct roots z_i (with $1 \leq r \leq n$) which have respective algebraic multiplicity m_i, for $i = 1, \ldots, r$, and moreover,

$$n = \sum_{i=1}^{r} m_i.$$

Example 45: The degree 8 polynomial

$$\begin{aligned} p(z) &= (z-1)^2(z-2)^4(z^2+1) \\ &= z^8 - 10z^7 + 42z^6 - 98z^5 + 145z^4 - 152z^3 + 120z^2 - 64z + 16. \end{aligned}$$

has distinct roots at 1, 2, i and $-i$ with multiplicities 2, 4, 1 and 1 respectively. Note that although there are only 2 real roots and a pair of complex conjugate roots, the sum of the multiplicities of the roots equals 8 (the order of the polynomial). Alternatively, in terms of Theorem 3.32 we would say that $p(z)$ has roots at 1, 1, 2, 2, 2, 2, i and $-i$.

Example 46: For each of the following polynomials, find all roots and hence factorise them in to real linear factors and real irreducible quadratic factors.

1. $p(z) = 2z^3 - 12z^2 + 22z - 12$.
2. $p(z) = z^4 - 4$. ⌟

3.15 Basic inequalities in $\mathbb{C}$

Theorem 3.33. *Solutions of the equation*

$$|z| = \alpha, \tag{3.23}$$

where $\alpha \in \mathbb{R}^+$, are represented in the Argand diagram by the points on a circle with centre at $z = 0$ and radius α.

Proof: Because both sides of equation (3.23) are positive real numbers, we can square each side:

$$|z| = \alpha \quad \Longleftrightarrow \quad |z|^2 = \alpha^2$$

$$\iff \quad z \cdot \overline{z} = \alpha^2.$$

If we substitute $z = x + iy$ into this equation, we get

$$\begin{aligned} |z| = \alpha \quad &\iff \quad (x+iy)(x-iy) = \alpha^2 \\ &\iff \quad x^2 + y^2 = \alpha^2, \end{aligned}$$

which represents a circle in the Argand diagram with centre at $z = 0$ and radius α, as required.

Corollary 3.34. *Solutions of the inequality*

$$|z| < \alpha,$$

where $\alpha \in \mathbb{R}^+$, are represented in the Argand diagram by the points inside a circle with centre at $z = 0$ and radius α.

Corollary 3.35. *Solutions of the inequality*

$$|z| > \alpha,$$

where $\alpha \in \mathbb{R}^+$, are represented in the Argand diagram by the points outside a circle with centre at $z = 0$ and radius α.

The proofs of Corollary 3.34 and Corollary 3.35 follow the same steps as the proof of Theorem 3.33 with appropriate inequalities replacing equalities.

Example 47: Graphically represent (sketch) the set of solutions of:

1. $|z| \leq 3$.
2. $|z| > 2$. ⌟

Theorem 3.36. *The modulus*

$$|\alpha - \beta|,$$

where $\alpha, \beta \in \mathbb{C}$, represents the length of the line segment in the Argand diagram connecting the points representing the complex numbers α and β.

Proof: If we write $\alpha = a + ib$ and $\beta = c + id$, then it follows that

$$|\alpha - \beta| \quad = \quad |(a+ib) - (c-id)| = |(a-c) - i(b-d)|$$

$$= \sqrt{(a-c)^2 + (c-d)^2}.$$

In the Argand diagram, this represents the length of the line segment connecting the points with coordinates $a+ib$ and $c+id$, i.e. the points representing $\alpha, \beta \in \mathbb{C}$, as required.

Corollary 3.37. *Solutions of the equation*

$$|z - z_0| = \alpha,$$

where $z_0 \in \mathbb{C}$ and $\alpha \in \mathbb{R}^+$, are represented in the Argand diagram by the points on a circle with centre at z_0 and radius α.

Corollary 3.38. *Solutions of the inequality*

$$|z - z_0| < \alpha,$$

where $z_0 \in \mathbb{C}$ and $\alpha \in \mathbb{R}^+$, are represented in the Argand diagram by the points inside a circle with centre at z_0 and radius α.

Corollary 3.39. *Solutions of the inequality*

$$|z - z_0| > \alpha,$$

where $z_0 \in \mathbb{C}$ and $\alpha \in \mathbb{R}^+$, are represented in the Argand diagram by the points outside a circle with centre at z_0 and radius α.

Example 48: Graphically represent (sketch) the set of solutions of:

1. $|z-1| < 3$.
2. $|z+1-2i| \geq 2$.

Chapter 4

Linear Equations and Matrices

▸ **Learning Outcomes** ◂ After the completion of the lectures and support sessions associated with this chapter you should be able to:

- add and multiply matrices;
- use matrices to solve systems of linear equations;
- prove basic results about matrices.

[2, p.602-612] and [4, Chapters SLE and M] contain alternative presentations of material in this chapter that you may find helpful.

It is likely that you have seen the equation of a line in $\mathbb{R}^2$ (in Cartesian form), specifically,

$$a_1x_1 + a_2x_2 = b, \tag{4.1}$$

with constants $a_1, a_2, b \in \mathbb{R}$ such that a_1 and a_2 are not both 0. A line is a 1-dimensional set (intuitively, this can be seen since every point on the line can be described by varying one parameter, say x_1 (since x_2 is then determined from x_1).

Similarly,in Chapter 1 we say that given a vector $\mathbf{n} = (a_1, a_2, a_3)$ and a real number $b \in \mathbb{R}$, then a plane in $\mathbb{R}^3$ is given by the vector equation $\mathbf{n} \cdot \mathbf{r} = b$. Therefore in $\mathbb{R}^3$, the equation

$$a_1x_1 + a_2x_2 + a_3x_3 = b \tag{4.2}$$

with constants $a_1, a_2, a_3, b \in \mathbb{R}$ such that a_1, a_2 and a_3 are not all equal to 0, defines a plane (a 2-dimensional set) in $\mathbb{R}^3$. Again, intuitively, this can be seen since every point on the plane can be described by varying two parameters, say x_1 and x_2 (since x_3 is then determined from x_1 and x_2).

Example 49:

1. In $\mathbb{R}^2$, the equation $x_1 = 0$ defines the x_2-axis.

2. In $\mathbb{R}^3$, the equation $x_3 = 1$ defines the plane parallel to the x_1x_2-plane and passing through the point $(x_1, x_2, x_3) = (0, 0, 1)$ as we saw in Chapter 1.

3. In general, in $\mathbb{R}^n$ for some $n \in \mathbb{N}$ $(n \geq 4)$, the equation

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b \tag{4.3}$$

with constants $a_i \in \mathbb{R}$ for $i = 1 \ldots n$, such that a_i are not all equal to 0, defines a $(n-1)$-dimensional **hyperplane** in $\mathbb{R}^n$.

4.1 Simultaneous linear equations

The equation, as given in (4.3) is known as a **linear equation** in the unknowns $x_1, x_2, \ldots, x_n$. Moreover, the **real constants** a_i in (4.3) are referred to as the coefficients of x_i (for each $i = 1, \ldots, n$) and b is also a real number. For instance, we would say, "the coefficient of x_2 is a_2". Moreover, x_i for $i = 1, \ldots, n$ are referred to as **real unknowns**. We can also define linear equations where the unknowns, the coefficients and the candidates for b are rational numbers, complex numbers and more generally members of any field $\mathbb{F}$. For the development in this chapter, we will consider real linear equations.

Definition 4.1. *A system of simultaneous linear equations in the unknowns $x_1, x_2, \ldots, x_n$ is:*

$$\begin{array}{ccccccccccc}
a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + & \cdots & + & a_{1n}x_n & = & b_1 \\
a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & + & \cdots & + & a_{2n}x_n & = & b_2 \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & \vdots & \vdots \\
a_{m1}x_1 & + & a_{m2}x_2 & + & a_{m3}x_3 & + & \cdots & + & a_{mn}x_n & = & b_m
\end{array}$$

with $b_i, a_{ij} \in \mathbb{R}$ for $i, j \in \mathbb{Z}$ with $1 \leq i \leq m$ and $1 \leq j \leq n$.

Remark 4.2. *The notation a_{11}, a_{mn} etc. is traditional. Here a_{11} represents the coefficient of x_1 is the first equation. Similarly, a_{mn} is the coefficient of x_n in the m-th equation. But what is a_{111}? It is ambiguous! A more precise notation is $a_{1,1}$ or $a_{m,n}$. Then we would write either $a_{11,1}$ or $a_{1,11}$ and we would have no possible confusion. Notice also that the notation for coefficients follows the rule*

$$a_{row,column}.$$

Example 50: The system of simultaneous linear equations given by:

$$\begin{array}{ccccccccc} x_1 & + & 2x_2 & + & 3x_3 & + & 4x_4 & = & 10 \\ x_1 & - & x_2 & + & x_3 & - & x_4 & = & 0 \end{array}$$

has $m = 2$ and $n = 4$, i.e. there are 2 equations in 4 unknowns.

Example 51: The system of simultaneous linear equations given by:

$$\begin{array}{ccccc} x_1 & + & x_2 & = & 1 \\ x_1 & + & 2x_2 & = & 3 \\ 2x_1 & + & 2x_2 & = & 1 \\ x_1 & - & x_2 & = & 0 \end{array}$$

has $m = 4$ and $n = 2$, i.e. there are 4 equations in 2 unknowns. Here, $a_{11} = 1$, $a_{12} = 1$, $a_{21} = 1$, $a_{22} = 2$, $a_{31} = 2$, $a_{32} = 2$, $a_{41} = 1$ and $a_{42} = -1$, with $b_1 = 1$, $b_2 = 3$, $b_3 = 1$ and $b_4 = 0$.

Definition 4.3. *The system of simultaneous linear equations:*

$$\begin{array}{ccccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & + & \cdots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots & \vdots & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & a_{m3}x_3 & + & \cdots & + & a_{mn}x_n & = & b_m \end{array}$$

has a solution

$$(x_1, x_2, \ldots, x_n) = (c_1, c_2, \ldots, c_n)$$

whenever the values $x_i = c_i$ for $i = 1, \ldots, n$ simultaneously satisfy all m equations above.

Definition 4.4. *The system of simultaneous linear equations:*

$$\begin{array}{ccccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + & \cdots & + & a_{1n}x_n & = & 0 \\ a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & + & \cdots & + & a_{2n}x_n & = & 0 \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots & \vdots & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & a_{m3}x_3 & + & \cdots & + & a_{mn}x_n & = & 0 \end{array}$$

is called ***homogeneous****. If a system of simultaneous linear equations is not homogeneous, then it's called* ***non-homogeneous****.*

Remark 4.5. *Homogeneous systems always have at least one solution called "the trivial solution", given by*

$$x_1 = x_2 = \cdots = x_n = 0.$$

Indeed, recalling Chapter 1 we know that if there are 3 unknowns, then each linear equation defines a plane which contains the origin $(0, 0, 0)$ and so this is a simultaneous solution when there are 3 unknowns.

We will show that for any such system of simultaneous linear equations, either:

- there exist no solutions;
- there exists a unique solution; or
- there exist infinitely many solutions.

4.2 Introduction to Matrices

The essential information in a system of simultaneous linear equations discussed previously is contained in the coefficients, which are naturally displayed as an array of numbers, for example,

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{pmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}. \tag{4.4}$$

Such an array is known as a **matrix**.

Example 52: Both

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \text{ and } \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

are matrices. Round brackets or square brackets can be used for matrices.

The **shape** or **dimension** of a matrix is measured by:

- the number of rows, denoted by m; and
- the number of columns, denoted by n.

In (4.4), we would state that "we have an $m \times n$ matrix", or, "the dimension of the matrix is $m \times n$". Remember that rows come first.

Definition 4.6. *Given $m, n \in \mathbb{N}$, we denote the set of all $m \times n$ matrices by $\mathcal{M}_{mn}(\mathbb{R})$.*

[1]

Example 53: $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ is a 2×1 matrix. $\begin{pmatrix} 5 & 6 & 7 & 8 \end{pmatrix}$ is a 1×4 matrix.

[1] More generally, we define the set $\mathcal{M}_{mn}(\mathcal{X})$ to be the set of $m \times n$ matrices with entries in the set $\mathcal{X}$. In this course, we will primarily consider $\mathcal{X} = \mathbb{R}$, but in later courses the sets of matrices $\mathcal{M}_{mn}(\mathbb{C})$ will be very important.

When the shape/dimension of a matrix $\mathbf{A}$ is understood from context, we often write

$$\mathbf{A} = [a_{ij}]$$

where a_{ij} is understood to be the entry in the i-th row and the j-th column of the matrix.

Example 54: What is the $(2,3)$ entry in $\mathbf{A} = [a_{ij}]$ given that $a_{ij} = i + j$?

Answer: The top left corner of the matrix is given as,

$$\begin{pmatrix} 2 & 3 & 4 & 5 & \cdots \\ 3 & 4 & \boxed{5} & 6 & \cdots \\ 4 & 5 & 6 & 7 & \cdots \\ 5 & 6 & 7 & 8 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

The answer is of course $a_{2,3} = 5$. **Note** again that when m and n are small (less than 10), we often write a_{23} to denote the $(2,3)$ entry of a matrix, whereas, if m or n are large (larger than 10), then we separate the subscripts with a comma, i.e $a_{11,3}$.

Definition 4.7. *Two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are equal if and only if:*

(i) Both have the same dimension - say $m \times n$.

(ii) $a_{ij} = b_{ij}$ for all $i, j \in \mathbb{Z}$ with $1 \leq i \leq m$ and $1 \leq j \leq n$.

4.3 Matrix Addition

Matrix addition is an internal binary operation on $\mathcal{M}_{mn}(\mathbb{R})$.

Definition 4.8. *Given two $m \times n$ matrices $\boldsymbol{A} = [a_{ij}] \in \mathcal{M}_{mn}(\mathbb{R})$ and $\boldsymbol{B} = [b_{ij}] \in \mathcal{M}_{mn}(\mathbb{R})$, we define $\boldsymbol{A} + \boldsymbol{B} \in \mathcal{M}_{mn}(\mathbb{R})$ to be the $m \times n$ matrix, given by:*

$$\boldsymbol{A} + \boldsymbol{B} = [a_{ij} + b_{ij}].$$

I.e. $\mathbf{A} + \mathbf{B}$ is the matrix where corresponding entries in $\mathbf{A}$ and $\mathbf{B}$ are added together,

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

Definition 4.9. *1. The matrix* $\boldsymbol{0}_{m\times n} \in \mathcal{M}_{mn}(\mathbb{R})$ *denotes the* $m \times n$ *matrix with all entries equal to zero.*

2. Given an $m \times n$ *matrix* $\boldsymbol{A} = [a_{ij}] \in \mathcal{M}_{mn}(\mathbb{R})$*, the* ***negative*** *of* A*,* $-\boldsymbol{A} \in \mathcal{M}_{mn}(\mathbb{R})$*, is the* $m \times n$ *matrix with entries* $[-a_{ij}]$*.*

I.e.

$$\mathbf{0}_{2\times 3} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

It also follows that

$$\mathbf{A} + (-\mathbf{A}) = (-\mathbf{A}) + \mathbf{A} = \mathbf{0}_{m\times n}.$$

In future we shall use $\mathbf{A} - \mathbf{B}$ for $\mathbf{A} + (-\mathbf{B})$.

4.4 Properties of Matrix Addition

The addition operations on two $m \times n$ matrices satisfies a number of *key properties*:

1. The set $\mathcal{M}_{mn}(\mathbb{R})$ is **closed** under addition, or, in other words, addition is an **internal** binary operation in the set $\mathcal{M}_{m\times n}(\mathbb{R})$:
$$\forall \mathbf{A}, \mathbf{B} \in \mathcal{M}_{mn}(\mathbb{R}), \mathbf{A} + \mathbf{B} \in \mathcal{M}_{mn}(\mathbb{R}).$$
2. Matrix addition is **associative** in $\mathcal{M}_{mn}(\mathbb{R})$: for all $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathcal{M}_{mn}(\mathbb{R})$,
$$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}).$$
3. The matrix $\mathbf{0}_{m\times n}$ is an **identity** in $\mathcal{M}_{mn}(\mathbb{R})$: for all $A \in \mathcal{M}_{mn}(\mathbb{R})$,
$$\mathbf{A} + \mathbf{0}_{m\times n} = \mathbf{0}_{m\times n} + \mathbf{A} = \mathbf{A}.$$
4. Given $A \in \mathcal{M}_{mn}(\mathbb{R})$, $-A \in \mathcal{M}_{mn}(\mathbb{R})$ satisfies
$$\mathbf{A} + (-\mathbf{A}) = (-\mathbf{A}) + \mathbf{A} = \mathbf{0}_{m\times n}.$$
5. Matrix addition is **commutative**: for all $\mathbf{A}, \mathbf{B} \in \mathcal{M}_{mn}(\mathbb{R})$,
$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}.$$

These 5 properties together show that $\mathcal{M}_{mn}(\mathbb{R})$ is an abelian group.

On the basis of these properties, one can prove many other properties about matrices in $\mathcal{M}_{mn}(\mathbb{R})$. For example, one can prove that the inverse of a matrix with respect to addition is unique.

Theorem 4.10. $(\mathcal{M}_{mn}(\mathbb{R}), +)$ *is an abelian group.*

4.5 Scalar Multiple of a Matrix

Definition 4.11. *Given a $m \times n$ matrix $\boldsymbol{A} = [a_{ij}]$ and $k \in \mathbb{R}$, the **scalar multiple** $k\boldsymbol{A}$ is the $m \times n$ matrix which has (i, j) entry ka_{ij}, i.e.*

$$k\boldsymbol{A} = [ka_{ij}].$$

So, if

$$\mathbf{A} = \begin{bmatrix} 9 & 11 \\ 3 & -1 \end{bmatrix}$$

then

$$2\mathbf{A} = \begin{bmatrix} 18 & 22 \\ 6 & -2 \end{bmatrix}.$$

Note that k can be any real number, i.e.

$$-\frac{1}{3}\mathbf{A} = \begin{bmatrix} -3 & -\frac{11}{3} \\ -1 & \frac{1}{3} \end{bmatrix}.$$

Save this theorem for your revision.

Theorem 4.12. *The scalar multiplication defined in Definition 4.11 together with matrix addition makes $\mathcal{M}_{mn}(\mathbb{R})$ into a vector space over $\mathbb{R}$.*

4.6 Matrix Multiplication

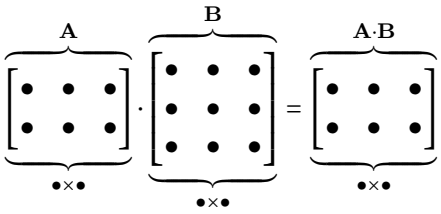
Definition 4.13. *Let $\boldsymbol{A} = [a_{ij}] \in \mathcal{M}_{mn}(\mathbb{R})$ have dimension $m \times n$ and $\boldsymbol{B} = [b_{ij}] \in \mathcal{M}_{np}(\mathbb{R})$ have dimension $n \times p$. The matrix product*

$$\boldsymbol{C} = [c_{ij}] = \boldsymbol{A} \cdot \boldsymbol{B} \in \mathcal{M}_{mp}(\mathbb{R})$$

where

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj} = a_{i1}b_{1j} + a_{i2}b_{2j} + a_{i3}b_{3j} + \cdots + a_{in}b_{nj}.$$

Note that the number of columns in $\mathbf{A}$ is equal to the number of rows in $\mathbf{B}$ in Definition 4.13. For example



(4.5)

Note in (4.5), $\mathbf{B} \cdot \mathbf{A}$ is not defined, since the matrix dimensions do not match for matrix multiplication. We now check the *key properties* for matrix multiplication.

1. When two matrices can be multiplied together, the result is another matrix, but because there is no guarantee that the product of any two given matrices exists, we cannot say that matrix multiplication is an **internal** operation over all matrices. If we consider the set $\mathcal{M}_{nn}(\mathbb{R})$, i.e. the set of all square $n \times n$ matrices for a given n, we can say that $\mathcal{M}_{nn}(\mathbb{R})$ is **closed** under matrix multiplication (or that matrix multiplication is an **internal** operation in this set) since any two matrices in $\mathcal{M}_{nn}(\mathbb{R})$ can be multiplied together, to make another $n \times n$ matrix.

2. For any three matrices that can be multiplied in a given sequence, matrix multiplication is **associative**, i.e.

$$(\mathbf{A} \cdot \mathbf{B}) \cdot \mathbf{C} = \mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{C}) = \mathbf{A} \cdot \mathbf{B} \cdot \mathbf{C}.$$

3. For any matrix, one can find a matrix that acts like an identity with respect to matrix multiplication.

4. When it makes sense, the distributive laws hold. For example, if $\mathbf{A} \in \mathcal{M}_{mn}(\mathbb{R})$ and $\mathbf{B}$ and $\mathbf{C}$ are in $\mathcal{M}_{np}(\mathbb{R})$, then

$$\mathbf{A} \cdot (\mathbf{B} + \mathbf{C}) = \mathbf{A} \cdot \mathbf{B} + \mathbf{A} \cdot \mathbf{C}.$$

This is called an *identity matrix*:

Definition 4.14. *The $n \times n$ **identity matrix** is:*

$$\boldsymbol{I}_n = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix} \in \mathcal{M}_{nn}(\mathbb{R}).$$

That is, the $n \times n$ matrix with (i, j) entry given by

$$\begin{cases} 1 & : \quad \text{when } i = j \\ 0 & : \quad \text{when } i \neq j. \end{cases}$$

For example:

$$\mathbf{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Additionally, given a 2×2 matrix, say

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

then it follows that

$$\mathbf{A} \cdot \mathbf{I}_2 = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \mathbf{I}_2 \cdot \mathbf{A}.$$

More generally, for any $n \times n$ matrix $\mathbf{A}$ we have,

$$\mathbf{A} \cdot \mathbf{I}_n = \mathbf{A} = \mathbf{I}_n \cdot \mathbf{A},$$

and for any $m \times n$ matrix $\mathbf{A}$ we have,

$$\mathbf{A} \cdot \mathbf{I}_n = \mathbf{A} = \mathbf{I}_m \cdot \mathbf{A}.$$

The last case clearly shows that there is not necessarily a single matrix that acts as the identity when multiplying to the left or to the right, so there is no identity in the set of all matrices with respect to matrix multiplication. There is a well defined identity in the set of $n \times n$ square matrices, $\mathcal{M}_{nn}(\mathbb{R})$, i.e. $\mathbf{I}_n$, so matrix multiplication has an **identity** in $\mathcal{M}_{nn}(\mathbb{R})$ (for each $n \in \mathbb{N}$).

5. It is clear that the zero matrix, $\mathbf{0}_{m\times n}$ cannot have an **inverse**, but do all other matrices have an inverse? We will discuss this briefly in the next section, and moreover devote an entire chapter to this issue (see Chapter 5).

6. We have seen earlier that matrix dimension may not allow us to calculate $\mathbf{B} \cdot \mathbf{A}$ even if the product $\mathbf{A} \cdot \mathbf{B}$ exists. Furthermore, even if $\mathbf{A}$ and $\mathbf{B}$ are both $n \times n$ matrices, then it need not be the case that $\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A}$. Hence, matrix multiplication is not **commutative**.

Example 55: Let $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. Then,

$$\mathbf{A} \cdot \mathbf{B} = \begin{bmatrix} 0 & 1 \\ 0 & 3 \end{bmatrix} \text{ and } \mathbf{B} \cdot \mathbf{A} = \begin{bmatrix} 3 & 4 \\ 0 & 0 \end{bmatrix}.$$

Hence matrix multiplication is not commutative in $\mathcal{M}_{nn}(\mathbb{R})$. An indication that other familiar rules do not hold is given by

Example 56: Let $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$. Then

$$\mathbf{A} \cdot \mathbf{B} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Some specific types of matrices, given in Definitions 4.15 and 4.16 will appear throughout your studies:

Definition 4.15. *An $n \times n$ matrix $A = [a_{ij}] \in \mathcal{M}_{nn}(\mathbb{R})$ is **diagonal** when:*

$$a_{i,j} = 0 \quad \text{for } i \neq j$$

i.e.

$$A = \begin{pmatrix} a_{1,1} & 0 & \cdots & \cdots & 0 \\ 0 & a_{2,2} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{n-1,n-1} & 0 \\ 0 & 0 & \cdots & 0 & a_{n,n} \end{pmatrix}.$$

Diagonal matrices are easy to work with, particularly the set of diagonal matrices is closed under both matrix addition and multiplication.

Example 57: If

$$\mathbf{A} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix},$$

then we have

$$\mathbf{A}^2 = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \cdot \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} = \begin{pmatrix} 4 & 0 \\ 0 & 9 \end{pmatrix}$$

and also that

$$\mathbf{A}^{10} = \begin{pmatrix} 2^{10} & 0 \\ 0 & 3^{10} \end{pmatrix}.$$

Consider how much harder it is to calculate $\mathbf{B}^{10}$ where

$$\mathbf{B} = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}.$$

Definition 4.16. *An **upper triangular matrix** $\mathbf{A} = [a_{ij}] \in \mathcal{M}_{nn}(\mathbb{R})$ is an $n \times n$ matrix with*

$$a_{ij} = 0 \quad \textit{for all } i > j$$

i.e.

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a_{22} & a_{23} & \cdots & a_{2n} \\ 0 & 0 & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & a_{nn} \end{pmatrix}.$$

*Similarly, a **lower triangular matrix** is a $n \times n$ matrix with:*

$$a_{ij} = 0 \quad \textit{for all } j > i$$

$$\begin{pmatrix} a_{11} & 0 & \cdots & 0 & 0 \\ a_{21} & a_{22} & 0 & \cdots & 0 \\ a_{31} & a_{32} & a_{33} & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{pmatrix}.$$

4.7 Matrix Inverses

Definition 4.17. *Let $\boldsymbol{A} \in \mathcal{M}_{nn}(\mathbb{R})$ be an $n \times n$ matrix. **If** there exists an $n \times n$ matrix $\boldsymbol{B}$ such that*

$$\boldsymbol{A} \cdot \boldsymbol{B} = \boldsymbol{I}_n = \boldsymbol{B} \cdot \boldsymbol{A}$$

*then $\boldsymbol{A}$ is said to be **invertible** and we write*

$$\boldsymbol{B} = \boldsymbol{A}^{-1}.$$

*$\boldsymbol{B}$ is called **the** inverse of $\boldsymbol{A}$.*

Note that it is also common to refer to invertible matrices as "**non-singular**" and matrices which are not invertible are then called "**singular**". To clarify, the inverse of a matrix $\mathbf{A}$, by definition, refers to the inverse with respect to matrix multiplication. To justify the "the" in Definition 4.17 (which implies uniqueness of the inverse) we prove:

Theorem 4.18. *Any $n \times n$ matrix has **at most one** inverse.*

Proof: Suppose that $\mathbf{A} \in \mathcal{M}_{nn}(\mathbb{R})$ and that $\mathbf{B}_1, \mathbf{B}_2 \in \mathcal{M}_{nn}(\mathbb{R})$ are both inverses of $\mathbf{A}$. Then

$$\mathbf{A} \cdot \mathbf{B}_1 = \mathbf{I}_n = \mathbf{B}_1 \cdot \mathbf{A} \tag{4.6}$$

and

$$\mathbf{A} \cdot \mathbf{B}_2 = \mathbf{I}_n = \mathbf{B}_2 \cdot \mathbf{A}. \tag{4.7}$$

Now we calculate that

$$\begin{aligned} \mathbf{B}_1 &= \mathbf{B}_1 \cdot \mathbf{I}_n \\ &= \mathbf{B}_1 \cdot \mathbf{A} \cdot \mathbf{B}_2 \quad \text{via (4.7)} \\ &= \mathbf{I}_n \cdot \mathbf{B}_2 \quad \text{via (4.6)} \\ &= \mathbf{B}_2, \end{aligned}$$

and hence $\mathbf{B}_1 = \mathbf{B}_2$. Therefore, there can be at most one inverse of an $n \times n$ matrix, as required.

For the first time, in the next example, we see that it is useful to be able to have inverses of elements in $\mathbb{R}$. We'll learn more about this in the shortly.

Example 58: For $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ is $\mathbf{B} = \begin{pmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{pmatrix}$ equal to $\mathbf{A}^{-1}$?

Answer: We should check if $\mathbf{B}$ satisfies the conditions in Definition 4.17 i.e., if

$$\mathbf{A} \cdot \mathbf{B} = \mathbf{I}_2 = \mathbf{B} \cdot \mathbf{A}.$$

Since

$$\begin{aligned} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \cdot \begin{pmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{pmatrix} &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \\ \begin{pmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \end{aligned}$$

we conclude that $\mathbf{B} = \mathbf{A}^{-1}$.

Note that not every matrix is invertible, for example, $\mathbf{0}_{2\times 2}$ is clearly not invertible.

Question: How was $\mathbf{A}^{-1}$ in Example 58 determined? Given a matrix $\mathbf{B}$ it is easy to check if it is equal to $\mathbf{A}^{-1}$ (simply check that it satisfies the conditions of Definition 4.17). But how would we have found it if we were not given it? We address this question in Chapter 5.

4.8 Matrices and Linear Equations

We can now see that the system of simultaneous linear equations:

$$\begin{array}{ccccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & + & \cdots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & \vdots & & & & \vdots & \vdots & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & a_{m3}x_3 & + & \cdots & + & a_{mn}x_n & = & b_m \end{array} \tag{4.8}$$

can be written, equivalently, in matrix form:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}, \tag{4.9}$$

or even more concisely as:

$$\mathbf{A} \cdot \mathbf{x} = \underline{b}, \tag{4.10}$$

where $\mathbf{A} = [a_{ij}]$, $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$.

We refer to $\mathbf{A}$ as the matrix associated with the corresponding system of simultaneous linear equations. It stores information about the coefficients.

The **augmented matrix** for the systems of simultaneous linear equations given in (4.8)-(4.10) has an extra column and is given by:

$$\left(\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} & b_m \end{array}\right). \tag{4.11}$$

The dimension of the augmented matrix is therefore $(m \times (n+1))$ and the final column stores the information contained on the right hand side of the (4.8) (or (4.9) or (4.10)).

4.9 Elementary Row Operations

There are a number of "allowable" things we can do to a system of simultaneous linear equations which *will not affect their solution*. For example, we might multiply every term in an equation by a non-zero real number. That is, the set

$$\{(x, y) \in \mathbb{R}^2 \ : \ 2x + 3y = 10\}$$

is equal to

$$\{(x, y) \in \mathbb{R}^2 \ : \ x + \frac{3}{2}y = 5\}.$$

Such manipulations translate to "allowable" operations on the rows of the augmented matrix. These are known as **elementary row operations** (abbreviated to EROs) and are listed below:

(i) interchange/switch two rows;

(ii) multiply a row by a non-zero constant; and

(iii) add a multiple of one row to another row.

Theorem 4.19. *EROs do not alter the solution set of a system of simultaneous linear equations.*

Proof: It is clear that if ERO (i) is applied to a system of simultaneous linear equations, then the solution set remains the same.

Now, observe that for $a, b, c, d \in \mathbb{R}$ with $c \neq 0$,

$$a = b \iff ac = bc, \tag{4.12}$$

$$a = b \iff a + d = b + d. \tag{4.13}$$

If follows from (4.12) and (4.13) respectively, that if ERO (ii) or ERO (iii) are applied to a system of simultaneous linear equations, then the solution set remains the same, as required.

Our strategy to solve systems of simultaneous linear equations will be to use EROs to systematically reduce the associated augmented matrix to a form that is easier to use.

Definition 4.20. *A matrix is in (row)* ***echelon form*** *when:*

- *all rows consisting of only zeros are at bottom of the matrix;*
- *the first non-zero number (in order from left to right) in any row is '1'; and*
- *successive non-zero rows begin with more zeros left of the '1' than the rows above.*

A matrix is in ***reduced echelon form*** *if it is in echelon form and the first non-zero entry in each row is the only non-zero entry in its column.*

Example 59: The matrices

$$\begin{bmatrix} 1 & \star & \star & \star & \star & \star & \star \\ 0 & 0 & 1 & \star & \star & \star & \star \\ 0 & 0 & 0 & 1 & \star & \star & \star \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & \star & \star & \star & \star & \star \\ 0 & 1 & \star & \star & \star & \star \\ 0 & 0 & 1 & \star & \star & \star \\ 0 & 0 & 0 & 1 & \star & \star \\ 0 & 0 & 0 & 0 & 1 & \star \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \text{ and } \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

are in row echelon form and

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

is in reduced echelon form.

4.10 Gaussian Elimination

Gaussian elimination, Gauss-Jordan elimination or simply **row reduction** are all names[2] given to the process of successive applications of EROs to transform a matrix to echelon form. It should be clear that the equations represented by an echelon matrix are much easier to solve by a process of "backwards substitution".

Thus we have a strategy for solving linear equations:

1. form the augmented matrix;
2. use ERO to reduce the augmented matrix to echelon form; and
3. solve the problem given in echelon form (if possible) by backwards substitution or by reducing to reduced echelon form.

Example 60: Determine the solution set for

$$\begin{array}{rcrcl} 2x & + & y & = & 3, \\ x & - & y & = & 3. \end{array} \tag{4.14}$$

Answer: The augmented matrix associated with the system of simultaneous linear equations is

$$\left(\begin{array}{cc|c} 2 & 1 & 3 \\ 1 & -1 & 3 \end{array}\right).$$

Applying the following EROs:

$$\boxed{r_1 \to r_1 - r_2} \qquad \left(\begin{array}{cc|c} 1 & 2 & 0 \\ 1 & -1 & 3 \end{array}\right)$$

$$\boxed{r_2 \to r_2 - r_1} \qquad \left(\begin{array}{cc|c} 1 & 2 & 0 \\ 0 & -3 & 3 \end{array}\right)$$

[2] If interested in the history of how this process got its name, see [3].

$$\boxed{r_2 \to r_2/(-3)} \qquad \left(\begin{array}{cc|c} 1 & 2 & 0 \\ 0 & 1 & -1 \end{array}\right)$$

demonstrates that the system of simultaneous linear equations in (4.14) is equivalent to

$$\begin{array}{rcrcl} x & + & 2y & = & 0, \\ & & y & = & -1. \end{array}$$

Therefore, $y = -1$ and so $x = 2$, and hence the solution set is $\{(2,-1)\}$.

Alternatively, we can take the EROs one step further and produce a matrix in reduced echelon form

$$\boxed{r_1 \to r_1 - 2r_2.} \qquad \left(\begin{array}{cc|c} 1 & 0 & 2 \\ 0 & 1 & -1 \end{array}\right)$$

and we read the solution to be $x = 2$ and $y = 1$.

Example 61: Determine the solution set for

$$\begin{array}{rcrcl} 2x & + & 2y & = & 4, \\ x & + & y & = & 2. \end{array} \tag{4.15}$$

Answer: The augmented matrix associated with the system of simultaneous linear equations is

$$\left(\begin{array}{cc|c} 2 & 2 & 4 \\ 1 & 1 & 2 \end{array}\right).$$

Applying the following EROs:

$$\boxed{r_1 \to \frac{1}{2}r_1} \qquad \left(\begin{array}{cc|c} 1 & 1 & 2 \\ 1 & 1 & 2 \end{array}\right)$$

$$\boxed{r_2 \to r_2 - r_1} \qquad \left(\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & 0 & 0 \end{array}\right)$$

demonstrates that the system of simultaneous linear equations in (4.15) is equivalent to

$$\begin{array}{rcrcl} x & + & y & = & 2, \\ 0 & + & 0 & = & 0. \end{array}$$

Therefore, y can be an arbitrary real number, say $y = k \in \mathbb{R}$ which implies that $x = 2 - k$. Hence, the solution set is the line

$$\{(2-k, k) : k \in \mathbb{R}\},$$

i.e. there are infinitely many solutions. [This means that the lines defined by $2x + 2y = 4$ and $x + y = 2$ are equal.]

Example 62: Determine the solution set for

$$\begin{array}{rcrcrcl} 2x_1 & + & 8x_2 & + & 6x_3 & = & 20, \\ 4x_1 & + & 2x_2 & - & 2x_3 & = & -2, \\ -6x_1 & + & 4x_2 & + & 10x_3 & = & 30. \end{array} \tag{4.16}$$

Answer: The augmented matrix associated with the system of simultaneous linear equations is

$$\left(\begin{array}{ccc|c} 2 & 8 & 6 & 20 \\ 4 & 2 & -2 & -2 \\ -6 & 4 & 10 & 30 \end{array}\right)$$

Applying the following EROs:

$$\boxed{r_1 \to \tfrac{1}{2}r_1} \qquad \left(\begin{array}{ccc|c} 1 & 4 & 3 & 10 \\ 4 & 2 & -2 & -2 \\ -6 & 4 & 10 & 30 \end{array}\right)$$

$$\boxed{r_2 \to r_2 - 4r_1} \qquad \left(\begin{array}{ccc|c} 1 & 4 & 3 & 10 \\ 0 & -14 & -14 & -42 \\ -6 & 4 & 10 & 30 \end{array}\right)$$

$$\boxed{r_3 \to r_3 + 6r_1} \qquad \left(\begin{array}{ccc|c} 1 & 4 & 3 & 10 \\ 0 & -14 & -14 & -42 \\ 0 & 28 & 28 & 90 \end{array}\right)$$

$$\boxed{r_3 \to r_3 + 2r_2} \qquad \left(\begin{array}{ccc|c} 1 & 4 & 3 & 10 \\ 0 & -14 & -14 & -42 \\ 0 & 0 & 0 & 6 \end{array}\right)$$

demonstrates that the system of simultaneous linear equations in (4.16) is equivalent to

$$\begin{array}{rcrcrcl} x_1 & + & 4x_2 & + & 3x_3 & = & 10, \\ & - & 14x_2 & - & 14x_3 & = & -42, \\ & & & & 0 & = & 6. \end{array}$$

By considering the last equation, it can be seen that there is no solution to the system of simultaneous linear equations, i.e. the solution set is $\varnothing$.

Note that in Example 62, although the final step is not in Echelon form, we can conclude at this stage that the system has no solution. Examples 60-62 illustrate the three possibilities that can arise - a solution set with one element, a solution set with infinitely many elements, or a solution set that is empty (i.e. the system cannot be solved).

4.11 Guidance for Reducing to Echelon Form

- Obtain a '1' at the top of the first (left to right) non-zero column by applying ERO(i) and then ERO(ii) to the augmented matrix, then use EROs (ii) and (iii) to create 0s in the rest of the column below this '1'.
- Then move on to the next column.
- Solve the system of simultaneous linear equations (if possible) by "backward substitution".
- Avoid build up of many fractions by leaving division throughout a row until the end.

Recall that systems of simultaneous linear equations may have:

(i) a unique solution;

(ii) no solution, if, for example one of the equations is $0 = c$ where $c \neq 0$; or

(iii) infinitely many solutions, if, for example an equation reduces to $0 = 0$.

Equations with solution sets of types (i) or (iii) are referred to as **consistent**, otherwise they are referred to as **inconsistent**.

Infinitely many solutions occur when there are fewer (consistent) equations than unknowns. Here, one or more variable(s) must be assigned an arbitrary real value. There is an element of choice here, and so equivalent solutions may appear in different guises.

Example 63: Determine the solution set for

$$x + 2y = 10.$$

Answer: Setting $y = k$ where $k \in \mathbb{R}$ is arbitrary, we have

$$x = 10 - 2k,$$

which gives the solution set as $\{(10 - 2k, k) : k \in \mathbb{R}\}$.

However, we could have chosen to set $x = t$ where $t \in \mathbb{R}$ is arbitrary, which gives $y = 5 - \frac{t}{2}$. Thus, the solution set is equivalently given by $\left\{\left(t, 5 - \frac{t}{2}\right) : t \in \mathbb{R}\right\}$.

Example 64: Determine the solution set for

$$\begin{array}{rcrcrcrcl} x_1 & + & 3x_2 & - & 5x_3 & + & x_4 & = & 4, \\ 2x_1 & + & 5x_2 & - & 2x_3 & + & 4x_4 & = & 6. \end{array} \tag{4.17}$$

Answer: The augmented matrix associated with the system of simultaneous linear equations is

$$\left(\begin{array}{cccc|c} 1 & 3 & -5 & 1 & 4 \\ 2 & 5 & -2 & 4 & 6 \end{array}\right).$$

Applying the following EROs:

$$\boxed{r_2 \to r_2 - 2r_1} \qquad \left(\begin{array}{cccc|c} 1 & 3 & -5 & 1 & 4 \\ 0 & -1 & 8 & 2 & -2 \end{array}\right)$$

$$\boxed{r_2 \to (-1)r_2} \qquad \left(\begin{array}{cccc|c} 1 & 3 & -5 & 1 & 4 \\ 0 & 1 & -8 & -2 & 2 \end{array}\right)$$

demonstrates that the system of simultaneous linear equations in (4.17) is equivalent to:

$$\begin{array}{rcrcrcl} x_1 & + & 3x_2 & - & 5x_3 & + & x_4 & = & 4, \\ & & x_2 & - & 8x_3 & - & 2x_4 & = & 2. \end{array}$$

These equations are consistent and have infinitely many of solutions. First set

$$x_4 = t \text{ and } x_3 = s$$

where $s, t \in \mathbb{R}$ are arbitrary. Then,

$$x_2 = 2 + 8s + 2t,$$

and hence,

$$\begin{aligned} x_1 &= 4 - 3x_2 + 5x_3 - x_4 \\ &= 4 - 3(2 + 8s + 2t) + 5s - t \\ &= -2 - 19s - 7t. \end{aligned}$$

Thus the solution set is

$$\{(-2 - 19s - 7t, 2 + 8s + 2t, s, t) : s, t \in \mathbb{R}\}.$$

Again, we could choose to compute the reduced echelon form:

$$\boxed{r_1 \to r_1 - 3r_2} \qquad \left(\begin{array}{cccc|c} 1 & 0 & 19 & 7 & -2 \\ 0 & 1 & -8 & -2 & 2 \end{array}\right)$$

and read of the solution from this.

Example 65: Find conditions upon a, b and c that ensure equations

$$\begin{array}{rcrcrcl} 2x & - & y & + & 3z & = & a \\ 3x & + & y & - & 5z & = & b \\ -5x & - & 5y & + & 21z & = & c \end{array} \tag{4.18}$$

are consistent.

Answer: The augmented matrix associated with the system of simultaneous linear equations is

$$\left(\begin{array}{ccc|c} 2 & -1 & 3 & a \\ 3 & 1 & -5 & b \\ -5 & -5 & 21 & c \end{array}\right).$$

Applying the following EROs:

$$\boxed{r_2 \to r_2 - r_1} \qquad \left(\begin{array}{ccc|c} 2 & -1 & 3 & a \\ 1 & 2 & -8 & -a+b \\ -5 & -5 & 21 & c \end{array}\right)$$

$$\boxed{r_1 \circlearrowleft r_2} \qquad \left(\begin{array}{ccc|c} 1 & 2 & -8 & -a+b \\ 2 & -1 & 3 & a \\ -5 & -5 & 21 & c \end{array}\right)$$

$$\boxed{\begin{array}{l} r_2 \to r_2 - 2r_1 \\ r_3 \to r_3 + 5r_1 \end{array}} \qquad \left(\begin{array}{ccc|c} 1 & 2 & -8 & -a+b \\ 0 & -5 & 19 & 3a-2b \\ 0 & 5 & -19 & -5a+5b+c \end{array}\right)$$

$$\boxed{r_3 \to r_3 + r_2} \qquad \left(\begin{array}{ccc|c} 1 & 2 & -8 & -a+b \\ 0 & -5 & 19 & 3a-2b \\ 0 & 0 & 0 & -2a+3b+c \end{array}\right)$$

demonstrates that the system of simultaneous linear equations in (4.18) is equivalent to

$$\begin{array}{rcrcrcl} x & + & 2y & - & 8z & = & -a+b, \\ & - & 5y & + & 19z & = & 3a-2b, \\ & & & & 0 & = & -2a+3b+c. \end{array}$$

Thus the equations are consistent if and only if $-2a+3b+c=0$. We can make this conclusion since if $-2a+3b+c \neq 0$ then the last equation cannot be satisfied. Moreover, if $-2a+3b+c=0$, then setting $z=k$ for $k \in \mathbb{R}$ arbitrary, it follows that

$$y = \frac{2b-3a+19k}{5},$$

and

$$\begin{aligned} x &= b-a+8k-2y \\ &= b-a+8k-\frac{4b-6a+38k}{5} \\ &= \frac{a+b+2k}{5}. \end{aligned}$$

Thus the solution set is

$$\left\{\left(\frac{a+b+2k}{5}, \frac{2b-3a+19k}{5}, k\right) : k \in \mathbb{R}\right\},$$

when $-2a+3b+c=0$, and $\varnothing$ when $-2a+3b+c \neq 0$.

Again we can produce the reduced echelon form:

$$\boxed{r_2 \to -r_2/5} \qquad \left(\begin{array}{ccc|c} 1 & 2 & -8 & -a+b \\ 0 & 1 & -19/5 & (2b-3a)/5 \\ 0 & 0 & 0 & -2a+3b+c \end{array}\right)$$

$$\boxed{r_1 \to r_1 - 2r_2,} \qquad \left(\begin{array}{ccc|c} 1 & 0 & -2/5 & (a+b)/5 \\ 0 & 1 & -19/5 & (2b-3a)/5 \\ 0 & 0 & 0 & -2a+3b+c \end{array}\right)$$

from which we can directly read the solution.

Note that in the above example, we answered the question (when is the system consistent?) without reducing the augmented matrix to its Echelon form.

4.12 Block Matrices*

For some problems involving matrices, it is helpful to represent the matrices in blocks. For example, for linear systems, the augmented matrix in (4.11) is a block matrix consisting of a $m \times n$ block (with entries a_{ij}) and a $m \times 1$ block (with entries b_j).[3] Moreover, in the Gaussian elimination algorithm (see Section 4.2), for $\mathbf{A} \in \mathcal{M}_{nn}(\mathbb{R})$ we manipulate an augmented matrix of dimension $n \times 2n$ to determine whether or not an inverse of $\mathbf{A}$ exists.

For another example, consider the matrix $\mathbf{A} \in \mathcal{M}_{22}(\mathbb{R})$ given by

$$\mathbf{A} = \begin{pmatrix} 2 & 3 \\ 1 & 1 \end{pmatrix} \tag{4.19}$$

which can be represented with the two $\mathcal{M}_{12}(\mathbb{R})$ matrices given by

$$\mathbf{A}_1 = \begin{pmatrix} 2 & 3 \end{pmatrix} \text{ and } \mathbf{A}_2 = \begin{pmatrix} 1 & 1 \end{pmatrix}, \tag{4.20}$$

specifically as

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{pmatrix}.$$

Similarly, $\mathbf{B} \in \mathcal{M}_{23}(\mathbb{R})$ given by

$$\mathbf{B} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \tag{4.21}$$

can be represented with the three $\mathcal{M}_{21}(\mathbb{R})$ matrices given by

$$\mathbf{B}_1 = \begin{pmatrix} 1 \\ 4 \end{pmatrix}, \quad \mathbf{B}_2 = \begin{pmatrix} 2 \\ 5 \end{pmatrix} \text{ and } \mathbf{B}_3 = \begin{pmatrix} 3 \\ 6 \end{pmatrix}, \tag{4.22}$$

[3] In these notes, simple block matrices related to the solution of linear systems, typically, are referred to as augmented matrices instead of block matrices.

specifically, as

$$\mathbf{B} = (\mathbf{B}_1 \mid \mathbf{B}_2 \mid \mathbf{B}_3).$$

When it is clear, the lines separating the blocks in a block representation of a matrix are omitted.

Example 66: Let $\mathbf{A} \in \mathcal{M}_{22}(\mathbb{R})$ be given by (4.19) and $\mathbf{B} \in \mathcal{M}_{23}(\mathbb{R})$ be given by (4.21) with (4.22). Establish that

$$\mathbf{A} \cdot \mathbf{B} = (\mathbf{A} \cdot \mathbf{B}_1 \mid \mathbf{A} \cdot \mathbf{B}_2 \mid \mathbf{A} \cdot \mathbf{B}_3),$$

i.e. that column j of $\mathbf{A} \cdot \mathbf{B}$ is $\mathbf{A} \cdot \mathbf{B}_j$.

Answer: Observe that

$$\mathbf{A} \cdot \mathbf{B} = \begin{pmatrix} 14 & 19 & 24 \\ 5 & 7 & 9 \end{pmatrix} \tag{4.23}$$

and

$$\mathbf{A} \cdot \mathbf{B}_1 = \begin{pmatrix} 14 \\ 5 \end{pmatrix}, \quad \mathbf{A} \cdot \mathbf{B}_2 = \begin{pmatrix} 19 \\ 7 \end{pmatrix}, \quad \mathbf{A} \cdot \mathbf{B}_3 = \begin{pmatrix} 24 \\ 9 \end{pmatrix}.$$

Therefore, $\mathbf{A} \cdot \mathbf{B} = (\mathbf{A} \cdot \mathbf{B}_1 \mid \mathbf{A} \cdot \mathbf{B}_2 \mid \mathbf{A} \cdot \mathbf{B}_3)$.

The general statement highlighted in Example 66 is given in

Proposition 4.21. *Let $\boldsymbol{A} \in \mathcal{M}_{mn}(\mathbb{R})$, $\boldsymbol{B} \in \mathcal{M}_{np}(\mathbb{R})$ and suppose $\boldsymbol{B}_j \in \mathcal{M}_{n1}(\mathbb{R})$ are the $j-th$ columns of $\boldsymbol{B}$ for $j = 1, \ldots, p$. Then,*

$$\boldsymbol{A} \cdot \boldsymbol{B} = \boldsymbol{A} \cdot (\boldsymbol{B}_1 \mid \boldsymbol{B}_2 \mid \ldots \mid \boldsymbol{B}_p) = (\boldsymbol{A} \cdot \boldsymbol{B}_1 \mid \boldsymbol{A} \cdot \boldsymbol{B}_2 \mid \ldots \mid \boldsymbol{A} \cdot \boldsymbol{B}_p).$$

Proof: Follows directly from Definition 4.13.

Similarly, we have

Example 67: Let $\mathbf{A} \in \mathcal{M}_{22}(\mathbb{R})$ be given by (4.19) with (4.20) and $\mathbf{B} \in \mathcal{M}_{23}(\mathbb{R})$ be given by (4.21). Establish that

$$\mathbf{A} \cdot \mathbf{B} = \begin{pmatrix} \mathbf{A}_1 \cdot \mathbf{B} \\ \mathbf{A}_2 \cdot \mathbf{B} \end{pmatrix},$$

i.e. that row i of $\mathbf{A} \cdot \mathbf{B}$ is $\mathbf{A}_i \cdot \mathbf{B}$.

Answer: Observe that

$$\mathbf{A}_1 \cdot \mathbf{B} = \begin{pmatrix} 14 & 19 & 24 \end{pmatrix} \quad \mathbf{A}_2 \cdot \mathbf{B} = \begin{pmatrix} 5 & 7 & 9 \end{pmatrix}.$$

Therefore, via (4.23), we have

$$\mathbf{A} \cdot \mathbf{B} = \begin{pmatrix} \mathbf{A}_1 \cdot \mathbf{B} \\ \mathbf{A}_2 \cdot \mathbf{B} \end{pmatrix}.$$

Similarly, Example 67 highlights the following result,

Proposition 4.22. *Let $\boldsymbol{A} \in \mathcal{M}_{mn}(\mathbb{R})$, $\boldsymbol{B} \in \mathcal{M}_{np}(\mathbb{R})$ and suppose $\boldsymbol{A}_i \in \mathcal{M}_{1n}(\mathbb{R})$ are the $i-th$ rows of $\boldsymbol{A}$ for $i = 1, \ldots, m$. Then,*

$$\boldsymbol{A} \cdot \boldsymbol{B} = \begin{pmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \\ \vdots \\ \boldsymbol{A}_m \end{pmatrix} \cdot \boldsymbol{B} = \begin{pmatrix} \boldsymbol{A}_1 \cdot \boldsymbol{B} \\ \boldsymbol{A}_2 \cdot \boldsymbol{B} \\ \vdots \\ \boldsymbol{A}_m \cdot \boldsymbol{B} \end{pmatrix}.$$

Proof: Follows directly from Definition 4.13.

Propositions 4.21 and 4.22 imply that for $\mathbf{A} \in \mathcal{M}_{mn}(\mathbb{R})$ and $\mathbf{B} \in \mathcal{M}_{nk}(\mathbb{R})$, each row/column in $\mathbf{A} \cdot \mathbf{B}$ is a sum of multiples of the rows/columns of $\mathbf{B}/\mathbf{A}$. This notion will be helpful to prove results in later chapters/practice questions.

Specifically, when we consider the general solution to linear systems of m equations in n unknowns, we will split a $m \times n$ matrix $\mathbf{A}$ into 4 blocks in the following form:

$$\mathbf{A} = \left(\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} \end{array}\right) \tag{4.24}$$

with $\mathbf{A}_{11}$ a $p \times p$ matrix, $\mathbf{A}_{12}$ a $p \times (n-p)$ matrix, $\mathbf{A}_{21}$ a $(m-p) \times p$ matrix, and $\mathbf{A}_{22}$ is a $(m-p) \times (n-p)$ matrix (with $1 \leq p < \min\{m, n\}$).

Example 68: For $\mathbf{A} \in \mathcal{M}_{45}(\mathbb{R})$ given by

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 0 & 6 & 7 & 8 & 9 \\ 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} = \left(\begin{array}{ccc|cc} 1 & 2 & 3 & 4 & 5 \\ 0 & 6 & 7 & 8 & 9 \\ 0 & 0 & 1 & 2 & 3 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right) = \left(\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} \end{array}\right),$$

we can represent $\mathbf{A}$ in the block structure detailed in (4.24) (setting $p = 3$), i.e. with,

$$\mathbf{A}_{11} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 6 & 7 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{A}_{12} = \begin{pmatrix} 4 & 5 \\ 8 & 9 \\ 2 & 3 \end{pmatrix}, \quad \mathbf{A}_{21} = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix} \text{ and } \mathbf{A}_{22} = \begin{pmatrix} 0 & 0 \end{pmatrix}.$$

Chapter 5

The Inverse of an Invertible Matrix

▸ **Learning Outcomes** ◂ After the completion of the lectures and support sessions associated with this chapter you should be able to:

- use Gaussian elimination to find the inverse of a square matrix when it exists;
- use elementary matrices to represent elementary row operations; and
- represent invertible matrices as a product of elementary matrices.

[2, p.602-612] and [4, Chapters SSLE and M] contain alternative presentations of material in this chapter that you may find helpful.

5.1 Introduction

Here we return to the question, first raised in Chapter 4, Section 4.7.

Recall from Definition 4.17 that an $n \times n$ matrix $\mathbf{A} \in \mathcal{M}_{nn}(\mathbb{R})$ is invertible with inverse $\mathbf{B}$ if and only if there exists an $n \times n$ matrix $\mathbf{B}$ that satisfies,

$$\mathbf{A} \cdot \mathbf{B} = \mathbf{I}_n = \mathbf{B} \cdot \mathbf{A}.$$

We saw in Theorem 4.18 that if the matrix $\mathbf{A}$ is invertible then it has a unique inverse and we denote this matrix by $\mathbf{A}^{-1}$.

Also recall that we have see that:

- not every matrix is invertible, for example $\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \in \mathcal{M}_{2\times 2}(\mathbb{R})$ or even $\mathbf{A} = \begin{pmatrix} 0 \end{pmatrix} \in \mathcal{M}_{1\times 1}(\mathbb{R})$.

- Definition 4.17 does not provide a means of finding $\mathbf{A}^{-1}$ when it exists, but it provides a means of deciding if a matrix $\mathbf{B}$ is equal to $\mathbf{A}^{-1}$.

In this chapter, we shall seek answers to the following questions for a general $n \times n$ matrix $\mathbf{A}$:

(i) Can we decide if $\mathbf{A}$ is invertible?

(ii) For an invertible matrix $\mathbf{A}$, is there a reasonably efficient routine which will allow $\mathbf{A}^{-1}$ to be calculated?

Theorem 5.1. *Suppose that $\boldsymbol{A}$ and $\boldsymbol{B}$ are invertible $n \times n$ matrices, then the matrix $\boldsymbol{A} \cdot \boldsymbol{B}$ is also invertible and*

$$(\boldsymbol{A} \cdot \boldsymbol{B})^{-1} = \boldsymbol{B}^{-1} \cdot \boldsymbol{A}^{-1}.$$

Proof: $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$ both exist, are $n \times n$ matrices, and satisfy:

$$\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{I}_n = \mathbf{A}^{-1} \cdot \mathbf{A}, \tag{5.1}$$

$$\mathbf{B} \cdot \mathbf{B}^{-1} = \mathbf{I}_n = \mathbf{B}^{-1} \cdot \mathbf{B}. \tag{5.2}$$

Therefore, (since matrix multiplication is associative)

$$\begin{aligned} (\mathbf{A} \cdot \mathbf{B}) \cdot (\mathbf{B}^{-1} \cdot \mathbf{A}^{-1}) &= \mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{B}^{-1}) \cdot \mathbf{A}^{-1} \\ &= \mathbf{A} \cdot \mathbf{I}_n \cdot \mathbf{A}^{-1} \qquad \text{via (5.2)} \\ &= \mathbf{A} \cdot \mathbf{A}^{-1} \\ &= \mathbf{I}_n \qquad \text{via (5.1).} \end{aligned} \tag{5.3}$$

Similarly,

$$(\mathbf{B}^{-1} \cdot \mathbf{A}^{-1}) \cdot (\mathbf{A} \cdot \mathbf{B}) = \mathbf{I}_n, \tag{5.4}$$

and so, from Definition 4.17, (5.3) and (5.4) we conclude that $(\mathbf{A} \cdot \mathbf{B})^{-1}$ exists and is given by

$$(\mathbf{B}^{-1} \cdot \mathbf{A}^{-1}),$$

as required.

Corollary 5.2. *Suppose that $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_m$ are all invertible $n \times n$ matrices. Then, $(\boldsymbol{A}_1 \cdot \boldsymbol{A}_2 \cdot \ldots \cdot \boldsymbol{A}_m)^{-1} = \boldsymbol{A}_m^{-1} \cdot \boldsymbol{A}_{m-1}^{-1} \cdot \ldots \cdot \boldsymbol{A}_1^{-1}$.*

To prove Corollary 5.2, use Theorem 5.1 and mathematical induction (see practice questions).

We now establish how inverse matrices can be used to solve systems of simultaneous linear systems of equations.

Theorem 5.3. *Let $\boldsymbol{A} \in \mathcal{M}_{nn}(\mathbb{R})$ be an $n \times n$ matrix. If $\boldsymbol{A}$ is invertible, then for any column vector $\mathbf{b} \in \mathcal{M}_{n\times 1}$, the system of simultaneous linear equations*

$$\boldsymbol{A} \cdot \mathbf{x} = \mathbf{b} \tag{5.5}$$

has a unique solution and this is given by $\mathbf{x} = \boldsymbol{A}^{-1} \cdot \mathbf{b}$.

Proof: Since $\mathbf{A}$ is invertible, $\mathbf{A}^{-1} \cdot \mathbf{b}$ exists. Substituting $\mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{b}$ into (5.5) gives

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{A} \cdot (\mathbf{A}^{-1} \cdot \mathbf{b}) = (\mathbf{A} \cdot \mathbf{A}^{-1}) \cdot \mathbf{b} = \mathbf{I}_n \cdot \mathbf{b} = \mathbf{b}. \tag{5.6}$$

Thus, via (5.6), $\mathbf{A}^{-1} \cdot \mathbf{b}$ is a solution to (5.5). Now, suppose that $\tilde{\mathbf{x}}$ satisfies $\mathbf{A} \cdot \tilde{\mathbf{x}} = \mathbf{b}$. Then,

$$\begin{aligned} \mathbf{A} \cdot \mathbf{x} = \mathbf{b} = \mathbf{A} \cdot \tilde{\mathbf{x}} &\implies \mathbf{A}^{-1} \cdot (\mathbf{A} \cdot \mathbf{x}) = \mathbf{A}^{-1} \cdot (\mathbf{A} \cdot \tilde{\mathbf{x}}) \\ &\implies (\mathbf{A}^{-1} \cdot \mathbf{A}) \cdot \mathbf{x} = (\mathbf{A}^{-1} \cdot \mathbf{A}) \cdot \tilde{\mathbf{x}} \\ &\implies \mathbf{I}_n \cdot \mathbf{x} = \mathbf{I}_n \cdot \tilde{\mathbf{x}} \\ &\implies \mathbf{x} = \tilde{\mathbf{x}}. \end{aligned}$$

Therefore, $\mathbf{A}^{-1} \cdot \mathbf{b}$ is the unique solution to (5.5), as required.

From Theorem 5.3 we also have the following results which guarantee that a matrix $\mathbf{A}$ is not invertible.

Corollary 5.4. *Let $\boldsymbol{A} \in \mathcal{M}_{nn}(\mathbb{R})$ be an $n \times n$ matrix and $\mathbf{b}$ be a column vector of length n. If the system of simultaneous linear equations*

$$\boldsymbol{A} \cdot \mathbf{x} = \mathbf{b} \tag{5.7}$$

has no solution, then $\boldsymbol{A}$ is not invertible, i.e. $\boldsymbol{A}^{-1}$ does not exist.

Proof: Suppose that $\mathbf{A}^{-1}$ exists. Then via Theorem 5.3, $\mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{b}$ is a solution to (5.7) which contradicts the condition that (5.7) has no solution. Therefore, we conclude that (since the supposition is incorrect) $\mathbf{A}^{-1}$ does not exist, as required. [1]

Corollary 5.5. *Let $\boldsymbol{A} \in \mathcal{M}_{nn}(\mathbb{R})$ be an $n \times n$ matrix and $\mathbf{b}$ be a column vector of length n. If the system of simultaneous linear equations*

$$\boldsymbol{A} \cdot \mathbf{x} = \mathbf{b} \tag{5.8}$$

has at least 2 distinct solutions, then $\boldsymbol{A}$ is not invertible and there exist infinitely many solutions to (5.8).

[1] An alternative proof of Corollary 5.4 is: Corollary 5.4 is the *contraposition* of Theorem 5.3! This type of proof is known as *proof by contraposition*, namely, (if $A \Rightarrow B$) then ($\neg B \Rightarrow \neg A$).

Proof: Suppose that $\mathbf{A}^{-1}$ exists. Then via Theorem 5.3, $\mathbf{x} = \mathbf{A}^{-1}\cdot\mathbf{b}$ is the unique solution to (5.8) which contradicts the condition that (5.8) has 2 distinct solutions. Therefore, $\mathbf{A}^{-1}$ does not exist.

Now suppose that $\mathbf{x}_1$ and $\mathbf{x}_2$ are 2 distinct solutions to (5.8). Then for $\lambda \in \mathbb{R}$,

$$\begin{aligned}\mathbf{A}\cdot(\mathbf{x}_1 + \lambda(\mathbf{x}_1 - \mathbf{x}_2)) &= \mathbf{A}\cdot(\mathbf{x}_1) + \mathbf{A}\cdot(\lambda(\mathbf{x}_1)) - \mathbf{A}\cdot(\lambda\mathbf{x}_2))\\ &= \mathbf{b} + \lambda(\mathbf{A}\cdot\mathbf{x}_1) - \lambda(\mathbf{A}\cdot\mathbf{x}_2)\\ &= \mathbf{b} + \lambda\mathbf{b} - \lambda\mathbf{b}\\ &= \mathbf{b}.\end{aligned}$$

Since $\mathbf{x}_1$ and $\mathbf{x}_2$ are distinct, it follows that for each $\lambda \in \mathbb{R}$, $\mathbf{x}_1 + \lambda(\mathbf{x}_1 - \mathbf{x}_2)$ is a distinct solution to (5.8) i.e. there are infinitely many solutions to (5.8), as required.

5.2 Gaussian Elimination Algorithm

If we wish to determine whether or not an $n \times n$ matrix $\mathbf{A}$ is invertible, and if invertible, determine $\mathbf{A}^{-1}$ explicitly, we can use the following **algorithm**.

Gaussian Elimination Algorithm:

1. Consider the augmented matrix of dimension $n \times 2n$, given by, $(\mathbf{A} \mid \mathbf{I}_n)$. Note that the line $\mid$ is merely used to provide a clear separation of the two $n \times n$ blocks in the augmented matrix.

2. Perform EROs on the augmented matrix to reduce the $n \times n$ block on the left hand side to reduced row echelon form (denoted $\text{REch}(\mathbf{A})$), i.e.

$$(\mathbf{A} \mid \mathbf{I}_n) \to (\text{REch}(\mathbf{A}) \mid \mathbf{B}).$$

3. (a) If $\text{REch}(\mathbf{A}) = \mathbf{I}_n$, then $\mathbf{B} = \mathbf{A}^{-1}$.
   (b) If $\text{REch}(\mathbf{A}) \neq \mathbf{I}_n$, then A is not invertible.

Note that for the algorithm defined above, all cases are covered (in step 3a and 3b). Before we justify the conclusions we can develop from the Gaussian Elimination Algorithm, we give 2 examples exhibiting an implementation of the algorithm. We also note that if, by applying EROs to the augmented matrix in the algorithm, the left hand side contains a row of zeros, we can immediately conclude that $\mathbf{A}$ is not invertible (to save time).

Example 69: Let

$$\mathbf{A} = \begin{pmatrix} 2 & 4 & 3 \\ 0 & 1 & -1 \\ 3 & 5 & 7 \end{pmatrix}.$$

Perform the Gaussian elimination algorithm to determine if $\mathbf{A}$ is invertible, and if $\mathbf{A}^{-1}$ exists, calculate it.

Answer: 1. We consider the augmented matrix of dimension 3×6, given by,

$$\left(\begin{array}{ccc|ccc} 2 & 4 & 3 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 1 & 0 \\ 3 & 5 & 7 & 0 & 0 & 1 \end{array}\right).$$

2. We perform EROs to reduce the left hand side of the augmented matrix to echelon form. Notice that the first "move" avoids creating rational entries early in the algorithm. A computer would not work this way.

$$\boxed{r_3 \to r_3 - r_1} \qquad \left(\begin{array}{ccc|ccc} 2 & 4 & 3 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 1 & 0 \\ 1 & 1 & 4 & -1 & 0 & 1 \end{array}\right)$$

$$\boxed{r_1 \to r_1 - r_3} \qquad \left(\begin{array}{ccc|ccc} 1 & 3 & -1 & 2 & 0 & -1 \\ 0 & 1 & -1 & 0 & 1 & 0 \\ 1 & 1 & 4 & -1 & 0 & 1 \end{array}\right)$$

$$\boxed{r_3 \to r_3 - r_1} \qquad \left(\begin{array}{ccc|ccc} 1 & 3 & -1 & 2 & 0 & -1 \\ 0 & 1 & -1 & 0 & 1 & 0 \\ 0 & -2 & 5 & -3 & 0 & 2 \end{array}\right)$$

$$\boxed{r_3 \to r_3 + 2r_2} \qquad \left(\begin{array}{ccc|ccc} 1 & 3 & -1 & 2 & 0 & -1 \\ 0 & 1 & -1 & 0 & 1 & 0 \\ 0 & 0 & 3 & -3 & 2 & 2 \end{array}\right)$$

$$\boxed{r_1 \to r_1 - 3r_2} \qquad \left(\begin{array}{ccc|ccc} 1 & 0 & 2 & 2 & -3 & -1 \\ 0 & 1 & -1 & 0 & 1 & 0 \\ 0 & 0 & 3 & -3 & 2 & 2 \end{array}\right)$$

$$\boxed{r_3 \to \tfrac{1}{3} r_3} \qquad \left(\begin{array}{ccc|ccc} 1 & 0 & 2 & 2 & -3 & -1 \\ 0 & 1 & -1 & 0 & 1 & 0 \\ 0 & 0 & 1 & -1 & \frac{2}{3} & \frac{2}{3} \end{array}\right)$$

$$\boxed{r_1 \to r_1 - 2r_3} \qquad \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 4 & -\frac{13}{3} & -\frac{7}{3} \\ 0 & 1 & -1 & 0 & 1 & 0 \\ 0 & 0 & 1 & -1 & \frac{2}{3} & \frac{2}{3} \end{array}\right)$$

$$\boxed{r_2 \to r_2 + r_3} \qquad \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 4 & -\frac{13}{3} & -\frac{7}{3} \\ 0 & 1 & 0 & -1 & \frac{5}{3} & \frac{2}{3} \\ 0 & 0 & 1 & -1 & \frac{2}{3} & \frac{2}{3} \end{array}\right)$$

3. Since the left hand side of the augmented matrix is, after EROs, equal to $\mathbf{I}_3$, we conclude that $\mathbf{A}$ is invertible, and the inverse is

$$\mathbf{A}^{-1} = \begin{pmatrix} 4 & -\frac{13}{3} & -\frac{7}{3} \\ -1 & \frac{5}{3} & \frac{2}{3} \\ -1 & \frac{2}{3} & \frac{2}{3} \end{pmatrix}.$$

Example 70: Let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 5 & 7 & 9 \end{pmatrix}.$$

Perform the Gaussian elimination algorithm to determine if $\mathbf{A}$ is invertible, and if $\mathbf{A}^{-1}$ exists, calculate it.

Answer: 1. We consider the augmented matrix of dimension 3×6, given by,

$$\left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 4 & 5 & 6 & 0 & 1 & 0 \\ 5 & 7 & 9 & 0 & 0 & 1 \end{array}\right).$$

2. We perform EROs to reduce the left hand side of the augmented matrix to echelon form.

$$\boxed{\begin{array}{l} r_2 \to r_2 - 4r_1 \\ r_3 \to r_3 - 5r_1 \end{array}} \qquad \left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -3 & -6 & -4 & 1 & 0 \\ 0 & -3 & -6 & -5 & 0 & 1 \end{array}\right)$$

$$\boxed{r_3 \to r_3 - r_2} \qquad \left(\begin{array}{ccc|ccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -3 & -6 & -4 & 1 & 0 \\ 0 & 0 & 0 & -1 & -1 & 1 \end{array}\right)$$

3. Since the left hand side of the augmented matrix, after EROs, contains a row of zeros, we conclude that $\mathbf{A}$ is not invertible.

In Example 69 one can directly verify against Definition 4.17 that the conclusion is correct i.e.

$$\begin{pmatrix} 2 & 4 & 3 \\ 0 & 1 & -1 \\ 3 & 5 & 7 \end{pmatrix} \cdot \begin{pmatrix} 4 & -\frac{13}{3} & -\frac{7}{3} \\ -1 & \frac{5}{3} & \frac{2}{3} \\ -1 & \frac{2}{3} & \frac{2}{3} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 4 & -\frac{13}{3} & -\frac{7}{3} \\ -1 & \frac{5}{3} & \frac{2}{3} \\ -1 & \frac{2}{3} & \frac{2}{3} \end{pmatrix} \cdot \begin{pmatrix} 2 & 4 & 3 \\ 0 & 1 & -1 \\ 3 & 5 & 7 \end{pmatrix}$$

i.e, that

$$\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{I}_n = \mathbf{A}^{-1} \cdot \mathbf{A}.$$

The conclusion in Example 70 is more difficult to directly check at this point, but will be addressed in the next sections.

5.3 Introduction to Elementary Matrices

We now introduce some 'machinery' to allow us to justify that the conclusions in the Gaussian Elimination algorithm (see 3.a and 3.b) are valid.

Definition 5.6. *A $n \times n$* ***elementary matrix*** *is any matrix obtained by performing* ***one*** *elementary row operation to the identity matrix $\boldsymbol{I}_n$.*

Recall the three types of EROs defined in Chapter 3, Section 8. Here are some examples:

$r_1 \circlearrowleft r_2$ $\qquad \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

$r_1 \to (-5) \times r_1$ $\qquad \begin{pmatrix} -5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

$r_1 \circlearrowleft r_3$ $\qquad \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$

$r_2 \to r_2 + 2r_1$ $\qquad \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

$r_1 \to \frac{1}{2} r_1$ $\qquad \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix}$

NOT ELEMENTARY $\qquad \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$

Note, that in the final example above, the matrix corresponds to two ERO applied simultaneously, specifically, $r_1 \circlearrowleft r_4$ and $r_2 \circlearrowleft r_3$.

5.4 Some Basic Properties of Elementary Matrices

(i) Each elementary matrix is invertible (see Proposition 5.10) and the inverse is easy to find (simply undo the appropriate row operation). For the last 3 elementary matrices in the previous section, we have:

$$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}^{-1} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \qquad (r_1 \circlearrowleft r_3 \text{ in } \mathbf{I}_3)$$

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad (r_2 \to r_2 - 2r_1 \text{ in } \mathbf{I}_3)$$

$$\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \qquad (r_1 \to 2r_1 \text{ in } \mathbf{I}_2)$$

(ii) Multiplying a matrix $\mathbf{A}$ on the left by an elementary matrix corresponds to performing the corresponding elementary row operation to $\mathbf{A}$.

For example, the second example above swaps rows 1 and 3, so multiplying by the associated matrix on the left of a general 3×3 matrix can be seen to switch the entries in rows 1 and 3:

$$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = \begin{pmatrix} g & h & i \\ d & e & f \\ a & b & c \end{pmatrix}.$$

Additionally, the fourth example in the previous section adds 2 times row 1 to row 2, so multiplying by the associated matrix on the left of a general 3×3 matrix can, similarly, be seen to add 2 times the entry in row 1 to the corresponding entry in row 2:

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = \begin{pmatrix} a & b & c \\ 2a+d & 2b+e & 2c+f \\ g & h & i \end{pmatrix}.$$

That is 2 times row 1 has been added to row 2.

Let's be more precise about the matrix entries. Let $\mathbf{E}_{ij}$ represent the $n \times n$ matrix which is defined to have every entry zero aside from having a 1 in the (i, j) place (ith row and jth column). Thus $\mathbf{E}_{ij} = (e_{rs})$ where $e_{rs} = 0$ for $(r, s) \neq (i, j)$ and $e_{ij} = 1$.

So, for example, the 4×4 matrix is

$$\mathbf{E}_{23} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Let $\mathbf{E}_{ij} = (e_{rs})$ and $\mathbf{E}_{\ell m} = (e^*_{rs})$. Then

$$\mathbf{E}_{ij}\mathbf{E}_{\ell m} = (b_{rs})$$

where

$$b_{rs} = \sum_{t=1}^{n} e_{rt}e^*_{ts}$$

and $\mathbf{E}_{\ell m}$. We first note that $b_{rs} = 0$ unless $r = i$ and $m = s$ and then we have $b_{im} = \sum_{t=1}^{n} e_{it}e^*_{tm}$ and this is 0 unless $t = j = \ell$ in which case it is 1. Hence

$$\mathbf{E}_{ij}\mathbf{E}_{\ell m} = \begin{cases} \mathbf{E}_{im} & j = \ell \\ 0 & j \neq \ell \end{cases}. \tag{5.9}$$

For $i \neq j$ and $\lambda \in \mathbb{R}$, set

$$\mathbf{E}_{(i,j,\lambda)} = \mathbf{I}_n + \lambda\mathbf{E}_{ij}.$$

For example, the 4×4 matrix

$$\mathbf{E}_{(2,4,\lambda)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & \lambda \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

For an arbitrary such matrix we calculate

$$\begin{aligned} \mathbf{E}_{(i,j,\lambda)} \cdot \mathbf{E}_{(i,j,-\lambda)} &= (\mathbf{I}_n + \lambda\mathbf{E}_{ij}) \cdot (\mathbf{I}_n - \lambda\mathbf{E}_{ij}) \\ &= \mathbf{I}_n + \lambda\mathbf{E}_{ij} - \lambda\mathbf{E}_{ij} + \lambda^2\mathbf{E}_{ij}^2 \\ &= \mathbf{I}_n. \end{aligned}$$

Replacing λ by $-\lambda$, we also see that $\mathbf{E}_{(i,j,-\lambda)} \cdot \mathbf{E}_{(i,j,\lambda)} = \mathbf{I}_n$. Hence $\mathbf{E}_{(i,j,\lambda)}$ is invertible.

Lemma 5.7. *Suppose that $\boldsymbol{A} \in \mathcal{M}_{nm}(\mathbb{R})$. Then, for $\lambda \in \mathbb{R}$, **left** multiplication by $\boldsymbol{E}_{(i,j,\lambda)}$ adds λ times the jth **row** of $\boldsymbol{A}$ to **row** i of $\boldsymbol{A}$. Furthermore, $\boldsymbol{E}^{-1}_{(i,j,\lambda)} = \boldsymbol{E}_{(i,j,-\lambda)}$.*

Assume that $\mu \in \mathbb{R}$ is non-zero. Let

$$\mathbf{E}_{[i,\mu]} = \mathbf{I}_n + (\mu - 1)\mathbf{E}_{ii}.$$

So $\mathbf{E}_{[i,\mu]} = \mathrm{diag}(1, \ldots, \mu, \ldots, 1)$ is a diagonal matrix with μ in the ith position. We see $\mathbf{E}^{-1}_{[i,\mu]} = \mathbf{E}_{[i,\mu^{-1}]}$ and so $\mathbf{E}_{[i,\mu]}$ is invertible.

Lemma 5.8. *Suppose that* $\boldsymbol{A} \in \mathcal{M}_{nm}(\mathbb{R})$. *Then, for non-zero* $\mu \in \mathbb{R}$, ***left*** *multiplication by* $\boldsymbol{E}_{[i,\mu]}$ ***scales*** *the* i*th* ***row*** *of* $\boldsymbol{A}$ *by* μ. *Furthermore,* $\boldsymbol{E}_{[i,\mu]}^{-1} = \boldsymbol{E}_{[i,\mu^{-1}]}$.

For $1 \leq i < j \leq n$, set

$$\mathbf{E}_{(i,j)} = \mathbf{I}_n - \mathbf{E}_{ii} - \mathbf{E}_{jj} + \mathbf{E}_{ij} + \mathbf{E}_{ji}.$$

For example, the 4×4 matrix

$$\mathbf{E}_{(2,4)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

Then, multiply out term by term in the second factor and using Equation 5.9 we get

$$\begin{aligned} \mathbf{E}_{(i,j)}^2 &= (\mathbf{I}_n - \mathbf{E}_{ii} - \mathbf{E}_{jj} + \mathbf{E}_{ij} + \mathbf{E}_{ji}) \cdot (\mathbf{I}_n - \mathbf{E}_{ii} - \mathbf{E}_{jj} + \mathbf{E}_{ij} + \mathbf{E}_{ji}) \\ &= (\mathbf{I}_n - \mathbf{E}_{ii} - \mathbf{E}_{jj} + \mathbf{E}_{ij} + \mathbf{E}_{ji}) - (\mathbf{E}_{ii} - \mathbf{E}_{ii} + \mathbf{E}_{ji}) \\ &\quad -(\mathbf{E}_{jj} - \mathbf{E}_{jj} + \mathbf{E}_{ij}) + (\mathbf{E}_{ij} - \mathbf{E}_{ij} + \mathbf{E}_{jj}) + (\mathbf{E}_{ji} - \mathbf{E}_{ji} + \mathbf{E}_{ii}) \\ &= \mathbf{I}_n. \end{aligned}$$

Hence $\mathbf{E}_{(i,j)}$ is invertible.

Lemma 5.9. *Suppose that* $\boldsymbol{A} \in \mathcal{M}_{nm}(\mathbb{R})$. *Then* ***left*** *multiplication by* $\boldsymbol{E}_{(i,j)}$ ***swaps*** ***row*** i *and* ***row*** j *in* $\boldsymbol{A}$. *Furthermore,* $\boldsymbol{E}_{(i,j)}^{-1} = \boldsymbol{E}_{(i,j)}$.

Elementary matrices are a useful algebraic device for implementing, through matrix multiplication on the left, elementary row operations. To establish the validity of the Gaussian elimination algorithm, the following proposition is helpful.

We have just seen that

Proposition 5.10. *Elementary matrices are invertible and the inverse of an elementary matrix is an elementary matrix.*

Proof: This follows from our previous discussion and Lemmas 5.7, 5.8 and 5.9.

5.5 Validity of the Gaussian Elimination Algorithm

To make the proof of the validity of conclusions from the Gaussian Elimination algorithm (GA) more transparent, we split the proof into 2 parts.

Proof of conclusion in 3(a) in the GA: Suppose that the elementary row operations $\rho_1, \rho_2, \ldots, \rho_m$ correspond to elementary matrices $\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_m$. So applying ρ_1 to $(\mathbf{A}|\mathbf{I}_n)$ replaces $\mathbf{A}$ bey $\mathbf{E}_1 \cdot \mathbf{A}$ and $\mathbf{I}_n$ by $\mathbf{E}_1$. So ρ_1 transforms

$$(\mathbf{A}|\mathbf{I}_n) \to (\mathbf{E}_1 \cdot \mathbf{A}|\mathbf{E}_1).$$

Now applying ρ_2 we have

$$(\mathbf{E}_1\mathbf{A}|\mathbf{E}_1) \to (\mathbf{E}_2 \cdot \mathbf{E}_1 \cdot \mathbf{A}|\mathbf{E}_2 \cdot \mathbf{E}_1)$$

continuing in this way finally yields

$$(\mathbf{E}_m \cdot \cdots \cdot \mathbf{E}_2 \cdot \mathbf{E}_1 \cdot \mathbf{A}|\mathbf{E}_m \cdot \cdots \cdot \mathbf{E}_2 \cdot \mathbf{E}_1)$$

Since these row operations reduce $(\mathbf{A}|\mathbf{I}_n)$ to $(\mathbf{I}_n|\mathbf{B})$, we have

$$\mathbf{E}_m \cdot \cdots \cdot \mathbf{E}_2 \cdot \mathbf{E}_1 \cdot \mathbf{A} = \mathbf{I}_n$$

and

$$\mathbf{B} = \mathbf{E}_m \cdot \cdots \cdot \mathbf{E}_2 \cdot \mathbf{E}_1.$$

In particular,

$$\mathbf{B} \cdot \mathbf{A} = \mathbf{I}_n. \tag{5.10}$$

Proposition 5.10, yields $\mathbf{E}_j$ is invertible for $1 \le j \le m$ and so Corollary 5.2 implies that $\mathbf{B}$ is invertible. Hence multiplying both sides of 5.10 by $\mathbf{B}^{-1}$ yields

$$\mathbf{B}^{-1} \cdot \mathbf{B} \cdot \mathbf{A} = \mathbf{B}^{-1}.$$

Hence $\mathbf{A} = \mathbf{B}^{-1}$ and so

$$\mathbf{A} \cdot \mathbf{B} = \mathbf{B}^{-1} \cdot \mathbf{B} = \mathbf{I}_n. \tag{5.11}$$

Together 5.10 and 5.11 imply $\mathbf{A}$ is invertible and its inverse in $\mathbf{B}$.

Proof of conclusion in 3(b) in GA algorithm: Suppose that for an $n \times n$ matrix $\mathbf{A}$ and after following steps 1 and 2 of the Gaussian Elimination Algorithm, $\mathbf{A}$ is in echelon form with a row of zeros in the bottom row. Therefore, there exist m elementary matrices $\mathbf{E}_1, \mathbf{E}_2 \ldots \mathbf{E}_m$ such that

$$\mathbf{E}_m \cdot \mathbf{E}_{m-1} \cdot \ldots \cdot \mathbf{E}_1 \cdot \mathbf{A} = \mathbf{C} \tag{5.12}$$

with $\mathbf{C}$ in reduced echelon form but is not $\mathbf{I}_n$. Then C has bottom row consisting of zeros. Let $\mathbf{b}$ be the column $n \times 1$ matrix with every entry 1. Then the system of equations

$$\mathbf{C} \cdot \mathbf{x} = \mathbf{b}$$

has **no solutions** as the bottom row reads $0 = 1$. Hence Corollary 5.7 implies that $\mathbf{C}$ is not invertible. However, if $\mathbf{A}$ is invertible, then Corollary 5.2 and 5.12 imply $\mathbf{C}$ is invertible. As this is not the case, we conclude $\mathbf{A}$ is not invertible.

Example 71: Show that the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 3 & 4 \\ 0 & 0 & 1 \\ 2 & 8 & 8 \end{pmatrix}$$

is invertible and determine its inverse $\mathbf{A}^{-1}$.

Answer: Start with

$$\left(\mathbf{A}\middle|\,\mathbf{I}_n\right) = \left(\begin{array}{ccc|ccc} 1 & 3 & 4 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 2 & 8 & 8 & 0 & 0 & 1 \end{array}\right).$$

Then since,

$\boxed{\rho_1 : r_3 \to r_3 - 2r_1}$ $\left(\begin{array}{ccc|ccc} 1 & 3 & 4 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & -2 & 0 & 1 \end{array}\right)$ $\mathbf{E}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix}$

$\boxed{\rho_2 : r_2 \circlearrowleft r_3}$ $\left(\begin{array}{ccc|ccc} 1 & 3 & 4 & 1 & 0 & 0 \\ 0 & 2 & 0 & -2 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{array}\right)$ $\mathbf{E}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$

$\boxed{\rho_3 : r_2 \to \frac{1}{2}r_2}$ $\left(\begin{array}{ccc|ccc} 1 & 3 & 4 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 & 1 & 0 \end{array}\right)$ $\mathbf{E}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}$

$\boxed{\rho_4 : r_1 \to r_1 - 3r_2}$ $\left(\begin{array}{ccc|ccc} 1 & 0 & 4 & 4 & 0 & -\frac{3}{2} \\ 0 & 1 & 0 & -1 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 & 1 & 0 \end{array}\right)$ $\mathbf{E}_4 = \begin{bmatrix} 1 & -3 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

$\boxed{\rho_5 : r_1 \to r_1 - 4r_3}$ $\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 4 & -4 & -\frac{3}{2} \\ 0 & 1 & 0 & -1 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 & 1 & 0 \end{array}\right)$ $\mathbf{E}_5 = \begin{bmatrix} 1 & 0 & -4 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

We conclude that

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

Example 72: Write $\mathbf{A}^{-1}$ and hence $\mathbf{A}$ from the previous example as a product of elementary matrices.

Answer: From the previous example we have that

$$\mathbf{A}^{-1} = \mathbf{E}_5 \cdot \mathbf{E}_4 \cdot \mathbf{E}_3 \cdot \mathbf{E}_2 \cdot \mathbf{E}_1.$$

Thus,

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

Furthermore (from Corollary 5.2)

$$\mathbf{A} = \mathbf{E}_1^{-1} \cdot \mathbf{E}_2^{-1} \cdot \mathbf{E}_3^{-1} \cdot \mathbf{E}_4^{-1} \cdot \mathbf{E}_5^{-1},$$

and hence,

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 3 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 4 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Theorem 5.11. *Every invertible matrix in $\mathcal{M}_{nn}(\mathbb{R})$ can be written as a product of a finite number of elementary matrices.*

Chapter 6

Determinants

▸ **Learning Outcomes** ◂ After the completion of the lectures and support sessions associated with this chapter you should be able to:

- define and calculate the determinant of an $n \times n$ matrix;
- cite and use the basic properties of a determinant;
- cite and use properties related to the determinant and basic row/column operations;
- cite and use properties of determinants of elementary matrices;
- cite and use properties related to the determinant and the inverse of a matrix;
- define and calculate the (i, j)-minor and (i, j)-cofactor of a matrix;
- define and calculate cofactor and adjoint matrices;
- relate the determinant of a square matrix to the existence of inverse matrices and solvability of linear systems of equations;
- solve systems of simultaneous linear equations using Cramer's rule; and
- prove a selection of theorems and corollaries in the notes.

6.1 Introduction

We have defined matrices and the basic operations involving matrices in Chapter 3. Here, we introduce the **determinant** of a square matrix. The determinant of a square matrix

is of particular importance in determining whether or not matrices are invertible and can be used to calculate the inverse of a square matrix (if it exists).

The definition of a determinant also allows the determination of closed-form representations for the solution to consistent systems of n equations in n unknowns (which have a unique solution). Hence, the determinant of a matrix contains fundamental information about square matrices.

Recall that the set of all $m \times n$ matrices with **real** entries is denoted by $\mathcal{M}_{mn}(\mathbb{R})$. In this chapter, we will primarily consider square matrices, i.e. $n \times n$ matrices. The set of all $n \times n$ matrices with real entries is written as $\mathcal{M}_{nn}(\mathbb{R})$ and we call the elements of $\mathcal{M}_{nn}(\mathbb{R})$ **square** $n \times n$ matrices. Matrices will be referred to as $\mathbf{A}$ or $[a_{ij}]$ throughout this chapter.

As motivation consider the homogeneous system of linear equations

$$\begin{array}{rcrcrcl} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & = & 0 \\ a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & = & 0 \\ a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & = & 0. \end{array} \tag{6.1}$$

Then we have seen in Theorem 5.5 that if the matrix $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$ is invertible then 6.1 has a unique solution. Since we know $(x_1, x_2, x_3) = (0, 0, 0)$ is a solution, if there is another solution say $(x_1, x_2, x_3) = (c_1, c_2, c_3)$ at least one of c_1, c_2 and c_3 must be non-zero and $\mathbf{A}$ must not be invertible. We may change notation and suppose that $c_1 \neq 0$. Then we have

$$\begin{array}{rcrcrcl} a_{11}c_1 & + & a_{12}c_2 & + & a_{13}c_3 & = & 0 \\ a_{21}c_1 & + & a_{22}c_2 & + & a_{23}c_3 & = & 0 \\ a_{31}c_1 & + & a_{32}c_2 & + & a_{33}c_3 & = & 0. \end{array} \tag{6.2}$$

Also assume that not all the coefficients a_{13}, a_{23} and a_{33} are 0 as otherwise we are in a degenerate case. From the first equation we obtain

$$a_{13}c_3 = -a_{11}c_1 - a_{12}c_2. \tag{6.3}$$

Multiplying equation two and three in 6.2 by a_{13} and substituting we get

$$\begin{aligned} 0 &= a_{21}a_{13}c_1 + a_{22}a_{13}c_2 - a_{23}(a_{11}c_1 + a_{12}c_2) \\ &= a_{21}a_{13}c_1 + a_{22}a_{13}c_2 - a_{23}a_{11}c_1 - a_{23}a_{12}c_2 \\ &= (a_{21}a_{13} - a_{23}a_{11})c_1 + (a_{22}a_{13} - a_{23}a_{12})c_2 \end{aligned} \tag{6.4}$$

and

$$\begin{aligned} 0 &= a_{31}a_{13}c_1 + a_{32}a_{13}c_2 - a_{33}(a_{11}c_1 + a_{12}c_2) \\ &= a_{31}a_{13}c_1 + a_{32}a_{13}c_2 - a_{33}a_{11}c_1 - a_{33}a_{12}c_2 \\ &= (a_{31}a_{13} - a_{33}a_{11})c_1 + (a_{32}a_{13} - a_{33}a_{12})c_2. \end{aligned} \tag{6.5}$$

Hence we may multiply 6.5 by $(a_{22}a_{13} - a_{23}a_{12})$ and substitute 6.4 to eliminate c_2 and obtain

$$\begin{aligned} 0 &= (a_{22}a_{13} - a_{23}a_{12})(a_{31}a_{13} - a_{33}a_{11})c_1 - (a_{32}a_{13} - a_{33}a_{12})(a_{21}a_{13} - a_{23}a_{11})c_1 \\ &= (a_{22}a_{13}a_{31}a_{13} - a_{23}a_{12}a_{31}a_{13} - a_{22}a_{13}a_{33}a_{11} + a_{23}a_{12}a_{33}a_{11} \\ &\quad -a_{32}a_{13}a_{21}a_{13} + a_{32}a_{13}a_{23}a_{11} + a_{33}a_{12}a_{21}a_{13} - a_{33}a_{12}a_{23}a_{11})c_1 \\ &= a_{13}(a_{22}a_{13}a_{31} - a_{23}a_{12}a_{31} - a_{22}a_{33}a_{11} - a_{32}a_{13}a_{21} + a_{32}a_{23}a_{11} + a_{33}a_{12}a_{21})c_1. \end{aligned}$$

Since $a_{13} \neq 0 \neq c_1$ we deduce that the long expression in the middle is zero. That is

$$a_{11}(a_{22}a_{33} - a_{32}a_{23}) - a_{12}(a_{33}a_{21} - a_{23}a_{31}) + a_{13}(a_{32}a_{21} - a_{22}a_{31}) = 0.$$

We define the determinant of the matrix A to be

$$a_{11}(a_{22}a_{33} - a_{32}a_{23}) - a_{12}(a_{33}a_{21} - a_{23}a_{31}) + a_{13}(a_{32}a_{21} - a_{22}a_{31})$$

and the observation we make is that the matrix **A** is not invertible implies the determinant of **A** is zero. You'll notice that each triple product in the formula for the determinant has one entry from each row and each column of **A** and there's something "odd" about the appearance of negative numbers in the description. The objective of this chapter is to make a more precise statement for all square matrices.

6.2 An excursion in to group theory

To define the determinant of a $n \times n$ matrix we require the following two definitions.[1]

Definition 6.1. *A* **permutation** *of* $\{1, \ldots, n\}$ *is a bijective function*

$$\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}.$$

The set of all permutations of $\{1, \ldots, n\}$ *is denoted by* S_n *and is called the* ***symmetric group*** *on* $\{1, \ldots, n\}$.

For instance $\sigma : \{1, 2, 3\} \to \{1, 2, 3\}$ given by

$$\sigma(1) = 2, \ \sigma(2) = 1 \text{ and } \sigma(3) = 3$$

is a permutation of $\{1, 2, 3\}$. We can represent a permutation $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ in two-row notation or sequence notation, denoted by

$$\sigma = \begin{pmatrix} 1 & 2 & \cdots & n-1 & n \\ \sigma(1) & \sigma(2) & \cdots & \sigma(n-1) & \sigma(n) \end{pmatrix} \text{ or } \sigma = [\sigma(1), \sigma(2), \ldots, \sigma(n-1), \sigma(n)]. \quad (6.6)$$

[1] Recall that a bijective function (also called a bijection) is a function that is both surjective and injective, i.e. 1-1 (see [15] for details).

We can also represent permutations in cycle notation (which are considered in 1AC [7]), albeit we do not do so in what follows.

Example 73: The permutation $\sigma : \{1,2,3,4\} \to \{1,2,3,4\}$ is equivalently written as

$$\sigma(x) = \begin{cases} 2 & \text{if } x = 1 \\ 4 & \text{if } x = 2 \\ 3 & \text{if } x = 3 \\ 1 & \text{if } x = 4 \end{cases} \iff \sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 3 & 1 \end{pmatrix} \iff \sigma = [2,4,3,1].$$

Definition 6.2. *The **identity permutation** in S_n is given by $\sigma_{Id} = [1,2,\ldots,n]$.*

Additionally, because $\sigma \in S_n$ is a bijection, σ has an inverse $\sigma^{-1} \in S_n$ which satisfies

$$\sigma(\sigma^{-1}(i)) = i = \sigma^{-1}(\sigma(i)) \quad \forall i \in \{1,\ldots,n\},$$

or equivalently,

$$\sigma^{-1} \circ \sigma = \sigma_{Id} = \sigma \circ \sigma^{-1}.$$

To find σ^{-1} given σ, we can write σ in 2 row notation, swap the bottom and top rows, and then swap columns so that the top row is in the form 1 2 3 ... n i.e.

$$\begin{aligned} \sigma = \begin{pmatrix} 1 & 2 & \cdots & n \\ \sigma(1) & \sigma(2) & \cdots & \sigma(n) \end{pmatrix} &\to \begin{pmatrix} \sigma(1) & \sigma(2) & \cdots & \sigma(n) \\ 1 & 2 & \cdots & n \end{pmatrix} \\ &\to \begin{pmatrix} 1 & 2 & \cdots & n \\ \sigma^{-1}(1) & \sigma^{-1}(2) & \cdots & \sigma^{-1}(n) \end{pmatrix} = \sigma^{-1}. \end{aligned} \tag{6.7}$$

Example 74: Find the inverse of $[1,3,4,2]$. Write

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 3 & 4 & 2 \end{pmatrix}$$

swap the top and bottom rows

$$\begin{pmatrix} 1 & 3 & 4 & 2 \\ 1 & 2 & 3 & 4 \end{pmatrix}.$$

Reorder the columns

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 2 & 3 \end{pmatrix}.$$

Definition 6.3. *Suppose that $n \geq 1$ is a natural number.*

1. *Suppose that $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ is a permutation in S_n. If $i, j \in \{1, \ldots, n\}$ with $i < j$ and $\sigma(i) > \sigma(j)$, then the pair (i, j) is called an **inversion** of σ.*

2. *Define the function $N : S_n \to \mathbb{N}_0$ by setting*

$$N(\sigma) = |\{(i, j) : (i, j) \text{ is an inversion of } \sigma\}|.$$

   So $N(\sigma)$ is the number of inversions in the permutation σ.

3. *A permutation $\sigma \in S_n$ is an **odd** permutation if and only if $N(\sigma)$ is odd. A permutation which is not odd is **even**.*

Example 75: The identity permutation in S_n is given by $\sigma_{Id} = [1, 2, \ldots, n]$. It is the only permutation in S_n which has no inversions i.e. $N(\sigma) = 0$ if and only if $\sigma = \sigma_{Id}$.

Example 76: For $\sigma = [2, 4, 3, 1] \in S_4$, there are 6 pairs (i, j) to check: $1 < 2, 1 < 3, 1 < 4, 2 < 3, 2 < 4$ and $3 < 4$. We have

(i, j)	$\sigma(i)$	$\sigma(j)$	*inversion*
$(1, 2)$	2	4	**no**
$(1, 3)$	2	3	**no**
$(1, 4)$	2	1	**yes**
$(2, 3)$	4	3	**yes**
$(2, 4)$	4	1	**yes**
$(3, 4)$	3	1	**yes**

Therefore,

$$N(\sigma) = |\{(1, 4), (2, 3), (2, 4), (3, 4)\}| = 4.$$

Hence $[2, 4, 3, 1]$ is an even permutation.

Moreover, given a permutation $\sigma \in S_n$, if you swap two adjacent terms to get a permutation $\tilde{\sigma}$, i.e.

$$\sigma = \begin{pmatrix} \cdots & i & i+1 & \cdots \\ \cdots & \sigma(i) & \sigma(i+1) & \cdots \end{pmatrix} \text{ and } \tilde{\sigma} = \begin{pmatrix} \cdots & i & i+1 & \cdots \\ \cdots & \sigma(i+1) & \sigma(i) & \cdots \end{pmatrix},$$

then

$$N(\tilde{\sigma}) = \begin{cases} N(\sigma) + 1 & \text{if } \sigma(i+1) > \sigma(i) \text{ (if } (i, i+1) \text{ is not an inversion)} \\ N(\sigma) - 1 & \text{if } \sigma(i+1) < \sigma(i) \text{ (if } (i, i+1) \text{ is an inversion).} \end{cases} \tag{6.8}$$

Using (6.8) we have the main result that we need about permutations.

Proposition 6.4. *Let $\sigma \in S_n$. Then $N(\sigma) = N(\sigma^{-1})$.*

Proof. For each $\theta \in \{1, \ldots, n\}$ let I_θ be the set of inversions of θ. Then

$$I_\sigma = \{(i,j) \ : \ i < j \text{ and } \theta(i) > \theta(j)\}.$$

For $(i,j) \in I_\sigma$, we have $\sigma(j) < \sigma(i)$ and

$$\sigma^{-1}(\sigma(j)) = j > i = \sigma^{-1}(\sigma(i)). \tag{6.9}$$

Therefore $(\sigma(j), \sigma(i)) \in I_{\sigma^{-1}}$ for each $(i,j) \in I_\sigma$.

As σ is a bijection, this shows that

$$N(\sigma) = |I_\sigma| = |\{(\sigma(j), \sigma(i)) \ : \ (i,j) \in I_\sigma\}| \leq |I_{\sigma^{-1}}| = N(\sigma^{-1})$$

for all $\sigma \in S_n$.

Hence, as $\sigma^{-1} \in S_n$, we have

$$N(\sigma^{-1}) \leq N((\sigma^{-1})^{-1}) = N(\sigma)$$

and thus $N(\sigma) = N(\sigma^{-1})$. □

Example 77: Suppose that $\sigma \in S_3$. Let $\mathbf{A} = (a_{ij}) \in \mathcal{M}_{33}(\mathbb{R})$. Show that

$$a_{\sigma(1),1}a_{\sigma(2),2}a_{\sigma(3),3} = a_{1,\sigma^{-1}(1)}a_{2,\sigma^{-1}(2)}a_{3,\sigma^{-1}(3)}.$$

We can certainly order numbers being multiplies together in the expression $a_{\sigma(1),1}a_{\sigma(2),2}a_{\sigma(3),3}$ in any order we want as $\mathbb{R}$ is commutative for multiplication. So we choose to order them with respect to rows rather than columns. Then the first term is $a_{1,x}$ where x is such that $\sigma(x) = 1$. So $x = \sigma^{-1}(1)$. That is the first term is $a_{1,\sigma^{-1}(1)}$. Arguing in this way we get

$$a_{\sigma(1),1}a_{\sigma(2),2}a_{\sigma(3),3} = a_{1,\sigma^{-1}(1)}a_{2,\sigma^{-1}(2)}a_{3,\sigma^{-1}(3)}.$$

6.3 Definition of the Determinant

After our excursion in to group theory, we now define the determinant of a $n \times n$ matrix as follows.

Definition 6.5. *Let $\boldsymbol{A} \in \mathcal{M}_{nn}(\mathbb{R})$. The determinant of $\boldsymbol{A} = [a_{ij}]$, denoted by det($\boldsymbol{A}$) or $|\boldsymbol{A}|$ is given by*

$$\sum_{\sigma \in S_n} (-1)^{N(\sigma)} \prod_{i=1}^{n} a_{\sigma(i)i}.$$

In Definition 6.5, note the following points:

- the sum is over all permutations in $\{1, \ldots, n\}$ (hence there are $n!$ terms in the sum);
- for each fixed σ in the sum, the a_{ij} terms in the product are in column i and row $\sigma(i)$ for $i = 1, \ldots, n$;
- for each fixed σ in the sum, the product of n terms contains exactly one entry from each column and each row of $\mathbf{A}$ (since σ is a bijective function);
- the determinant of a matrix is a **real number** (more generally is an element of the field where the matrix entries come from);
- the determinant of $\mathbf{A}$ only exists if $\mathbf{A}$ is a square matrix, i.e. an $n \times n$ matrix; and
- $\det : \mathcal{M}_{nn}(\mathbb{R}) \to \mathbb{R}$ is a function.

Example 78: Consider $\mathbf{A} = [a_{ij}] \in \mathcal{M}_{22}(\mathbb{R})$. Since S_2 contains only 2 permutations given by $\sigma_1 = [1, 2]$ and $\sigma_2 = [2, 1]$, with $N(\sigma_1) = 0$ and $N(\sigma_2) = 1$, it follows that

$$\begin{aligned}\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} &= \sum_{\sigma \in S_2} (-1)^{N(\sigma)} \prod_{i=1}^{2} a_{\sigma(i)i} \\ &= (-1)^{N([1,2])} a_{11}a_{22} + (-1)^{N([2,1])} a_{21}a_{12} \\ &= (-1)^0 a_{11}a_{22} + (-1)^1 a_{21}a_{12} \\ &= a_{11}a_{22} - a_{12}a_{21}.\end{aligned}$$

Moreover, consider $\mathbf{A} = [a_{ij}] \in \mathcal{M}_{33}(\mathbb{R})$. Since S_3 contains only 6 permutations given by

$$\sigma_1 = [1, 2, 3],\ \sigma_2 = [1, 3, 2],\ \sigma_3 = [2, 1, 3],\ \sigma_4 = [2, 3, 1],\ \sigma_5 = [3, 1, 2] \text{ and } \sigma_6 = [3, 2, 1],$$

with

$$N(\sigma_1) = 0,\ N(\sigma_2) = 1,\ N(\sigma_3) = 1,\ N(\sigma_4) = 2,\ N(\sigma_5) = 2 \text{ and } N(\sigma_6) = 3,$$

it follows that

$$\begin{aligned}\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} &= \sum_{\sigma \in S_3} (-1)^{N(\sigma)} \prod_{i=1}^{3} a_{\sigma(i)i} \\ &= (-1)^{N([1,2,3])} a_{11}a_{22}a_{33} + (-1)^{N([1,3,2])} a_{11}a_{32}a_{23} \\ &\quad + (-1)^{N([2,1,3])} a_{21}a_{12}a_{33} + (-1)^{N([2,3,1])} a_{21}a_{32}a_{13} \\ &\quad + (-1)^{N([3,1,2])} a_{31}a_{12}a_{23} + (-1)^{N([3,2,1])} a_{31}a_{22}a_{13} \\ &= (-1)^0 a_{11}a_{22}a_{33} + (-1)^1 a_{11}a_{32}a_{23} \\ &\quad + (-1)^1 a_{21}a_{12}a_{33} + (-1)^2 a_{21}a_{32}a_{13}\end{aligned}$$

$$+ (-1)^2 a_{31}a_{12}a_{23} + (-1)^3 a_{31}a_{22}a_{13}$$
$$= a_{11}(a_{22}a_{33} - a_{32}a_{23}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}).$$

Note that this is consistent with Examples 12 and 14 from Chapter 1.

Example 79: Suppose that $\mathbf{A}$ is a 13×13 with random integer entries between 1 and 10. Assume that in your head you can multiply 13 numbers between 1 and 10 together in 1 second and add numbers together in no time at all. Then to calculate the determinant of $\mathbf{A}$ using Definition 6.5 would take you approximately 197 Years, 5 Months and 2 Weeks. Your great, great, great, great grandchildren would have to inherit the exam question. We have to do better that this!

Remark 6.6. *Geometrically determinants are related to volumes. We remarked in Chapter 1 that the area of a parallelepiped is given by a scalar triple product and we saw that was described as a determinant in Example 16 in Chapter 1.*

6.4 The transpose matrix and its determinant

We encountered the **transpose** of a matrix in the practice questions. We repeat its definition here:

Definition 6.7. *The* ***transpose matrix****,* $\boldsymbol{A}^T = [b_{ij}] \in \mathcal{M}_{nm}(\mathbb{R})$*, of* $\boldsymbol{A} = [a_{ij}] \in \mathcal{M}_{mn}(\mathbb{R})$ *is the matrix with*

$$b_{ij} = a_{ji}$$

for $i = 1, \ldots, m$ *and* $j = 1, \ldots, n$*.*

Notice that in the above example $\mathbf{A}$ has shape $n \times m$ and $\mathbf{A}^T$ has shape $m \times n$.

Example 80: Suppose that $A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 6 & 7 & 8 & 6 \end{pmatrix}$. Then

$$A^T = \begin{pmatrix} 1 & 6 \\ 2 & 7 \\ 3 & 8 \\ 4 & 6 \end{pmatrix}.$$

A property illustrated in one of the exercises was:

Theorem 6.8. *Given matrices* $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{M}_{nn}(\mathbb{R})$*, then*

$$(\boldsymbol{A} \cdot \boldsymbol{B})^T = \boldsymbol{B}^T \cdot \boldsymbol{A}^T.$$

Proof. Write $\mathbf{A} = [a_{ij}]$, $\mathbf{B} = [b_{ij}]$, $\mathbf{C} = [c_{ij}] = (\mathbf{A} \cdot \mathbf{B})^T$ and $\mathbf{D} = [d_{ij}] = \mathbf{A} \cdot \mathbf{B}$ and $\mathbf{E} = [e_{ij}] = \mathbf{B}^T \cdot \mathbf{A}^T$. Then

$$c_{ij} = d_{ji} = \sum_{k=1}^{n} a_{jk} b_{ki}$$

On the other hand,

$$e_{ij} = \sum_{k=1}^{n} b_{ki} a_{jk} = \sum_{k=1}^{n} a_{jk} b_{ki}.$$

Hence the matrices C and E are equal. □

Using mathematical induction, one can easily deduce the following corollary

Corollary 6.9. *Given matrices $\boldsymbol{A}_i \in \mathcal{M}_{nn}(\mathbb{R})$, with $1 \leq i \leq k$, then*

$$(\boldsymbol{A}_1 \cdot \boldsymbol{A}_2 \cdot \ldots \cdot \boldsymbol{A}_{k-1} \cdot \boldsymbol{A}_k)^T = \boldsymbol{A}_k^T \cdot \boldsymbol{A}_{k-1}^T \cdot \ldots \cdot \boldsymbol{A}_2^T \cdot \boldsymbol{A}_1^T.$$

Proof. Use Theorem 6.8 and the principle of mathematical induction. □

Importantly, we can determine the following result concerning the determinant of the transpose of a square matrix.

Theorem 6.10. *Let $\boldsymbol{A} \in \mathcal{M}_{nn}(\mathbb{R})$. Then*

$$\det(\boldsymbol{A}) = \det(\boldsymbol{A}^T).$$

Proof. We start with two observations:

First, because every element of S_n has a unique inverse

$$S_n = \{\sigma : \sigma \in S_n\} = \{\sigma^{-1} \mid \sigma \in S_n\}. \tag{6.10}$$

Let $\theta \in S_n$. Then

$$\prod_{i=1}^{n} a_{\theta(i),i} = \prod_{i=1}^{n} a_{i,\theta^{-1}(i)}. \tag{6.11}$$

Hence using Theorem 6.4 yields

$$\begin{aligned} \det(\mathbf{A}) &= \sum_{\sigma \in S_n} (-1)^{N(\sigma)} \prod_{i=1}^{n} a_{\sigma(i),i} \text{ by Definition 6.5} \\ &= \sum_{\sigma^{-1} \in S_n} (-1)^{N(\sigma^{-1})} \prod_{i=1}^{n} a_{\sigma^{-1}(i),i} \text{ by 6.10} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{\sigma^{-1}\in S_n} (-1)^{N(\sigma)} \prod_{i=1}^{n} a_{i,\sigma(i)} \text{ by Theorem 6.4 and 6.11} \\
&= \sum_{\sigma\in S_n} (-1)^{N(\sigma)} \prod_{i=1}^{n} a_{i,\sigma(i)} \text{ by 6.10} \\
&= \det(\mathbf{A}^T).
\end{aligned}
$$

□

6.5 Determinants and row/column operations

Theorem 6.11. *Suppose that* $\boldsymbol{A} = [a_{ij}] \in \mathcal{M}_{nn}(\mathbb{R})$*, with* $n \geq 2$*. Assume that* $\boldsymbol{A}$ *has two equal columns or two equal rows. Then* $\det(\boldsymbol{A}) = 0$*.*

Proof. Suppose that columns r and s of $\mathbf{A} = [a_{ij}]$ are equal and assume without loss of generality that $r < s$. This means that

$$a_{i,r} = a_{i,s}$$

for all $1 \leq i \leq n$. For $\sigma \in S_n$ write

$$\sigma = \begin{pmatrix} 1 & 2 & \dots & r & \dots & s & \dots & n \\ \sigma(1) & \sigma(2) & \dots & \sigma(r) & \dots & \sigma(s) & \dots & \sigma(n) \end{pmatrix}$$

and define

$$\overline{\sigma} = \begin{pmatrix} 1 & 2 & \dots & r & \dots & s & \dots & n \\ \sigma(1) & \sigma(2) & \dots & \sigma(s) & \dots & \sigma(r) & \dots & \sigma(n) \end{pmatrix}.$$

Then $N(\overline{\sigma}) = N(\sigma) \pm 1$ by repeated use of 6.8 and so

$$(-1)^{N(\sigma)} + (-1)^{N(\overline{\sigma})} = 0. \tag{6.12}$$

Indeed, we transform σ to $\overline{\sigma}$. Let $k = s - r$. Swap $\sigma(r)$ and $\sigma(r+1)$ on the bottom row of σ.

$$\begin{pmatrix} 1 & 2 & \dots & r & r+1 & \dots & s & \dots & n \\ \sigma(1) & \sigma(2) & \dots & \sigma(r+1) & \sigma(r) & \dots & \sigma(s) & \dots & \sigma(n) \end{pmatrix}$$

Now swap $\sigma(r)$ and $\sigma(r+2)$ and so on until we swap $\sigma(r)$ and $\sigma(s)$. This requires k swaps. We have

$$\begin{pmatrix} 1 & 2 & \dots & r & r+1 & \dots & s-2 & s-1 & s & \dots & n \\ \sigma(1) & \sigma(2) & \dots & \sigma(r+1) & \sigma(r+2) & \dots & \sigma(s-1) & \sigma(s) & \sigma(r) & \dots & \sigma(n) \end{pmatrix}$$

Now we do it in reverse. Swap $\sigma(s)$ and $\sigma(s-1)$ on the bottom row and continue until we swap $\sigma(s)$ and $\sigma(r+1)$. This requires $k-1$ swaps. Once we have done this, we have

obtained $\overline{\sigma}$. In total, we have performed $2k-1$ swaps of adjacent images and $2k-1$ is odd. Thus $N(\sigma)$ and $N(\overline{\sigma})$ are not of the same parity by 6.8. Hence 6.12 holds.

Since columns r and s are the same and using the definition o $\overline{\sigma}$ we get,

$$a_{\sigma(r),r} = a_{\sigma(r),s} = a_{\overline{\sigma}(s),s},$$

$$a_{\sigma(s),s} = a_{\sigma(s),r} = a_{\overline{\sigma}(r),r}$$

and for all $i \notin \{r, s\}$,

$$a_{\sigma(i),i} = a_{\overline{\sigma}(i),i}.$$

We conclude that

$$\prod_{i=1}^{n} a_{\sigma(i),i} = \prod_{i=1}^{n} a_{\overline{\sigma}(i),i}$$

and so

$$(-1)^{N(\sigma)} \prod_{i=1}^{n} a_{\sigma(i),i} + (-1)^{N(\overline{\sigma})} \prod_{i=1}^{n} a_{\overline{\sigma}(i),i} = 0.$$

Hence by pairing σ with $\overline{\sigma}$, we obtain

$$\det(\mathbf{A}) = 0.$$

The result for matrices with two equal rows follows as $\det(\mathbf{A}^T) = \det(\mathbf{A})$ by Theorem 6.10. This means that we can always transform a matrix with two equal rows into one with two equal columns and keep the determinant unchanged. □

Theorem 6.12. *For any matrix $\boldsymbol{A} = [a_{ij}] \in \mathcal{M}_{nn}(\mathbb{R})$, with $n \geq 2$, if the matrix $\boldsymbol{P}$ is obtained from $\boldsymbol{A}$ by multiplying column j_1 (or row i_1) of $\boldsymbol{A}$ by $\lambda \in \mathbb{R}$, then*

$$\det(\boldsymbol{P}) = \lambda \det(\boldsymbol{A}).$$

Proof. Observe that for any $\sigma \in S_n$ that $\prod_{i=1}^{n} p_{\sigma(i)i}$ contains exactly 1 entry from column j_1 of $\mathbf{P}$. Since

$$p_{ij} = \begin{cases} a_{ij} & \text{if } j \neq j_1 \\ \lambda a_{ij} & \text{if } j = j_1 \end{cases},$$

for $i, j = 1, \ldots, n$, it follows that

$$\begin{aligned} |\mathbf{P}| &= \sum_{\sigma \in S_n} (-1)^{N(\sigma)} \prod_{i=1}^{n} p_{\sigma(i)i} \\ &= \sum_{\sigma \in S_n} (-1)^{N(\sigma)} \lambda \left(\prod_{i=1}^{n} a_{\sigma(i)i} \right) \\ &= \lambda \sum_{\sigma \in S_n} (-1)^{N(\sigma)} \prod_{i=1}^{n} a_{\sigma(i)i} \end{aligned}$$

$$= \lambda|\mathbf{A}|.$$

The result for multiplication of row i_1 by λ follows from the argument above and Theorem 6.10. □

Corollary 6.13. *If the matrix* $\boldsymbol{A} = [a_{ij}] \in \mathcal{M}_{nn}(\mathbb{R})$ *has a row of zeros or a column of zeros, then*

$$\det(\boldsymbol{A}) = 0.$$

Proof. Because of Theorem 6.10, it suffices to prove the result for the case when $\mathbf{A}$ has a row of zeros. Suppose that the jth row of $\mathbf{A}$ consists of zeros. Let $\lambda \in \mathbb{R}$ be such that $\lambda \neq 1$. Then multiplying row j be λ gives us a new matrix $\mathbf{P}$ and of course we know that $\mathbf{P} = \mathbf{A}$. By Theorem 6.12 we have

$$|\mathbf{A}| = \lambda|\mathbf{P}| = \lambda|\mathbf{A}|.$$

Since $\lambda \neq 1$, we conclude that $|\mathbf{A}| = 0$. □

Corollary 6.14. *Consider the matrix* $\boldsymbol{A} = [a_{ij}] \in \mathcal{M}_{nn}(\mathbb{R})$. *Then*

$$\det(\lambda \boldsymbol{A}) = \lambda^n \det(\boldsymbol{A}).$$

Proof. Starting with $\mathbf{A}$ we multiply the first row by λ and obtain a matrix $\mathbf{A}_1$ with determinant $\lambda \det(\mathbf{A})$ by Theorem 6.12. Now multiply row 2 of $\mathbf{A}_1$ by λ to obtain a matrix $\mathbf{A}_2$ with determinant $\lambda^2 \det(\mathbf{A})$. After n steps, we have a matrix $\mathbf{A}_n$ with $\det(\mathbf{A}_n) = \lambda^n \det(\mathbf{A})$. Since $\mathbf{A}_n = \lambda\mathbf{A}$, we have proved the result.

We can also prove the result directly as follows: observe that

$$\begin{aligned} |\lambda\mathbf{A}| &= \sum_{\sigma \in S_n} (-1)^{N(\sigma)} \prod_{i=1}^{n} (\lambda a_{\sigma(i)i}) \\ &= \sum_{\sigma \in S_n} (-1)^{N(\sigma)} \left(\lambda^n \prod_{i=1}^{n} a_{\sigma(i)i} \right) \\ &= \lambda^n \sum_{\sigma \in S_n} (-1)^{N(\sigma)} \prod_{i=1}^{n} a_{\sigma(i)i} = \lambda^n |\mathbf{A}|, \end{aligned}$$

as required. □

Corollary 6.15. *Consider the matrix* $\boldsymbol{A} = [a_{ij}] \in \mathcal{M}_{nn}(\mathbb{R})$. *Then,*

$$\det(-\boldsymbol{A}) = (-1)^n \det(\boldsymbol{A}).$$

Proof. Follows from Corollary 6.14 with $\lambda = -1$. □

Theorem 6.16. *Consider $\boldsymbol{A} = [a_{ij}] \in \mathcal{M}_{nn}(\mathbb{R})$, with $n \geq 2$ and the matrix $\boldsymbol{P} = [p_{ij}]$ which is obtained from $\boldsymbol{A}$ by adding λ times column j_1 of $\boldsymbol{A}$ to column j_2 of $\boldsymbol{A}$ (or λ times row i_1 of $\boldsymbol{A}$ to row i_2 of $\boldsymbol{A}$), where $\lambda \in \mathbb{R}$. Then*

$$\det(\boldsymbol{P}) = \det(\boldsymbol{A}).$$

Proof. Since

$$p_{ij} = \begin{cases} a_{ij} & \text{if } j \neq j_2 \\ a_{ij} + \lambda a_{ij_1} & \text{if } j = j_2, \end{cases}$$

we have

$$\begin{aligned} |\mathbf{P}| &= \sum_{\sigma \in S_n} (-1)^{N(\sigma)} \left(\prod_{i \in \{1,\ldots,n\} \setminus \{j_2\}} a_{\sigma(i)i} \right) (a_{\sigma(j_2)j_2} + \lambda a_{\sigma(j_2)j_1}) \\ &= \left(\sum_{\sigma \in S_n} (-1)^{N(\sigma)} \prod_{i=1}^{n} a_{\sigma(i)i} \right) + \lambda \left(\sum_{\sigma \in S_n} (-1)^{N(\sigma)} \left(\prod_{i \in \{1,\ldots,n\} \setminus \{j_2\}} a_{\sigma(i)i} \right) a_{\sigma(j_2)j_1} \right) \\ &= |\mathbf{A}| + \lambda \left(\sum_{\sigma \in S_n} (-1)^{N(\sigma)} \left(\prod_{i \in \{1,\ldots,n\} \setminus \{j_2\}} a_{\sigma(i)i} \right) a_{\sigma(j_2)j_1} \right). \end{aligned} \tag{6.13}$$

Consider the matrix $\mathbf{Q}$ with column j_1 equal to column j_2, given by

$$q_{ij} = \begin{cases} a_{ij} & \text{if } j \neq j_2 \\ a_{ij_1} & \text{if } j = j_2. \end{cases}$$

Then, $\det(\mathbf{Q}) = 0$ by Theorem thes2.12-1. Hence

$$0 = |\mathbf{Q}| = \sum_{\sigma \in S_n} (-1)^{N(\sigma)} \prod_{i=1}^{n} q_{\sigma(i)i} = \sum_{\sigma \in S_n} (-1)^{N(\sigma)} \left(\prod_{i \in \{1,\ldots,n\} \setminus \{j_2\}} a_{\sigma(i)i} \right) (a_{\sigma(j_2)j_1}). \tag{6.14}$$

Substitution of (6.14) into (6.13) yields $|\mathbf{P}| = |\mathbf{A}|$. The result for rows follows from the above argument and Theorem 6.10. □

Theorem 6.17. *For any matrix $\boldsymbol{A} = [a_{ij}] \in \mathcal{M}_{nn}(\mathbb{R})$, with $n \geq 2$, if the matrix $\boldsymbol{P} = [p_{ij}]$ is obtained from $\boldsymbol{A}$ by swapping columns j_1 and j_2 (or rows i_1 and i_2) of $\boldsymbol{A}$, then*

$$\det(\boldsymbol{P}) = -\det(\boldsymbol{A}).$$

Proof. For $1 \le k \le n$, let's write $\mathbf{r}_k = (a_{k1} \dots a_{kn})$ for the kth row of $\mathbf{A}$. Then

$$\mathbf{A} = \begin{pmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_i \\ \vdots \\ \mathbf{r}_j \\ \vdots \\ \mathbf{r}_n \end{pmatrix}.$$

We now use Theorem 6.16 for the second, third and fifth equality and Theorem 6.12 for the second last equality and obtain

$$\begin{aligned} |\mathbf{A}| &= \left|\begin{pmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_i \\ \vdots \\ \mathbf{r}_j \\ \vdots \\ \mathbf{r}_n \end{pmatrix}\right| = \left|\begin{pmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_i + \mathbf{r}_j \\ \vdots \\ \mathbf{r}_j \\ \vdots \\ \mathbf{r}_n \end{pmatrix}\right| = \left|\begin{pmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_i + \mathbf{r}_j \\ \vdots \\ \mathbf{r}_j - (\mathbf{r}_i + \mathbf{r}_j) \\ \vdots \\ \mathbf{r}_n \end{pmatrix}\right| = \left|\begin{pmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_i + \mathbf{r}_j \\ \vdots \\ -\mathbf{r}_i \\ \vdots \\ \mathbf{r}_n \end{pmatrix}\right| \\ &= \left|\begin{pmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_i + \mathbf{r}_j - \mathbf{r}_i \\ \vdots \\ -\mathbf{r}_i \\ \vdots \\ \mathbf{r}_n \end{pmatrix}\right| = \left|\begin{pmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_j \\ \vdots \\ -\mathbf{r}_i \\ \vdots \\ \mathbf{r}_n \end{pmatrix}\right| = -\left|\begin{pmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_j \\ \vdots \\ \mathbf{r}_i \\ \vdots \\ \mathbf{r}_n \end{pmatrix}\right| = -|\mathbf{P}|. \end{aligned}$$

This establishes the result for exchanged rows and using Theorem 6.10, the result follows for columns. □

6.6 Row/Column expansion of the determinant

Recall that for $\mathbf{A} = [a_{ij}] \in \mathcal{M}_{nn}(\mathbb{R})$ each of the $n!$ terms in the sum

$$|\mathbf{A}| = \sum_{\sigma \in S_n} (-1)^{N(\sigma)} \prod_{i=1}^{n} a_{\sigma(i)i}$$

has exactly one entry from each row and column of $\mathbf{A}$ (in the product). Therefore, we can bring the entry in row one to the front of each product $\prod_{i=1}^{n} a_{\sigma(i)i}$ and the group them by columns to write

$$|\mathbf{A}| = \sum_{j=1}^{n} a_{1j} C_{1j}(\mathbf{A}) \tag{6.15}$$

for some cofactor $C_{1j}(\mathbf{A})$ of the term a_{1j}. We say that (6.15) is the expansion of $|\mathbf{A}|$ along row 1.

The main results in the section give a representation for $C_{1j}(\mathbf{A})$ and demonstrate that $|\mathbf{A}|$ can be represented as an expansion along any column or row. To begin, we have

Definition 6.18. *A* ***submatrix*** *of* $\boldsymbol{A} \in \mathcal{M}_{mn}(\mathbb{R})$ *is obtained from* $\boldsymbol{A}$ *by deleting* $\tilde{m}$ *rows from* $\boldsymbol{A}$ *or* $\tilde{n}$ *columns from* $\boldsymbol{A}$ *with* $0 \leq \tilde{m} \leq m-1$, $0 \leq \tilde{n} \leq n-1$ *and* $\max\{\tilde{m}, \tilde{n}\} \geq 1$.

Example 81: Let

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{pmatrix}.$$

By deleting the first row and column we obtain the submatrix

$$\begin{pmatrix} 6 & 7 & 8 \\ 10 & 11 & 12 \end{pmatrix}.$$

Definition 6.19. *The* (i,j)***-minor*** *of a matrix* $\boldsymbol{A} \in \mathcal{M}_{nn}(\mathbb{R})$, *written* $M_{ij}(\boldsymbol{A})$, *is the determinant of the* $(n-1) \times (n-1)$*-submatrix of* $\boldsymbol{A}$ *obtained by deleting row* i *and column* j *of* $\boldsymbol{A}$.

Example 82: For

$$\mathbf{A} = \begin{pmatrix} 2 & -6 & 1 \\ 3 & 1 & -4 \\ -1 & -2 & 5 \end{pmatrix}$$

find $M_{12}(\mathbf{A})$ and $M_{31}(\mathbf{A})$.

We now have

Lemma 6.20. *Let* $\boldsymbol{A} = [a_{ij}] \in \mathcal{M}_{nn}(\mathbb{R})$ *with* $n \geq 2$. *Then*

$$|\boldsymbol{A}| = \sum_{j=1}^{n} a_{1j}(-1)^{j-1} M_{1j}(\boldsymbol{A}).$$

Proof. We first find the cofactor of a_{11} in the expansion (6.15). In the $n!$ terms in sum in $|\mathbf{A}|$, the terms which contain a_{11} have a permutation that satisfies $\sigma(1) = 1$. By considering the map

$$- : \{\sigma \in S_n \ : \ \sigma(1) = 1\} \to S_{n-1}$$

given by

$$\overline{\sigma}(i) = \sigma(i+1) - 1 \quad \forall i = 1, \dots, n-1 \tag{6.16}$$

it follows that – is a bijection. Also, since $\sigma(1) = 1$, it follows that

$$N(\overline{\sigma}) = N(\sigma) \quad \forall \sigma \in S_n. \tag{6.17}$$

To find the cofactor of a_{11}, we consider the submatrix $\overline{\mathbf{A}} = [\overline{a}_{ij}]$ which is obtained by deleting row 1 and column 1 from $\mathbf{A}$, and consider

$$\begin{aligned} a_{11}C_{11}(\mathbf{A}) &= \sum_{\substack{\sigma \in S_n \\ \sigma(1)=1}} (-1)^{N(\sigma)} \prod_{i=1}^{n} a_{\sigma(i)i} \\ &= a_{11} \sum_{\substack{\sigma \in S_n \\ \sigma(1)=1}} (-1)^{N(\sigma)} \prod_{i=2}^{n} a_{\sigma(i)i} \\ &= a_{11} \sum_{\overline{\sigma} \in S_{n-1}} (-1)^{N(\overline{\sigma})} \prod_{i=1}^{n-1} \overline{a}_{\overline{\sigma}(i)i} \\ &= a_{11} |\overline{\mathbf{A}}| \\ &= a_{11} M_{11}(\mathbf{A}). \end{aligned} \tag{6.18}$$

To determine $C_{1j}(\mathbf{A})$ we essentially repeat the argument above. In the $n!$ terms in sum in $|\mathbf{A}|$, the terms which contain a_{1j} have a permutation that satisfies $\sigma(j) = 1$. Similarly to (6.16), we consider the bijection $\_ : \{\sigma \in S_n \ : \ \sigma(j) = 1\} \to S_{n-1}$ given by

$$\overline{\sigma}(i) = \begin{cases} \sigma(i) - 1 & i < j \\ \sigma(i+1) - 1 & i \geq j \end{cases} \tag{6.19}$$

and note that

$$N(\sigma) = N(\overline{\sigma}) + (j-1).$$

The $j-1$ appears since there are $j-1$ inversions present in σ which take into account the fact that $\sigma(j) = 1$, which are not present in $\overline{\sigma}$. Indeed (ℓ, j) is an inversion for σ for all $1 \leq \ell < j$. Therefore, by considering the submatrix $\overline{\mathbf{A}} = [\overline{a}_{ij}]$ obtained from $\mathbf{A}$ by deleting row 1 and column j, we have

$$\begin{aligned} a_{1j}C_{1j}(\mathbf{A}) &= \sum_{\substack{\sigma \in S_n \\ \sigma(j)=1}} (-1)^{N(\sigma)} \prod_{i=1}^{n} a_{\sigma(i)i} \\ &= a_{1j} \sum_{\substack{\sigma \in S_n \\ \sigma(j)=1}} (-1)^{N(\sigma)} \prod_{\substack{1 \leq i \leq n \\ i \neq j}} a_{\sigma(i)i} \\ &= a_{1j}(-1)^{j-1} \sum_{\overline{\sigma} \in S_{n-1}} (-1)^{N(\overline{\sigma})} \prod_{i=1}^{n-1} \overline{a}_{\overline{\sigma}(i)i} \end{aligned}$$

$$= a_{1j}(-1)^{j-1}|\overline{\mathbf{A}}|$$
$$= a_{1j}M_{1j}(\mathbf{A}). \tag{6.20}$$

Hence, via (6.18) and (6.20), we have

$$|\mathbf{A}| = \sum_{j=1}^{n} a_{1j}(-1)^{j-1}M_{1j}(\mathbf{A}),$$

as required. □

Definition 6.21. *The* (i,j)***-cofactor*** *of a matrix* $\boldsymbol{A} \in \mathcal{M}_{nn}(\mathbb{R})$*, written as* $C_{ij}(\boldsymbol{A})$*, is defined as*

$$C_{ij}(\boldsymbol{A}) = (-1)^{i+j}M_{ij}(\boldsymbol{A}).$$

An alternative reference to the (i,j)-cofactor of $\mathbf{A}$ is the term "cofactor of the (i,j) entry, a_{ij}, of $\mathbf{A}$" and denoted as A_{ij}.

The number $(-1)^{i+j}$, which is equal to 1 or -1 when $i+j$ is even or odd respectively, is called the *sign of the* (i,j) *position*. In a square matrix, the sign corresponding to the (i,j) position is as follows:

$$\begin{pmatrix} + & - & + & - & \cdots \\ - & + & - & + & \cdots \\ + & - & + & - & \cdots \\ - & + & - & + & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Example 83: For

$$\mathbf{A} = \begin{pmatrix} 2 & -6 & 1 \\ 3 & 1 & -4 \\ -1 & -2 & 5 \end{pmatrix}$$

find $C_{12}(\mathbf{A})$ and $C_{31}(\mathbf{A})$. Recall Example 82.

One can show that the determinant can also be found in terms of the entries of any row and their cofactors or alternatively the entries of any column and their cofactors.

Theorem 6.22. *For any matrix* $\boldsymbol{A} = [a_{ij}] \in \mathcal{M}_{nn}(\mathbb{R})$,

$$\det(\boldsymbol{A}) = \sum_{j=1}^{n} a_{ij} C_{ij}(\boldsymbol{A}) = a_{i1}C_{i1}(\boldsymbol{A}) + a_{i2}C_{i2}(\boldsymbol{A}) + \ldots + a_{in}C_{in}(\boldsymbol{A}),$$

for any i *such that* $1 \leq i \leq n$ *and*

$$\det(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ij} C_{ij}(\boldsymbol{A}) = a_{1j}C_{1j}(\boldsymbol{A}) + a_{2j}C_{2j}(\boldsymbol{A}) + \ldots + a_{nj}C_{nj}(\boldsymbol{A}),$$

for any j *such that* $1 \leq j \leq n$.

Sketch of the proof: Suppose we consider an expansion of row i. Then by swapping successive rows, we can define a new matrix $\mathbf{B}$ with row i of $\mathbf{A}$ in row 1 and every other row in the previous order. This requires $i-1$ swaps of successive rows. Using Lemma 6.20 (and following steps in its proof) it follows that

$$\begin{aligned} |\mathbf{A}| &= (-1)^{i-1}|\mathbf{B}| \\ &= (-1)^{i-1}\left(\sum_{j=1}^{n} a_{ij}(-1)^{j-1}M_{ij}(\mathbf{A})\right) \\ &= \sum_{j=1}^{n} a_{ij}(-1)^{i+j}M_{ij}(\mathbf{A}) \\ &= \sum_{j=1}^{n} a_{ij}C_{ij}(\mathbf{A}). \end{aligned} \tag{6.21}$$

The result for column expansions follows from (6.21) with $|\mathbf{A}^T| = |\mathbf{A}|$.

The sum of products of entries with their cofactor is often referred to as an **expansion** of $\det(\mathbf{A})$ in terms of entries of row i (or column j) and their cofactors.

This means we can choose which row or column to use for the expansion needed to calculate the determinant. In practice, this means we can make use of the possible presence of zero values in a row or column.

Example 84: Calculate $|\mathbf{A}|$ when

$$\mathbf{A} = \begin{pmatrix} 2 & 5 & 0 \\ 35 & 41 & 13 \\ 4 & 11 & 0 \end{pmatrix}.$$

Corollary 6.23. *If* $\boldsymbol{I}_n$ *is the identity matrix in* $\mathcal{M}_{nn}(\mathbb{R})$, *then*

$$\det(\boldsymbol{I}_n) = 1 \quad \forall n \in \mathbb{N}.$$

Proof. Use induction on n together with the first row expansion of $\det(\mathrm{I}_n)$. □

6.7 Triangular matrices

There are a number of corollaries related to triangular matrices that follow from Theorem 6.22.

Corollary 6.24. *Let $n \in \mathbb{N}$ with $n \geq 2$. If $\boldsymbol{A} = [a_{ij}] \in \mathcal{M}_{nn}(\mathbb{R})$ is an upper triangular matrix, i.e. $a_{ij} = 0$ when $i > j$, then*

$$\det(\boldsymbol{A}) = \prod_{i=1}^{n} a_{ii} = a_{11}a_{22}\dots a_{nn}.$$

If $\boldsymbol{A} = [a_{ij}] \in \mathcal{M}_{nn}(\mathbb{R})$ is a lower triangular matrix, i.e. $a_{ij} = 0$ when $i < j$, then

$$\det(\boldsymbol{A}) = a_{11}a_{22}\dots a_{nn}.$$

Proof. Consider the statement $P(n)$ for $n \in \mathbb{N} \setminus \{1\}$ given by:

$$\text{If } \mathbf{A} = [a_{ij}] \in \mathcal{M}_{nn}(\mathbb{R}) \text{ is an upper triangular matrix then } \det(\mathbf{A}) = a_{11}a_{22}\dots a_{nn}. \tag{6.22}$$

$P(2)$ is true since

$$\begin{vmatrix} a_{11} & a_{12} \\ 0 & a_{22} \end{vmatrix} = a_{11}a_{22} - 0a_{12} = a_{11}a_{22}. \tag{6.23}$$

Now, assume $P(k)$ is true for some $k \in \mathbb{N} \setminus \{1\}$, i.e. for any upper triangular matrix $\mathbf{A} \in \mathcal{M}_{nn}(\mathbb{R})$, we have

$$|\mathbf{A}| = \begin{vmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1k} \\ 0 & a_{22} & a_{23} & \cdots & a_{2k} \\ 0 & 0 & a_{33} & \cdots & a_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{kk} \end{vmatrix} = a_{11}a_{22}\dots a_{kk}. \tag{6.24}$$

By expanding along row $k+1$, we have,

$$\begin{aligned} \begin{vmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1(k+1)} \\ 0 & a_{22} & a_{23} & \cdots & a_{2(k+1)} \\ 0 & 0 & a_{33} & \cdots & a_{3(k+1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{(k+1)(k+1)} \end{vmatrix} &= a_{(k+1)(k+1)}(-1)^{k+1+k+1} \begin{vmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1k} \\ 0 & a_{22} & a_{23} & \cdots & a_{2k} \\ 0 & 0 & a_{33} & \cdots & a_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{kk} \end{vmatrix} \\ &= a_{(k+1)(k+1)}a_{11}a_{22}a_{33}\cdots a_{kk} \\ &= a_{11}a_{22}a_{33} \cdots a_{kk}a_{(k+1)(k+1)}. \end{aligned}$$

and hence

$$P(k) \text{ is true} \implies P(k+1) \text{ is true}$$

for all $k \in \mathbb{N} \setminus \{1\}$. From the principle of mathematical induction, (Theorem B.2) we conclude that $P(n)$ given by (6.22) is true for all $n \in \mathbb{N} \setminus \{1\}$, as required.

The proof for lower triangular matrices follows from the result for upper triangular matrices and Theorem 6.10 (since the transpose of a lower triangular matrix is an upper triangular matrix). □

Corollary 6.25. *If $\boldsymbol{D} = [d_{ij}] \in \mathcal{M}_{nn}(\mathbb{R})$ is a diagonal matrix, i.e. $a_{ij} = 0$ when $i \neq j$, then*

$$\det(\boldsymbol{D}) = \prod_{i=1}^{n} d_{ii} = d_{11}d_{22}\dots d_{nn}.$$

Proof. Since a diagonal matrix is a special case of both an upper triangular matrix and a lower triangular matrix, the result follows from the Corollary 6.24. □

Example 85: Find the determinant of the following matrices:

$$\mathbf{A}_1 = \begin{pmatrix} 2 & 5 & 0 & 7 \\ 0 & 5 & 13 & 23 \\ 0 & 0 & -2 & 37 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{A}_2 = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 3 & 11 & 0 & 0 \\ 9 & -5 & -3 & 0 \\ 7 & -8 & 11 & -1 \end{pmatrix} \text{ and } \mathbf{A}_3 = \begin{pmatrix} 2 & 5 & 0 & -13 \\ 35 & 41 & 0 & 29 \\ 4 & 11 & 0 & -17 \\ 11 & -9 & 0 & 89 \end{pmatrix}.$$

6.8 Calculating determinants

The results in the previous sections will be useful when calculating determinants by hand or by computer. In particular:

- We can swap two rows or two columns of a matrix $\mathbf{A}$ to obtain matrix $\mathbf{B}$, but we must change the sign of the determinant, i.e. $|\mathbf{B}| = -|\mathbf{A}|$.
- We can multiply a single row (or a single column) of $\mathbf{A}$ with a real number λ to obtain $\mathbf{B}$ but we must change the determinant by multiplying it by the same number i.e. $|\mathbf{B}| = \lambda|\mathbf{A}|$.
- We can add the multiple of a row of $\mathbf{A}$ to another row of $\mathbf{A}$ (or add the multiple of a column of $\mathbf{A}$ to another column of $\mathbf{A}$) without changing the value of the determinant.
- To calculate the determinant of $\mathbf{A}$, we can use row (column) operations to convert the matrix to a triangular matrix which has a determinant which is given by Corollary 6.24.

Example 86: Find the determinant of the following matrices:

$$\mathbf{A}_1 = \begin{pmatrix} 2 & 5 & 0 & 7 \\ 0 & 0 & -2 & 37 \\ 0 & 5 & 13 & 23 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{A}_2 = \begin{pmatrix} 2 & 3 & 5 & 4 \\ 0 & 1 & 10 & -1 \\ 1 & -2 & 15 & 0 \\ 0 & 0 & -5 & -1 \end{pmatrix} \text{ and } \mathbf{A}_3 = \begin{pmatrix} 2 & 5 & 9 & -13 \\ 3 & 41 & 47 & 29 \\ 4 & 11 & 19 & -17 \\ 11 & -9 & 13 & 89 \end{pmatrix}.$$

In more general cases, we may need to expand matrices along a row or column, in terms of determinants of matrices of smaller sizes until we are left with determinants of 2×2 matrices.

Example 87: Calculate

$$\begin{vmatrix} 0 & -1 & 3 & 4 \\ 1 & 3 & 5 & 2 \\ 2 & 1 & 9 & 6 \\ 3 & 2 & 4 & 6 \end{vmatrix}.$$

You are likely to encounter the following type of problem frequently later in your degree programme.

Example 88: Find all real values of λ, if any, for which $|\mathbf{A} - \lambda\mathbf{I}_3| = 0$, with

$$\mathbf{A} = \begin{pmatrix} 2 & 2 & 1 \\ 1 & 3 & 1 \\ 1 & 2 & 2 \end{pmatrix}.$$

The polynomial in λ given by $|\mathbf{A}-\lambda\mathbf{I}_3| = 0$ is referred to as the **characteristic polynomial** of $\mathbf{A}$. The zeros of this polynomial are referred to as the **eigenvalues** of $\mathbf{A}$.

6.9 Determinants of elementary matrices

Since row operations can be achieved by multiplying a matrix with an appropriate elementary matrix on the left, we should deduce the determinants of elementary matrices.

Lemma 6.26. *Let $\boldsymbol{E}_{(i,j)} \in \mathcal{M}_{nn}(\mathbb{R})$ be the elementary matrix corresponding to the swapping of rows i and j. Then,*

$$\det(\boldsymbol{E}_{(i,j)}) = -1.$$

Proof. $\mathbf{E}_{(i,j)}$ is obtained from the identity matrix by swapping rows i and j, hence, via Theorem 6.17 and Corollary 6.23,

$$\det(\mathbf{E}_{(i,j)}) = -\det(\mathbf{I}_n) = -1.$$

□

Lemma 6.27. *Let $\boldsymbol{E}_{[i,\lambda]} \in \mathcal{M}_{nn}(\mathbb{R})$ be the elementary matrix corresponding to the multiplication of row i with $\lambda \in \mathbb{R} \setminus \{0\}$. Then,*

$$\det(\boldsymbol{E}_{[i,\lambda]}) = \lambda.$$

Proof. $\mathbf{E}_{[i,\lambda]}$ is obtained from the identity matrix by multiplying row i by λ, hence, via Theorem 6.12 and Corollary 6.23,

$$\det(\mathbf{E}_{[i,\lambda]}) = \lambda \det(\mathbf{I}_n) = \lambda.$$

□

Lemma 6.28. *Let $\boldsymbol{E}_{(i,j,\lambda)} \in \mathcal{M}_{nn}(\mathbb{R})$ be the elementary matrix corresponding to adding λ times row j to row i, where λ is non-zero real number. Then,*

$$\det(\boldsymbol{E}_{(i,j,\lambda)}) = 1.$$

Proof. $\mathbf{E}_{(i,j,\lambda)}$ is obtained from the identity matrix by multiplying row j by λ and then adding this result to row i, hence via Theorem 6.16 and Corollary 6.23,

$$\det(\mathbf{E}_{(i,j,\lambda)}) = \det(\mathbf{I}_n) = 1.$$

□

Notably, elementary matrices have non-zero determinants.

Example 89: Find the determinants of the following matrices defined by elementary row operations:

$$\mathbf{E}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \ \mathbf{E}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } \mathbf{E}_3 = \begin{pmatrix} 1 & 0 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Theorem 6.29. *Given a matrix $\boldsymbol{A} \in \mathcal{M}_{nn}(\mathbb{R})$ and an elementary matrix $\boldsymbol{E} \in \mathcal{M}_{nn}(\mathbb{R})$, then*

$$\det(\boldsymbol{E} \cdot \boldsymbol{A}) = \det(\boldsymbol{E}) \cdot \det(\boldsymbol{A}).$$

Proof. We split the proof into 3 cases, for each type of elementary row operation.

1. Consider the elementary matrix $\mathbf{E}_{(i,j)}$ which corresponds to swapping row i and row j. Then $\mathbf{B} = \mathbf{E}_{(i,j)} \cdot \mathbf{A}$ is the matrix obtained from $\mathbf{A}$ by swapping rows i and j. Hence, via Lemma 6.26, $\det(\mathbf{E}_{(i,j)}) = -1$ and via Theorem 6.17,

$$\begin{aligned} \det(\mathbf{E}_{(i,j)} \cdot \mathbf{A}) &= \det(\mathbf{B}) \\ &= -\det(\mathbf{A}) \\ &= \det(\mathbf{E}_{(i,j)}) \cdot \det(\mathbf{A}). \end{aligned}$$

2. Consider the elementary matrix $\mathbf{E}_{[i,\lambda]}$ which corresponds to multiplying row i with the non-zero real number λ. Then $\mathbf{B} = \mathbf{E}_{[i,\lambda]} \cdot \mathbf{A}$ is the matrix obtained from $\mathbf{A}$ by multiplying row i by λ. Hence, via Lemma 6.27, $\det(\mathbf{E}_{[i,\lambda]}) = \lambda$ and via Theorem 6.12,

$$\begin{aligned} \det(\mathbf{E}_{[i,\lambda]} \cdot \mathbf{A}) &= \det(\mathbf{B}) \\ &= \lambda \det(\mathbf{A}) \\ &= \det(\mathbf{E}_{[i,\lambda]}) \cdot \det(\mathbf{A}). \end{aligned}$$

3. Consider the elementary matrix $\mathbf{E}_{(i,j,\lambda)}$ which corresponds to multiplying row j with the non-zero real number λ and adding this result to row i. Then $\mathbf{B} = \mathbf{E}_{ij}(\lambda) \cdot \mathbf{A}$ is the matrix obtained from $\mathbf{A}$ by adding to row i, row j multiplied by λ. Hence, via Lemma 6.28, $\det(\mathbf{E}_{(i,j,\lambda)}) = 1$ and via Theorem 6.16,

$$\begin{aligned} \det(\mathbf{E}_{ij}(\lambda) \cdot \mathbf{A}) &= \det(\mathbf{B}) \\ &= \det(\mathbf{A}) \\ &= \det(\mathbf{E}_{(i,j,\lambda)}) \cdot \det(\mathbf{A}). \end{aligned}$$

□

Example 90: Evaluate the determinant of $\mathbf{E} \cdot \mathbf{A}$ where

$$\mathbf{E} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } \mathbf{A} = \begin{pmatrix} 1 & 2 & -1 \\ -1 & 1 & 1 \\ 2 & 0 & -1 \end{pmatrix}.$$

Corollary 6.30. *Given a matrix* $\boldsymbol{A} \in \mathcal{M}_{nn}(\mathbb{R})$ *and elementary matrices* $\boldsymbol{E}_i \in \mathcal{M}_{nn}(\mathbb{R})$, *with* $1 \leq i \leq k$ *then*

$$\det(\boldsymbol{E}_k \cdot \boldsymbol{E}_{k-1} \cdot \ldots \cdot \boldsymbol{E}_1 \cdot \boldsymbol{A}) = \det(\boldsymbol{E}_k) \cdot \det(\boldsymbol{E}_{k-1}) \ldots \cdot \det(\boldsymbol{E}_1) \cdot \det(\boldsymbol{A}).$$

Proof. Consider the statement $P(k)$ for $k \in \mathbb{N}$ given by

$$\det(\mathbf{E}_k \cdot \mathbf{E}_{k-1} \cdot \ldots \cdot \mathbf{E}_1 \cdot \mathbf{A}) = \det(\mathbf{E}_k) \cdot \det(\mathbf{E}_{k-1}) \ldots \cdot \det(\mathbf{E}_1) \cdot \det(\mathbf{A}), \tag{6.25}$$

where $\mathbf{E}_i \in \mathcal{M}_{nn}(\mathbb{R})$ are elementary matrices and $\mathbf{A} \in \mathcal{M}_{nn}(\mathbb{R})$. Via Theorem 6.29

$$\det(\mathbf{E}_1 \cdot \mathbf{A}) = \det(\mathbf{E}_1) \cdot \det(\mathbf{A}),$$

and hence,

$$P(1) \text{ is true.} \tag{6.26}$$

Now assume that the statement $P(m)$ is true, i.e.

$$\det(\mathbf{E}_m \cdot \mathbf{E}_{m-1} \cdot \ldots \cdot \mathbf{E}_1 \cdot \mathbf{A}) = \det(\mathbf{E}_m) \cdot \det(\mathbf{E}_{m-1}) \ldots \cdot \det(\mathbf{E}_1) \cdot \det(\mathbf{A}), \tag{6.27}$$

is true for any elementary matrices $\mathbf{E}_i$ for $1 \leq i \leq m$ and $\mathbf{A} \in \mathcal{M}_{nn}(\mathbb{R})$. Then via Theorem 6.29 and (6.27),

$$\begin{aligned} \det(\mathbf{E}_{m+1} \cdot \mathbf{E}_m \cdot \ldots \cdot \mathbf{E}_1 \cdot \mathbf{A}) &= \det(\mathbf{E}_{m+1} \cdot (\ \mathbf{E}_m \cdot \ldots \cdot \mathbf{E}_1 \cdot \mathbf{A})) \\ &= \det(\mathbf{E}_{m+1}) \cdot \det(\mathbf{E}_m \cdot \ldots \cdot \mathbf{E}_1 \cdot \mathbf{A}) \\ &= \det(\mathbf{E}_{m+1}) \cdot \det(\mathbf{E}_m) \cdot \ldots \cdot \det(\mathbf{E}_1) \cdot \det(\mathbf{A}). \end{aligned} \tag{6.28}$$

Therefore, via (6.27) and (6.28)

$$P(m) \text{ is true} \implies P(m+1) \text{ is true} \tag{6.29}$$

for all $m \in \mathbb{N}$. Finally, via (6.26), (6.29) and the principle of mathematical induction (Theorem B.1), the statement $P(k)$ given by (6.25) is true for all $k \in \mathbb{N}$, as required. □

Corollary 6.31. *Given elementary matrices $\boldsymbol{E}_i \in \mathcal{M}_{nn}(\mathbb{R})$, with $1 \leq i \leq k$ then*

$$\det(\boldsymbol{E}_k \cdot \boldsymbol{E}_{k-1} \cdot \ldots \cdot \boldsymbol{E}_1) = \det(\boldsymbol{E}_k) \cdot \det(\boldsymbol{E}_{k-1}) \cdot \ldots \cdot \det(\boldsymbol{E}_1).$$

Proof. This follows from Corollary 6.30 by choosing $\mathbf{A} = \mathbf{I}_n$. □

Corollaries 6.30 and 6.31 are important to allow us to prove that $|\mathbf{A} \cdot \mathbf{B}| = |\mathbf{A}||\mathbf{B}|$ for $\mathbf{A}, \mathbf{B} \in \mathcal{M}_{nn}(\mathbb{R})$.

6.10 Two major theorems about determinants

We can now generalize the observation we made in the introduction.

Theorem 6.32. *Suppose that* $\boldsymbol{A} \in \mathcal{M}_{nn}(\mathbb{R})$. *Then*

1. $\boldsymbol{A}$ *is not invertible if and only if* $\det(\boldsymbol{A}) = 0$.
2. $\boldsymbol{A}$ *is invertible if and only if* $\det(\boldsymbol{A}) \neq 0$.

Proof. Via the Gaussian Elimination algorithm there exist elementary row operations which convert $\mathbf{A}$ into a matrix which is in reduced echelon form $\mathbf{U}$. In particular, $\mathbf{U}$ an upper triangular matrix $\mathbf{U}$ i.e.

$$\mathbf{U} = \mathbf{E}_k \cdot \mathbf{E}_{k-1} \cdot \ldots \cdot \mathbf{E}_2 \cdot \mathbf{E}_1 \cdot \mathbf{A}. \tag{6.30}$$

Using Corollary 6.30 and (6.30), we have

$$\begin{aligned} \det(\mathbf{U}) &= \det(\mathbf{E}_k \cdot \mathbf{E}_{k-1} \cdot \ldots \cdot \mathbf{E}_2 \cdot \mathbf{E}_1 \cdot \mathbf{A}) \\ &= \det(\mathbf{E}_k) \cdot \det(\mathbf{E}_{k-1}) \cdot \ldots \cdot \det(\mathbf{E}_2) \cdot \det(\mathbf{E}_1) \cdot \det(\mathbf{A}). \end{aligned} \tag{6.31}$$

Since the determinant of an elementary matrix is either $-1, 1$ or $\lambda \neq 0$ (follows from Lemmas 6.26-6.28), the equation in (6.31) yields

$$\det(\mathbf{A}) = 0 \text{ if and only if } \det(\mathbf{U}) = 0.$$

We know from the Gaussian Elimination Algorithm that $\mathbf{A}$ is not invertible, if and only if there is a row of zeros in the bottom row of $\mathbf{U}$.

If $\mathbf{U}$ has a row of zeros, then $\det(\mathbf{U}) = 0$ by Corollary 6.13. If $\mathbf{U}$ does not have a row of zeros, then, as $\mathbf{U}$ is in reduced echelon form $\mathbf{U} = \mathbf{I}_n$ and so $\mathbf{U}$ has determinant 1.

It follows that $\det(\mathbf{U}) = 0$ if and only if $\mathbf{A}$ is invertible. As $\det(\mathbf{A}) = 0$ if and only if $\det(\mathbf{U}) = 0$ we have proved the result. □

Theorem 6.33. ***(Determinant product theorem)*** *Let* $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{M}_{nn}(\mathbb{R})$. *Then,*

$$\det(\boldsymbol{A} \cdot \boldsymbol{B}) = \det(\boldsymbol{A}) \cdot \det(\boldsymbol{B}).$$

Proof. We consider two cases, when $\mathbf{A}$ is invertible, and when $\mathbf{A}$ is not invertible.

1. Suppose that $\mathbf{A}$ is invertible. Then Theorem 5.11 implies that

   $$\mathbf{A} = \mathbf{E}_1 \cdot \cdots \cdot \mathbf{E}_k$$

   is a product of elementary matrices. Therefore

   $$\mathbf{A} \cdot \mathbf{B} = \mathbf{E}_1 \cdot \mathbf{E}_2 \cdot \ldots \cdot \mathbf{E}_{k-1} \cdot \mathbf{E}_k \cdot \mathbf{B}.$$

Thus Corollary 6.30

$$\begin{aligned}\det(\mathbf{A}\cdot\mathbf{B}) &= \det(\mathbf{E}_1\cdot\mathbf{E}_2\cdot\ldots\cdot\mathbf{E}_{k-1}\cdot\mathbf{E}_k\cdot\mathbf{B})\\ &= \det(\mathbf{E}_1)\cdot\det(\mathbf{E}_2)\cdot\ldots\cdot\det(\mathbf{E}_{k-1})\cdot\det(\mathbf{E}_k)\cdot\det(\mathbf{B})\\ &= \det(\mathbf{E}_1\cdot\mathbf{E}_2\cdot\ldots\cdot\mathbf{E}_{k-1}\cdot\mathbf{E}_k)\cdot\det(\mathbf{B})\\ &= \det(\mathbf{A})\cdot\det(\mathbf{B}).\end{aligned}$$

2. When $\mathbf{A}$ is not invertible then a sequence of elementary row operations can convert $\mathbf{A}$ into an upper triangular matrix with a row of zeros on the bottom row. So, we have

$$\mathbf{U} = \mathbf{E}_k\cdot\mathbf{E}_{k-1}\cdot\ldots\cdot\mathbf{E}_2\cdot\mathbf{E}_1\cdot\mathbf{A},$$

which means that

$$\mathbf{A} = \mathbf{E}_1^{-1}\cdot\mathbf{E}_2^{-1}\cdot\ldots\cdot\mathbf{E}_{k-1}^{-1}\cdot\mathbf{E}_k^{-1}\cdot\mathbf{U}. \tag{6.32}$$

Hence, via (6.32),

$$\begin{aligned}\mathbf{A}\cdot\mathbf{B} &= \mathbf{E}_1^{-1}\cdot\mathbf{E}_2^{-1}\cdot\ldots\cdot\mathbf{E}_{k-1}^{-1}\cdot\mathbf{E}_k^{-1}\cdot\mathbf{U}\cdot\mathbf{B}\\ &= \mathbf{E}_1^{-1}\cdot\mathbf{E}_2^{-1}\cdot\ldots\cdot\mathbf{E}_{k-1}^{-1}\cdot\mathbf{E}_k^{-1}\cdot\mathbf{C},\end{aligned} \tag{6.33}$$

where $\mathbf{C} = \mathbf{U}\cdot\mathbf{B}$ also has a row of zeros on the bottom row. Therefore, (via Corollary 6.13) $\det(\mathbf{C}) = 0$, and hence, via Corollary 6.30 and (6.33),

$$\begin{aligned}\det(\mathbf{A}\cdot\mathbf{B}) &= \det(\mathbf{E}_1^{-1}\cdot\mathbf{E}_2^{-1}\cdot\ldots\cdot\mathbf{E}_{k-1}^{-1}\cdot\mathbf{E}_k^{-1}\cdot\mathbf{C})\\ &= \det(\mathbf{E}_1^{-1})\cdot\det(\mathbf{E}_2^{-1})\cdot\ldots\cdot\det(\mathbf{E}_{k-1}^{-1})\cdot\det(\mathbf{E}_k^{-1})\cdot\det(\mathbf{C})\\ &= 0.\end{aligned}$$

Since $\mathbf{A}$ is not invertible, $\det(\mathbf{A}) = 0$ so that in this case

$$\det(\mathbf{A}\cdot\mathbf{B}) = \det(\mathbf{A})\cdot\det(\mathbf{B}),$$

which completes the proof.

□

Example 91: Determine $\det(\mathbf{A}\cdot\mathbf{B})$ with

$$\mathbf{A} = \begin{pmatrix}0 & 1 & 3\\ 1 & 0 & 5\\ 1 & 1 & 1\end{pmatrix} \text{ and } \mathbf{B} = \begin{pmatrix}1 & 1 & 0\\ -3 & 1 & -1\\ 1 & 1 & 1\end{pmatrix}.$$

Corollary 6.34. *Let $\boldsymbol{A}_i \in \mathcal{M}_{nn}(\mathbb{R})$, with $1 \le i \le k$. Then,*

$$\det(\boldsymbol{A}_1\cdot\boldsymbol{A}_2\cdot\ldots\cdot\boldsymbol{A}_{k-1}\cdot\boldsymbol{A}_k) = \det(\boldsymbol{A}_1)\cdot\det(\boldsymbol{A}_2)\cdot\ldots\cdot\det(\boldsymbol{A}_{k-1})\cdot\det(\boldsymbol{A}_k).$$

Proof. Consider the statement $P(k)$ for $k \in \mathbb{N} \setminus \{1\}$ given by

$$\det(\mathbf{A}_1 \cdot \mathbf{A}_2 \ldots \mathbf{A}_{k-1} \cdot \mathbf{A}_k) = \det(\mathbf{A}_1) \cdot \det(\mathbf{A}_2) \cdot \ldots \cdot \det(\mathbf{A}_{k-1}) \cdot \det(\mathbf{A}_k). \tag{6.34}$$

Using Theorem 6.33, we have

$$\det(\mathbf{A}_1 \cdot \mathbf{A}_2) = \det(\mathbf{A}_1) \cdot \det(\mathbf{A}_2),$$

and hence

$$P(2) \text{ is true.} \tag{6.35}$$

Now assume that $P(m)$ is true for some $m \in \mathbb{N} \setminus \{1\}$, i.e.,

$$\det(\mathbf{A}_1 \cdot \mathbf{A}_2 \cdot \ldots \cdot \mathbf{A}_{m-1} \cdot \mathbf{A}_m) = \det(\mathbf{A}_1) \cdot \det(\mathbf{A}_2) \ldots \det(\mathbf{A}_{m-1}) \cdot \det(\mathbf{A}_m), \tag{6.36}$$

for any $\mathbf{A}_i \in \mathcal{M}_{nn}(\mathbb{R})$. Then, via Theorem 6.33 and (6.36) it follows that

$$\begin{aligned}
\det(\mathbf{A}_1 \cdot \mathbf{A}_2 \cdot \ldots \cdot \mathbf{A}_m \cdot \mathbf{A}_{m+1}) &= \det((\mathbf{A}_1 \cdot \mathbf{A}_2 \cdot \ldots \cdot \mathbf{A}_m) \cdot \mathbf{A}_{m+1}) \\
&= \det(\mathbf{A}_1 \cdot \mathbf{A}_2 \cdot \ldots \cdot \mathbf{A}_m) \cdot \det(\mathbf{A}_{m+1}) \\
&= (\det(\mathbf{A}_1) \cdot \det(\mathbf{A}_2) \cdot \ldots \cdot \det(\mathbf{A}_{m-1}) \cdot \det(\mathbf{A}_m)) \cdot \det(\mathbf{A}_{m+1}) \\
&= \det(\mathbf{A}_1) \cdot \det(\mathbf{A}_2) \cdot \ldots \cdot \det(\mathbf{A}_m) \cdot \det(\mathbf{A}_{m+1}). \qquad (6.37)
\end{aligned}$$

Therefore, it follows from (6.36) and (6.37) that

$$P(m) \text{ is true} \Longrightarrow P(m+1) \text{ is true} \tag{6.38}$$

for all $m \in \mathbb{N} \setminus \{1\}$. Therefore, via (6.35) and (6.38), via the principle of mathematical induction (Theorem B.2), it follows that statement $P(k)$ given by (6.34) is true for all $k \in \mathbb{N} \setminus \{1\}$, as required. □

Corollary 6.35. *Let $\boldsymbol{A} \in \mathcal{M}_{nn}(\mathbb{R})$. Then, for $k \in \mathbb{N}$,*

$$\det(\boldsymbol{A}^k) = \det(\boldsymbol{A})^k.$$

Proof. Follows from Corollary 6.34 with $\mathbf{A}_i = \mathbf{A}$ for $1 \leq i \leq k$. □

Example 92: Suppose that

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 3 \\ 1 & 0 & 5 \\ 1 & 1 & 1 \end{pmatrix}.$$

Find $\det(\mathbf{A}^k)$ for $k \in \mathbb{N}$.

A very important property links the determinant of a matrix with whether or not a matrix is invertible.

Theorem 6.36. *A matrix* $\boldsymbol{A} \in \mathcal{M}_{nn}(\mathbb{R})$ *is invertible if and only if* $\det(\boldsymbol{A}) \neq 0$. *Furthermore,*

$$\det(\boldsymbol{A}^{-1}) = \frac{1}{\det(\boldsymbol{A})}.$$

Proof. We know from Theorem 6.32 that $\mathbf{A}$ is invertible if and only if $\det(\mathbf{A}) \neq 0$.

If $\mathbf{A}$ is invertible then $\mathbf{A}^{-1}$ exists and $\mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{I}_n$. Therefore,

$$\det(\mathbf{A} \cdot \mathbf{A}^{-1}) = \det(\mathbf{I}_n) = 1. \tag{6.39}$$

Hence, via Theorem 6.33 and (6.39),

$$\det(\mathbf{A} \cdot \mathbf{A}^{-1}) = \det(\mathbf{A}) \cdot \det(\mathbf{A}^{-1}) = 1.$$

So $\det(\mathbf{A}) \neq 0$ and

$$\det(\mathbf{A}^{-1}) = \frac{1}{\det(\mathbf{A})}.$$

□

Corollary 6.37. *A* ***homogeneous*** *system* $\boldsymbol{A} \cdot \boldsymbol{x} = \boldsymbol{0}$ *of* n *linear equations in* n *unknowns has a non-trivial solution if and only if* $\det(\boldsymbol{A}) = 0$.

Proof. If $\det(\mathbf{A}) \neq 0$ then via Theorem 6.36, $\mathbf{A}$ is invertible and the only possible solution to the homogeneous linear system of equations, is the trivial one. This follows since,

$$\mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{A} \cdot \mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{0} = \mathbf{0}.$$

Therefore, a non-trivial solution to a homogeneous linear system of equations can only exist if $\det(\mathbf{A}) = 0$.

To show that when $\det(\mathbf{A}) = 0$ there exist a non-trivial solution, we examine the outcome of the sequence of elementary row operations transforming $\mathbf{A}$ into its echelon form. When $\det(\mathbf{A}) = 0$, we can reduce $\mathbf{A}$ using ERO to an upper triangular matrix with $m \geq 1$ rows of zeros at the bottom. Since the RHS of a homogeneous system of equations only contains zeros, this means that m entries of $\mathbf{x}$ remain undetermined by the system of equations, and hence, a non-trivial solution exists.

Therefore, a non-trivial solution to $\mathbf{A} \cdot \mathbf{x} = \mathbf{0}$ exists if and only if $\mathbf{A}$ is not invertible, which is the case if and only if $\det(\mathbf{A}) = 0$, as required. □

6.11 Cofactor and adjoint matrices

Definition 6.38. *The **cofactor matrix**, $\boldsymbol{C}(\boldsymbol{A}) = [c_{ij}]$ of $\boldsymbol{A} = [a_{ij}] \in \mathcal{M}_{nn}(\mathbb{R})$ is the matrix in $\mathcal{M}_{nn}(\mathbb{R})$ with (i,j)-th entry given by the cofactor of a_{ij} in $\boldsymbol{A}$, i.e.*

$$c_{ij} = C_{ij}(\boldsymbol{A}).$$

Definition 6.39. *The **adjoint matrix**, $\mathrm{adj}(\boldsymbol{A})$ of $\boldsymbol{A} = [a_{ij}] \in \mathcal{M}_{nn}(\mathbb{R})$ is the transpose of the cofactor matrix of $\boldsymbol{A}$, i.e.*

$$\mathrm{adj}(\boldsymbol{A}) = (\boldsymbol{C}(\boldsymbol{A}))^T.$$

Example 93: Determine the cofactor matrix and the adjoint matrix of

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 2 \\ 3 & 4 & 5 \\ 7 & 8 & 10 \end{pmatrix}.$$

Example 94: Determine the cofactor matrix and the adjoint matrix of

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

Answer: From the definition, it follows that

$$\begin{aligned} \mathbf{C}(\mathbf{A}) &= \begin{pmatrix} a_{22} & -a_{21} \\ -a_{12} & a_{11} \end{pmatrix} \\ \mathrm{adj}(\mathbf{A}) &= \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}. \end{aligned} \tag{6.40}$$

Observe that for $\mathbf{A} \in \mathcal{M}_{22}(\mathbb{R})$,

$$\begin{aligned} \mathbf{A} \cdot \mathrm{adj}(\mathbf{A}) &= \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \cdot \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix} \\ &= \begin{pmatrix} a_{11}a_{22} + a_{12}(-a_{21}) & a_{11}(-a_{12}) + a_{12}a_{11} \\ a_{21}a_{22} + a_{22}(-a_{21}) & a_{21}(-a_{12}) + a_{22}a_{11} \end{pmatrix} \\ &= \begin{pmatrix} a_{11}a_{22} - a_{12}a_{21} & 0 \\ 0 & a_{11}a_{22} - a_{12}a_{21} \end{pmatrix} \end{aligned}$$

$$= (a_{11}a_{22} - a_{12}a_{21})\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$$

$$= \det(\mathbf{A})\mathbf{I}_2.$$

Similarly, one finds that $\text{adj}(\mathbf{A})\cdot\mathbf{A} = \det(\mathbf{A})\mathbf{I}_2$. In full generality, this type of result allows us to represent $\mathbf{A}^{-1}$ (for invertible $\mathbf{A} \in \mathcal{M}_{nn}(\mathbb{R})$) solely in terms of $\text{adj}(\mathbf{A})$ and $\det(\mathbf{A})$.

Theorem 6.40. *For a matrix* $\boldsymbol{A} \in \mathcal{M}_{nn}(\mathbb{R})$ *,*

$$\boldsymbol{A}\cdot\text{adj}(\boldsymbol{A}) = \text{adj}(\boldsymbol{A})\cdot\boldsymbol{A} = \det(\boldsymbol{A})\boldsymbol{I}_n.$$

In particular, for $\det(\boldsymbol{A}) \neq 0$*,*

$$\boldsymbol{A}^{-1} = \frac{1}{\det(\boldsymbol{A})}\text{adj}(\boldsymbol{A}).$$

Proof. 1. First we show that $\mathbf{A}\cdot\text{adj}(\mathbf{A}) = \det(\mathbf{A})\mathbf{I}_n$. If $\text{adj}(\mathbf{A}) = [f_{ij}]$, then $f_{ij} = C_{ji}(\mathbf{A})$. Denote $\mathbf{Q} = \mathbf{A}\cdot\text{adj}(\mathbf{A})$ with $\mathbf{Q} = [q_{ij}]$, then for $i = 1, \ldots, n$,

$$\begin{aligned} q_{ii} &= \sum_{k=1}^{n} a_{ik}f_{ki} \\ &= \sum_{k=1}^{n} a_{ik}C_{ik}(\mathbf{A}) \\ &= \det(\mathbf{A}). \end{aligned} \tag{6.41}$$

Also, for $i, j = 1, \ldots, n$ with $i \neq j$, we find that

$$\begin{aligned} q_{ij} &= \sum_{k=1}^{n} a_{ik}f_{kj} \\ &= \sum_{k=1}^{n} a_{ik}C_{jk}(\mathbf{A}). \end{aligned} \tag{6.42}$$

The RHS of (6.42) is the sum of entries in row i multiplied with the cofactors of entries in row j. Therefore, the RHS of (6.42) can be seen to be the determinant of a matrix where row j has been replaced by row i, i.e. a matrix where row i and j are identical. Such a matrix has determinant equal to 0 (via Corollary **??**) and so

$$q_{ij} = 0. \tag{6.43}$$

We conclude from (6.43) that $\mathbf{A}\cdot\text{adj}(\mathbf{A})$ is a diagonal matrix, and via (6.41), all diagonal entries are equal to $\det(\mathbf{A})$. Hence,

$$\mathbf{A}\cdot\text{adj}(\mathbf{A}) = \det(\mathbf{A})\mathbf{I}_n.$$

2. To show that $\text{adj}(\mathbf{A}) \cdot \mathbf{A} = \det(\mathbf{A})\mathbf{I}_n$, we follow a similar proof, but using columns of $\mathbf{A}$ instead of rows.

It then follows that provided $\det(\mathbf{A}) \neq 0$,

$$\left(\frac{1}{\det(\mathbf{A})}\text{adj}(\mathbf{A})\right) \cdot \mathbf{A} = \mathbf{A} \cdot \left(\frac{1}{\det(\mathbf{A})}\text{adj}(\mathbf{A})\right) = \mathbf{I}_n,$$

and since the inverse matrix of $\mathbf{A}$ is unique (see Theorem 4.18),

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\text{adj}(\mathbf{A}),$$

as required. □

Example 95: Using Theorem 6.40 find the inverse (when it exists) of

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

Answer: Observe that $\det(\mathbf{A}) = a_{11}a_{22} - a_{12}a_{21}$ and via (6.40),

$$\text{adj}(\mathbf{A}) = \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}.$$

Therefore, it follows from Theorem 6.40 that provided $\det(\mathbf{A}) = a_{11}a_{22} - a_{12}a_{21} \neq 0$, then

$$\mathbf{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}. \tag{6.44}$$

Example 96: Use (6.44) to find the inverse of $\mathbf{A}$, given by

$$\mathbf{A} = \begin{pmatrix} 5 & 6 \\ 8 & 7 \end{pmatrix}.$$

Theorem 6.40 yields a closed form representation for the inverse of a matrix. It does not generally provide the best method to find the inverse in practice but has significant implications for further theoretical results.

6.12 Cramer's rule

Theorem 6.41. ***(Cramer's rule)*** *If $\boldsymbol{A} \in \mathcal{M}_{nn}(\mathbb{R})$ is invertible, then the (unique) solution of the system $\boldsymbol{A} \cdot \boldsymbol{x} = \boldsymbol{b}$ of n linear equations in n unknowns, where*

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \text{ and } \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix},$$

is given by

$$x_1 = \frac{\det(\boldsymbol{A}_1)}{\det(\boldsymbol{A})},\ x_2 = \frac{\det(\boldsymbol{A}_2)}{\det(\boldsymbol{A})},\ \ldots, x_n = \frac{\det(\boldsymbol{A}_n)}{\det(\boldsymbol{A})}.$$

For each $k = 1, 2, \ldots n$, the matrix $\boldsymbol{A}_k$ is obtained from $\boldsymbol{A}$ by replacing the entries in column k of $\boldsymbol{A}$ by the entries in the column vector $\boldsymbol{b}$.

Proof. If $\mathbf{A}$ is an invertible matrix, then via Theorem **??** the system $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$ has a unique solution given by $\mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{b}$. Using the formula for the inverse obtained in Theorem 6.40 we find that

$$\begin{aligned} \mathbf{x} &= \mathbf{A}^{-1} \cdot \mathbf{b} \\ &= \left(\frac{\text{adj}(\mathbf{A})}{\det(\mathbf{A})} \right) \cdot \mathbf{b} \\ &= \frac{1}{\det(\mathbf{A})} \left(\text{adj}(\mathbf{A}) \cdot \mathbf{b} \right). \end{aligned}$$

From Definition 6.39 the adjoint matrix is of the form

$$\text{adj}(\mathbf{A}) = \begin{pmatrix} C_{11}(\mathbf{A}) & C_{21}(\mathbf{A}) & \ldots & C_{n1}(\mathbf{A}) \\ C_{12}(\mathbf{A}) & C_{22}(\mathbf{A}) & \ldots & C_{n2}(\mathbf{A}) \\ \vdots & \vdots & \vdots & \vdots \\ C_{1n}(\mathbf{A}) & C_{2n}(\mathbf{A}) & \ldots & C_{nn}(\mathbf{A}) \end{pmatrix}$$

so that the result for the k-th entry in $\mathbf{x}$ is given by

$$\begin{aligned} x_k &= \frac{1}{\det(\mathbf{A})} \sum_{i=1}^{n} C_{ik}(\mathbf{A}) b_i \\ &= \frac{1}{\det(\mathbf{A})} \left(b_1 C_{1k}(\mathbf{A}) + b_2 C_{2k}(\mathbf{A}) + \ldots + b_n C_{nk}(\mathbf{A}) \right). \end{aligned}$$

Observe that the sum

$$b_1 C_{1k}(\mathbf{A}) + b_2 C_{2k}(\mathbf{A}) + \ldots + b_n C_{nk}(\mathbf{A})$$

has the form of an expansion for a determinant, with the difference that the cofactors belong to column k. So this can be seen as the expansion for the determinant (around column k) of the matrix

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1,k-1} & b_1 & a_{1,k+1} & \dots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2,k-1} & b_2 & a_{2,k+1} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{n,k-1} & b_n & a_{n,k+1} & \dots & a_{nn} \end{pmatrix}$$

This holds for all values of $k = 1, \dots, n$, as required. □

Again, this is not the most efficient way to calculate the solution to a system of linear equations by hand (in general), but having a closed form representation for the solution is of significant theoretical importance.

Example 97: Solve the system of equations below using Cramer's rule:

$$\begin{aligned} 2x_1 - 3x_2 + 4x_3 &= 5, \\ 3x_1 + 2x_2 + x_3 &= 7, \\ 5x_1 + x_2 - 2x_3 &= -3. \end{aligned}$$

6.13 Eigenvalues of square matrices

Eigenvalues and eigenvectors of square matrices can be used to infer information about matrices and the objects that they may represent. Eigenvalues and eigenvectors of square matrices will appear frequently in mathematical theory developed in courses later in your degree programme (and highlighted in this course in the practice questions).

Definition 6.42. *Let $\boldsymbol{A} \in \mathcal{M}_{nn}(\mathbb{R})$. The polynomial in λ of degree n given by*

$$|\boldsymbol{A} - \lambda \boldsymbol{I}_n| = 0,$$

*is referred to as the **characteristic polynomial** of $\boldsymbol{A}$. The zeros of the characteristic polynomial of $\boldsymbol{A}$ are referred to as the **eigenvalues** of $\boldsymbol{A}$.*

Via Theorem 3.32 and Definition 6.42, it follows that every matrix $\mathbf{A} \in \mathcal{M}_{nn}(\mathbb{R})$ has m distinct eigenvalues in $\mathbb{C}$ $(1 \leq m \leq n)$. Note that the eigenvalues of $\mathbf{A} \in \mathcal{M}_{nn}(\mathbb{R})$ are in $\mathbb{C}$ and not necessarily in $\mathbb{R}$.

Example 98: Find the characteristic polynomial, and hence, the eigenvalues of the following matrices:

$$\mathbf{A} = \begin{pmatrix} 2 & 0 \\ 3 & 3 \end{pmatrix} \text{ and } \mathbf{B} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

For each real eigenvalue[2] of $\mathbf{A} \in \mathcal{M}_{nn}(\mathbb{R})$, we can also introduce

Definition 6.43. *Let $\boldsymbol{A} \in \mathcal{M}_{nn}(\mathbb{R})$ and suppose that $\lambda \in \mathbb{R}$ is an eigenvalue of $\boldsymbol{A}$. Then a non-zero vector $\boldsymbol{x} \in \mathcal{M}_{n1}(\mathbb{R})$ (or vector in $\mathbb{R}^n$) is an eigenvector for $\boldsymbol{A}$ corresponding to eigenvalue λ if*

$$\boldsymbol{A} \cdot \boldsymbol{x} = \lambda \boldsymbol{x}.$$

Using results in Chapter 9, one can establish that eigenvectors (as described above) exist.

Example 99: Find all eigenvectors for:

$$\mathbf{A} = \begin{pmatrix} 2 & 0 \\ 3 & 3 \end{pmatrix}.$$

[2] A similar definition can be given for complex eigenvalues.

Chapter 7

Vector Spaces

▸ **Learning Outcomes** ◂ After the completion of the lectures and support sessions associated with this chapter you should be able to:

- state and verify the definitions of a vector space;
- define and identify subspaces of a vector space;
- define linear (in)dependence, a spanning set and a basis, and, apply these concepts on given vector spaces or subspaces and the row space and column space of matrices;
- cite and apply given properties relating to linear (in)dependence, spanning sets and bases;
- define and use the dimension of a vector space;
- cite and apply given properties related to the dimension of vector spaces and subspaces, including row rank, column rank and rank of matrices;
- cite an apply rank-related properties to systems of simultaneous linear equations, determinants and existence of inverse matrices;
- define and calculate the coordinates of a vector; and
- prove a selection of theorems and corollaries in the notes.

[4, p.197-255] contains an alternative presentation of material in this chapter that you may find helpful.

7.1 Introduction

In this chapter, we introduce important mathematical structures related to sets with a given operation. These are initially formulated in a general way, but we will soon focus on a limited number of examples. It is important to understand that the structures introduced are common to a variety of mathematical entities and that their abstract formulation is a powerful way of developing mathematical theory.

One can motivate the theory of vector spaces by looking at systems of linear equations from different viewpoints. First, we presented systems of equations in association with their geometrical interpretation. Namely,

$$\begin{cases} a_{11}x + a_{12}y + a_{13}z = b_1 \\ a_{21}x + a_{22}y + a_{23}z = b_2 \\ a_{31}x + a_{32}y + a_{33}z = b_3, \end{cases} \tag{7.1}$$

can be seen as the representation of the intersection of 3 planes in 3-dimensional space, Π_1, Π_2 and Π_3, with equations

$$\begin{aligned} \Pi_1 &: a_{11}x + a_{12}y + a_{13}z = b_1, \\ \Pi_2 &: a_{21}x + a_{22}y + a_{23}z = b_2, \\ \Pi_3 &: a_{31}x + a_{32}y + a_{33}z = b_3. \end{aligned}$$

A second viewpoint is obtained by writing the system of linear equations in (7.1) in matrix notation,

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b}, \tag{7.2}$$

with

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}, \ \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} \text{ and } \mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

by considering it as an equation within the set of matrices where the objective is to find the 3×1 matrix $\mathbf{x}$ (or length 3 vector) which satisfies this equation.

A third viewpoint is obtained by rewriting the system of linear equations in (7.1) as

$$x \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} + y \begin{pmatrix} a_{12} \\ a_{22} \\ a_{32} \end{pmatrix} + z \begin{pmatrix} a_{13} \\ a_{23} \\ a_{33} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}, \tag{7.3}$$

and considering it as the problem to identify the coefficients x, y, and z such that the combination of the *vectors* (or 3×1 matrices)

$$\begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix}, \begin{pmatrix} a_{12} \\ a_{22} \\ a_{32} \end{pmatrix} \text{ and } \begin{pmatrix} a_{13} \\ a_{23} \\ a_{33} \end{pmatrix},$$

yields the *vector*

$$\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}.$$

It is the last interpretation, which regards the matrix A as a function mapping vectors to vectors, requires the theory of vector spaces introduced in this chapter.

7.2 Definition of a vector space

Until now, we only discussed binary operations, i.e. operations involving two elements from within a given set. But we have already encountered operations which don't involve two elements from the same set, i.e. in expressions like $\lambda \cdot \mathbf{v}$, the "product" of a real number with a vector, which results in a vector. Another example is $\lambda \cdot \mathbf{A}$, the "product" of a real number with a matrix, resulting in a matrix. These are examples of a **scalar multiplication**. Scalar multiplication typically uses numbers from a field, in the previous examples $(\mathbb{R}, +, \cdot)$. Other fields can also be used, for instance $(\mathbb{C}, +, \cdot)$.

Let us now consider a set V with an internal binary operation $\oplus$ and a scalar multiplication with elements from a field $(F, +, \cdot)$. Then a **vector space** is defined a follows:

Definition 7.1. *A set non-empty V with a binary operation $\oplus$ and a scalar multiplication $\odot$ with elements from a field $(F, +, \cdot)$ is called a* ***vector space*** *over F when the following properties are satisfied.*

1. *$(V, \oplus)$ is an abelian group with identity $\boldsymbol{0}$;*
2. *V is closed under the scalar multiplication $\odot$:*

$$\forall\ \boldsymbol{a} \in V,\ \forall\ \lambda \in F :\ \lambda \odot \boldsymbol{a} \in V;$$

3. *Distributivity for scalar multiplication with respect to $\oplus$ in V:*

$$\forall\ \boldsymbol{a}, \boldsymbol{b} \in V,\ \forall\ \lambda \in F :\ \lambda \odot (\boldsymbol{a} \oplus \boldsymbol{b}) = (\lambda \odot \boldsymbol{a}) \oplus (\lambda \odot \boldsymbol{b});$$

4. *Distributivity for scalar multiplication with respect to $+$ in F:*

$$\forall\ \boldsymbol{a} \in V,\ \forall\ \lambda, \nu \in F :\ (\lambda + \nu) \odot \boldsymbol{a} = (\lambda \odot \boldsymbol{a}) \oplus (\nu \odot \boldsymbol{a});$$

5. *Mixed associativity for $\cdot$ and $\odot$:*

$$\forall\ \boldsymbol{a} \in V,\ \forall\ \lambda, \nu \in F :\ \lambda \odot (\nu \odot \boldsymbol{a}) = (\lambda \cdot \nu) \odot \boldsymbol{a};\ \textit{and}$$

6. *The identity of $(F \setminus \{0\}, \cdot)$ is also the identity for scalar multiplication:*

$$\forall\ \boldsymbol{a} \in V :\ 1 \odot \boldsymbol{a} = \boldsymbol{a}.$$

Note that the definition of a vector space involves four different operations: the internal binary operation ' $\oplus$ ', often referred to as an addition, the scalar multiplication ' $\odot$ ' and the two binary operations of the associated field, ' $+$ ' and ' $\cdot$ '. In the remainder of this course, the symbols $\oplus$ and $\odot$ will be replaced by the more usual $+$ and $\cdot$, with the multiplication symbol often omitted. But it should be clear from the context which of the four operations is being referred to.

One refers to vector spaces over the field of real numbers, $(\mathbb{R}, +, \cdot)$, as **real vector spaces**. Let us consider explicitly some examples of real vector spaces.

7.2.1 E^2: Vectors in the plane

First consider the set vectors in the plane, E^2, with the vector addition and scalar multiplication as defined in Chapter 5. We will use the familiar $+$ notation for the vector addition (instead of $\oplus$), and omit the multiplication sign ($\odot$) for the scalar multiplication. Then:

1. $(E^2, +)$ is an abelian group, with identity $\mathbf{0}$:
   (a) E^2 is closed under + since $\forall\ \mathbf{a}, \mathbf{b} \in E^2$, $\mathbf{a} + \mathbf{b} \in E^2$;
   (b) + is associative since $\forall\ \mathbf{a}, \mathbf{b}, \mathbf{c} \in E^2$, $(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c}) = \mathbf{a} + \mathbf{b} + \mathbf{c}$;
   (c) + has an identity since $\forall\ \mathbf{a} \in E^2$, $\mathbf{0} \in E^2$ and $\mathbf{0} + \mathbf{a} = \mathbf{a} + \mathbf{0} = \mathbf{a}$;
   (d) Existence of inverse. Since $\forall\ \mathbf{a} \in E^2$, $\exists\ \mathbf{b} \in E^2$ such that $\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a} = \mathbf{0}$;
   (e) + is commutative since $\forall\ \mathbf{a}, \mathbf{b} \in E^2$, $\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}$;
2. E^2 is closed under scalar multiplication since
$$\forall\ \mathbf{a} \in E^2,\ \forall\ \lambda \in \mathbb{R},\ \lambda\mathbf{a} \in E^2;$$
3. Distributivity for scalar multiplication with respect to addition in E^2,
$$\forall\ \mathbf{a}, \mathbf{b} \in E^2,\ \forall\ \lambda \in \mathbb{R},\ \lambda(\mathbf{a} + \mathbf{b}) = (\lambda\mathbf{a}) + (\lambda\mathbf{b});$$
4. Distributivity for scalar multiplication with respect to + in $\mathbb{R}$,
$$\forall\ \mathbf{a} \in E^2,\ \forall\ \lambda, \nu \in \mathbb{R},\ (\lambda + \nu)\mathbf{a} = (\lambda\mathbf{a}) + (\nu\mathbf{a});$$
5. Mixed associativity for multiplication in $\mathbb{R}$ and scalar multiplication,
$$\forall\ \mathbf{a} \in E^2,\ \forall\ \lambda, \nu \in \mathbb{R},\ \lambda(\nu\mathbf{a}) = (\lambda\nu)\mathbf{a};\ \text{and}$$
6. The identity of $(\mathbb{R} \setminus \{0\}, \cdot)$ is also the identity for scalar multiplication,
$$\forall\ \mathbf{a} \in E^2,\ 1\mathbf{a} = \mathbf{a}.$$

One can also verify that the set of vectors in three dimensions, E^3 forms a vector space over the field $(\mathbb{R}, +, \cdot)$. The term *vector space* derives from the fact that it describes the properties of the set of vectors with addition and scalar multiplication over the field of real numbers. In general, the elements of a vector space will be referred to as **vectors**.

7.2.2 $\mathbb{R}^3$: Points in 3-dimensional space

As a second example, consider the set of ordered triples of real numbers, $\mathbb{R}^3$. Addition is then defined as

$$(x_1,\ x_2,\ x_3) + (y_1,\ y_2,\ y_3) = (x_1 + y_1,\ x_2 + y_2,\ x_3 + y_3),$$

and scalar multiplication with a real number as

$$\lambda(x_1,\ x_2,\ x_3) = (\lambda x_1,\ \lambda x_2,\ \lambda x_3).$$

Then $\mathbb{R}^3$ is a real vector space (in full detail, we mean $(\mathbb{R}^3, +)$ with the field $(\mathbb{R}, +, \cdot)$) since:

1. $(\mathbb{R}^3, +)$ is an abelian group, with identity $(0,\ 0,\ 0)$. This follows from:
   (a) $\mathbb{R}^3$ is closed under + since $\forall\ (x_1,\ x_2,\ x_3),\ (y_1,\ y_2,\ y_3) \in \mathbb{R}^3$ we have
   $$(x_1,\ x_2,\ x_3) + (y_1,\ y_2,\ y_3) \in \mathbb{R}^3;$$
   (b) + is associative since $\forall\ (x_1,\ x_2,\ x_3),\ (y_1,\ y_2,\ y_3),\ (z_1,\ z_2,\ z_3) \in \mathbb{R}$ we have
   $$\begin{aligned}((x_1,\ x_2,\ x_3) + (y_1,\ y_2,\ y_3)) + (z_1,\ z_2,\ z_3) &= \\ (x_1,\ x_2,\ x_3) + ((y_1,\ y_2,\ y_3) + (z_1,\ z_2,\ z_3)) &= \\ (x_1,\ x_2,\ x_3) + (y_1,\ y_2,\ y_3) + (z_1,\ z_2,\ z_3);\end{aligned}$$
   (c) + has an identity since $\forall\ (x_1,\ x_2,\ x_3) \in \mathbb{R}^3,\ (0,\ 0,\ 0) \in \mathbb{R}^3$ we have
   $$(0,\ 0,\ 0) + (x_1,\ x_2,\ x_3) = (x_1,\ x_2,\ x_3) + (0,\ 0,\ 0) = (x_1,\ x_2,\ x_3);$$
   (d) Existence of inverse. Since $\forall\ (x_1,\ x_2,\ x_3) \in \mathbb{R}^3,\ \exists\ (y_1,\ y_2,\ y_3) \in \mathbb{R}^3$ such that
   $$(x_1,\ x_2,\ x_3) + (y_1,\ y_2,\ y_3) = (y_1,\ y_2,\ y_3) + (x_1,\ x_2,\ x_3) = (0,\ 0,\ 0);$$
   (e) + is commutative since $\forall\ (x_1,\ x_2,\ x_3),\ (y_1,\ y_2,\ y_3) \in \mathbb{R}^3$ we have
   $$(x_1,\ x_2,\ x_3) + (y_1,\ y_2,\ y_3) = (y_1,\ y_2,\ y_3) + (x_1,\ x_2,\ x_3);$$
2. $\mathbb{R}^3$ is closed under scalar multiplication since
   $$\forall\ (x_1,\ x_2,\ x_3) \in \mathbb{R}^3,\ \forall\ \lambda \in \mathbb{R},\ \text{we have } \lambda(x_1,\ x_2,\ x_3) \in \mathbb{R}^3;$$
3. Distributivity for scalar multiplication with respect to addition in $\mathbb{R}^3$. Since
   $$\forall\ (x_1,\ x_2,\ x_3),\ (y_1,\ y_2,\ y_3) \in \mathbb{R}^3,\ \forall\ \lambda \in \mathbb{R} \text{ we have}$$
   $$\lambda((x_1,\ x_2,\ x_3) + (y_1,\ y_2,\ y_3)) = (\lambda(x_1,\ x_2,\ x_3)) + (\lambda(y_1,\ y_2,\ y_3));$$
4. Distributivity for scalar multiplication with respect to + in $\mathbb{R}$. Since
   $$\forall\ (x_1,\ x_2,\ x_3) \in \mathbb{R}^3,\ \forall\ \lambda, \nu \in \mathbb{R} \text{ we have}$$
   $$(\lambda + \nu)(x_1,\ x_2,\ x_3) = (\lambda(x_1,\ x_2,\ x_3)) + (\nu(x_1,\ x_2,\ x_3));$$
5. Mixed associativity for multiplication in $\mathbb{R}$ and scalar multiplication. Since
   $\forall\ (x_1,\ x_2,\ x_3) \in \mathbb{R}^3,\ \forall\ \lambda, \nu \in \mathbb{R}$ we have $\lambda(\nu(x_1,\ x_2,\ x_3)) = (\lambda\nu)(x_1,\ x_2,\ x_3)$; and
6. The identity of $(\mathbb{R} \setminus \{0\}, \cdot)$ is also the identity for scalar multiplication. Since
   $$\forall\ (x_1,\ x_2,\ x_3) \in \mathbb{R}^3 \text{ we have } 1(x_1,\ x_2,\ x_3) = (x_1,\ x_2,\ x_3).$$

This result for $\mathbb{R}^3$ can easily be extended to the set of n-tuples ($n \in \mathbb{N}$), $\mathbb{R}^n$, with elements of the form $(x_1,\ x_2,\ \ldots,\ x_n)$.

Previously we used calculations in the vector space $\mathbb{R}^3$ to solve problems posed in terms of vectors in three dimensions, i.e. elements of the vector space E^3. One can do this because the two sets are both vector spaces with similar properties. The link between a vector, its components and an ordered triple is expressed in the following notation:

$$\mathbf{a} = \vec{OA} = x_1\mathbf{i} + x_2\mathbf{j} + x_3\mathbf{k} = (x_1,\ x_2,\ x_3).$$

7.3 Properties of vector spaces

We now have introduced an abstract mathematical structure, a vector space, which is characterised by a number of given properties. We have seen examples of vector spaces, and more will be introduced in the practice questions and in later years. We now establish general results which hold in any arbitrary vector space.

We have already seen that the identity of a binary operation is unique, hence

Lemma 7.2. *The zero vector (identity),* $\mathbf{0}$*, in a vector space is unique.*

Proof: From Definition 7.1 point 1, $\oplus$ is a binary operation with identity $\mathbf{0}$. Therefore, via Theorem 2.6, $\mathbf{0}$ is unique, as required.

Lemma 7.3. *In a real vector space,* V*, the inverse with respect to addition in* V*, of a vector* $\mathbf{v}$ *is unique.*

Proof: Assume that $\mathbf{v} \in V$ has two inverses (negatives), $\hat{\mathbf{v}}$ and $\tilde{\mathbf{v}}$. We have

$$(\hat{\mathbf{v}} + \mathbf{v}) + \tilde{\mathbf{v}} = \mathbf{0} + \tilde{\mathbf{v}} = \tilde{\mathbf{v}},$$

and

$$\hat{\mathbf{v}} + (\mathbf{v} + \tilde{\mathbf{v}}) = \hat{\mathbf{v}} + \mathbf{0} = \hat{\mathbf{v}}.$$

Hence as $+$ is associative,

$$\hat{\mathbf{v}} = (\hat{\mathbf{v}} + \mathbf{v}) + \tilde{\mathbf{v}} = \hat{\mathbf{v}} + (\mathbf{v} + \tilde{\mathbf{v}}) = \tilde{\mathbf{v}},$$

and hence, both inverses are identical. Since $\mathbf{v} \in V$ has an inverse with respect to vector addition, it follows that the inverse is unique, as required.

Lemma 7.4. *Consider* m *vectors in a vector space,* $\mathbf{v}_1,\ \mathbf{v}_2, \ldots,\ \mathbf{v}_m \in V$*, with* $m \geq 2$ *and* $\lambda \in \mathbb{F}$*, then*

$$\lambda(\mathbf{v}_1 + \mathbf{v}_2 + \ldots + \mathbf{v}_m) = \lambda\mathbf{v}_1 + \lambda\mathbf{v}_2 + \ldots + \lambda\mathbf{v}_m.$$

Proof: We use the principle of mathematical induction. Consider the statement $P(m)$ for $m \in \mathbb{N} \setminus \{1\}$ given by

$$\lambda(\mathbf{v}_1 + \mathbf{v}_2 + \ldots + \mathbf{v}_m) = \lambda\mathbf{v}_1 + \lambda\mathbf{v}_2 + \ldots + \lambda\mathbf{v}_m. \tag{7.4}$$

Note that

$$P(2) \text{ is true} \tag{7.5}$$

since V is a vector space (see Definition 7.1 point 3). Now assume that $P(q)$ is true for some $q \in \mathbb{N} \setminus \{1\}$ i.e.

$$\lambda(\mathbf{v}_1 + \mathbf{v}_2 + \ldots + \mathbf{v}_q) = \lambda\mathbf{v}_1 + \lambda\mathbf{v}_2 + \ldots + \lambda\mathbf{v}_q \tag{7.6}$$

is true. Then, via (7.5) and (7.6), respectively (and Definition 7.1 point 3)

$$\begin{aligned} \lambda(\mathbf{v}_1 + \mathbf{v}_2 + \ldots + \mathbf{v}_{q+1}) &= \lambda((\mathbf{v}_1 + \mathbf{v}_2 + \ldots + \mathbf{v}_q) + \mathbf{v}_{q+1}) \\ &= \lambda(\mathbf{v}_1 + \mathbf{v}_2 + \ldots + \mathbf{v}_q) + \lambda\mathbf{v}_{q+1} \\ &= \lambda\mathbf{v}_1 + \lambda\mathbf{v}_2 + \ldots + \lambda\mathbf{v}_q + \lambda\mathbf{v}_{q+1}. \end{aligned} \tag{7.7}$$

So, via (7.6) and (7.7),

$$P(q) \text{ is true} \Longrightarrow P(q+1) \text{ is true} \tag{7.8}$$

for any $q \in \mathbb{N} \setminus \{1\}$. Finally, via (7.5) and (7.8), from the principle of mathematical induction (see Theorem B.2), $P(m)$ given by (7.4) is true for all $m \in \mathbb{N} \setminus \{1\}$, as required.

Similarly,

Lemma 7.5. *Consider a vector $\mathbf{v} \in V$, with V a vector space, and m real numbers, $\lambda_1, \lambda_2, \ldots, \lambda_m \in \mathbb{F}$, with $m \geq 2$. Then,*

$$(\lambda_1 + \lambda_2 + \ldots + \lambda_m)\mathbf{v} = \lambda_1\mathbf{v} + \lambda_2\mathbf{v} + \ldots + \lambda_m\mathbf{v}.$$

Proof: This can again be proved by mathematical induction. Consider the statement $P(m)$ for $m \in \mathbb{N} \setminus \{1\}$, given by

$$(\lambda_1 + \lambda_2 + \ldots + \lambda_m)\mathbf{v} = \lambda_1\mathbf{v} + \lambda_2\mathbf{v} + \ldots + \lambda_m\mathbf{v}. \tag{7.9}$$

Note that

$$P(2) \text{ is true} \tag{7.10}$$

since V is a vector space (via Definition 7.1 point 4). Now assume that P(q) is true for some $q \in \mathbb{N} \setminus \{1\}$ i.e.

$$(\lambda_1 + \lambda_2 + \ldots + \lambda_q)\mathbf{v} = \lambda_1\mathbf{v} + \lambda_2\mathbf{v} + \ldots + \lambda_q\mathbf{v}. \tag{7.11}$$

is true. Then, via (7.10) and (7.11), respectively (and Definition 7.1 point 4)

$$\begin{aligned} (\lambda_1 + \lambda_2 + \ldots + \lambda_{q+1})\mathbf{v} &= ((\lambda_1 + \lambda_2 + \ldots + \lambda_q) + \lambda_{q+1})\mathbf{v} \\ &= (\lambda_1 + \lambda_2 + \ldots + \lambda_q)\mathbf{v} + \lambda_{q+1}\mathbf{v} \\ &= \lambda_1\mathbf{v} + \lambda_2\mathbf{v} + \ldots + \lambda_q\mathbf{v} + \lambda_{q+1}\mathbf{v}. \end{aligned} \tag{7.12}$$

So, via (7.11) and (7.12),

$$P(q) \text{ is true} \Longrightarrow P(q+1) \text{ is true} \tag{7.13}$$

for any $q \in \mathbb{N} \setminus \{1\}$. Finally, via (7.10) and (7.13), from the principle of mathematical induction (Theorem B.2), $P(m)$ given by (7.9) is true for all $m \in \mathbb{N} \setminus \{1\}$, as required.

Theorem 7.6. *Assume that V is a vector space $\mathbb{F}$.*

1. *If $\mathbf{v} \in V$ and $\mathbf{v} = \mathbf{v} + \mathbf{v}$, then $\mathbf{v} = \mathbf{0}$.*
2. *If $\mathbf{v} \in V$, then $0\mathbf{v} = \mathbf{0}$;*
3. *If $\lambda \in \mathbb{F}$, then $\lambda\mathbf{0} = \mathbf{0}$;*
4. *If $\lambda \in \mathbb{F}$ and $\mathbf{v} \in V$, then $\lambda\mathbf{v} = \mathbf{0}$ if and only if either $\lambda = 0$ or $\mathbf{v} = \mathbf{0}$; and*
5. *If $\lambda \in \mathbb{F}$ and $\mathbf{v} \in V$ then $(-\lambda)\mathbf{v} = \lambda(-\mathbf{v}) = -(\lambda\mathbf{v})$. Here $(-\mathbf{v})$ denotes the additive inverse element of $\mathbf{v}$.*

Proof:

1. Since every element in V has an inverse under vector addition, there exists an additive inverse $(-\mathbf{v}) \in V$ of $\mathbf{v}$. Suppose that $\mathbf{v} = \mathbf{v} + \mathbf{v}$. Then

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

   This proves (1).

2. For any $\mathbf{v} \in V$, we have $0\mathbf{v} = (0 + 0)\mathbf{v} = 0\mathbf{v} + 0\mathbf{v}$. Hence $0\mathbf{v} = \mathbf{0}$ by (1).

3. We have

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

   So the result follows from (1).

4. Suppose that $\lambda\mathbf{v} = \mathbf{0}$ and $\lambda \neq 0$. Then, there exists an inverse $\nu \in \mathbb{R}$ of λ (with respect to multiplication in the scalar field) such that $\lambda\nu = 1$. Hence,

$$\nu(\lambda\mathbf{v}) \quad = \quad (\nu\lambda)\mathbf{v} = 1\mathbf{v} = \mathbf{v}.$$

   Since $\lambda\mathbf{v} = \mathbf{0}$, the left hand side of the equation is the zero vector (via point 3). Therefore, $\nu(\lambda\mathbf{v}) = \mathbf{0}$ which implies that $\mathbf{v} = \mathbf{0}$.

   Alternatively, supposing that $\lambda\mathbf{v} = \mathbf{0}$ and $\lambda = 0$ satisfies the conclusion. The if and only if result is completed by points 2 and 3 above.

5. Let $\lambda \in \mathbb{R}$. Note here that $-\mathbf{v} \in V$ is the additive inverse of $\mathbf{v}$. We have

$$(-\lambda)\mathbf{v} + \lambda\mathbf{v} = (-\lambda + \lambda)\mathbf{v} = 0\mathbf{v} = \mathbf{0}$$

   by (2) and so $(-\lambda)\mathbf{v} = -(\lambda\mathbf{v})$.

   Similarly,

$$\lambda(-\mathbf{v}) + \lambda\mathbf{v} = \lambda(\mathbf{v} + (-\mathbf{v})) = \lambda\mathbf{0} = \mathbf{0}$$

by 3. So $\lambda(-\mathbf{v}) = -(\lambda\mathbf{v})$. It follows that

$$(-\lambda)\mathbf{v} = \lambda(-\mathbf{v}) = -(\lambda\mathbf{v}),$$

as required. □

Corollary 7.7. *Let V be a vector space over $\mathbb{F}$ and $\mathbf{v} \in V$. Then $(-1)\mathbf{v} = -\mathbf{v}$.*

Proof: Follows from Theorem 7.6 point 5 with $\lambda = 1$.

Observe that since

(i) vector spaces over $\mathbf{F}$ are closed under addition,

(ii) there is an identity for addition, i.e. $\mathbf{0}$,

(iii) every element $\mathbf{v}$ of a vector space has an inverse $-\mathbf{v}$,

we can define a new binary operation, called **subtraction** as

Definition 7.8. *A subtraction is a binary operation on the real vector space V defined as follow:*

$$\mathbf{u} - \mathbf{v} = \mathbf{u} + (-\mathbf{v}) \quad \forall \mathbf{u}, \mathbf{v} \in V$$

with $(-\mathbf{v})$ denoting the inverse of $\mathbf{v}$.

Example 100: Simplify the following expressions:

1. $5\mathbf{v} - 7\mathbf{w} + 3\mathbf{v} - 9\mathbf{v} + \mathbf{w}$,
2. $2(2,4,5) - 6(1,3,5) + (4,8,10) + 5(1,3,5) - 4(2,4,5)$. ⌟

7.4 Subspaces of vector spaces

Consider the vector space $\mathbb{R}^3$ of all ordered triples of real numbers. This can be visualised as the set of all coordinates of points in a three-dimensional space, or the components of three-dimensional (position) vectors. Consider the subset

$$U = \{(x_1,\ x_1,\ x_1) :\ x_1 \in \mathbb{R}\}.$$

Then $U \subset \mathbb{R}^3$. Also, we can verify that $(U, +)$ is an abelian group:

1. U is closed under +: $\forall\ (x_1,\ x_1,\ x_1),\ (x_2,\ x_2,\ x_2) \in U$:

$$(x_1,\ x_1,\ x_1) + (x_2,\ x_2,\ x_2) \in U;$$

2. + is associative: $\forall\ (x_1,\ x_1,\ x_1),\ (x_2,\ x_2,\ x_2),\ (x_3,\ x_3,\ x_3) \in \mathbb{R}^3$:

$$\begin{aligned}((x_1,\ x_1,\ x_1) + (x_2,\ x_2,\ x_2)) + (x_3,\ x_3,\ x_3) &= \\ (x_1,\ x_1,\ x_1) + ((x_2,\ x_2,\ x_2) + (x_3,\ x_3,\ x_3)) &= \\ (x_1,\ x_1,\ x_1) + (x_2,\ x_2,\ x_2) + (x_3,\ x_3,\ x_3);&\end{aligned}$$

3. + has an identity: $\forall\ (x_1,\ x_1,\ x_1) \in U,\ (0,\ 0,\ 0) \in U$:

$$(0,\ 0,\ 0) + (x_1,\ x_1,\ x_1) = (x_1,\ x_1,\ x_1) + (0,\ 0,\ 0) = (x_1,\ x_1,\ x_1);$$

4. Existence of inverse: $\forall\ (x_1,\ x_1,\ x_1) \in U,\ \exists\ (x_2,\ x_2,\ x_2) \in U$:

$$(x_1,\ x_1,\ x_1) + (x_2,\ x_2,\ x_2) = (x_2,\ x_2,\ x_2) + (x_1,\ x_1,\ x_1) = (0,\ 0,\ 0);$$

5. + is commutative: $\forall\ (x_1,\ x_1,\ x_1),\ (x_2,\ x_2,\ x_2) \in U$:

$$(x_1,\ x_1,\ x_1) + (x_2,\ x_2,\ x_2) = (x_2,\ x_2,\ x_2) + (x_1,\ x_1,\ x_1).$$

Note that in the five requirements for $(U, +)$ to be an abelian group, associativity and commutativity are obviously satisfied since these properties are satisfied by any set of elements of $\mathbb{R}^3$ operated on by the binary operation +.

The fact that U is closed under addition can be derived from the definition of the sum in $\mathbb{R}^3$:

$$(x_1,\ x_1,\ x_1) + (x_2,\ x_2,\ x_2) = (x_1 + x_2, x_1 + x_2, x_1 + x_2),$$

and since $x_1 + x_2 \in \mathbb{R}$, the sum is an element of U.

Also, the identity $(0, 0, 0) \in U$, by choosing $x_1 = 0 \in \mathbb{R}$, and since $(0, 0, 0)$ is the identity in $\mathbb{R}^3$ it will also be the identity in U. So the only thing that needed to be checked here is that the identity is an element of the subset U.

Every element in U has an inverse in $(\mathbb{R}^3, +)$, but we need to establish that that inverse is also an element of U. We find that

$$-(x_1,\ x_1,\ x_1) = (-x_1,\ -x_1,\ -x_1) \in U.$$

So the inverse of an element in U is also an element in U.

If we consider the properties of a vector space established in the previous section, then all but one will hold in U since they hold for all real numbers and all sets of vectors in

$\mathbb{R}^3$. The only one that needs to be checked, is whether or not the set U is closed under scalar multiplication:

$$\forall\ \lambda \in \mathbb{R}, \forall\ (x_1,\ x_1,\ x_1) \in U:\ \lambda(x_1,\ x_1,\ x_1) \in U.$$

This can be verified using the definition of scalar multiplication in $\mathbb{R}^3$,

$$\lambda(x_1,\ x_1,\ x_1) = (\lambda x_1,\ \lambda x_1,\ \lambda x_1),$$

where $\lambda x_1 \in \mathbb{R}$.

So we see that U satisfies all conditions of a real vector space with binary operation $+$ in $\mathbb{R}^3$, and hence, is a vector space in its own right. We will call U a **subspace** of $\mathbb{R}^3$. Geometrically, all points of the form $(x_1,\ x_1,\ x_1)$ with $x_1 \in \mathbb{R}$ constitute a line through the origin and in the direction given by the vector with components $(1, 1, 1)$.

Let us now extend the idea of a subspace from this specific example.

Definition 7.9. *A subset U of a vector space V over $\mathbb{F}$ is called a **subspace** of V if U is non-empty and U is itself a vector space over $\mathbb{F}$ under the same operations of addition and scalar multiplication of V as a vector space over $\mathbb{F}$. We write $U \leq V$ to indicate that U is a subspace of V and not just a subset of V.*

Example 101: The simplest subspaces that satisfy Definition 7.9 are:

(i) $\{\mathbf{0}\}$ is a subspace of V;

(ii) V is a subspace of V. ⌟

These two examples are known as **improper** subspaces. All other subspaces of V are known as **proper** subspaces, if they exist. How do we recognise a subspace?

Theorem 7.10. *A subset U of a vector space V over $\mathbb{F}$ is a subspace of V if and only if U is non-empty and*

1. *$\mathbf{u} + \mathbf{v} \in U$ for all $\mathbf{u}, \mathbf{v} \in U$; and*
2. *$\lambda\mathbf{u} \in U$ for all $\lambda \in \mathbb{F},\ \mathbf{u} \in U$.*

Equivalently, U is a subspace of V if and only if U is closed under addition and scalar multiplication i.e. $\lambda\mathbf{u} + \mu\mathbf{v} \in U$ for all $\mathbf{u}, \mathbf{v} \in U$ and $\lambda, \mu \in \mathbb{F}$.

Proof. Most conditions for a vector space apply to any set of vectors in V and any real numbers and hence automatically apply to the elements of the subset U. Aside from the two conditions in the theorem statement above, the only conditions that are not automatically satisfied are:

1. $\mathbf{0} \in U$,
2. If $\mathbf{u} \in U$ then $-\mathbf{u} \in U$ for all $\mathbf{u} \in U$.

Both of these follow from condition 2 and Theorem 7.6. If one chooses an element $\mathbf{u} \in U$ with $\lambda = 0 \in \mathbb{R}$, then $0\mathbf{u} = \mathbf{0} \in U$. Similarly, since $-\mathbf{u} = (-1)\mathbf{u}$, for any $\mathbf{u} \in U$, choosing $\lambda = -1$ establishes that $-\mathbf{u} = (-1)\mathbf{u} \in U$. Hence, all conditions of a vector space are satisfied if the subset is closed under both addition and scalar multiplication, as required. □

When determining if a subset of a vector space, is a vector space, it is useful to quickly check whether or not the identity $\mathbf{0}$ is in the subset. If the additive identity is not an element of the subset, then the subset is not a subspace. If the additive identity is an element of the subset, then we can check whether the subset is closed under addition and scalar multiplication to conclude that the set is a subspace.

Lemma 7.11. *A subset U of a vector space V over $\mathbb{F}$ is a subspace of V if and only if U is non-empty and $\mathbf{u} + \lambda\mathbf{v} \in U$ for all $\mathbf{u}, \mathbf{v} \in U$ and $\lambda \in \mathbb{F}$.*

Proof. It is easy to see that if U is a subspace, then U is non-empty and $\mathbf{u} + \lambda\mathbf{v} \in U$.

Suppose that $\mathbf{u} + \lambda\mathbf{v} \in U$ for all $\mathbf{u}, \mathbf{v} \in U$ and $\lambda \in \mathbb{F}$. Then taking $\lambda = 1$, we have

$$\mathbf{u} + \mathbf{v} \in U$$

for all $\mathbf{u}$ and $\mathbf{v} \in U$. Similarly, taking $\mathbf{u} = \mathbf{0}$ we have

$$\lambda\mathbf{v} \in U$$

for all $\mathbf{v} \in V$ and $\lambda \in \mathbb{F}$. Hence U is a subspace by Theorem 7.10. □

Example 102: Check whether the following subsets U are subspaces of the real vector spaces V:

1. $V = \mathbb{R}^4$; $U = \{(x_1,\ x_2,\ x_3,\ x_4) \in \mathbb{R}^4 |\ x_1, x_2, x_3, x_4 \in \mathbb{Z}\ \}$;
2. $V = \mathbb{R}^5$; $U = \{(x_1,\ x_2,\ x_3,\ x_4,\ x_5) \in \mathbb{R}^5 |\ x_1 + 2x_2 + 3x_3 + 2x_4 + x_5 = 1\}$; and
3. $V = \mathbb{R}^3$; $U = \{(x_1,\ x_2,\ x_3) \in \mathbb{R}^3 |\ 3x_1 - x_2 - 2x_3 = 0\}$. ⌟

Definition 7.12. *Consider two subspaces U and W of a vector space V over $\mathbb{F}$. The **intersection** $U \cap W$, of U and W is defined by*

$$U \cap W := \{\mathbf{u} \in V | \ \mathbf{u} \in U \text{ and } \mathbf{u} \in W\}.$$

*The **sum** $U + W$, of U and W is defined by*

$$U + W := \{\mathbf{v} \in V | \ \mathbf{v} = \mathbf{u} + \mathbf{w}, \text{ with } \mathbf{u} \in U \text{ and } \mathbf{w} \in W\}.$$

Theorem 7.13. *Consider a vector space V over $\mathbb{F}$. Let U and W be subspaces of V. Then $U \cap W$ and $U + W$ are both subspaces of V.*

Proof: See practice questions.

7.5 Spanning set

We have seen how any vector in E^3 can be written in the form $\mathbf{v} = a\mathbf{i} + b\mathbf{j} + c\mathbf{k}$. This can be extended to general vector spaces.

Definition 7.14. *A vector $\mathbf{v} \in V$, with V a vector space over $\mathbb{F}$, is a **linear combination** of the vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k \in V$ if $\mathbf{v}$ can be written in the form*

$$\mathbf{v} = \lambda_1\mathbf{u}_1 + \lambda_2\mathbf{u}_2 + \ldots + \lambda_k\mathbf{u}_k = \sum_{i=1}^{k} \lambda_i\mathbf{v}_i$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k \in \mathbb{F}$.

Notice that λ_i can be zero and there may be more than one choice for $\lambda_1, \lambda_2, \ldots, \lambda_k$.

Example 103: Show that the vector $(0, 1, -2) \in \mathbb{R}^3$ is a linear combination of the vectors $(-2, 1, 0), (1, 3, 4)$ and $(3, -2, 1)$.

All possible linear combinations of a given set of vectors can be used to define a subset of the vector space. We call this subset the **span**:

Definition 7.15. *If $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k \in V$, with V a vector space over $\mathbb{F}$, then the subset of V consisting of all possible linear combinations of $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k$ is called their **span** and is denoted by*

$$\text{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\} = \{\lambda_1\mathbf{u}_1 + \lambda_2\mathbf{u}_2 + \ldots + \lambda_k\mathbf{u}_k \mid \lambda_1, \lambda_2, \ldots, \lambda_k \in \mathbb{F}\}.$$

*If $U = \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$, then we say that $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ is a **spanning set** for U or that $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ **spans** U.*

We will also use the notation

$$\langle u_1, \ldots, u_k \rangle$$

to denote the span of the set $\{u_1, \ldots, u_k\} \subseteq V$. Also if X is a set of vectors, we will write

$$\langle X \rangle$$

for the set of vectors spanned by X.

Example 104: In $\mathbb{R}^3$, determine the span of $\{(1,0,1),(0,0,1)\}$.

What properties does such a spanning set have?

Theorem 7.16. *Suppose that $U = \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$, where for a vector space V over $\mathbb{F}$, we have $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k \in V$. Then,*

1. *$\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k \in U$,*
2. *U is a subspace of V, and*
3. *U is the **smallest** subspace of V that contains $\mathbf{u}_1, \mathbf{u}_2 \ldots, \mathbf{u}_k$ in the sense that if $\tilde{U}$ is another subspace of V which contains $\mathbf{u}_1, \mathbf{u}_2 \ldots, \mathbf{u}_k$, then $U \subseteq \tilde{U}$.*

Proof:

1. $\mathbf{u}_i \in U$ for all $i = 1, 2, \ldots, k$ since

$$\mathbf{u}_i = 0\mathbf{u}_1 + 0\mathbf{u}_2 + \ldots + 1\mathbf{u}_i + \ldots + 0\mathbf{u}_k.$$

2. We first show that the span is closed under addition. Consider $\mathbf{v}_1, \mathbf{v}_2 \in U$, with

$$\begin{aligned} \mathbf{v}_1 &= \lambda_1\mathbf{u}_1 + \lambda_2\mathbf{u}_2 + \ldots + \lambda_k\mathbf{u}_k, \\ \mathbf{v}_2 &= \nu_1\mathbf{u}_1 + \nu_2\mathbf{u}_2 + \ldots + \nu_k\mathbf{u}_k. \end{aligned}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k, \nu_1, \nu_2, \ldots, \nu_k \in \mathbb{F}$. Then,

$$\begin{aligned} \mathbf{v}_1 + \mathbf{v}_2 &= (\lambda_1\mathbf{u}_1 + \lambda_2\mathbf{u}_2 + \ldots + \lambda_k\mathbf{u}_k) + (\nu_1\mathbf{u}_1 + \nu_2\mathbf{u}_2 + \ldots + \nu_k\mathbf{u}_k) \\ &= \lambda_1\mathbf{u}_1 + \lambda_2\mathbf{u}_2 + \ldots + \lambda_k\mathbf{u}_k + \nu_1\mathbf{u}_1 + \nu_2\mathbf{u}_2 + \ldots + \nu_k\mathbf{u}_k \\ &= (\lambda_1 + \nu_1)\,\mathbf{u}_1 + (\lambda_2 + \nu_2)\,\mathbf{u}_2 + \ldots + (\lambda_k + \nu_k)\,\mathbf{u}_k \\ &\in U. \end{aligned} \tag{7.14}$$

since $\lambda_1 + \nu_1,\ \lambda_2 + \nu_2, \ldots,\ \lambda_k + \nu_k \in \mathbb{R}$. Similarly, for $\alpha \in \mathbb{F}$,

$$\begin{aligned} \alpha\mathbf{v}_1 &= \alpha\left(\lambda_1\mathbf{u}_1 + \lambda_2\mathbf{u}_2 + \ldots + \lambda_k\mathbf{u}_k\right) \\ &= \alpha\lambda_1\mathbf{u}_1 + \alpha\lambda_2\mathbf{u}_2 + \ldots + \alpha\lambda_k\mathbf{u}_k \end{aligned}$$

$$\begin{aligned} &= (\alpha\lambda_1)\,\mathbf{u}_1 + (\alpha\lambda_2)\,\mathbf{u}_2 + \ldots + (\alpha\lambda_k)\,\mathbf{u}_k \\ &\in U, \end{aligned} \tag{7.15}$$

since $\alpha\lambda_1,\ \alpha\lambda_2, \ldots,\ \alpha\lambda_k \in \mathbb{F}$. Hence (7.14) and (7.15) show that U is closed under addition and scalar multiplication, and therefore, Theorem 7.10 yields U is a subspace of V.

3. Assume $\tilde{U}$ is a subspace of V which contains the vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k$. Then for any $\lambda_1, \lambda_2, \ldots, \lambda_k \in \mathbb{R}$, we have

$$\lambda_1\mathbf{u}_1, \lambda_2\mathbf{u}_2, \ldots, \lambda_k\mathbf{u}_k \in \tilde{U},$$

since as a subspace, $\tilde{U}$ is closed under scalar multiplication. Additionally, because $\tilde{U}$ is also closed under addition,

$$\lambda_1\mathbf{u}_1 + \lambda_2\mathbf{u}_2 + \ldots + \lambda_k\mathbf{u}_k \in \tilde{U},$$

for all $\lambda_1,\ \lambda_2, \ldots,\ \lambda_k \in \mathbb{F}$. Hence, all elements in $\operatorname{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\} = U$ are also in $\tilde{U}$ and hence $U \subseteq \tilde{U}$.

This completes the proof, as required.

In fact, a very nice way to define the span of a set of vectors $X \subseteq V$ is as the intersection of all subspaces of V which contain X. That is

$$\langle X \rangle = \bigcap_{X \subseteq U \leq V} U.$$

With this definition Theorem 7.13 implies that $\langle X \rangle$ is a subspace. Notice that this way of defining the span does not require the notion of linear combinations. It's a good definition for proofs, but not for actual calculations.

Example 105: In $\mathbb{R}^4$, determine $\operatorname{span}\{(1,0,0,0), (0,1,0,0)\}$ and compare this with the subspace $\tilde{U} \subset \mathbb{R}^4$, with

$$\tilde{U} = \{(x_1, x_2, x_3, x_4) \in \mathbb{R}^4 \mid x_4 = 0\}.$$

Now, how can we determine a spanning set for a given subspace $U \subset V$? First, we will need to find a set of vectors such that every vector in U can be written as a linear combination of these vectors, i.e.,

$$U \subseteq \operatorname{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}.$$

If we can choose each of these vectors so that they also belong to U, i.e. $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k \in U$, then

$$\operatorname{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\} \subseteq U,$$

and hence

$$U = \operatorname{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}.$$

So to find a spanning set of U, we need to find a set of vectors in U such that every vector in U can be written as a linear combination of elements of the set.

Example 106: Find a spanning set for the subspace of $\mathbb{R}^3$ given by

$$U = \{(x_1, x_2, x_3) \in \mathbb{R}^3 \mid 3x_1 - x_2 - 2x_3 = 0\}$$

of $\mathbb{R}^3$.

7.6 Linear (In)dependence

Consider the vectors $\mathbf{a} = 2\mathbf{i} + \mathbf{j}$, $\mathbf{b} = 4\mathbf{i} + 2\mathbf{j}$ and $\mathbf{c} = 2\mathbf{i}$, then one can easily verify that

$$2\mathbf{a} - \mathbf{b} = \mathbf{0}.$$

However there is only one linear combination of $\mathbf{a}$ and $\mathbf{c}$ that equals $\mathbf{0}$:

$$\begin{aligned} \alpha\mathbf{a} + \beta\mathbf{c} &= 2\alpha\mathbf{i} + \alpha\mathbf{j} + 2\beta\mathbf{i} \\ &= 2(\alpha + \beta)\mathbf{i} + \alpha\mathbf{j}, \end{aligned}$$

which equals $\mathbf{0}$ if and only if $\beta = \alpha = 0$. The following definition addresses this observation:

Definition 7.17. *A set of vectors $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ in a vector space V over $\mathbb{F}$ is said to be **linearly independent** if and only if for $\lambda_1, \ldots, \lambda_k \in \mathbb{F}$,*

$$\lambda_1\mathbf{u}_1 + \lambda_2\mathbf{u}_2 + \ldots + \lambda_k\mathbf{u}_k = \mathbf{0} \iff \lambda_1 = \lambda_2 = \ldots = \lambda_k = 0.$$

This means that $\{\mathbf{u}\}$ with $\mathbf{u} \neq \mathbf{0}$ is linearly independent. The linear combination

$$0\mathbf{u}_1 + 0\mathbf{u}_2 + \ldots + 0\mathbf{u}_k$$

is called the **trivial linear combination**. To complement Definition 7.17, we have:

Definition 7.18. *A set of vectors $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ in a vector space V is **linearly dependent** if there is a non-trivial linear combination which is equal to the zero vector, i.e.*

$$\lambda_1\mathbf{u}_1 + \lambda_2\mathbf{u}_2 + \ldots + \lambda_k\mathbf{u}_k = \mathbf{0} \not\Longrightarrow \lambda_i = 0 \text{ for all } i = 1, \ldots, k.$$

According to Definition 7.18, the set $\{\mathbf{0}\}$ in V is a linearly dependent set.

Example 107: Show that the set $S = \{(0,0,1),(0,1,1),(1,1,1)\}$ in $\mathbb{R}^3$ is linearly independent.

Example 108: Show that the set $S = \{(2,-3,0,1),(-1,0,0,1),(0,-2,0,2)\}$ of vectors in $\mathbb{R}^4$ is linearly dependent.

Example 109: Show that if one of the vectors in the set $S = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ of k vectors in a vector space V is the zero vector $\mathbf{0}$ of V, then S is a linearly dependent set.

Before we consider the related theory in full generality, it is worth considering sets that contain 2 vectors (see practice questions).

Theorem 7.19. *Consider a vector space V over $\mathbb{F}$ and assume that $V \neq \{\mathbf{0}\}$. Then any set of finitely many non-zero vectors in V which spans V contains a linearly independent subset which spans V.*

Proof: Consider a set of k non-zero vectors which spans V,

$$T = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}.$$

Choose a subset S of T with as few elements as possible so that S spans V. That is $V = \langle S \rangle$. If S is a linearly independent set, then the conclusion is satisfied. Hence we assume that S is linearly dependent and aim to find a contradiction. Plainly if S consists of one non-zero vector, then S is linearly independent and is S is not empty as otherwise $V = \{\mathbf{0}\}$. Hence $k \geq 2$. Since S is linearly dependent, there exists $\lambda_1, \ldots, \lambda_k \in \mathbb{F}$ not all zero, such that

$$\lambda_1 \mathbf{u}_1 + \cdots + \lambda_k \mathbf{u}_k = 0.$$

Pick $\ell \in \{1, \ldots, k\}$ such that $\lambda_\ell \neq 0$. Then

$$-\lambda_\ell \mathbf{u}_\ell = \lambda_1 \mathbf{u}_1 + \cdots + \lambda_{\ell-1}\mathbf{u}_{\ell-1} + 0\mathbf{u}_\ell + \lambda_{\ell+1}\mathbf{u}_{\ell+1} + \cdots + \lambda_k \mathbf{u}_k.$$

multiplying by $-\lambda_\ell^{-1}$ yields

$$\mathbf{u}_\ell = -\lambda_\ell^{-1}\lambda_1 \mathbf{u}_1 + \ldots \lambda_\ell^{-1}\lambda_{\ell-1}\mathbf{u}_{\ell-1} + 0\mathbf{u}_\ell + \lambda_\ell^{-1}\lambda_{\ell+1}\mathbf{u}_{\ell+1} + \cdots - \lambda_\ell^{-1}\lambda_k \mathbf{u}_k.$$

To simplify the expression let's write $\mu_j = -\lambda_\ell^{-1}\lambda_{\ell-1} \in \mathbb{F}$ for $1 \leq j \leq k$ and $j \neq \ell$. So

$$\mathbf{u}_\ell = \mu_1 \mathbf{u}_1 + \ldots \mu_{\ell-1}\mathbf{u}_{\ell-1} + 0\mathbf{u}_\ell + \mu_{\ell+1}\mathbf{u}_{\ell+1} + \ldots + \mu_k \mathbf{u}_k.$$

Let $\mathbf{v} \in V$. Then, as S spans V, there exists $\alpha_1, \alpha_2, \ldots \alpha_k \in \mathbb{F}$ such that

$$\mathbf{v} = \alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2 + \ldots + \alpha_{\ell-1}\mathbf{u}_{\ell-1} + \alpha_\ell \mathbf{u}_\ell + \ldots + \alpha_k \mathbf{u}_k$$

$$
\begin{aligned}
&= \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \ldots + \alpha_{\ell-1}\mathbf{u}_{\ell-1} + \\
&\quad \alpha_\ell(\mu_1\mathbf{u}_1 + \mu_2\mathbf{u}_2 + \ldots + \mu_{\ell-1}\mathbf{u}_{\ell-1}) + \ldots + \alpha_k\mathbf{u}_k \\
&= (\alpha_1 + \alpha_\ell\mu_1)\mathbf{u}_1 + (\alpha_2 + \alpha_\ell\mu_2)\mathbf{u}_2 + \ldots + \\
&\quad (\alpha_{\ell-1} + \alpha_\ell\mu_{\ell-1})\mathbf{u}_{\ell-1} + \alpha_{\ell+1}\mathbf{u}_{\ell+1} + \ldots + \alpha_k\mathbf{u}_k.
\end{aligned}
$$

Hence, $S \setminus \{\mathbf{u}_\ell\}$ spans V and this contradicts the minimal choice of S. If this set of $k-1$ vectors is linearly independent, the property is satisfied. This proves the theorem.

Note that, to remove the word 'finitely' in Theorem 7.19, we require a definition for linear combinations of infinitely many vectors to extend Definition 7.14 (which will be considered in courses later in your degree programme). We can now formulate the **fundamental theorem**:

Theorem 7.20. *Let S be a set of k vectors in a vector space V which spans V. Let T be a linearly independent set of m vectors in V. Then,*

$$m \leq k.$$

Alternatively, Theorem 7.20 states that the number of vectors in a linearly independent set in a vector space V cannot exceed the number of vectors in a spanning set of V.

Proof: Assume that $k < m$ (to obtain a contradiction). We denote the spanning set S to be,

$$S = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\},$$

and the linearly independent set T as

$$T = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_m\}.$$

Therefore

$$\lambda_1\mathbf{t}_1 + \lambda_2\mathbf{t}_2 + \ldots + \lambda_m\mathbf{t}_m = \mathbf{0} \iff \lambda_i = 0 \text{ for } i = 1, \ldots, m. \tag{7.16}$$

Theorem 7.19 states that if the vectors in S are linearly dependent, then there is a subset S_1 of S with $\bar{k}$ elements ($\bar{k} \leq k$) which is also a spanning set and contains linearly independent vectors. If we can prove that the assumption $k < m$ is false for the smaller spanning set S_1, then it also is false for the larger spanning set S.

Since S_1 is a spanning set, every vector in T can be written as a linear combination of the vectors in S_1, i.e.

$$\mathbf{t}_i = \alpha_{1i}\mathbf{u}_1 + \alpha_{2i}\mathbf{u}_2 + \ldots + \alpha_{\bar{k}i}\mathbf{u}_{\bar{k}}, \tag{7.17}$$

for $i = 1, 2, \ldots, m$. Substitution of $\mathbf{t}$ in (7.17) into (7.16) yields

$$
\begin{aligned}
&\lambda_1(\alpha_{11}\mathbf{u}_1 + \alpha_{21}\mathbf{u}_2 + \ldots + \alpha_{\bar{k}1}\mathbf{u}_{\bar{k}}) + \\
&\lambda_2(\alpha_{12}\mathbf{u}_1 + \alpha_{22}\mathbf{u}_2 + \ldots + \alpha_{\bar{k}2}\mathbf{u}_{\bar{k}}) +
\end{aligned}
$$

$$\vdots$$
$$\lambda_m(\alpha_{1m}\mathbf{u}_1 + \alpha_{2m}\mathbf{u}_2 + \ldots + \alpha_{\bar{k}m}\mathbf{u}_{\bar{k}}) = \mathbf{0} \tag{7.18}$$

$$\begin{aligned} \iff \quad & (\lambda_1\alpha_{11} + \lambda_2\alpha_{12} + \ldots + \lambda_m\alpha_{1m})\mathbf{u}_1 + \\ & (\lambda_1\alpha_{21} + \lambda_2\alpha_{22} + \ldots + \lambda_m\alpha_{2m})\mathbf{u}_2 + \\ & \quad \vdots \\ & (\lambda_1\alpha_{\bar{k}1} + \lambda_2\alpha_{\bar{k}2} + \ldots + \lambda_m\alpha_{\bar{k}m})\mathbf{u}_{\bar{k}} = \mathbf{0}. \end{aligned} \tag{7.19}$$

Since the spanning set S_1 is linearly independent, each of the coefficients in the sum on the LHS of (7.19) has to be zero, yielding the following system of $\bar{k}$ equations in m unknowns:

$$\begin{cases} \alpha_{11}\lambda_1 + \alpha_{12}\lambda_2 + \ldots + \alpha_{1m}\lambda_m = 0 \\ \alpha_{21}\lambda_1 + \alpha_{22}\lambda_2 + \ldots + \alpha_{2m}\lambda_m = 0 \\ \quad \vdots \qquad\quad \vdots \qquad\qquad\quad \vdots \quad \vdots \\ \alpha_{\bar{k}1}\lambda_1 + \alpha_{\bar{k}2}\lambda_2 + \ldots + \alpha_{\bar{k}m}\lambda_m = 0 \end{cases} \tag{7.20}$$

Since, by the assumption $\bar{k} \leq k < m$, there are fewer equations than unknowns in (7.20), so we can add equations of our choice to generate a system of m equations in m unknowns. By doing this, we can always make the determinant of the coefficient matrix in (7.20) equal to zero (for example, by using an already existing equation, hence generating two identical rows in the matrix of coefficients). Such a system has a non-trivial solution via Corollary 6.37, and hence, we can solve (7.18) with at least one $\lambda_i \neq 0$.

Therefore, via (7.17)-(7.20), we have shown that there exists a linear combination

$$\lambda_1\mathbf{t}_1 + \lambda_2\mathbf{t}_2 + \ldots + \lambda_m\mathbf{t}_m = \mathbf{0},$$

which has at least one $\lambda_i \neq 0$. This contradicts the condition that T is linearly independent. Finally, we conclude that the assumption $m > k$ is false, and hence $m \leq k$, as required.

7.7 Basis

The ideas of a spanning set and linear independence can be combined to define *a basis*, a very important notion in the study of vector spaces.

Definition 7.21. *A set of vectors* $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ *in a vector space* V *is called a* ***basis*** *of* V *if*

(i) $V = \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$*; and*

(ii) $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ *is a linearly independent set.*

Definition 7.22. *A vector space V is called **finite-dimensional** if there exists a basis of V which contains a finite number of vectors. A vector space which is not finite-dimensional is called **infinite-dimensional**.*

Example 110: $\{(1,0),(0,1)\}$ is a basis of $\mathbb{R}^2$.

Example 111: $\{\mathbf{i},\mathbf{j},\mathbf{k}\}$ is a basis of E^3.

We primarily consider finite-dimensional vector spaces in this module[1].

Theorem 7.23. *Any two bases of a finite-dimensional vector space V contain the same number of vectors.*

Proof: Consider two bases[2], S and T of V, with m and k number of vectors respectively. Both S and T are spanning sets as well as linearly independent. Since S is a spanning set and T a set of k linearly independent vectors, the fundamental theorem (Theorem 7.20) states that $k \leq m$. Similarly, since T is a spanning set and S a set of m linearly independent vectors, again, the fundamental theorem states that $m \leq k$. Hence, $k = m$, as required.

Definition 7.24. *The number of vectors in any basis of a finite-dimensional vector space V is called the **dimension** of V, also written as* $\dim(V)$.

These definitions require clarification in one special case. The subset $\{\mathbf{0}\}$ of any vector space forms a subspace, but has no basis since $\{\mathbf{0}\}$ is not linearly independent. For completeness, we say that the dimension of $\{\mathbf{0}\}$ is zero.

A proper subspace U of a finite-dimensional vector space V will be also finite-dimensional.

Example 112: Show that the set of vectors $\{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n\}$ with

$$\begin{aligned}\mathbf{e}_1 &= (1,0,0,\ldots,0),\\ \mathbf{e}_2 &= (0,1,0,\ldots,0),\\ &\vdots\\ \mathbf{e}_n &= (0,0,0,\ldots,1),\end{aligned}$$

is a basis of $\mathbb{R}^n$.

The basis of $\mathbb{R}^n$ introduced in the previous example is referred to as the **standard ordered basis** of $\mathbb{R}^n$, since the vectors are listed in a particular order.

Example 113: Find a basis of the subspace

$$U = \{(x_1,x_2,x_3) \in \mathbb{R}^3 |\ 3x_1 - x_2 - 2x_3 = 0\},$$

of $\mathbb{R}^3$.

[1] Infinite dimensional vector spaces appear in [15] and the practice questions.

[2] "bases" is the plural of basis.

7.8 More properties of vector space bases

Theorem 7.25. *Consider a vector space V and a basis $B = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ of V. Then any vector $\mathbf{v} \in V$ can be written as a linear combination of the vectors of B in exactly one way.*

Proof: Since $\mathbf{v} \in V$, $\mathbf{v}$ can be written as a linear combination of the basis vectors in at least one way thus we have

$$\mathbf{v} = \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \ldots + \alpha_n\mathbf{u}_n$$

where $\alpha_i \in \mathbb{F}$ for $1 \leq i \leq n$. Suppose that there exist $\beta_1, \ldots, \beta_n$ such that

$$\mathbf{v} = \beta_1\mathbf{u}_1 + \beta_2\mathbf{u}_2 + \ldots + \beta_n\mathbf{u}_n.$$

Then subtracting one from the other we obtain

$$(\alpha_1 - \beta_1)\mathbf{u}_1 + (\alpha_2 - \beta_2)\mathbf{u}_2 + \ldots + (\alpha_n - \beta_n)\mathbf{u}_n = \mathbf{0}.$$

Since the vectors in a basis B are linearly independent, it follows that $\alpha_i = \beta_i$, for all values of i. Hence, $\mathbf{v}$ can be written as a linear combination of the basis vectors in exactly one way.

Definition 7.26. *Consider the vector space V with an ordered basis $B = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ and a vector $\mathbf{v} \in V$ with*

$$\mathbf{v} = \lambda_1\mathbf{u}_1 + \lambda_2\mathbf{u}_2 + \ldots + \lambda_n\mathbf{u}_n,$$

*where $\lambda_i \in \mathbb{R}$ for all values of i. Then the uniquely determined real numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$ (in the given order) are known as the **coordinates** of $\mathbf{v}$ with respect to the basis B. We call $(\lambda_1, \lambda_2, \ldots, \lambda_n)$ the coordinate vector of $\mathbf{v}$ with respect to the basis B.*

Example 114:

(i) Show that $B = \{(1,0,1), (0,0,1), (0,2,0)\}$ is an ordered basis of $\mathbb{R}^3$.

(ii) Determine the coordinates of the vector $(2,1,3) \in \mathbb{R}^3$ with respect to the basis B. ⌟

Theorem 7.27. *Let V be a vector space of dimension $n \in \mathbb{N}$ and assume that S is a set of vectors in V. Then:*

(i) if $|S| > n$, then S is linearly dependent; and

(ii) If $|S| < n$, then $V \neq \text{span}(S)$.

Proof: (i) Assume $|S| = k > n$. If S is linearly independent, then $n < k \leq n$ by Theorem 7.20 which is impossible. Hence S is linearly dependent.

(ii) Assume $|S| = k < n$ and S spans V. Then Theorem 7.19 implies that S contains a basis for V. Hence$n \leq k < n$ which is nonsense.

Lemma 7.28. *Consider a linearly independent set of k vectors in a vector space V, $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$, and a vector $\mathbf{v} \in V$. Then the set*

$$\{\mathbf{v}, \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$$

is linearly independent if and only if

$$\mathbf{v} \notin \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}.$$

Proof: Suppose that $S = \{\mathbf{v}, \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ is linearly independent. If $\mathbf{v} \in \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$, Then, by definition we can write

$$\lambda_1\mathbf{u}_1 + \lambda_2\mathbf{u}_2 + \ldots + \lambda_k\mathbf{u}_k = \mathbf{v}$$

and this implies

$$\mathbf{v} - \lambda_1\mathbf{u}_1 - \lambda_2\mathbf{u}_2 - \ldots - \lambda_k\mathbf{u}_k = \mathbf{0}$$

which means that S is linearly dependent and contradicts our supposition. So, if $S = \{\mathbf{v}, \mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ is linearly independent, then

$$\mathbf{v} \notin \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}.$$

Conversely, suppose that $\mathbf{v} \notin \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ and that $S = \{\mathbf{v}, \mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is linearly dependent. Then there are $\lambda_i, \lambda \in \mathbb{F}$ not all zero such that

$$\lambda\mathbf{v} + \lambda_1\mathbf{u}_1 + \lambda_2\mathbf{u}_2 + \ldots + \lambda_k\mathbf{u}_k = \mathbf{0}.$$

If $\lambda = 0$, then $\lambda_i = 0$ for $1 \leq i \leq k$ as S is linearly independent. So we may suppose that $\lambda \neq 0$. Then,

$$\begin{aligned} & \mathbf{v} + \frac{\lambda_1}{\lambda}\mathbf{u}_1 + \frac{\lambda_2}{\lambda}\mathbf{u}_2 + \ldots + \frac{\lambda_k}{\lambda}\mathbf{u}_k = \mathbf{0} \\ \iff \quad & \mathbf{v} = -\frac{\lambda_1}{\lambda}\mathbf{u}_1 - \frac{\lambda_2}{\lambda}\mathbf{u}_2 - \ldots - \frac{\lambda_k}{\lambda}\mathbf{u}_k \in U. \end{aligned}$$

This contradicts the supposition that $\mathbf{v} \notin \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$. Hence if $\mathbf{v} \notin \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$, then S is linearly independent.

Theorem 7.29. *Let V be a vector space over $\mathbb{F}$ and assume that $V \neq 0$ is spanned by ℓ vectors ($\ell \in \mathbb{N}$). Then:*

(i) each linearly independent set of vectors is part of a basis of V;

(ii) if $X \subseteq V$ spans V, then there exists $B \subseteq X$ such that B is a basis for V; and

(iii) V has a basis and $\dim(V) \leq \ell$.

Proof: Since V is spanned by ℓ elements, we have $\dim V \leq \ell$ by Theorem 7.19 and by Theorem 7.27 any linearly independent subset of V has at most ℓ elements.

(i) Suppose that there exists a linearly independent subset of V which does not extend to a basis. From amongst all possibilities choose $S = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ with a maximal number of vectors. Then S is linearly independent and $k \leq \ell$. If $V = \langle S \rangle$, then$S$ is a basis, a contraction as we chose it not to be part of a basis. Therefore S does not span V, choose any vector $\mathbf{u}_{k+1} \in V$ such that $\mathbf{u}_{k+1} \notin \text{span}(S)$. Then the set $S \cup \{\mathbf{u}_{k+1}\}$ is linearly independent by Lemma 7.28. The maximal choice of S implies $S \cup \{\mathbf{u}_{k+1}\}$ is part of a basis. But then so is S, a contradiction. We conclude that such an S could not be chosen. Hence (i) holds.

(ii) Consider a set $S = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ of k vectors which spans V. Then Theorem 7.19 states that S contains a subset which also spans V but is linearly independent. This subset is a basis of V.

(iii) From (ii), we conclude that as V is spanned by ℓ vectors, $\dim(V) \leq \ell$.

Example 115: Find a basis of the subspace $U \subset \mathbb{R}^3$ which is spanned by $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ with

$$\mathbf{u}_1 = (0, 1, 1),\ \mathbf{u}_2 = (2, 4, 4) \text{ and } \mathbf{u}_3 = (1, 1, 1).$$

Theorem 7.30. *Let V be a vector space V of dimension $n \geq 1$ and let $S = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ be a set of n vectors in V.*

(i) If S spans V, then S is a basis for V.

(ii) If S is linearly independent, then S is a basis for V.

Proof: (i) Assume S spans V. Then by Theorem 7.29(ii) there is a subset of S which is a basis of V. Since every basis of V contains exactly n vectors, this subset must be S itself.

(ii) Assume S is a linearly independent set of vectors in V. Then Theorem 7.29(i) implies $S \subseteq B$ where B is a basis for V. Hence $n = |S| \leq |B| = n$ and so $S = B$.

So, if we know the dimension n of a vector space, then we do not need to check both conditions for a basis in Definition 7.21 in the case of a set of n vectors. It is usually easier to show that a set of n vectors is linearly independent than it is to show that it is a spanning set.

Example 116: Verify that the set

$$U = \{(1,2,1,3),\ (2,5,3,9)\} \subset \mathbb{R}^4$$

is linearly independent and extend it to a basis of $\mathbb{R}^4$.

Theorem 7.31. *Consider a real vector space V of dimension $n \in \mathbb{N}$. Let U and W be subspaces of V. Then,*

(i) U and W are finite-dimensional with $\dim(U) \leq n$ and $\dim(W) \leq n$;

(ii) any basis of U and any basis of W can be extended to a basis of V; and

(iii) if $U \subseteq W$ and $\dim(U) = \dim(W)$ then $U = W$.

Proof:

(i) Let S be a linearly independent subset of U. Then Theorem 7.29(i) implies $|S| \leq n$. Hence $\dim(U) \leq n$. The same is true for W.

(ii) Since any basis of U is a linearly independent set in V, it is part of a basis of V via Theorem 7.29. The same justification establishes the result for W.

(iii) Suppose $U \subseteq W$ and $\dim(U) = \dim(W)$. Let B be basis of U. Then $|B| = \dim(U) = \dim(W)$. Hence B is a basis for W. Therefore

$$U = \langle B \rangle = W.$$

7.9 Vector spaces arising from linear ODEs*

Consider the 2nd order ordinary differential equation (ODE)

$$u'' + u = 0 \quad \text{on } \mathbb{R} \tag{7.21}$$

for $u : \mathbb{R} \to \mathbb{R}$. Notably, there are 2 (there are many more) distinct solutions $\mathbf{u}_1, \mathbf{u}_2 : \mathbb{R} \to \mathbb{R}$ to the ODE in (7.21) given by

$$\mathbf{u}_1(x) = \sin(x) \text{ and } \mathbf{u}_2(x) = \cos(x) \quad \forall x \in \mathbb{R}. \tag{7.22}$$

Now, consider the set

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

Also consider the binary operation $\oplus$ given by

$$(g \oplus f)(x) = g(x) + f(x) \quad \forall x \in \mathbb{R},\ f, g \in V$$

and scalar multiplication $\odot$ with elements in the field $\mathbb{R}, +, \cdot$ defined as

$$(\lambda \odot f)(x) = \lambda f(x) \quad \forall x \in \mathbb{R},\ \lambda \in \mathbb{R},\ f \in V.$$

Then via results considered in [15], $(V, \oplus)$ is a real vector space (of infinite dimension). Since $\mathbf{u}_1, \mathbf{u}_2 \in V$, it follows from (7.22) that

$$\begin{aligned} U &= \text{span}\{\mathbf{u}_1, \mathbf{u}_2\} \\ &= \{f \in V :\ f = \lambda_1 \mathbf{u}_1 + \lambda_2 \mathbf{u}_2,\ \lambda_1, \lambda_2 \in \mathbb{R}\} \\ &= \{f \in V :\ f(x) = \lambda_1 \sin(x) + \lambda_2 \cos(x),\ \lambda_1, \lambda_2 \in \mathbb{R},\ \forall x \in \mathbb{R}\}. \end{aligned}$$

is a 2-dimensional subspace of V. This follows from Theorem 7.10 since U is closed under the scalar multiplcation and vector space addition operations in V, and Definition 7.24 since

$$\lambda_1 \sin(x) + \lambda_2 \cos(x) = 0 \quad \forall x \in \mathbb{R} \implies \begin{cases} \lambda_1 \sin(0) + \lambda_2 \cos(0) = 0 \\ \lambda_1 \sin\left(\frac{\pi}{2}\right) + \lambda_2 \cos\left(\frac{\pi}{2}\right) = 0 \end{cases} \implies \begin{cases} \lambda_2 = 0 \\ \lambda_1 = 0. \end{cases}$$

i.e. $\mathbf{u}_1$ and $\mathbf{u}_2$ are linearly independent elements of U, and hence, form a basis of U.

Now, since the ODE in (7.21) is linear, any linear combination of solutions to the ODE is also a solution to the ODE. Thus, U is contained in the general solution set for the ODE in (7.21). In fact, U is the general solution set for the ODE in (7.21) since U is a solution set that is 2-dimensional and the ODE is of order 2 with constant coefficients.

Now, if we wish to solve initial value problems for the ODE in (7.21), for example, find u that satisfies (7.21) and

$$\begin{cases} u(0) = a \\ u'(0) = b, \end{cases}$$

for some $a, b \in \mathbb{R}$, then we would solve

$$\begin{cases} \lambda_1 \sin(0) + \lambda_2 \cos(0) = a \\ \lambda_1 \cos(0) - \lambda_2 \sin(0) = b \end{cases} \iff \begin{pmatrix} \sin(0) & \cos(0) \\ \cos(0) & -\sin(0) \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix} \iff \begin{cases} \lambda_1 = b \\ \lambda_2 = a. \end{cases} \tag{7.23}$$

Using this idea, we can see that solutions to boundary value problems for the ODE in (7.21) can be found by simply finding a suitable element of the subspace U (which involves solving a system of simultaneous linear equations). In general, for 'nice' linear ODE of order n on $\mathbb{R}$, there exist n linearly independent solutions (in V) which can be used as

basis vectors that span the solution set to the ODE. One can then use these bases to solve boundary value problems associated with ODEs.

As an exercise, find the solution to the ODE in (7.21) such that u satisfies

$$\begin{cases} u(x^*) = a \\ u'(x^*) = b, \end{cases}$$

for fixed $x^* \in \mathbb{R}$ (hint - consider the inverse of a 2×2 matrix like that in (7.23)).

7.10 The dimension of a sum of subspaces

Theorem 7.32. *Let V be a vector space and U and W be finite dimensional subspaces of V. Then,*

$$\dim(U + W) = \dim(U) + \dim(W) - \dim(U \cap W).$$

Proof: Since $U \cap W$ is a subspace of V, consider a basis $\{\mathbf{i}_1, \ldots, \mathbf{i}_q\}$ of $U \cap W$, where q is the dimension of $U \cap W$. Then $\{\mathbf{i}_1, \ldots, \mathbf{i}_q\}$ is a set of linearly independent vectors in U, so we can extend it to obtain a basis of U, $\{\mathbf{i}_1, \ldots, \mathbf{i}_q, \mathbf{u}_1, \ldots, \mathbf{u}_p\}$, where $p + q$ is the dimension of U. Similarly, $\{\mathbf{i}_1, \ldots, \mathbf{i}_q\}$ is a set of linearly independent vectors in W, so we can extend it to obtain a basis of W, $\{\mathbf{i}_1, \ldots, \mathbf{i}_q, \mathbf{w}_1, \ldots, \mathbf{w}_r\}$, where $r + q$ is the dimension of W. Set

$$B = \{\mathbf{i}_1, \ldots, \mathbf{i}_q, \mathbf{u}_1, \ldots, \mathbf{u}_p, \mathbf{w}_1, \ldots, \mathbf{w}_r\}.$$

Then

$$|B| = p + q + r = \dim(U) + \dim(W) - \dim(U \cap W).$$

Hence to prove the theorem we show that B is a basis for $U + W$.

Now, by definition, any vector in $U + W$ can be written as $\mathbf{u} + \mathbf{w}$, with $\mathbf{u} \in U$ and $\mathbf{w} \in W$. Each of these two vectors can uniquely be written as a linear combination of the vectors in the basis of U and W respectively:

$$\begin{aligned} \mathbf{u} &= \alpha_1 \mathbf{i}_1 + \cdots + \alpha_q \mathbf{i}_q + \alpha_{q+1} \mathbf{u}_1 + \cdots + \alpha_{q+p} \mathbf{u}_p, \\ \mathbf{w} &= \beta_1 \mathbf{i}_1 + \cdots + \beta_q \mathbf{i}_q + \beta_{q+1} \mathbf{w}_1 + \cdots + \beta_{q+r} \mathbf{w}_r, \end{aligned} \tag{7.24}$$

with $\alpha_i, \beta_j \in \mathbb{R}$ for $1 \le i \le q + p$ and $1 \le j \le q + r$. Therefore,

$$\begin{aligned} \mathbf{u} + \mathbf{w} &= \alpha_1 \mathbf{i}_1 + \cdots + \alpha_q \mathbf{i}_q + \alpha_{q+1} \mathbf{u}_1 + \cdots + \alpha_{q+p} \mathbf{u}_p + \\ &\quad \beta_1 \mathbf{i}_1 + \cdots + \beta_q \mathbf{i}_q + \beta_{q+1} \mathbf{w}_1 + \cdots + \beta_{q+r} \mathbf{w}_r \\ &= \gamma_1 \mathbf{i}_1 + \cdots + \gamma_q \mathbf{i}_q + \alpha_{q+1} \mathbf{u}_1 + \cdots + \alpha_{q+p} \mathbf{u}_p + \beta_{q+1} \mathbf{w}_1 + \cdots + \beta_{q+r} \mathbf{w}_r, \end{aligned} \tag{7.25}$$

with $\gamma_j = \alpha_j + \beta_j$ for $i = 1, 2, \ldots, q$. Hence B spans $U + W$.

To establish that B is linearly independent, consider a linear combination of the vectors in B which equals the zero vector:

$$\mathbf{0} = \alpha_1\mathbf{i}_1 + \cdots + \alpha_q\mathbf{i}_q + \gamma_1\mathbf{u}_1 + \cdots + \gamma_p\mathbf{u}_p + \beta_1\mathbf{w}_1 + \cdots + \beta_r\mathbf{w}_r. \tag{7.26}$$

Equation (7.26) can be rewritten as

$$-\gamma_1\mathbf{u}_1 - \cdots - \gamma_p\mathbf{u}_p = \alpha_1\mathbf{i}_1 + \cdots + \alpha_q\mathbf{i}_q + \beta_1\mathbf{w}_1 + \cdots + \beta_r\mathbf{w}_r. \tag{7.27}$$

The vector on the left-hand side of (7.27) is an element of the subspace U and the vector on the right-hand side of (7.27) is a vector in the subspace W, hence, this vector must be in the intersection of U and W and can therefore be written as a unique linear combination of the basis vectors in $U \cap W$. So, for some $\mathbf{d} \in U \cap W$, we have,

$$\begin{aligned} \mathbf{d} &= -\gamma_1\mathbf{u}_1 - \cdots - \gamma_p\mathbf{u}_p, \\ \mathbf{d} &= \tau_1\mathbf{i}_1 + \cdots + \tau_q\mathbf{i}_q. \end{aligned}$$

However, $\{\mathbf{i}_1, \ldots, \mathbf{i}_q, \mathbf{u}_1, \ldots, \mathbf{u}_p\}$ is a basis for U and $\mathbf{d}$ has two expressions as a linear combination of elements in this basis. This contradicts Theorem 7.25 unless

$$\gamma_1 = \cdots = \gamma_p = \tau_1 = \cdots = \tau_q = 0.$$

Hence

$$\begin{aligned} \mathbf{0} &= \alpha_1\mathbf{i}_1 + \cdots + \alpha_q\mathbf{i}_q + \gamma_1\mathbf{u}_1 + \cdots + \gamma_p\mathbf{u}_p + \beta_1\mathbf{w}_1 + \cdots + \beta_r\mathbf{w}_r \\ &= \alpha_1\mathbf{i}_1 + \cdots + \alpha_q\mathbf{i}_q + \beta_1\mathbf{w}_1 + \cdots + \beta_r\mathbf{w}_r \end{aligned}$$

Since $\{\mathbf{i}_1, \ldots, \mathbf{i}_q, \mathbf{w}_1, \ldots, \mathbf{w}_r\}$, is linearly independent we have $\beta_1 = \cdots = \beta_r = 0$ and we have demonstrated that B is linearly independent. This proves the claim.

7.11 Row space, column space, row rank and column rank of a matrix

Consider a matrix $\mathbf{A} \in \mathcal{M}_{mn}(\mathbb{F})$. Each row is an ordered set of n numbers, and hence can be considered as a vector in the vector space $\mathbb{F}^n$. Hence, the matrix $\mathbf{A}$ defines an ordered set of m vectors in $\mathbb{F}^n$. We can then consider the subspace of $\mathbb{F}^n$ which is spanned by this set of vectors.

Definition 7.33. *Consider a matrix $\boldsymbol{A} \in \mathcal{M}_{mn}(\mathbb{F})$ and let $\mathbf{u}_i$ denote the vector in $\mathbb{F}^n$ associated with the i-th row in $\boldsymbol{A}$. Then the **row space** of $\boldsymbol{A}$, denoted by* $\mathrm{row}(\boldsymbol{A})$ *is the subspace of $\mathbb{F}^n$ given by*

$$\mathrm{row}(\boldsymbol{A}) = \mathrm{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m\}.$$

Similarly, one can consider the n columns in $\mathbf{A}$ as vectors in $\mathbb{F}^m$ and we can define the column space of a matrix:

Definition 7.34. *Consider a matrix $\boldsymbol{A} \in \mathcal{M}_{mn}(\mathbb{F})$ and let $\mathbf{u}_i$ denote the vector in $\mathbb{F}^m$ associated with the i-th column in $\boldsymbol{A}$. Then the **column space** of $\boldsymbol{A}$, denoted by $\text{col}(\boldsymbol{A})$ is the subspace of $\mathbb{F}^m$ given by*

$$\text{col}(\boldsymbol{A}) = \text{span}\{\mathbf{u}_1^T, \mathbf{u}_2^T, \ldots, \mathbf{u}_n^T\}.$$

Example 117: Determine the row space and column space of

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & -3 & -1 \\ -2 & -6 & 5 & -2 \\ 3 & -5 & -1 & 2 \end{pmatrix}.$$

The dimensions of the row space and column space are of particular importance. Therefore we introduce:

Definition 7.35. *Consider $\boldsymbol{A} \in \mathcal{M}_{mn}(\mathbb{F})$. The **row rank** of $\boldsymbol{A}$ is given by*

$$\dim(\text{row}(\boldsymbol{A})),$$

*and the **column rank** of $\boldsymbol{A}$ is given by*

$$\dim(\text{col}(\boldsymbol{A})).$$

Theorem 7.36. *Consider matrices $\boldsymbol{A} \in \mathcal{M}_{mn}(\mathbb{F})$, $\boldsymbol{M} \in \mathcal{M}_{mm}(\mathbb{F})$ and $\boldsymbol{N} \in \mathcal{M}_{nn}(\mathbb{F})$. Then,*

(i) $\text{row}(\boldsymbol{M} \cdot \boldsymbol{A}) \subseteq \text{row}(\boldsymbol{A})$ *where* $\text{row}(\boldsymbol{M} \cdot \boldsymbol{A}) = \text{row}(\boldsymbol{A})$ *if $\boldsymbol{M}$ is invertible; and*

(ii) $\text{col}(\boldsymbol{A} \cdot \boldsymbol{N}) \subseteq \text{col}(\boldsymbol{A})$ *where* $\text{col}(\boldsymbol{A} \cdot \boldsymbol{N}) = \text{col}(\boldsymbol{A})$ *if $\boldsymbol{N}$ is invertible.*

Proof:

(i) Definition 4.13 implies that each row of $\mathbf{M} \cdot \mathbf{A}$ is a linear combination of rows of $\mathbf{A}$. Thus, $\text{row}(\mathbf{M} \cdot \mathbf{A}) \subseteq \text{row}(\mathbf{A})$. If $\mathbf{M}^{-1}$ exists, then

$$\text{row}(\mathbf{A}) = \text{row}(\mathbf{M}^{-1} \cdot \mathbf{M} \cdot \mathbf{A}) \subseteq \text{row}(\mathbf{M} \cdot \mathbf{A}) \subseteq \text{row}(\mathbf{A}),$$

which implies that $\text{row}(\mathbf{M} \cdot \mathbf{A}) = \text{row}(\mathbf{A})$.

(ii) We apply (i) and use the transpose.

$$\begin{aligned} \text{col}(\mathbf{A}\cdot\mathbf{N}) &= \text{row}((\mathbf{A}\cdot\mathbf{N})^T) = \text{row}(\mathbf{N}^T\cdot\mathbf{A}^T) \\ &\subseteq \text{row}(\mathbf{A}^T) = \text{col}(\mathbf{A}) \end{aligned}$$

and if $\mathbf{N}$ is invertible, then so is $\mathbf{N}^T$ and we obtain equality.

Note that in the proof of Theorem 7.36, we used that in the matrix $\mathbf{A}\cdot\mathbf{B}$, each row is a linear combination of the rows of $\mathbf{B}$ and each column in $\mathbf{A}\cdot\mathbf{B}$ is a linear combination of the columns of $\mathbf{A}$.

Corollary 7.37. *Consider matrices $\boldsymbol{A}\in\mathcal{M}_{mn}(\mathbb{F})$, $\boldsymbol{M}\in\mathcal{M}_{mm}(\mathbb{F})$ and $\boldsymbol{N}\in\mathcal{M}_{nn}(\mathbb{F})$, with $\boldsymbol{M}$ and $\boldsymbol{N}$ invertible matrices. Then*

(i) row rank of $\boldsymbol{M}\cdot\boldsymbol{A}$ = the row rank of $\boldsymbol{A}$; and

(ii) column rank of $\boldsymbol{A}\cdot\boldsymbol{N}$ = the column rank of $\boldsymbol{A}$.

Proof: This follows from the conclusion of Theorem 7.36, namely that

$$\text{row}(\mathbf{M}\cdot\mathbf{A}) = \text{row}(\mathbf{A}) \text{ and } \text{col}(\mathbf{A}\cdot\mathbf{N}) = \text{col}(\mathbf{A})$$

when $\mathbf{M}$ and $\mathbf{N}$ are invertible, respectively.

Lemma 7.38. *Suppose that $\mathbf{P}\in\mathcal{M}_{mm}(\mathbb{F})$ is invertible and $\mathbf{A}\in\mathcal{M}_{m,n}(\mathbb{F})$. Then $\mathbf{A}$ and $\mathbf{PA}$ have the same column rank.*

Proof: We will come back to this after we have seen linear transformations.

Lemma 7.39. *Let $\boldsymbol{R}\in\mathcal{M}_{mn}(\mathbb{F})$. Assume that $\mathbf{R}$ is in echelon form. Then the row rank of $\mathbf{R}$ equals the column rank of $\mathbf{R}$ is equal to the number of non-zero rows in $\mathbf{R}$.*

Proof: Consider the echelon matrix $\mathbf{R}$. It will contain a number of non-zero rows and possibly a number of zero rows. The non-zero rows are of a specific form, e.g.,

$$\begin{pmatrix} 1 & a & b & c & \dots & d \\ 0 & 1 & e & f & \dots & g \\ 0 & 0 & 0 & 1 & \dots & h \end{pmatrix}.$$

The first row cannot be written as a linear combination of the lower rows, since none of these has a non-zero entry in the first column. Similarly, the second row cannot be

written as a linear combination of the others below. Hence, the non-zero rows in the echelon matrix are linearly independent. Therefore, the non-zero rows in the echelon matrix span the row space of $\mathbf{R}$ and form a linearly independent set, hence they form a basis of row$(\mathbf{R})$. The dimension of these row spaces is given by the number of vectors in a basis, hence if there are r non-zero rows in the matrix in echelon form, the row rank of $\mathbf{R}$ = row rank of $\mathbf{A}$ = r.

It is also straight forward to see the columns corresponding to the positions were the leading ones sit in $\mathbf{R}$ form a basis for the column space of $\mathbf{R}$. Hence the column rank of $\mathbf{R}$ is also r.

Theorem 7.40. *Consider the matrix $\boldsymbol{A} \in \mathcal{M}_{mn}(\mathbb{F})$ and let $\mathbf{R}$ be an echelon form of $\mathbf{A}$ obtained by elementary row operations. Then* row$(\mathbf{A})$ = row$(\mathbf{R})$ *and the rows of $\mathbf{R}$ are a basis for* row$(\mathbf{A})$.

Proof: Any elementary row operation on the matrix $\mathbf{A}$ is equivalent to the multiplication (on the left) with an elementary matrix. So, if $\mathbf{R}$ is matrix $\mathbf{A}$ in echelon form, obtained after performing a sequence of elementary row operations, we can write

$$\mathbf{R} = \mathbf{E}_q \cdot \mathbf{E}_{q-1} \cdot \ldots \cdot \mathbf{E}_1 \cdot \mathbf{A} = \mathbf{P} \cdot \mathbf{A},$$

where $\mathbf{P} = \mathbf{E}_q \cdot \mathbf{E}_{q-1} \cdot \ldots \cdot \mathbf{E}_1$. Because every elementary matrix is invertible, $\mathbf{P}$ is invertible and hence using Theorem 7.36

$$\text{row}(\mathbf{R}) = \text{row}(\mathbf{P} \cdot \mathbf{A}) = \text{row}(\mathbf{A}).$$

Now use Lemma 7.39 to finish the proof.

Theorem 7.40 can be used to verify whether or not a set of vectors in a vector space $\mathbb{F}^n$ is linearly independent. Consider the set of vectors $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$. If we construct a matrix $\mathbf{A}$ such that row i is given by the elements in the vector $\mathbf{u}_i$, then

$$\text{row}(\mathbf{A}) = \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}.$$

By reducing the matrix $\mathbf{A}$ to echelon form, we can determine the row rank of $\mathbf{A}$. If the row rank of $\mathbf{A}$, m, is equal to k, then the k vectors span a vector space of dimension k and hence have to be linearly independent. If, on the other hand, $m < k$, then the k vectors span a vector space with dimension less then k and hence are linearly dependent.

Corollary 7.41. *Let $\mathbf{A} \in \mathcal{M}_{nn}(\mathbb{F})$. Then $\mathbf{A}$ is invertible if and only if $\mathbf{A}$ has row rank n.*

Proof: This follows as $\mathbf{A}$ is invertible if and only if the reduced echelon form of $\mathbf{A}$ is the identity matrix $\mathbf{I}_n$ (see Section 5.5).

Example 118: Verify that the set of vectors in $\mathbb{F}^3$,

$$U = \{(0,0,1),(0,1,1),(1,1,1)\}$$

is linearly independent.

Example 119: Investigate whether or not the set of vectors $\{\mathbf{u}_1,\ \mathbf{u}_2,\ \mathbf{u}_3\}$ in $\mathbb{F}^4$ is linearly independent, where

$$\mathbf{u}_1 = (1,-3,0,2),\ \mathbf{u}_2 = (0,-3,0,3),\ \mathbf{u}_3 = (0,-2,0,2).$$

Theorem 7.42. *For any matrix* $\boldsymbol{A} \in \mathcal{M}_{mn}(\mathbb{F})$,

row rank of $\boldsymbol{A}$ = *column rank of* $\boldsymbol{A}$.

Proof: Let $\mathbf{R}$ be an echelon matrix obtained from $\mathbf{A}$. Then $\mathbf{R} = \mathbf{PA}$ where $\mathbf{P}$ is a product of elementary matrices and is invertible. By Lemma 7.38 and Theorem 7.40, A has the same row rank and column rank as $\mathbf{R}$. By Lemma 7.39, the row rank of $\mathbf{R}$ equals the column rank of $\mathbf{R}$. This proves the result.

Definition 7.43. *For any matrix* $\boldsymbol{A} \in \mathcal{M}_{mn}(\mathbb{F})$,

rank of $\boldsymbol{A}$ = *row rank of* $\boldsymbol{A}$ = *column rank of* $\boldsymbol{A}$.

It is, however, worthwhile to remember that when $m \neq n$, $\text{row}(\mathbf{A}) \subseteq \mathbb{F}^n$, and $\text{col}(\mathbf{A}) \subseteq \mathbb{F}^m$. So the row space and column space are not generally identical, even if their dimensions are the same.

To determine the rank of a matrix, one can transform it into echelon form, using elementary row operations, and then count the number of non-zero rows in the matrix in echelon form.

Proposition 7.44. *For any matrix* $\boldsymbol{A} \in \mathcal{M}_{mn}(\mathbb{F})$,

rank of $\boldsymbol{A} \leq \min(m,n)$.

Proof: Since $\text{row}(\mathbf{A})$ and $\text{col}(\mathbf{A})$ are spanned by a set of m and n vectors respectively, it follows that $\text{rank}(\mathbf{A}) \leq \min\{m,n\}$, as required.

7.12 Systems of linear equations

We can now consider properties of systems[3] of linear equations from a new perspective. Consider the system of equations

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1n}x_n & = & d_1, \\ a_{21}x_1 & + & a_{22}x_2 & + & \ldots & + & a_{2n}x_n & = & d_2, \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \ldots & + & a_{mn}x_n & = & d_m, \end{array} \tag{7.28}$$

which can be written as,

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{d}, \tag{7.29}$$

where $\mathbf{A} = [a_{ij}]$,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \text{ and } \mathbf{d} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_m \end{bmatrix}.$$

Alternatively, the system of equations can also be written in the form

$$x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + \ldots + x_n\mathbf{v}_n = \mathbf{d},$$

where

$$\mathbf{v}_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix}, \ \mathbf{v}_2 = \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix}, \ \ldots, \ \mathbf{v}_n = \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix}. \tag{7.30}$$

So the vector $\mathbf{v}_j$ represents the j-th column in the matrix $\mathbf{A}$.

Example 120: Express

$$\begin{array}{ccccccccc} x_1 & - & 3x_2 & + & 4x_3 & - & 2x_4 & = & 5, \\ & & 2x_2 & + & 5x_3 & + & x_4 & = & 2, \\ & & x_2 & - & 3x_3 & & & = & 4. \end{array}$$

in the equivalent forms in (7.29) and (7.30).

Theorem 7.45. *The system of equations $\boldsymbol{A} \cdot \mathbf{x} = \mathbf{d}$ of m equations in n unknowns has at least one solution if and only if*

$$\mathbf{d}^T \in \text{col}(\boldsymbol{A}).$$

[3]Note that we have stopped writing simultaneous ... in general from now onwards we don't.

Proof: If $\mathbf{d} \in \text{col}(\mathbf{A})$, then there exists at least one set on real numbers $\lambda_1,\ \lambda_2,\ \ldots, \lambda_n$ such that

$$\lambda_1 \mathbf{v}_1^T + \lambda_2 \mathbf{v}_2^T + \ldots + \lambda_n \mathbf{v}_n^T = \mathbf{d}^T.$$

Hence, the system $\mathbf{A} \cdot \mathbf{x} = \mathbf{d}$ has at least one solution.

If the system has at least one solution, e.g.,

$$\mathbf{x} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix},$$

then

$$\mu_1 \mathbf{v}_1^T + \mu_2 \mathbf{v}_2^T + \ldots + \mu_n \mathbf{v}_n^T = \mathbf{d}^T.$$

Hence $\mathbf{d}^T \in \text{col}(\mathbf{A})$, as required.

Corollary 7.46. *The system of equations $\boldsymbol{A} \cdot \mathbf{x} = \mathbf{d}$ of m equations in n unknowns has at least one solution if and only if*

$$\text{col}([\boldsymbol{A} \mid \mathbf{d}]) = \text{col}(\boldsymbol{A}),$$

where $[\boldsymbol{A} \mid \mathbf{d}]$ is the augmented m by $n+1$ matrix obtained by adding a column, consisting of the elements of $\mathbf{d}$ to the right of the matrix $\boldsymbol{A}$.

Proof: This is equivalent to Theorem 7.45.

Theorem 7.47. *A system of equations $\boldsymbol{A} \cdot \mathbf{x} = \mathbf{d}$ of m equations in n unknowns has at least one solution if and only if the coefficient matrix $\boldsymbol{A}$ and the augmented matrix $[\boldsymbol{A} \mid \mathbf{d}]$ have the same rank.*

Proof: Via Corollary 7.46 the system has a solution if and only if $\text{col}([\mathbf{A} \mid \mathbf{d}]) = \text{col}(\mathbf{A})$, which implies that the (column) rank of $\mathbf{A}$ is equal to the (column) rank of $[\mathbf{A} \mid \mathbf{d}]$ and via Theorem 7.42, rank$[\mathbf{A} \mid \mathbf{d}]$ =rank$(\mathbf{A})$.

Conversely, assume that rank$[\mathbf{A} \mid \mathbf{d}]$ =rank$(\mathbf{A})$. Note that $\text{col}(\mathbf{A}) \subseteq \text{col}([\mathbf{A} \mid \mathbf{d}])$. In Theorem 7.31, we have seen that two subspaces with the same dimension and where one is a subset of the other have to be equal, so $\text{col}(\mathbf{A}) = \text{col}([\mathbf{A} \mid \mathbf{d}])$ and hence the system has a solution, as required.

Theorem 7.48. *A system of equations $\boldsymbol{A} \cdot \mathbf{x} = \mathbf{d}$ of m equations in n unknowns has a unique solution if the coefficient matrix $\boldsymbol{A}$ and the augmented matrix $[\boldsymbol{A} \mid \mathbf{d}]$ have rank equal to n.*

Proof: Since the rank of a matrix is also the column rank, the fact that the rank($\mathbf{A}$) = n means that the set of n vectors formed by the columns of $\mathbf{A}$ is a basis for the column space. Hence, $\mathbf{d}$ can be expressed as a linear combination of these vectors, uniquely, i.e. the expansion

$$\mathbf{d} = \lambda_1\mathbf{v}_1 + \lambda_2\mathbf{v}_2 + \ldots + \lambda_n\mathbf{v}_n,$$

is unique. Therefore there is a unique solution $\mathbf{x} = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ to the system of linear equations, as required.

If rank($\mathbf{A}$) (and rank($[\mathbf{A} \mid \mathbf{d}]$)) is r with $0 < r < n$, then the system has infinitely many solutions. One can see that there will then be $n - r$ degrees of freedom, i.e. $n - r$ of the unknowns can be chosen without restriction.

Example 121: For all real values of α, find all solutions, if any, of the system of linear equations

$$\begin{array}{rcrcrcl} x & + & 3y & + & 2z & = & 1, \\ 2x & + & 7y & + & \alpha z & = & 5, \\ (\alpha-3)x & + & y & - & 10z & = & 11. \end{array}$$

Chapter 8

Linear Transformations

▸ **Learning Outcomes** ◂ After the completion of the lectures and support sessions associated with this chapter you should be able to:

- define a linear transformation between two vector spaces;
- verify that a mapping between two vector spaces is a linear transformation;
- cite and use a number of properties of linear transformations;
- define and identify the kernel and image of a linear transformation;
- represent a linear transformation by a matrix;
- define and use coordinate transformations;
- apply the composition of two linear transformations and its equivalent on their matrix representations;
- define and use transition matrices; and

[4, p.323-384] contains an alternative presentation of material in this chapter that you may find helpful.

8.1 Introduction

In this chapter, we will study mappings (functions) between vector spaces, represent these mappings by matrices and study mappings from a vector space to itself. Such mappings encompass coordinate transformations when different bases are considered within the same vector space.

8.2 Definition

Consider two real vector spaces V and W. One can then consider a relationship between the elements of the first vector space and elements of the second. For example, one can consider a mapping between vectors in E^2 and ordered pairs of real numbers in $\mathbb{R}^2$. It is particularly important to define mappings that treat addition in the vector space and scalar multiplication with entries in the field, in a consistent way.

Definition 8.1. *Let V and W be vectir spaces over $\mathbb{F}$. A function $T : V \to W$ is called a **linear transformation** if:*

(i) $T(\mathbf{v}_1 + \mathbf{v}_2) = T(\mathbf{v}_1) + T(\mathbf{v}_2)$ *for all* $\mathbf{v}_1, \mathbf{v}_2 \in V$*; and*

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

Note that the vector spaces V and W in Definition 8.1 can be equal. A linear transformation is also referred to as a *linear mapping* or a *homomorphism*. Notice that the addition and scalar multiplication on both sides of the identities in Definition 8.1 refer to operations in different vector spaces. Specifically, on the LHS and RHS of the equations in (i) and (ii), the operations are in V and W, respectively.

In other words, if T is a linear transformation, then the image of the sum of two vectors in V is the sum (in W) of the images of these two vectors, and, the image of a scalar multiple of a vector in V is the scalar multiple (in W) of the image of the vector.

Example 122: Show that $T : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$T((x, y)) = (y, x) \quad \forall (x, y) \in \mathbb{R}^2$$

is a linear transformation and give its geometric interpretation when $\mathbb{R}^2$ is represented by the set of all points in a plane.

Example 123: Show that $O : V \to W$ given by

$$O(\mathbf{v}) = \mathbf{0} \quad \forall \mathbf{v} \in V$$

is a linear transformation (known as the **zero transformation**.)

Example 124: Show that $I_V : V \to V$ given by

$$I_V(\mathbf{v}) = \mathbf{v} \quad \forall \mathbf{v} \in V$$

is a linear transformation (known as the **identity transformation**.)

8.3 Properties of Linear Transformations

Theorem 8.2. *Consider vector spaces V and W over $\mathbb{F}$ and let $T : V \to W$ be a linear transformation. Then:*

(i) $T(\mathbf{0}) = \mathbf{0} \in W$;

(ii) $T(-\mathbf{v}) = -T(\mathbf{v})$ *for all* $\mathbf{v} \in V$ *with* $-\mathbf{v} \in V$ *the additive inverse of* $\mathbf{v}$; *and*

(iii) $T(\lambda_1\mathbf{v}_1 + \lambda_2\mathbf{v}_2 + \ldots + \lambda_k\mathbf{v}_k) = \lambda_1 T(\mathbf{v}_1) + \lambda_2 T(\mathbf{v}_2) + \ldots + \lambda_k T(\mathbf{v}_k)$ *for all* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k \in V$, $\lambda_1, \lambda_2, \ldots, \lambda_k \in \mathbb{F}$.

Proof: The proof is an exercise. However, we note that the proofs of (i)-(iii) follow almost immediately from Definition 8.1 (mathematical induction should also be used for (iii)).

Thus, a necessary condition for a function $T : V \to W$ to be a linear transformation is that $T(\mathbf{0}) = \mathbf{0}$.

Theorem 8.3. *Consider two real vector spaces V and W. Let $B = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ be a basis of V with $\dim(V) = n$. Let $T_1 : V \to W$ and $T_2 : V \to W$ be linear transformations such that*

$$T_1(\mathbf{v}_i) = T_2(\mathbf{v}_i) \quad \forall \mathbf{v}_i \in B. \tag{8.1}$$

Then $T_1 = T_2$.

Proof: Since $B = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is a basis for V, it follows that for each $\mathbf{v} \in V$, there exists $\lambda_i \in \mathbb{R}$ for $i = 1, \ldots, n$ such that

$$\mathbf{v} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i. \tag{8.2}$$

Since $T_1(\mathbf{v}_i) = T_2(\mathbf{v}_i)$ for each $i = 1, \ldots, n$, it follows that

$$\begin{aligned} T_1(\mathbf{v}) &= T_1\left(\sum_{i=1}^{n} \lambda_i \mathbf{v}_i\right) \quad \text{(via (8.2))} \\ &= \sum_{i=1}^{n} \lambda_i T_1(\mathbf{v}_i) \quad \text{(via Theorem 8.2)} \\ &= \sum_{i=1}^{n} \lambda_i T_2(\mathbf{v}_i) \quad \text{(via (8.1))} \\ &= T_2\left(\sum_{i=1}^{n} \lambda_i \mathbf{v}_i\right) \quad \text{(via Theorem 8.2)} \end{aligned}$$

$$= T_2(\mathbf{v}),$$

as required.

So if two linear transformations map all vectors of a basis into the same images, then the linear transformations are the same. Therefore a linear transformation $T : V \to W$ is known if the images under T of the vectors in a basis of V are known.

Theorem 8.4. *Consider two real vector spaces V and W. Let $B = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ be a basis of V with $\dim(V) = n$. Let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n$ be n (not necessarily distinct) vectors in W. Then **there exists** a **unique** linear transformation $T : V \to W$ such that*

$$T(\mathbf{v}_1) = \mathbf{w}_1,\ T(\mathbf{v}_2) = \mathbf{w}_2, \ldots, T(\mathbf{v}_n) = \mathbf{w}_n.$$

Proof: We define the transformation using the basis for V. For all $\mathbf{v} \in V$, there exist $\mu_i \in \mathbb{R}$ such that[1]

$$\mathbf{v} = \sum_{i=1}^{n} \mu_i \mathbf{v}_i. \tag{8.3}$$

Define $T : V \to W$, using (8.3), to be

$$T(\mathbf{v}) = \sum_{i=1}^{n} \mu_i \mathbf{w}_i \quad \forall \mathbf{v} \in V. \tag{8.4}$$

Firstly, note that by (8.3) and (8.4),

$$T(\mathbf{v}_i) = \mathbf{w}_i \tag{8.5}$$

since

$$\mathbf{v}_j = \sum_{i=1}^{n} \mu_i \mathbf{v}_i \iff \mu_i = \begin{cases} 1 & ; i = j \\ 0 & ; i \neq j. \end{cases}$$

Now, for $\mathbf{u}, \mathbf{w} \in V$ we can write $\mathbf{u} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i$ and $\mathbf{w} = \sum_{i=1}^{n} \beta_i \mathbf{v}_i$. Hence for $\lambda \in \mathbb{F}$,

$$\lambda\mathbf{u} + \mathbf{w} = \sum_{i=1}^{n} \lambda\alpha_i \mathbf{v}_i + \sum_{i=1}^{n} \beta_i \mathbf{v}_i = \sum_{i=1}^{n} (\lambda\alpha_i + \beta_i)\mathbf{v}_i.$$

it follows from (8.3) that

$$\begin{aligned} T(\lambda\mathbf{u} + \mathbf{w}) &= T\left(\sum_{i=1}^{n} (\lambda\alpha_i + \beta_i)\mathbf{v}_i\right) \\ &= \sum_{i=1}^{n} (\lambda\alpha_i + \beta_i)\mathbf{w}_i \end{aligned}$$

[1] These are the coordinates of $\mathbf{v}$ with respect to the basis B.

$$\begin{aligned} &= \lambda \sum_{i=1}^{n} \alpha_i \mathbf{w}_i + \sum_{i=1}^{n} \beta_i \mathbf{w}_i \\ &= \lambda T(\mathbf{u}) + T(\mathbf{w}). \end{aligned}$$

Therefore, T is a linear transformation from V to W. In addition, we've seen that T maps $\mathbf{v}_i$ to $\mathbf{w}_i$ directly from (8.5). Finally, the transformation in (8.4) is unique by Theorem 8.3 since $T(\mathbf{v}_i)$ is defined for all basis vectors $\mathbf{v}_i \in B$. This completes the proof, as required.

Example 125: Consider the basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}$ of $\mathbb{R}^4$ with

$$\mathbf{v}_1 = (1,1,1,1),\ \mathbf{v}_2 = (0,1,1,1),\ \mathbf{v}_3 = (0,0,1,1) \text{ and } \mathbf{v}_4 = (0,0,0,1),$$

and its property

$$(x_1, x_2, x_3, x_4) = x_1\mathbf{v}_1 + (x_2 - x_1)\mathbf{v}_2 + (x_3 - x_2)\mathbf{v}_3 + (x_4 - x_3)\mathbf{v}_4$$

for all $(x_1, x_2, x_3, x_4) \in \mathbb{R}^4$. Find an expression for $T : \mathbb{R}^4 \to \mathbb{R}^3$, such that

$$T(\mathbf{v}_1) = (1,1,0),\ T(\mathbf{v}_2) = (0,1,1),\ T(\mathbf{v}_3) = (1,0,1) \text{ and } T(\mathbf{v}_4) = (1,1,1).$$

Consider a linear transformation $T : U \to V$ and a linear transformation $S : V \to W$. Then for a vector in U, map it by T to V and then use S to map it to W. This is called a composition of two mappings.

Definition 8.5. *Let U, V and W be vector spaces over $\mathbb{F}$ and $T : U \to V$ and $S : V \to W$ be linear transformations. Then the composition of T and S is defined by the mapping $S \circ T : U \to W$, with*

$$S \circ T = S(T(\mathbf{u})) \quad \forall \mathbf{u} \in U.$$

Theorem 8.6. *Let U, V and W be vector spaces over $\mathbb{F}$ and $T : U \to V$ and $S : V \to W$ be linear transformations. Then the composition of T and S, denoted by $S \circ T : U \to W$, is a linear transformation from U to W.*

Proof:[2] Observe that $(S \circ T) : U \to W$. Additionally,

(i) For all $\mathbf{u}_1, \mathbf{u}_2 \in U$ we have

$$\begin{aligned} (S \circ T)(\mathbf{u}_1 + \mathbf{u}_2) &= S(T(\mathbf{u}_1 + \mathbf{u}_2)) \\ &= S(T(\mathbf{u}_1) + T(\mathbf{u}_2)) \\ &= S(T(\mathbf{u}_1)) + S(T(\mathbf{u}_2)) \\ &= (S \circ T)(\mathbf{u}_1) + (S \circ T)(\mathbf{u}_2). \end{aligned}$$

[2] As an exercise, write this proof in 1 paragraph by instead considering $(S \circ T)(\lambda\mathbf{u}_1 + \mathbf{u}_2)$ directly for $\mathbf{u}_1, \mathbf{u}_2 \in U$ and $\lambda \in \mathbb{R}$.

(ii) For all $\lambda \in \mathbb{R}$ and $\mathbf{u} \in U$ we have

$$(S \circ T)(\lambda \mathbf{u}) = S(T(\lambda \mathbf{u})) = S(\lambda T(\mathbf{u})) = \lambda S(T(\mathbf{u})) = \lambda (S \circ T)(\mathbf{u}).$$

Therefore, $(S \circ T)$ is a linear transformation from U to W, as required.

8.4 Kernel and image

The **kernel** and the **image** of a linear transformation $T : V \to W$ are important subspaces associated with T. The kernel is a subspace of V, while the image is a subspace of W.

Definition 8.7. *Suppose that $T : V \to W$ is a linear transformation. Then the **kernel**, is the subset of V defined by*

$$\ker(T) = \{\mathbf{v} \in V \mid \ T(\mathbf{v}) = \mathbf{0}\}$$

*and the **image** of T is the subset of W defined by*

$$\mathrm{im}(T) = \{T(\mathbf{v}) \mid \mathbf{v} \in V\}.$$

Theorem 8.8. *Suppose that V and W are vector spaces over $\mathbb{F}$ and let $T : V \to W$ be a linear transformation. Then $\ker(T)$ is a subspace of V and $\mathrm{im}(T)$ is a subspace of W.*

Proof: We show that $\ker(T)$ is a subspace of V and leave the second part as an exercise. Since $T(\mathbf{0}) = \mathbf{0}$, the kernel is a non-empty set. Let $\mathbf{u}, \mathbf{w} \in \ker(T)$ and $\lambda \in \mathbb{F}$. Then

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

Hence $\lambda \mathbf{u} + \mathbf{w} \in \ker(T)$ and therefore $\ker(T)$ is a subspace by the subspace test.

Lemma 8.9. *Suppose that V and W are vector spaces over $\mathbb{F}$ and let $T : V \to W$ be a linear transformation. Then*

(i) T is onto (surjective) if and only if $\mathrm{im}(T) = W$.

(ii) T is one to one (injective) if and only if $\ker(T) = \{\mathbf{0}\}$.

(iii) if $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis for V, then $\mathrm{im}(T) = \mathrm{span}(T(\mathbf{v}_1), \ldots, T(\mathbf{v}_n))$.

Proof: Part (i) is just the definition of a map being onto. Part (ii) is more interesting. Suppose that $\ker(T) = \{\mathbf{0}\}$ and that $\mathbf{v}, \mathbf{u} \in V$ have the property that $T(\mathbf{v}) = T(\mathbf{u})$. Then $T(\mathbf{v}) - T(\mathbf{u}) = \mathbf{0}$. Since T is a linear transformation, $T(\mathbf{v}) - T(\mathbf{u}) = T(\mathbf{v} - \mathbf{u})$ and so $\mathbf{v} - \mathbf{u} \in \ker(T) = \{0\}$. Hence $\mathbf{v} = \mathbf{u}$ and T is one to one. Suppose that T is one to one. Then $\mathbf{0}$ is the unique element that maps to $\mathbf{0}$ and so $\ker(T) = \{\mathbf{0}\}$.

Sometimes $\ker(T)$ is also referred to as the **nullspace** of T. It can be that $\ker(T) = \{\mathbf{0}\}$, but this is not always true. For example, $\ker(O) = V$, where O is the zero transformation (see Example 123). The dimension of $\ker(T)$ is called the **nullity** of the linear transformation T. The subspace $\text{im}(T)$ is also called the **range** of T or the **image space** of T. The dimension of the image of T is called the **rank** of the linear transformation T.

Example 126: Find a basis of $\ker(T)$ and $\text{im}(T)$ where $T : \mathbb{R}^4 \to \mathbb{R}^3$ is the linear transformation given by

$$T((x_1, x_2, x_3, x_4)) = (x_1 + x_2 + 2x_3 + x_4,\ 2x_1 + x_2 + x_3 - x_4,\ 3x_1 - x_2 - 6x_3 - 9x_4)$$

for all $(x_1, x_2, x_3, x_4) \in \mathbb{R}^4$, and hence determine the nullity of T and the rank of T.

For the linear transformation used in the previous two exercises, we find that the nullity of T is 2 and the rank of T is 2. Hence, the nullity of T + the rank of T is the dimension of the domain of T, i.e. $\mathbb{R}^4$. This holds more generally:

Theorem 8.10. ***(rank-nullity)*** *Consider two real vector spaces V and W, with $\dim(V) = n$. Let $T : V \to W$ be a linear transformation. Then,*

$$\textit{rank of } T + \textit{nullity of } T = n.$$

We delay the proof of the rank-nullity theorem until next year.

8.5 Matrix representation

Every linear transformation from a real vector space to another real vector space can be described by a matrix. Such a representation, requires specified bases in both vector spaces. Since the coordinates of a vector with respect to a basis generally differ from the coordinates of the same vector with respect to a different basis, there is no unique matrix representation valid for all bases in a vector space!

Definition 8.11. *Let V and W be vector spaces over $\mathbb{F}$, and suppose that $B_V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ and $B_W = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m\}$ are ordered bases of V and W respectively. Assume that*

$$R : V \to W$$

is a linear transformation.
Then the matrix of R corresponding to the ordered bases B_V and B_W is the $m \times n$ matrix in which the j-th column is given by the transpose of the coordinate vector of $R(\mathbf{v}_j)$ with respect to B_W.

To emphasise the dependence of the matrix on the choice of basis, often the notation ${}_{B_V}M_{B_W}(R)$, $M_{B_V B_W}(R)$ or $M\{R; B_V, B_W\}$ is used to denote the matrix representing the linear transformation R from V to W corresponding to the bases B_V of V and B_W of W. Thus if

$$\begin{array}{rcccccccc} R(\mathbf{v}_1) & = & a_{11}\mathbf{w}_1 & + & a_{21}\mathbf{w}_2 & + & \ldots & + & a_{m1}\mathbf{w}_m, \\ R(\mathbf{v}_2) & = & a_{12}\mathbf{w}_1 & + & a_{22}\mathbf{w}_2 & + & \ldots & + & a_{m2}\mathbf{w}_m, \\ \vdots & & \vdots & & \vdots & & & & \vdots \\ R(\mathbf{v}_n) & = & a_{1n}\mathbf{w}_1 & + & a_{2n}\mathbf{w}_2 & + & \ldots & + & a_{mn}\mathbf{w}_m, \end{array}$$

the matrix of R corresponding to the basis B_V in V and B_W in W is given by

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}.$$

Hence, we can also write,

$$\begin{pmatrix} R(\mathbf{v}_1) \\ R(\mathbf{v}_2) \\ \vdots \\ R(\mathbf{v}_n) \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} & \ldots & a_{m1} \\ a_{12} & a_{22} & \ldots & a_{m2} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & \ldots & a_{mn} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_m \end{pmatrix} = \mathbf{A}^T \cdot \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_m \end{pmatrix}.$$

Suppose that

$$\mathbf{v} = \lambda_1\mathbf{v}_1 + \lambda_2\mathbf{v}_2 + \cdots + \lambda_n\mathbf{v}_n$$

so that $\mathbf{v}$ has coordinate vector $(\lambda_1, \lambda_2, \ldots, \lambda_n)$ with respect to B_V and

$$R(\mathbf{v}) = \mu_1\mathbf{w}_1 + \mu_2\mathbf{w}_2 + \cdots + \mu_m\mathbf{w}_m$$

has coordinate vector $(\mu_1, \mu_2, \ldots, \mu_m)$ with respect to B_W. Then

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} = A \cdot \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix}.$$

Example 127: Consider the linear transformation $T : \mathbb{R}^3 \to \mathbb{R}^2$ given by

$$T((x_1,\ x_2,\ x_3)) = (3x_1 - x_2 - 4x_3,\ 2x_1 + 3x_2 - x_3)$$

for all $(x_1, x_2, x_3) \in \mathbb{R}^3$. Find the matrix of T corresponding to

(i) the standard ordered bases of $\mathbb{R}^3$ and $\mathbb{R}^2$ respectively,

(ii) the ordered basis $\{(1,1,0),(1,-1,0),(1,1,1)\}$ of $\mathbb{R}^3$ and the ordered basis $\{(1,2),(2,1)\}$ of $\mathbb{R}^2$.

Typically, if we change one of the two bases, the matrix of T will change, but it will always be a $m \times n$ matrix.

One exception to the matrix changing is the zero transformation, which for whatever basis in V and whatever basis in W is always represented by a $m \times n$ zero matrix. Indeed, whatever bases are used,

$$O(\mathbf{v}_j) = 0\mathbf{w}_1 + 0\mathbf{w}_2 + \ldots + 0\mathbf{w}_m.$$

Example 128: Let V be a real vector space with $\dim(V) = n$ and ordered basis B. Find the matrix representing the identity transformation $I_V : V \to V$ with respect to the basis B in both the domain and co-domain.

Example 129: Let $B = \{(1,1),(1,3)\}$ be the ordered basis for the domain of identity transformation $I_V : \mathbb{R}^2 \to \mathbb{R}^2$. The basis for the co-domain is the standard ordered basis of $\mathbb{R}^2$. Determine the matrix representing of the identity transformation with respect to these two bases.

Theorem 8.12. *Suppose that V and W are finite dimensional vector spaces over $\mathbb{F}$ and $T : V \to W$ is a linear transformation. Let $\boldsymbol{A}$ be the matrix representing T with respect to a basis B_V in V and basis B_W in W. Then*

$$\textit{rank of } T = \textit{rank of } \boldsymbol{A}.$$

Proof: The image of T is spanned by the image of the elements of $B_V = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. Thus $\mathrm{im}(T)$ is spanned by the coordinates corresponding to the columns of $\mathbf{A}$. Thus $\dim T = \dim \mathrm{col}(\mathbf{A})$. Since the column rank of $\mathbf{A}$ is the rank of $\mathbf{A}$, this proves the claim.

8.6 Transformation of coordinates

Can we use the matrix representation to calculate the image of a vector from a linear transformation? To do this, we need to work with the **coordinates** of that vector with respect to a given basis.

Theorem 8.13. *Suppose that V and W are vector spaces over $\mathbb{F}$ and*

$$R : V \to W$$

be a linear transformation. Let $B_V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, and $B_W = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m\}$ be ordered basis of V and W respectively and $\mathbf{A}$ be the matrix representing R with respect to the bases B_V and B_W. Assume that $\mathbf{v} \in V$, $\mathbf{w} = R(\mathbf{v}) \in W$ and

(i) $\mathbf{a} = (\lambda_1, \ldots, \lambda_n)$ *are the coordinates of* $\mathbf{v}$ *with respect to* B_V.

(ii) $\mathbf{b} = (\mu_1, \ldots, \mu_m)$ *are the coordinates of* $\mathbf{w}$ *with respect to* B_W.

Then

$$\mathbf{b}^T = \boldsymbol{A} \cdot \mathbf{a}^T.$$

Proof: By the definition of coordinates

$$\mathbf{v} = \lambda_1\mathbf{v}_1 + \lambda_2\mathbf{v}_2 + \ldots + \lambda_n\mathbf{v}_n$$

and

$$R(\mathbf{v}) = \mathbf{w} = \mu_1\mathbf{w}_1 + \mu_2\mathbf{w}_2 + \ldots + \mu_m\mathbf{w}_m. \tag{8.6}$$

Now,

$$\begin{aligned}
R(\mathbf{v}) &= R(\lambda_1\mathbf{v}_1 + \lambda_2\mathbf{v}_2 + \ldots + \lambda_n\mathbf{v}_n) \\
&= \lambda_1 R(\mathbf{v}_1) + \lambda_2 R(\mathbf{v}_2) + \ldots + \lambda_n R(\mathbf{v}_n) \\
&= \lambda_1(a_{11}\mathbf{w}_1 + a_{21}\mathbf{w}_2 + \ldots + a_{m1}\mathbf{w}_m) \\
&\quad +\lambda_2(a_{12}\mathbf{w}_1 + a_{22}\mathbf{w}_2 + \ldots + a_{m2}\mathbf{w}_m) \\
&\quad \vdots \\
&\quad +\lambda_n(a_{1n}\mathbf{w}_1 + a_{2n}\mathbf{w}_2 + \ldots + a_{mn}\mathbf{w}_m) \\
&= (\lambda_1 a_{11} + \lambda_2 a_{12} + \ldots + \lambda_n a_{1n})\mathbf{w}_1 \\
&\quad +(\lambda_1 a_{21} + \lambda_2 a_{22} + \ldots + \lambda_n a_{2n})\mathbf{w}_2 \\
&\quad \vdots \\
&\quad +(\lambda_1 a_{m1} + \lambda_2 a_{m2} + \ldots + \lambda_n a_{mn})\mathbf{w}_m.
\end{aligned} \tag{8.7}$$

Since $R(\mathbf{v})$ is expressed uniquely as a linear combination of the vectors in the basis B_W, it follows from (8.7) that

$$\begin{aligned}
\mu_1 &= \lambda_1 a_{11} + \lambda_2 a_{12} + \ldots + \lambda_n a_{1n}, \\
\mu_2 &= \lambda_1 a_{21} + \lambda_2 a_{22} + \ldots + \lambda_n a_{2n}, \\
&\vdots \\
\mu_m &= \lambda_1 a_{m1} + \lambda_2 a_{m2} + \ldots + \lambda_n a_{mn},
\end{aligned}$$

or, in matrix form,

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix},$$

which completes the proof, as required.

Example 130: Consider the linear transformation $T : \mathbb{R}^4 \to \mathbb{R}^3$, with matrix representation

$$\begin{pmatrix} 1 & 2 & -3 & 1 \\ 1 & -3 & 1 & -2 \\ 2 & 1 & -3 & 4 \end{pmatrix}$$

with respect to the ordered basis

$$B_V = \{(1, 2, -2, -1),\ (2, 1, -1, 4),\ (3, 0, -3, 2),\ (3, 7, -5, 1)\},$$

in $\mathbb{R}^4$ and the ordered basis

$$B_W = \{(1, 2, -3),\ (1, 3, 2),\ (2, -1, -5)\}$$

in $\mathbb{R}^3$. Determine the image $T(\mathbf{v}) \in \mathbb{R}^3$ of the vector $\mathbf{v} \in \mathbb{R}^4$ with coordinates $(5, -1, 4, 2)$ relative to B_V.

The result highlighted in the previous example is a general one:

Theorem 8.14. *Suppose that U, V and W are vector spaces over $\mathbb{F}$ and $T : U \to V$ and $S : V \to W$ are linear transformations. Consider the ordered bases B_U, B_V and B_W of U, V and W respectively. Let $\dim(U) = n$, $\dim(V) = m$ and $\dim(W) = p$. Assume that*

(i) $\boldsymbol{B} \in \mathcal{M}_{mn}(\mathbb{F})$ be the matrix representing the linear transformation T with respect to the bases B_U and B_V,

(ii) $\boldsymbol{A} \in \mathcal{M}_{pm}(\mathbb{F})$ is the matrix representing the linear transformation S with respect to the bases B_V and B_W.

Then $\boldsymbol{C} = \boldsymbol{A} \cdot \boldsymbol{B} \in \mathcal{M}_{pn}(\mathbb{F})$ is the matrix representing the composition of T and S, $S \circ T : U \to W$, with respect to the bases B_U and B_W.

Proof: We have

$$\mathbf{u} = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \in U. \tag{8.8}$$

So that the coordinate vector for $\mathbf{u}$ is $(\lambda_1, \ldots, \lambda_n)$ with respect to B_U. Let $(\mu_1, \ldots, \mu_m)$ be the coordinate vector for $\mathbf{v} = T(\mathbf{u})$ with respect to B_V. Then

$$\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix} = \mathbf{B} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix}.$$

The coordinate vectors for $S(\mathbf{w})$ with respect to B_W is then $(\tau_1, \ldots, \tau_p)$ where

$$\begin{pmatrix} \tau_1 \\ \vdots \\ \tau_p \end{pmatrix} = \mathbf{A} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix}.$$

Hence

$$\begin{pmatrix} \tau_1 \\ \vdots \\ \tau_p \end{pmatrix} = \mathbf{A}(\mathbf{B} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix}) = (\mathbf{A} \cdot \mathbf{B}) \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix}.$$

Hence, the matrix representing $S \circ T$ with respect to B_U in the domain and B_W in the co-domain is $\mathbf{C} = \mathbf{A} \cdot \mathbf{B}$, as required.

Example 131: Consider the two ordered bases B and $\tilde{B}$ in $\mathbb{R}^2$, with

$$B = \{(1,1),\ (1,2)\} \text{ and } \tilde{B} = \{(2,3),\ (3,2)\}.$$

Consider the identity transformation $1_{\mathbb{R}^2} : \mathbb{R}^2 \to \mathbb{R}^2$. Then,

(i) determine the matrix representing $\mathbf{A}_1$ of $1_{\mathbb{R}^2}$ with respect to the basis B (in the domain) and $\tilde{B}$ (in the co-domain),

(ii) determine the matrix representing $\mathbf{A}_2$ of $1_{\mathbb{R}^2}$ with respect to the basis $\tilde{B}$ (in the domain) and B (in the co-domain),

(iii) evaluate the matrix products $\mathbf{A}_1 \cdot \mathbf{A}_2$ and $\mathbf{A}_2 \cdot \mathbf{A}_1$.

Theorem 8.15. *Suppose that V is a vector space over $\mathbb{F}$, with $\dim(V) = n$. Consider two ordered bases B and $\tilde{B}$ for V. Let*

(i) the $n \times n$ matrix $\boldsymbol{A}_1$ be the matrix representing the identity transformation 1_V with respect to the bases B in the domain and $\tilde{B}$ in the co-domain,

(ii) the $n \times n$ matrix $\boldsymbol{A}_2$ be the matrix representing the identity transformation 1_V with respect to the bases $\tilde{B}$ in the domain and B in the co-domain.

Then,

$$\boldsymbol{A}_1 \cdot \boldsymbol{A}_2 = \boldsymbol{A}_2 \cdot \boldsymbol{A}_1 = \boldsymbol{I}_n.$$

Proof: See practise questions.

This implies that both matrix representations $\mathbf{A}_1$ and $\mathbf{A}_2$ for the identity transformation $1_V : V \to V$ are invertible and that $\mathbf{A}_2 = \mathbf{A}_1^{-1}$, i.e. $\mathbf{A}_2$ is the unique inverse of $\mathbf{A}_1$.

8.7 Transition matrices; change of basis matrices

Consider two ordered bases B and $\tilde{B}$ in the same vector space V. Any vector can be written uniquely as a linear combination of vectors in each basis, yielding two different sets of coordinates. Like with a linear transformation, the relationship between these two sets of coordinates is fully determined if we know how the vectors in one basis are expressed in terms of the other.

Definition 8.16. *Let V be a vector space over $\mathbb{F}$ with $\dim(V) = n$ and bases B and $\tilde{B}$ given by*

$$B = \{\mathbf{u}_1,\ \mathbf{u}_2, \ldots,\ \mathbf{u}_n\} \text{ and } \tilde{B} = \{\mathbf{v}_1,\ \mathbf{v}_2, \ldots,\ \mathbf{v}_n\}.$$

*The **matrix of transition** from the ordered basis B to the ordered basis $\tilde{B}$ is the $n \times n$ matrix $\mathbf{P}$ which has j-th column the transpose of the coordinate vector of $u_j \in B$ written with respect to the basis $\tilde{B}$.*

So, if

$$\begin{aligned} I_V(\mathbf{u}_1) = \mathbf{u}_1 &= p_{11}\mathbf{v}_1 + p_{21}\mathbf{v}_2 + \ldots + p_{n1}\mathbf{v}_n, \\ I_V(\mathbf{u}_2) = \mathbf{u}_2 &= p_{12}\mathbf{v}_1 + p_{22}\mathbf{v}_2 + \ldots + p_{n2}\mathbf{v}_n, \\ &\vdots \\ I_V(\mathbf{u}_n) = \mathbf{u}_n &= p_{1n}\mathbf{v}_1 + p_{2n}\mathbf{v}_2 + \ldots + p_{nn}\mathbf{v}_n, \end{aligned}$$

the transition matrix $\mathbf{P}$ is given by

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix}. \tag{8.9}$$

The relationship between the vectors of the new basis and the vectors of the old basis can then be written in short form as

$$\begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_n \end{pmatrix} = \mathbf{P}^T \cdot \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_n \end{pmatrix}.$$

Example 132: Consider the ordered bases $B = \{(1,1),\ (1,2)\}$ and $\tilde{B} = \{(2,3),\ (3,2)\}$ in $\mathbb{R}^2$. Determine the matrix of transition from B to $\tilde{B}$.

Comparing this result with the result from the last exercise in the previous section, we can see that the transition matrix of B to $\tilde{B}$ is equal to the matrix representing the identity transformation $1_{\mathbb{R}^2}$ relative to the basis B in the domain and $\tilde{B}$ in the co-domain. Again, this is an example of a more general result.

Theorem 8.17. *Let V be a vector space over $\mathbb{F}$ of dimension n and let B and $\tilde{B}$ be ordered bases for V. Let the $n \times n$ matrix $\boldsymbol{P}$ (as in (8.9)) be the matrix representation of the identity transformation 1_V with respect to the bases B in the domain and $\tilde{B}$ in the co-domain. Then $\boldsymbol{P}$ is the matrix of transition in V from the ordered basis B to the ordered basis $\tilde{B}$.*

Proof: Follows immediately from Definition 8.16 and the definition of the identity transformation.

This immediately implies that a transition matrix $\mathbf{P}$ from the ordered basis B to the ordered basis $\tilde{B}$ is invertible and that $\mathbf{P}^{-1}$ is the matrix of transition from the ordered basis $\tilde{B}$ to the ordered basis B.

Now if $(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is the coordinate vector $\mathbf{w}$ in V with respect to the basis $\tilde{B} = \{\mathbf{v}_1,\ \mathbf{v}_2,\ \ldots,\ \mathbf{v}_n\}$, we have that

$$\mathbf{w} = \lambda_1\mathbf{v}_1 + \lambda_2\mathbf{v}_2 + \ldots + \lambda_n\mathbf{v}_n.$$

Similarly, let $(\mu_1, \mu_2, \ldots, \mu_n)$ the coordinate vector of $\mathbf{w}$ in V with respect to the basis $B = \{\mathbf{u}_1,\ \mathbf{u}_2,\ \ldots,\ \mathbf{u}_n\}$, or

$$\mathbf{w} = \mu_1\mathbf{u}_1 + \mu_2\mathbf{u}_2 + \ldots + \mu_n\mathbf{u}_n.$$

Since $1_V(\mathbf{w}) = \mathbf{w}$, the transformation of coordinates formula in Theorem 8.13, we have,

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix} \cdot \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}.$$

This can now be read as stating that the coordinates of a vector with respect to the "new" basis $\tilde{B}$ can be found by the matrix product of the transition matrix from B to $\tilde{B}$ with the column vector containing the coordinates with respect to the "old" basis B.

Please note that not all textbooks/sources will follow this definition of a transition matrix. Some[3] will define the transition matrix as being $\mathbf{P}^{-1}$.

Example 133: Consider the linear transformations $T : U \to V$ and $S : V \to W$ with $U = \mathbb{R}^2$, $V = \mathbb{R}^4$ and $W = \mathbb{R}^3$ with

$$\begin{aligned} T((x_1, x_2)) &= (x_1 + 2x_2, -2x_1 + 3x_2, 3x_1 - 2x_2, -x_1 - x_2), \\ S((x_1, x_2, x_3, x_4)) &= (2x_1 + x_2 + x_3 - x_4, 3x_1 - x_3 + x_4, x_1 - 4x_2 + 2x_3) \end{aligned}$$

[3] So if using textbooks or alternative sources for additional support, please remember this.

for all $(x_1, x_2) \in \mathbb{R}^2$ and $(x_1, x_2, x_3, x_4) \in \mathbb{R}^4$ with respect to standard ordered bases. Consider the following bases for $\mathbb{R}^2$, $\mathbb{R}^4$ and $\mathbb{R}^3$ respectively:

$$\begin{aligned} B_U &= \{(1,1),\ (1,-1)\}, \\ B_V &= \{(1,0,0,0),\ (1,1,0,0),\ (1,1,1,0),\ (1,1,1,1)\}, \\ B_W &= \{(0,0,1),\ (0,1,1),\ (1,1,1)\}. \end{aligned}$$

Determine the matrix representation of T with respect to the bases B_U and B_V, the matrix representation of S with respect to the bases B_V and B_W, and, the matrix representation of $S \circ T$ with respect to the bases B_U and B_W.

Chapter 9

Conics

▸ **Learning Outcomes** ◂ After the completion of the lectures and support sessions associated with this chapter you should be able to:

- list the types of conics and their properties;
- sketch (draw) a conic from an equation;
- derive the standard equation of a conic; and
- understand and prove key properties satisfied by conics.

[2, p.458-468] contains an alternative presentation of material in this chapter that you may find helpful.

9.1 Introduction

Conic sections are the family of curves defined by the intersection of a cone and a plane in $\mathbb{R}^3$ (see Section 9.5 for the details). This intersection can take different forms according to the angle the intersecting plane makes with the cone. The standard conic sections comprise the circle, the ellipse, the parabola and the hyperbola. Degenerate cases are obtained when the plane intersects the vertex of the cone. These so-called *degenerate conics* can be either a single point or a single line or pair of intersecting lines. The various possible non-degenerate conic sections are depicted in Figure 9.1. We will now discuss these conic sections in more detail.

9.1.1 Intersection of a cone and a plane

Quadratic equations in 2 variables occur frequently in applications of mathematics, and hence, it is useful to classify their solution structure. We proceed with this classification

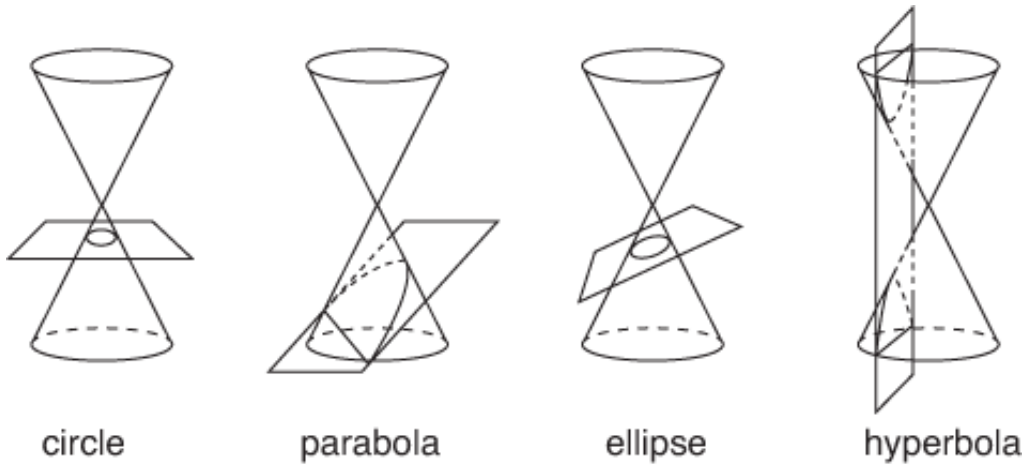


Figure 9.1: The various conic sections, as depicted in [1].

in the following subsections, giving details at the end of the chapter.

Definition 9.1. *A (real) quadratic equation in 2 variables is given by,*

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0,$$

for $A, B, C, D, E, F \in \mathbb{R}$.

Consider a cone with its vertex at the origin in 3-dimensional space and axis of symmetry along the z-axis. Its equation is given by

$$x^2 + y^2 = k^2 z^2,$$

for some $k > 0$. The general equation of a plane in 3-dimensional space is

$$ax + by + cz + d = 0$$

for $a, b, c, z \in \mathbb{R}$. If we can eliminate z using the equation of the plane (i.e. $c \neq 0$[1]) we get

$$x^2 + y^2 = k^2 \left(\frac{-d - ax - by}{c} \right)^2,$$

[1]When $c = 0$ we can substitute x or y instead of z and derive a similar equation in 2 variables.

which rearranges to a quadratic equation in 2 variables, with

$$\begin{array}{lll} A = a^2 - \frac{c^2}{k^2}, & B = 2ab, & C = b^2 - \frac{c^2}{k^2}, \\ D = 2ad, & E = 2bd, & F = d^2. \end{array} \tag{9.1}$$

In standard form, as in Theorems 9.3, 9.14 and 9.24, we note that the quadratic equation in 2 variables describing the ellipse and hyperbola have B, D and E equal to 0, whereas for a parabola, B, C, D and F equal 0.

9.1.2 Rotation of the coordinate system

In this subsection, we establish that via a change of coordinates (a rotation), we can eliminate the xy term in a quadratic equation in 2 variables. We achieve this by setting the transformed coefficient of $\tilde{x}\tilde{y}$, namely $\tilde{B}$, equal to 0. This quadratic equation, with no $\tilde{x}\tilde{y}$ term, can then be simplified further by a translation (change of origin of the coordinate system).

Consider a coordinate system which is rotated by an angle α, which has the following transformation formulae,

$$\left\{ \begin{array}{rcl} \tilde{x} &=& x\cos(\alpha) + y\sin(\alpha), \\ \tilde{y} &=& -x\sin(\alpha) + y\cos(\alpha), \end{array} \right. \qquad \left\{ \begin{array}{rcl} x &=& \tilde{x}\cos(\alpha) - \tilde{y}\sin(\alpha), \\ y &=& \tilde{x}\sin(\alpha) + \tilde{y}\cos(\alpha), \end{array} \right. \tag{9.2}$$

See Figure 9.2 for a geometric interpretation of the new coordinate system.

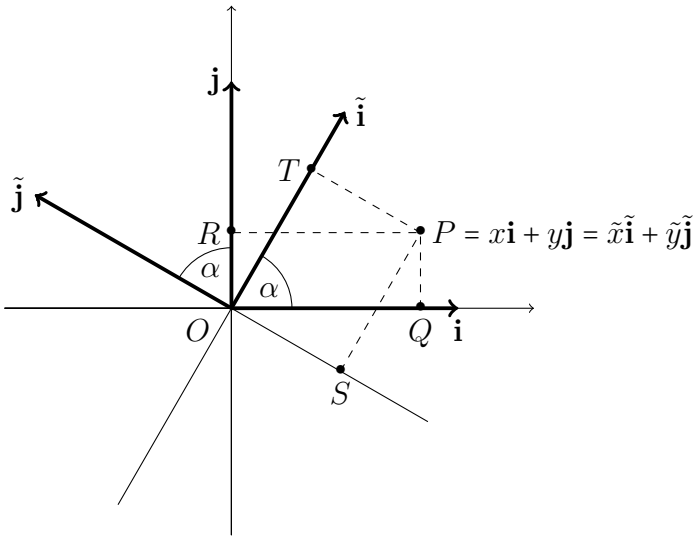


Figure 9.2: Sketch of the rotated coordinate system. Here, that $\vec{OQ} = x\mathbf{i}$, $\vec{OR} = y\mathbf{j}$, $\vec{OT} = \tilde{x}\mathbf{i}$ and $\vec{OS} = \tilde{y}\mathbf{j}$.

Note that the coordinate relations in (9.2) can be given as,

$$\begin{pmatrix}\tilde{x}\\ \tilde{y}\end{pmatrix} = \begin{pmatrix}\cos(\alpha) & \sin(\alpha)\\ -\sin(\alpha) & \cos(\alpha)\end{pmatrix}\cdot\begin{pmatrix}x\\ y\end{pmatrix}, \quad \begin{pmatrix}x\\ y\end{pmatrix} = \begin{pmatrix}\cos(\alpha) & -\sin(\alpha)\\ \sin(\alpha) & \cos(\alpha)\end{pmatrix}\cdot\begin{pmatrix}\tilde{x}\\ \tilde{y}\end{pmatrix} \tag{9.3}$$

The 2×2 matrices in (9.3) are referred to as **rotation matrices**.

Using (9.2), the quadratic equation in Definition 9.1 can be transformed into the equation

$$\begin{aligned}&A\left(\tilde{x}\cos(\alpha)-\tilde{y}\sin(\alpha)\right)^2 + B\left(\tilde{x}\cos(\alpha)-\tilde{y}\sin(\alpha)\right)\left(\tilde{x}\sin(\alpha)+\tilde{y}\cos(\alpha)\right) +\\ &C\left(\tilde{x}\sin(\alpha)+\tilde{y}\cos(\alpha)\right)^2 + D\left(\tilde{x}\cos(\alpha)-\tilde{y}\sin(\alpha)\right) +\\ &E\left(\tilde{x}\sin(\alpha)+\tilde{y}\cos(\alpha)\right) + F = 0,\end{aligned}$$

which can be rewritten as

$$\begin{aligned}&\left(A\cos^2(\alpha)+B\sin(\alpha)\cos(\alpha)+C\sin^2(\alpha)\right)\tilde{x}^2 +\\ &\left(-2A\sin(\alpha)\cos(\alpha)+B\cos^2(\alpha)-B\sin^2(\alpha)+2C\sin(\alpha)\cos(\alpha)\right)\tilde{x}\tilde{y} +\\ &\left(A\sin^2(\alpha))-B\sin(\alpha)\cos(\alpha)+C\cos^2(\alpha)\right)\tilde{y}^2 +\\ &\left(D\cos(\alpha)+E\sin(\alpha)\right)\tilde{x} +\\ &\left(-D\sin(\alpha)+E\cos(\alpha)\right)\tilde{y} + F = 0.\end{aligned}$$

The $\tilde{x}\tilde{y}$ term vanishes when

$$B\cos(2\alpha)+(C-A)\sin(2\alpha) = 0,$$

or when α satisfies one of the equations

$$\cot(2\alpha) = \frac{A-C}{B} \quad \text{or} \quad \tan(2\alpha) = \frac{B}{A-C}.$$

9.1.3 Translation of the coordinate system

If we have the quadratic equation in the form

$$\tilde{A}\tilde{x}^2+\tilde{C}\tilde{y}^2+\tilde{D}\tilde{x}+\tilde{E}\tilde{y}+\tilde{F} = 0, \tag{9.4}$$

we can eliminate the linear terms in $\tilde{x}$ and $\tilde{y}$ (assuming $\tilde{A},\tilde{C}\neq 0$) by either translating the coordinate system (or by completing the squares in the expression). The equation then becomes,

$$\tilde{A}\left(\tilde{x}+\frac{\tilde{D}}{2\tilde{A}}\right)^2+\tilde{C}\left(\tilde{y}+\frac{\tilde{E}}{2\tilde{C}}\right)^2+F-\frac{\tilde{D}^2}{4\tilde{A}}-\frac{\tilde{E}^2}{4\tilde{C}} = 0.$$

A translation of the coordinate system to coordinates $(\hat{x},\hat{y})$, such that the new origin is at

$$\hat{O}\left(-\frac{\tilde{D}}{2\tilde{A}},-\frac{\tilde{E}}{2\tilde{C}}\right),$$

i.e.

$$\begin{cases} \hat{x} &= \tilde{x} + \frac{\tilde{D}}{2\tilde{A}} \\ \hat{y} &= \tilde{y} + \frac{\tilde{E}}{2\tilde{C}}, \end{cases}$$

then yields a standard form of the quadratic equation in 2 variables:

$$\tilde{A}(\hat{x})^2 + \tilde{C}(\hat{y})^2 = -\Delta, \tag{9.5}$$

with

$$\Delta = F - \frac{\tilde{D}^2}{4\tilde{A}} - \frac{\tilde{E}^2}{4\tilde{C}}. \tag{9.6}$$

From (9.4)-(9.6) and the standard equations for conics which follow, we will classify solutions to quadratic equations in 2 variables (see end of the chapter).

9.2 Parabola

Definition 9.2. *A **parabola** is the set of all points in the plane that are equidistant from a given fixed point and fixed line in the plane, that do not intersect.*

- The fixed point is called the *focus* of the parabola.
- The fixed line is known as the *directrix* of the parabola.
- The distance from a point P to a line is the distance between P and the point of intersection between the line and the perpendicular to this line through P (this is the same as the distance from the point P to the nearest point on the line).

If we remove the words “that do not intersect" from Definition 9.2, then what shapes would be defined to be parabolas (see practice questions).

We will now derive the standard equation for a parabola, using a convenient coordinate system.

9.2.1 The equation of a parabola

Theorem 9.3. *The **standard equation** of a parabola with **focus** at the point F with coordinates $(0, p)$ and **directrix** $y = -p$, with $p > 0$, is given by*

$$x^2 = 4py.$$

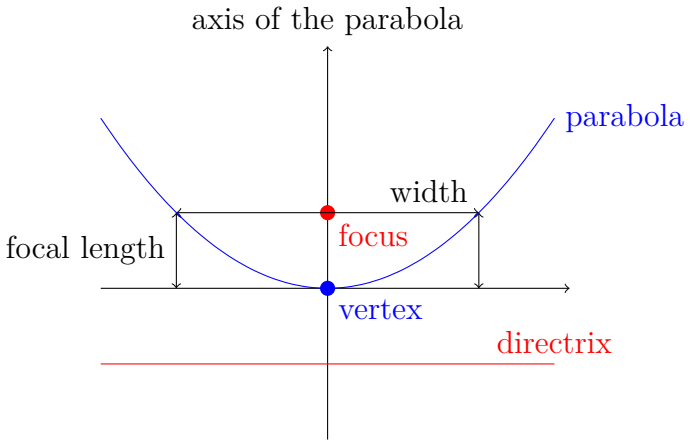


Figure 9.3: Sketch of main terms related to a parabola.

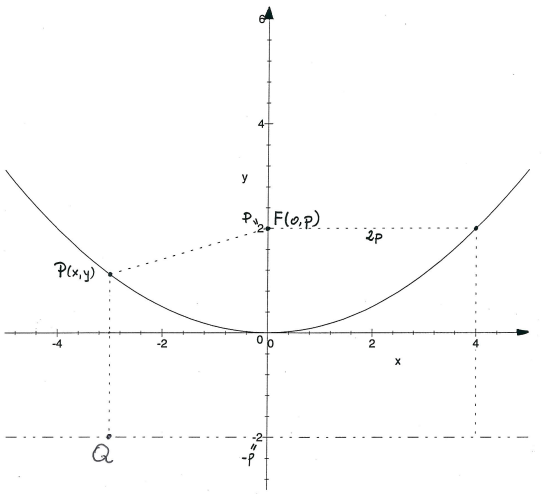


Figure 9.4: The standard equation of a parabola.

The parabola, focus and directrix are illustrated in Fig 9.3.

Proof: The distance from the point $P(x, y)$ to the focus $F(0, p)$ is then given by

$$|\vec{PF}| = \sqrt{x^2 + (y-p)^2}. \tag{9.7}$$

The distance from P to the directrix is given by the distance from P to the point $Q(x, -p)$ which is the intersection of the directrix with the perpendicular line through P. The distance from P to the directrix is thus given by

$$|\vec{PQ}| = \sqrt{(y+p)^2}. \tag{9.8}$$

The points on the parabola are then described by setting $|\vec{PF}| = |\vec{PQ}|$ or, via (9.7) and (9.8)

$$\begin{aligned} & \sqrt{x^2 + (y-p)^2} = \sqrt{(y+p)^2} \\ \iff \quad & x^2 + (y-p)^2 = (y+p)^2 \\ \iff \quad & x^2 + y^2 - 2py + p^2 = y^2 + 2py + p^2 \\ \iff \quad & x^2 = 4py. \end{aligned} \tag{9.9}$$

The equation in (9.9), is often referred to as the *standard equation* of a parabola and can be written as

$$y = \frac{x^2}{4p}. \tag{9.10}$$

The parabola in (9.10) is obviously symmetric with respect to the y-axis, which is called the *axis of the parabola*. The point where the parabola crosses its axis, midway between the focus and the directrix, is called the *vertex of the parabola*. Here, the parabola has its vertex at the origin. Notice that p is the distance from the vertex to the focus and is known as the *focal length*. The *width* of the parabola at the focus is equal to $4p$.

Example 134: The equation

$$y = \frac{x^2}{8},$$

represents a parabola with the y-axis as the axis of the parabola and focal length $p = 2$. It is depicted in Figure 9.5.

The focal length p dictates the shape of the parabola. When p is small, the parabola is narrow, but when p is large, the parabola opens up wide. This is also illustrated in Figure 9.5.

9.2.2 Parabolas with vertex at the origin

The equation for differently orientated parabolas can be obtained in the same way, starting from the definition. Alternatively, one can use a coordinate transformation to obtain the equation. Recall that the coordinates $(\tilde{x}, \tilde{y})$ in a coordinate system which is rotated anti-clockwise by an angle θ with respect to the standard coordinate system with coordinates (x, y) (see Figure 9.2) is linked to the original coordinates by the transformation formulae given in (9.2) and (9.3).

We now consider four special cases.

Axis equal to the positive x-axis

Theorem 9.4. *The equation of a parabola with focus at the point F with coordinates $(p, 0)$ and directrix $x = -p$, with $p > 0$, is given by*

$$y^2 = 4px.$$

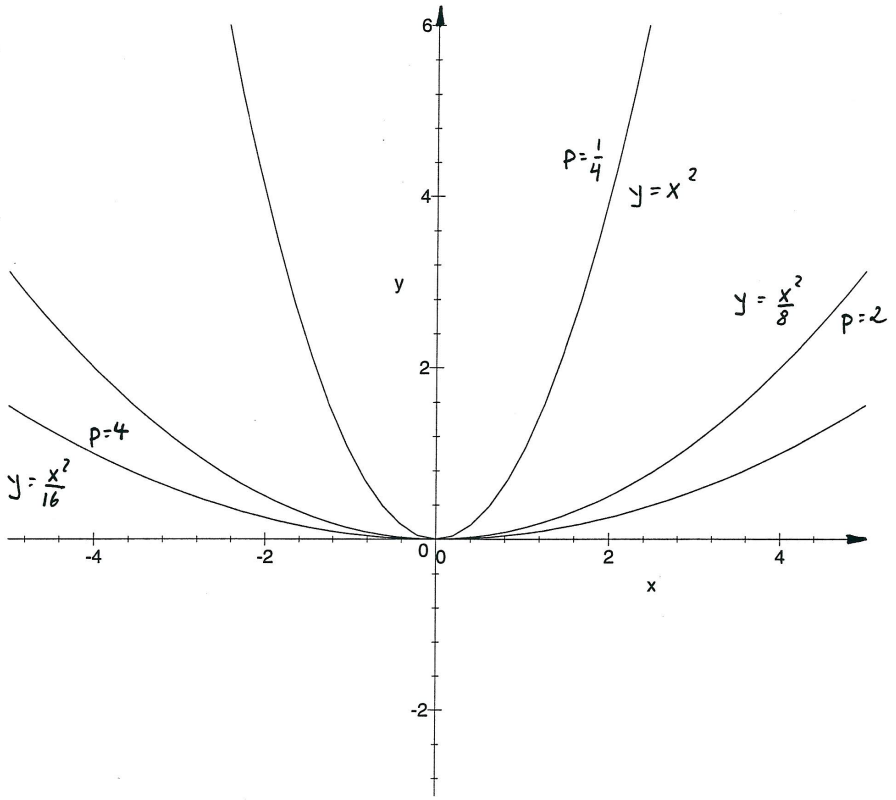


Figure 9.5: The parabolas with equations $y = x^2/8$, $y = x^2$ and $y = x^2/16$.

Proof: The graph of the parabola is identical to the graph of the parabola with standard equation in a coordinate system which is rotated by $\alpha = -\pi/2$. In this rotated coordinate system, the equation for the parabola is given by

$$\tilde{y} = \frac{\tilde{x}^2}{4p},$$

while the coordinates are linked by the transformation formula

$$\begin{cases} \tilde{x} &= x\cos(-\pi/2) + y\sin(-\pi/2) = -y, \\ \tilde{y} &= -x\sin(-\pi/2) + y\cos(-\pi/2) = x. \end{cases}$$

The equation in the original coordinate system is then given by

$$x = \frac{(-y)^2}{4p},$$

or equivalently,

$$y^2 = 4px, \tag{9.11}$$

as required.

Figure 9.6 depicts the parabola given in equation (9.11) with $p = 2$.

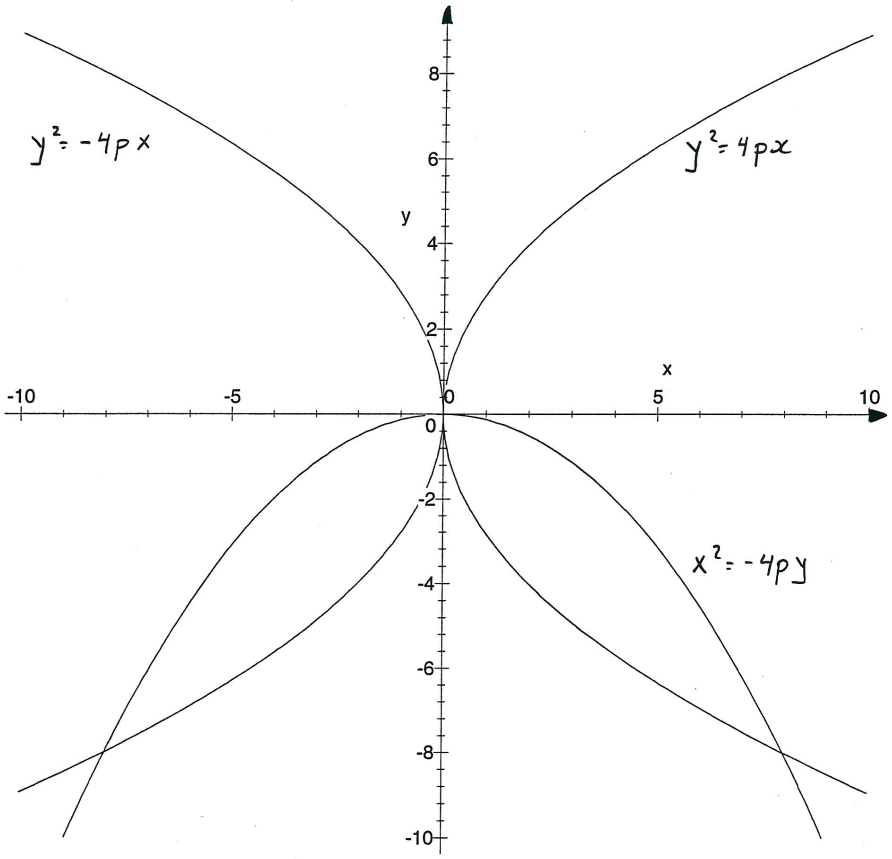


Figure 9.6: Parabolas from Theorems 9.4-9.6 plotted with $p = 2$.

Axis equal to the negative x-axis

Theorem 9.5. *The equation of a parabola with focus at the point F with coordinates $(-p, 0)$ and directrix $x = p$, with $p > 0$, is given by*

$$y^2 = -4px.$$

Proof: In this case, the standard equation can be used in a coordinate system which is rotated through an angle $\alpha = \pi/2$, leading to the transformation formula

$$\begin{cases} \tilde{x} &= x\cos(\pi/2) + y\sin(\pi/2) = y, \\ \tilde{y} &= -x\sin(\pi/2) + y\cos(\pi/2) = -x. \end{cases}$$

so that the equation in the original coordinate system is given by

$$-x = \frac{y^2}{4p},$$

or equivalently,

$$y^2 = -4px, \tag{9.12}$$

as required.

Figure 9.6 depicts the parabola given in (9.12) for $p = 2$.

Axis equal to the negative y-axis

Theorem 9.6. *The equation of a parabola with focus at the point F with coordinates $(0, -p)$ and directrix $y = p$, with $p > 0$, is given by*

$$x^2 = -4py.$$

Proof: We find the equation by considering a coordinate system rotated through an angle $\alpha = \pi$, so that

$$\begin{cases} \tilde{x} &= x\cos(\pi) + y\sin(\pi) = -x, \\ \tilde{y} &= -x\sin(\pi) + y\cos(\pi) = -y. \end{cases}$$

This leads to the equation

$$-y = \frac{(-x)^2}{4p},$$

or equivalently

$$y = -\frac{x^2}{4p}, \tag{9.13}$$

as required.

Again, the parabola given in (9.13) is plotted in Figure 9.6 for $p = 2$.

Axis equal to the bisectrix in the first quadrant

The **bisectrix** of the first quadrant (region where $x, y > 0$) is the line that divides the first quadrant in the xy-plane into 2 *isometric* (same shape) regions.

Theorem 9.7. *The equation of a parabola with vertex at $(0, 0)$ and directrix $y = -x - \sqrt{2}p$, with $p > 0$ is given by*

$$(x - y)^2 = 4p\sqrt{2}(x + y).$$

Proof: The equation can now be deduced by considering a coordinate system which is rotated through an angle $\alpha = -\pi/4$.

$$\begin{cases} \tilde{x} &= x\cos(-\pi/4) + y\sin(-\pi/4) = \frac{\sqrt{2}}{2}(x-y), \\ \tilde{y} &= -x\sin(-\pi/4) + y\cos(-\pi/4) = \frac{\sqrt{2}}{2}(x+y). \end{cases}$$

In the $(\tilde{x}, \tilde{y})$ coordinate system, the equation of the parabola is given by

$$4p\tilde{y} = \tilde{x}^2,$$

with p the focal distance, i.e. the distance from the origin to the point F, and with vertex at $(0,0)$. This yields the equation

$$\frac{\sqrt{2}}{2}(x+y) = \frac{1}{2}\frac{(x-y)^2}{4p},$$

or equivalently,

$$(x-y)^2 = 4p\sqrt{2}(x+y), \tag{9.14}$$

as required.

If we expand the equation in (9.14) we get

$$\frac{1}{8p}x^2 - \frac{1}{4p}xy + \frac{1}{8p}y^2 - \frac{\sqrt{2}}{2}x - \frac{\sqrt{2}}{2}y = 0.$$

This illustrates the fact that the equation of a parabola in an arbitrary position is not a simple relationship between x and y. Notice that this equation is quadratic in both x and y and does contain a term in xy! The parabola given in (9.14) is depicted in Figure 9.7 for $p = 0.5$.

9.2.3 Shifting a parabola

Theorem 9.8. *The equation of a parabola with axis parallel to the positive y-axis (and pointing in the same direction), vertex at the point P with coordinates (q, s) and focal length p, with $p > 0$, is given by*

$$(x-q)^2 = 4p(y-s).$$

Proof: Consider a parabola with equation

$$x^2 = 4py,$$

with vertex at the origin, focal length p and axis (of the parabola) given by the y-axis. The parabola we seek, has the same form,

$$\tilde{x}^2 = 4p\tilde{y}, \tag{9.15}$$

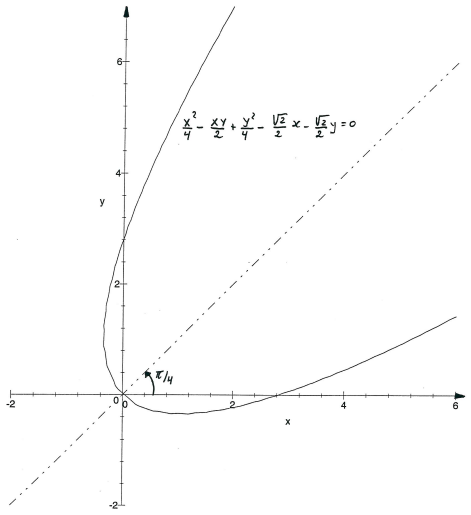


Figure 9.7: A parabola with axis $y = x$.

where the transformation formula is given a *translation*[2] ,

$$\begin{cases} \tilde{x} &= x - q, \\ \tilde{y} &= y - s. \end{cases} \tag{9.16}$$

Note that this transformation merely shifts the origin of the coordinate system without rotating or stretching the axes. Substituting (9.16) into (9.15) gives

$$(x - q)^2 = 4p(y - s), \tag{9.17}$$

as required.

Observe that we can rewrite (9.17) as

$$y = ax^2 - 2aqx + (aq^2 + s), \tag{9.18}$$

with $a = \frac{1}{4p}$ so that the RHS of (9.18) is a quadratic in x.

From (9.18), we can easily find the equation of a parabola with vertex in a given point, if the equation of a similar parabola with vertex at the origin is known. In particular, via Theorem 9.6, the parabola with vertex at $P(q, s)$ and axis parallel to the negative y-axis, is characterised by

$$y - s = -a(x - q)^2, \tag{9.19}$$

with $a = \frac{1}{4p}$. Similarly, via Theorem 9.4, a parabola with axis parallel to the positive x-axis (and pointing in the same direction) and with vertex at $P(q, s)$, is described by the equation

$$(y - s)^2 = 4p(x - q).$$

[2]The word translation here means a shift of any combination of left/right/up/down.

Moreover, via Theorem 9.5, a parabola that has the axis parallel to the negative x-axis and vertex at $P(q, s)$ is described by the equation

$$(y - s)^2 = -4p(x - q).$$

9.2.4 The graph with equation $y = ax^2 + bx + c$

We found in (9.18) and (9.19) that the equation of a parabola with vertical axis parallel to the y-axis can be written in the general form (with $a \neq 0$),

$$y = ax^2 + bx + c. \tag{9.20}$$

Let us now consider whether every equation of the form of equation (9.20) represents a parabola. Therefore, we rewrite the general equation (9.20) as

$$y = a\left(x + \frac{b}{2a}\right)^2 + \left(c - \frac{b^2}{4a}\right),$$

or

$$y - \left(c - \frac{b^2}{4a}\right) = a\left(x + \frac{b}{2a}\right)^2.$$

This can be recognised as the equation of a parabola with vertex at

$$P\left(-\frac{b}{2a}\ ,\ c - \frac{b^2}{4a}\right),$$

axis $x = -\frac{b}{2a}$, and focal length $p = \frac{1}{4a}$.

9.2.5 The tangent to a parabola

Consider the intersection of a line with equation

$$y = mx + q, \tag{9.21}$$

and the parabola with equation

$$4px = y^2. \tag{9.22}$$

The points of intersection are found by solving the nonlinear system of equations

$$\begin{cases} y &= mx + q, \\ 4px &= y^2. \end{cases}$$

When we use y in (9.21) and substitute it for y in (9.22), we find a quadratic equation for the x-coordinates of the intersection points:

$$m^2x^2 + (2mq - 4p)x + q^2 = 0.$$

The discriminant of this equation[3] is given by

$$D = (2mq - 4p)^2 - 4m^2q^2 = 16p(p - mq), \tag{9.23}$$

so that when $D > 0$ the line has two points of intersection. When $D < 0$, the line does not intersect the parabola, and when $D = 0$, the two points of intersection coincide. In this case, we say that the line is a *tangent*[4] to the parabola at the specified point.

Let us now assume that $D = 0$ or equivalently via (9.23) that $p = mq$, and that the line $y = mx + q$ is the tangent to the parabola at the point $P(x_1, y_1)$. Since P is on the parabola, $4px_1 = y_1^2$, and

$$\begin{aligned}
& y_1 - mx_1 - q = 0 \\
\iff \quad & 4py_1 - my_1^2 - 4pq = 0 \\
\iff \quad & 4py_1 - \frac{p}{q}y_1^2 - 4pq = 0 \\
\iff \quad & 4pq^2 - 4qpy_1 + py_1^2 = 0 \\
\iff \quad & p\left(2q - y_1\right)^2 = 0 \\
\iff \quad & q = \frac{y_1}{2}.
\end{aligned} \tag{9.24}$$

It then follows from (9.24) that (for $y_1 \neq 0$)

$$m = \frac{p}{q} = \frac{2p}{y_1}.$$

The equation of the tangent is thus given by (for $y_1 \neq 0$)

$$y = \frac{2p}{y_1}x + \frac{y_1}{2}, \tag{9.25}$$

which can be rewritten as

$$y_1 y = 2px + \frac{y_1^2}{2},$$

or equivalently[5]

$$y_1 y = 2p\left(x + x_1\right). \tag{9.26}$$

It is now straightforward to write the equation of the normal to the parabola at a point. Since the normal is perpendicular to the tangent, its gradient is given by

$$m = -\frac{y_1}{2p},$$

[3] This only exists if $m \neq 0$. If $m = 0$ we have a line parallel to the axis of the parabola that intersects/crosses the parabola once at (x_1, y_1) which is not the tangent line to the parabola at (x_1, y_1).

[4] Compare this approach to finding a tangent to a general curve (see [15] or [14]). You should consider why the approach we adopt here works? Additionally, note that one can find the tangent in specific problems via implicit differentiation with more ease, however, to prove properties satisfied by parabola, this idea of a tangent line is helpful.

[5] Note that this equation holds for any $y_1 \in \mathbb{R}$, since if $y_1 = 0$, $x = 0$ is in fact the tangent to the parabola

so that the equation of the normal through a point $P(x_1, y_1)$ on the parabola is given by

$$y - y_1 = -\frac{y_1}{2p}(x - x_1).$$

Figure 9.8 depicts the parabola $3x = y^2$ and the tangent and normal at the point $P(2, 2\sqrt{a})$ on the parabola.

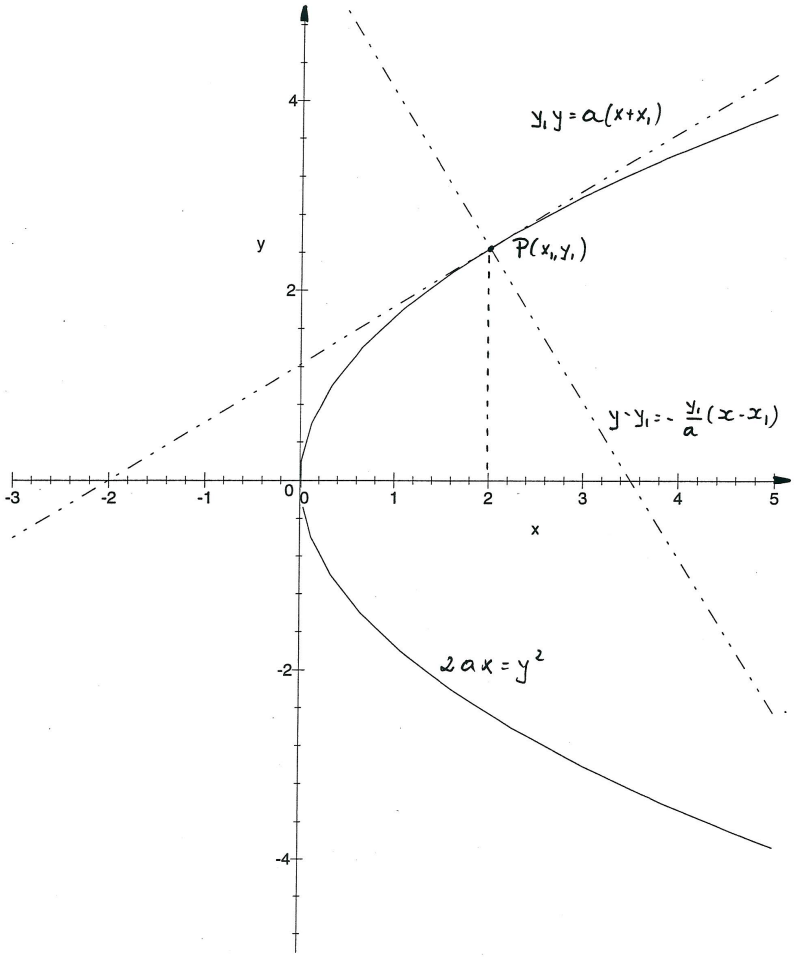


Figure 9.8: Tangent and normal to a parabola at $P(2, 2\sqrt{a})$.

9.2.6 Reflective property

Theorem 9.9. *The tangent and the normal at a point of a parabola are the bisectors of the angles defined by the line PF, with F the focus, and the line through P parallel to the axis of the parabola.*

This is illustrated in Figure 9.9. Note that the bisector of an angle is the line which divides an angle into two equal angles.

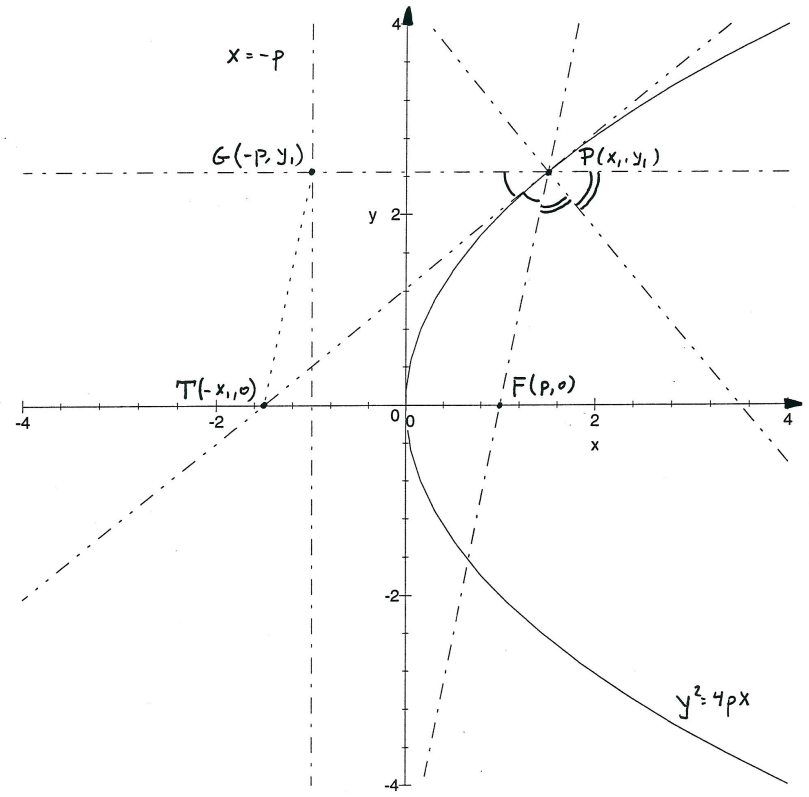


Figure 9.9: Reflective property of a parabola.

Proof: To prove this property, we first define the point G as the intersection of the directrix of the parabola and the line parallel to the axis of the parabola through P. The point T is defined as the intersection of the tangent at P and the axis of the parabola. When we can prove that the quadrangle $PFTG$ is a rhombus, then we know that its diagonal PT bisects the angle between the sides PF and PG. It then also follows that the normal bisects the adjacent angle.

So all we have to do is to prove that $PFTG$ is a rhombus. By construction, we know that $\vec{TF}$ is parallel to $\vec{PG}$. When we can show that $|\vec{TF}| = |\vec{GP}|$, then we have shown that the quadrangle $PFTG$ is a parallelogram. First choose a coordinate system in which the origin is located at the vertex of the parabola and the x-axis coincides with the axis of the parabola. The equation of the parabola is then given by

$$y^2 = 4px,$$

with focus at $F(p, 0)$. The directrix is given by $x = -p$. We assume that the point P has coordinates $P(x_1, y_1)$. The point G then has coordinates $G(-p, y_1)$ since it is on the

directrix and the line PG is parallel to the x-axis. The coordinates of T can be found as the intersection of the axis, $y = 0$, with the tangent in P, namely in (9.26)

$$y_1 y = 2p(x + x_1),$$

so $T(-x_1, 0)$. The distances $|\vec{TF}|$ and $|\vec{GP}|$ are then given by

$$|\vec{TF}| = |\, p + x_1 \,|, \tag{9.27}$$
$$|\vec{GP}| = |\, -p - x_1 \,| = |\, p + x_1 \,|\,. \tag{9.28}$$

Identities (9.27) and (9.28) prove that $|\vec{TF}| = |\vec{GP}|$, or that the quadrangle $PFTG$ is a parallelogram. It will be a rhombus when in addition

$$|\vec{PG}| = |\vec{PF}|.$$

This equality follows from the fact that P is a point of the parabola with focus at F (see Definition 9.2). Indeed, $|\vec{PG}|$ is the distance from P to the directrix and $|\vec{PF}|$ is the distance from P to the focal point.

Therefore, we have proved that the quadrangle $PFTG$ is a rhombus. It follows that the tangent PT is the bisector of the inner angle in P defined by PF and PG. Since the normal is perpendicular to the tangent, it must also bisect the adjacent angle, which completes the proof, as required.

A consequence of Theorem 9.9, the reflection property, is that any ray of light which falls onto a parabolic mirror (see Figure 9.8) parallel to its axis will be reflected into the focal point.

9.2.7 Alternative equations

Theorem 9.10. ***Parametric equations*** *for a parabola with focus at the point F with coordinates $(0, p)$ and directrix $y = -p$, with $p > 0$, are*

$$\begin{cases} x = 2pt, \\ y = pt^2, \end{cases} \qquad t \in \mathbb{R}$$

Proof: This can easily be verified by substituting these expressions in the standard equation in Theorem 9.3.

When using polar coordinates, the following two results are useful.

Theorem 9.11. *The* ***polar equation*** *of a parabola with focus at $O(0,0)$ and directrix $y = -p$, with $p > 0$, is*

$$r(\theta) = \frac{p}{1 - \sin(\theta)}.$$

Proof: Given the focus and directrix of the parabola are at $(0,0)$ and on $y=-p$ respectively, it follows that the parabola has vertex at $(0,-\frac{p}{2})$ and focal length $\frac{p}{2}$. Since the axis of the parabola is parallel to (and in the same direction as) the y-axis, via Theorem 9.8, the equation of the parabola is

$$x^2 = 4\left(\frac{p}{2}\right)\left(y+\frac{p}{2}\right) \iff x^2 - 2py - p^2 = 0. \tag{9.29}$$

Substituting $x = r\cos(\theta)$ and $y = r\sin(\theta)$ into (9.29) gives,

$$\cos^2(\theta)r^2 - 2p\sin(\theta)r - p^2 = 0.$$

The solutions to the quadratic equation (in r) are given by

$$\begin{aligned} r &= \frac{2p\sin(\theta) \pm \sqrt{4p^2\sin^2(\theta) + 4p^2\cos^2(\theta)}}{2\cos^2(\theta)} \\ &= \frac{p(\sin(\theta)\pm 1)}{\cos^2(\theta)}. \end{aligned} \tag{9.30}$$

Since the polar coordinate r is non-negative, this implies that we must add the second term on the RHS of (9.30) and hence,

$$\begin{aligned} r &= \frac{p(\sin(\theta)+1)}{\cos^2(\theta)} \\ &= \frac{p(\sin(\theta)+1)}{1-\sin^2(\theta)} \\ &= \frac{p(\sin(\theta)+1)}{(1-\sin(\theta))(1+\sin(\theta))} \\ &= \frac{p}{1-\sin(\theta)}. \end{aligned}$$

Note that although we can only cancel terms in the last step of the calculation when $\theta \neq -\frac{\pi}{2}$ (to avoid dividing by zero), the result still holds when $\theta = -\frac{\pi}{2}$, as required.

Theorem 9.12. *The **polar equation** of a parabola with focus at $O(0,0)$ and directrix $x=-p$, with $p>0$, is*

$$r(\theta) = \frac{p}{1-\cos(\theta)}.$$

Proof: As in Theorem 9.4, the graph of the parabola is identical to the graph of the parabola with standard equation (given in polar form in Theorem 9.11) in a coordinate system which is rotated by $\alpha = -\pi/2$. In the rotated coordinate system with $\tilde{r} = r$ and $\tilde{\theta} = \theta + \frac{\pi}{2}$, it follows from Theorem 9.11 that the equation of the parabola is given by

$$\tilde{r} = \frac{p}{1-\sin(\tilde{\theta})} \iff r = \frac{p}{1-\sin(\theta+\frac{\pi}{2})} = \frac{p}{1-\cos(\theta)},$$

as required.

9.3 Ellipse

Definition 9.13. *An ellipse is the set of points in a plane whose distances from two distinct fixed points F_1 and F_2 in the plane have a constant sum $2a$ with $2a > |\vec{F_1F_2}|$.*

- The two fixed points are the *foci* of the ellipse.
- The midpoint between the two foci is called the *centre* of the ellipse.
- The line through the foci of an ellipse is the *focal axis* (or *major axis*) of the ellipse. The line perpendicular to the focal axis through the centre of the ellipse is the *minor axis*[6].
- The points where the focal axis intersects the ellipse are the *vertices* of the ellipse.

An ellipse can easily be drawn using a piece of string fixed at the two foci. A proper choice of the coordinate system will lead to the standard equation of an ellipse. We will also consider how the standard equation changes in a translated or rotated coordinate system. We will derive the equation for the tangent and normal at a point of the ellipse and discuss the reflective properties of ellipses.

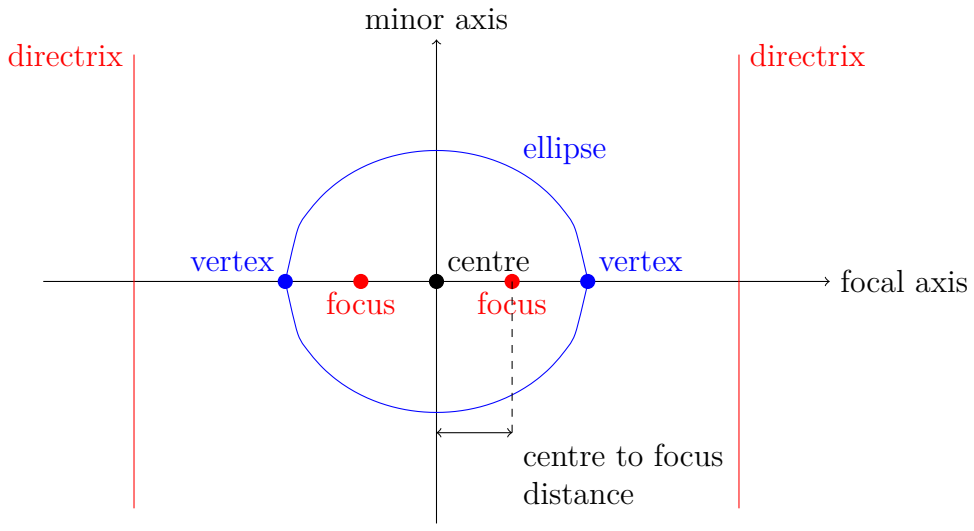


Figure 9.10: Sketch of main terms related to a ellipse.

[6]Sometimes, we refer to the major axis of an ellipse as the distance between the 2 intersection points of the ellipse and its 'major axis'. Similarly, we sometimes refer to the minor axis of an ellipse as the distance between the 2 intersection points of the ellipse and the ' minor axis'.

9.3.1 The standard equation of an ellipse

Theorem 9.14. *The **standard equation** of an ellipse with foci at $F_1(c,0)$ and $F_2(-c,0)$, with $c>0$, is given by*

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1,$$

where $2a=|\vec{PF_1}|+|\vec{PF_2}|$ (the sum of the distances of a point on the ellipse to both foci) and $b^2=a^2-c^2>0$.

This implies that the centre of the in ellipse in standard form is located at the origin. This ellipse is illustrated in Figure 9.11.

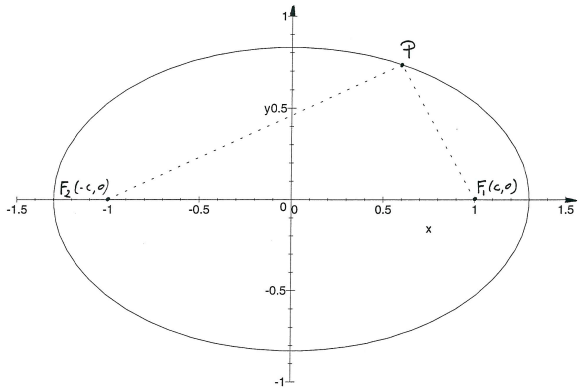


Figure 9.11: The standard equation of the ellipse.

Proof: Since $|\vec{F_1F_2}|=2c$ and $2a=|\vec{PF_1}|+|\vec{PF_2}|$, it follows that $a>c$. A point $P(x,y)$ is on the ellipse (defined by F_1, F_2 and a) when the sum of the distances $|\vec{PF_1}|$ and $|\vec{PF_2}|$ is $2a$, or equivalently,

$$|\vec{PF_1}|+|\vec{PF_2}|=2a$$

$$\iff \sqrt{(x-c)^2+y^2}+\sqrt{((x+c)^2+y^2}=2a \tag{9.31}$$

$$\iff (x-c)^2+y^2+2\sqrt{((x-c)^2+y^2)\,((x+c)^2+y^2)}$$
$$+(x+c)^2+y^2=4a^2. \tag{9.32}$$

Note that (9.32) is equivalent to (9.31) since both sides of equation (9.31) are positive. Now, equation (9.32) can be rewritten as,

$$2\left(x^2+c^2+y^2\right)+2\sqrt{\left((x-c)^2+y^2\right)\left((x+c)^2+y^2\right)}=4a^2$$

$$\iff \sqrt{\left((x-c)^2+y^2\right)\left((x+c)^2+y^2\right)}= 2a^2-\left(x^2+c^2+y^2\right)$$

$$\implies \left((x-c)^2+y^2\right)\left((x+c)^2+y^2\right)= 4a^4-4a^2\left(x^2+y^2+c^2\right)+\left(x^2+c^2+y^2\right)^2 \tag{9.33}$$

$$\iff (x-c)^2(x+c)^2+\left((x-c)^2+(x+c)^2\right)y^2+y^4= 4a^4-4a^2\left(x^2+y^2+c^2\right)+x^4+y^4+c^4+2x^2y^2+2x^2c^2+2c^2y^2$$

$$\iff x^4-2x^2c^2+c^4+2x^2y^2+2c^2y^2+y^4= 4a^4-4a^2\left(x^2+y^2+c^2\right)+x^4+y^4+c^4+2x^2y^2+2x^2c^2+2c^2y^2$$

$$\iff 4a^4-4a^2\left(x^2+y^2+c^2\right)+4x^2c^2=0$$

$$\iff \left(a^2-c^2\right)x^2+a^2y^2=a^2\left(a^2-c^2\right). \tag{9.34}$$

Since

$$b^2=a^2-c^2>0, \tag{9.35}$$

we can express equation (9.34) as

$$b^2x^2+a^2y^2=a^2b^2,$$

or equivalently,

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1. \tag{9.36}$$

Notice that we have only shown that all points, for which the sum of their distance to the two foci is constant, satisfy equation (9.36) (see the step in (9.33)). Before we can state that equation (9.36) is the standard equation of an ellipse, we need to show that all points satisfying this equation indeed are on the ellipse. Therefore, assume that the point $Q(x_1,y_1)$ satisfies (9.36), i.e.

$$\frac{x_1^2}{a^2}+\frac{y_1^2}{b^2}=1, \tag{9.37}$$

then

$$|\vec{QF_1}| = \sqrt{(x_1-c)^2+y_1^2}$$

$$= \sqrt{x_1^2 - 2cx_1 + c^2 + b^2 - \frac{b^2x_1^2}{a^2}} \quad \text{(via (9.37))}$$

$$= \sqrt{\frac{a^2-b^2}{a^2}x_1^2 - 2cx_1 + c^2 + b^2}$$

$$= \sqrt{\frac{c^2}{a^2}x_1^2 - 2cx_1 + a^2} \quad \text{(via (9.35))}$$

$$= \sqrt{\left(\frac{cx_1}{a} - a\right)^2}. \tag{9.38}$$

Since $a > c$ and $-a \leq x_1 \leq a$ (since $Q(x_1, y_1)$ satisfies (9.36)), it follows that

$$-a < \frac{cx_1}{a} < a, \tag{9.39}$$

and therefore,

$$|\vec{QF_1}| = a - \frac{cx_1}{a}. \tag{9.40}$$

Similarly, the distance from Q to the second focus is given by

$$|\vec{QF_2}| = \sqrt{\frac{c^2x_1^2}{a^2} + 2cx_1 + a^2} = \sqrt{\left(\frac{cx_1}{a} + a\right)^2}. \tag{9.41}$$

Thus, via (9.39),

$$|\vec{QF_2}| = a + \frac{cx_1}{a}. \tag{9.42}$$

Finally, via (9.40) and (9.42)

$$|\vec{QF_1}| + |\vec{QF_2}| = a - \frac{cx_1}{a} + a + \frac{cx_1}{a} = 2a,$$

which shows that any point, at which the coordinates satisfy (9.36) must be on the ellipse, as required.

We therefore call equation (9.36) the *standard equation* of the ellipse. This ellipse intersects the y-axis at the points $B_1(0, b)$ and $B_2(0, -b)$, as for $x = 0$, we have that

$$\frac{y^2}{b^2} = 1 \iff y = \pm b.$$

The number c is the *centre-to-focus distance* of the ellipse.

Example 135: The ellipse with equation

$$\frac{x^2}{9} + \frac{y^2}{4} = 1,$$

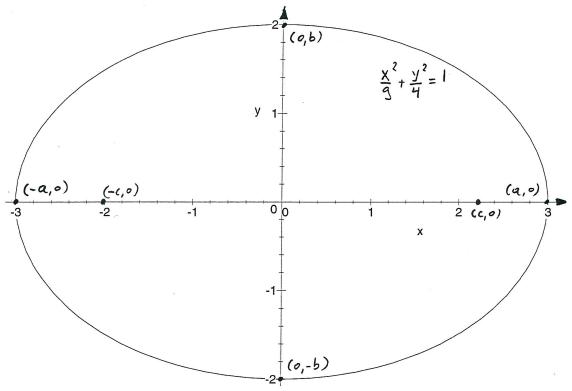


Figure 9.12: The ellipse with equation $\frac{x^2}{9} + \frac{y^2}{4} = 1$.

is depicted in Figure 9.12.

In this example notice that $c < a$ and $b < a$. If we remove the condition in Definition 9.13 that F_1 and F_2 are distinct, then we can show that if $c = 0$ (foci at the same point) and $a > 0$ then $b = a$ and we can obtain the equation of a circle in standard form (see practice questions).

9.3.2 Ellipse with a vertical major axis

Theorem 9.15. *The equation of an ellipse with foci at $F_1(0, c)$ and $F_2(0, -c)$, where $c > 0$, is given by*

$$\frac{x^2}{b^2} + \frac{y^2}{a^2} = 1,$$

with a and b as in Theorem 9.14.

Proof: To obtain the standard equation of an ellipse with its foci on the y-axis, we write the standard equation in Theorem 9.14 in a coordinate system which is rotated by $\alpha = \pi/2$:

$$\frac{\tilde{x}^2}{a^2} + \frac{\tilde{y}^2}{b^2} = 1,$$

with

$$\begin{cases} \tilde{x} &= x\cos(\pi/2) + y\sin(\pi/2) = y, \\ \tilde{y} &= -x\sin(\pi/2) + y\cos(\pi/2) = -x. \end{cases}$$

This yields the standard equation in the original coordinates,

$$\frac{x^2}{b^2} + \frac{y^2}{a^2} = 1.$$

Now the vertices have coordinates $(0, a)$ and $(0, -a)$. The endpoints of the minor axis have coordinates $(-b, 0)$ and $(b, 0)$. The foci are located at $F_1(0, c)$ and $F_2(0, -c)$ with $c^2 = a^2 - b^2$, as required.

Example 136: The ellipse with equation

$$x^2 + \frac{y^2}{4} = 1,$$

is plotted as in Figure 9.13.

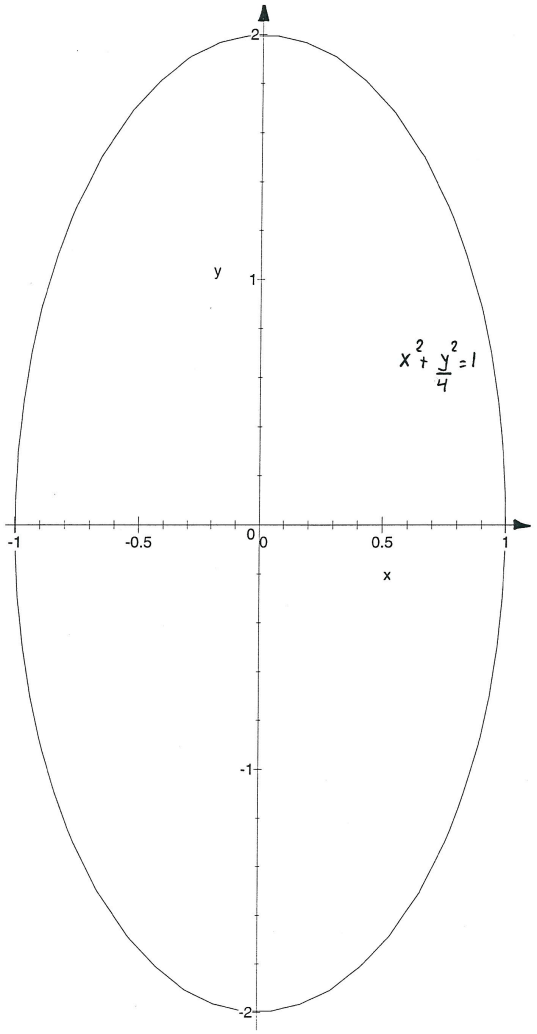


Figure 9.13: The ellipse with equation $x^2 + \frac{y^2}{4} = 1$.

One can recognise the orientation of the major axis by identifying the smallest coefficient in the equation: when the coefficient of x^2 is smallest, the major axis is horizontal.

When the coefficient of y^2 is smallest, the major axis is vertical. Of course both of these statements only apply if there is no xy term present in the equation defining the ellipse.

9.3.3 Ellipse with centre at $P(q, s)$

Theorem 9.16. *The equation of the ellipse with centre at $P(q, s)$, with centre-to-focus distance $c > 0$ and with major axis parallel to the x-axis, is given by*

$$\frac{(x-q)^2}{a^2} + \frac{(y-s)^2}{b^2} = 1,$$

with a and b as in Theorem 9.14.

Proof: To obtain the equation of an ellipse with centre at a given point $P(q, s)$, we consider the standard equation in a shifted coordinate system with origin at $P(q, s)$:

$$\frac{\tilde{x}^2}{a^2} + \frac{\tilde{y}^2}{b^2} = 1,$$

with

$$\begin{cases} \tilde{x} &= x - q, \\ \tilde{y} &= y - s. \end{cases}$$

Substituting the expressions for $\tilde{x}$ and $\tilde{y}$ yields the equation for the ellipse in the original coordinates,

$$\frac{(x-q)^2}{a^2} + \frac{(y-s)^2}{b^2} = 1,$$

as required.

Example 137: Figure 9.14 depicts the ellipse with equation

$$\frac{(x-1.3)^2}{3} + \frac{(y-2.2)^2}{1.8} = 1. \tag{9.43}$$

This ellipse has centre at $P(1.3, 2.2)$, and the major axis is horizontal.

We can rewrite the equation of an ellipse with centre at $P(q, s)$ as

$$b^2x^2 - 2b^2qx + a^2y^2 - 2a^2sy + \left(b^2q^2 + a^2s^2 - a^2b^2\right) = 0,$$

which is of the form

$$Ax^2 + Cy^2 + Dx + Ey + F = 0 \tag{9.44}$$

(the letter B is associated with the xy-term, so here $B = 0$). An equation of the form in (9.44) can be reduced to the standard equation for an ellipse by completing the squares in x and y.

The ellipse defined by equation (9.43) and plotted in Figure 9.14, can equivalently be defined by the equation,

$$1.8x^2 - 4.68x + 3y^2 - 13.2y + 12.162 = 0.$$

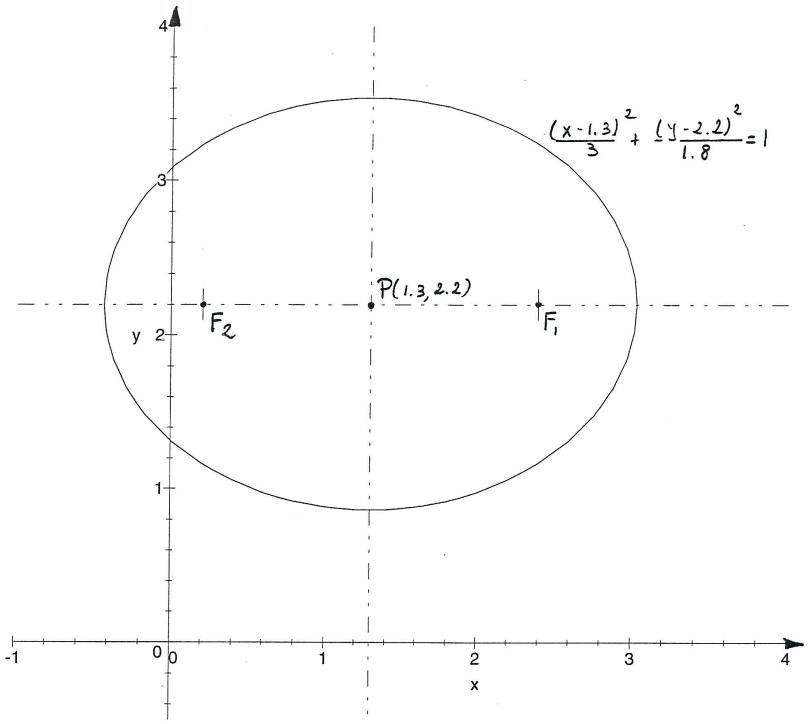


Figure 9.14: The ellipse with centre at $P(1.3, 2.2)$.

9.3.4 Ellipse with a tilted major axis

The equation of an ellipse with major axis neither horizontal or vertical can be obtained by writing the standard equation in a rotated coordinate system. Consider the coordinate system rotated by an angle α. The standard equation in this system is

$$\frac{\tilde{x}^2}{a^2} + \frac{\tilde{y}^2}{b^2} = 1,$$

with

$$\begin{cases} \tilde{x} &= x\cos(\alpha) + y\sin(\alpha), \\ \tilde{y} &= -x\sin(\alpha) + y\cos(\alpha). \end{cases}$$

The equation for the ellipse in the original coordinate system is

$$\frac{(x\cos(\alpha) + y\sin(\alpha))^2}{a^2} + \frac{(-x\sin(\alpha) + y\cos(\alpha))^2}{b^2} = 1,$$

which can be rewritten as

$$\left(\frac{\cos^2(\alpha)}{a^2} + \frac{\sin^2(\alpha)}{b^2}\right)x^2 + \left(\frac{\sin^2(\alpha)}{a^2} + \frac{\cos^2(\alpha)}{b^2}\right)y^2 +$$

$$2\cos(\alpha)\sin(\alpha)\left(\frac{1}{a^2}-\frac{1}{b^2}\right)xy = 1. \tag{9.45}$$

The equation defining the rotated ellipse typically gains a term in xy (see (9.45)) when the major axis is not parallel to the x-axis or the y-axis.

Example 138: The ellipse with major axis rotated anti-clockwise (from standard form) by $\alpha = \pi/6$ radians, with focal length $c = 2$, and length of the major axis equal to 6 (or $a = 3$), can be written as

$$\begin{aligned}
1 &= \left(\frac{\cos^2(\pi/6)}{9}+\frac{\sin^2(\pi/6)}{5}\right)x^2+\left(\frac{\sin^2(\pi/6)}{9}+\frac{\cos^2(\pi/6)}{5}\right)y^2+ \\
&\quad 2\cos(\pi/6)\sin(\pi/6)\left(\frac{1}{9}-\frac{1}{5}\right)xy, \\
&= \left(\frac{3/4}{9}+\frac{1/4}{5}\right)x^2+\left(\frac{1/4}{9}+\frac{3/4}{5}\right)y^2+ \\
&\quad 2\left(\frac{\sqrt{(}3)}{2}\right)\left(\frac{1}{2}\right)\left(-\frac{4}{45}\right)xy, \\
&= \frac{2x^2}{15}+\frac{8}{45}y^2-\frac{2\sqrt{3}xy}{45}.
\end{aligned}$$

This ellipse is depicted in Figure 9.15.

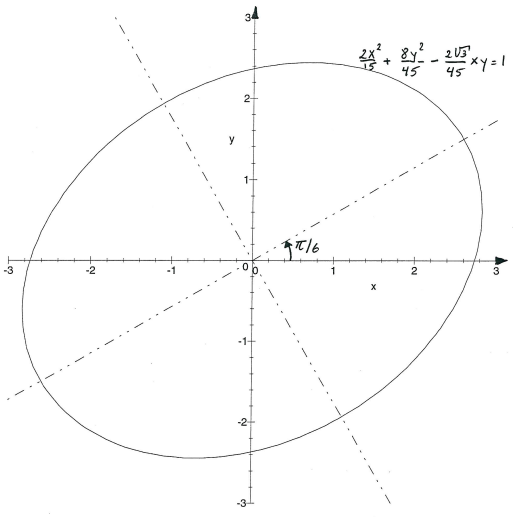


Figure 9.15: The ellipse with tilted major axis.

9.3.5 Eccentricity

Definition 9.17. *The* ***eccentricity*** *of an ellipse (when expressed in a coordinate system in standard form) is defined to be,*

$$e = \frac{c}{a} = \frac{\sqrt{a^2 - b^2}}{a}.$$

An ellipse can be almost circular or very flat. Its shape is controlled by the *eccentricity* e of the ellipse. From Definition 9.17 and Definition 9.13, it is clear that $0 < e < 1$. For very small eccentricity, $a \approx b$ and the ellipse is "close to" a circle. For an eccentricity close to 1 ($b \ll a$) the ellipse is very flat.

Figure 9.16 illustrates this by depicting two ellipses with different eccentricities.

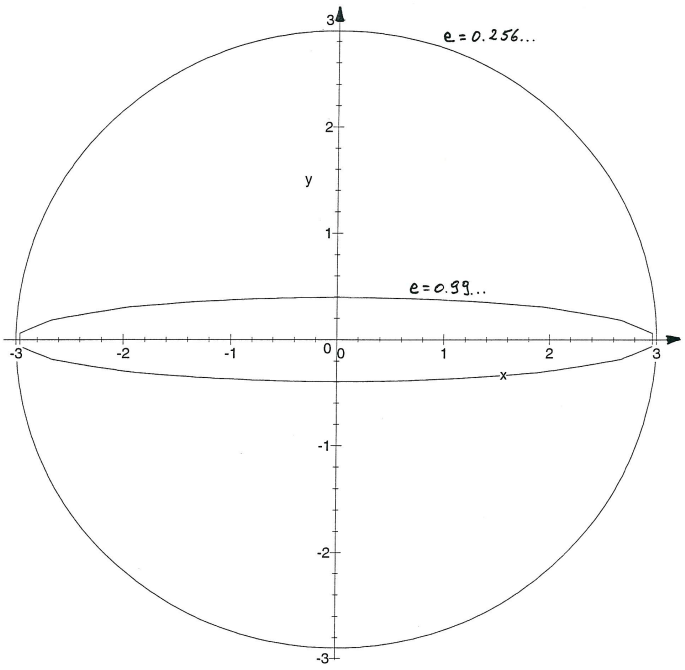


Figure 9.16: Two ellipses with eccentricities $e \approx 0.99$ and $e \approx 0.26$ respectively.

The first ellipse is given by

$$\frac{x^2}{9} + \frac{y^2}{8.41} = 1,$$

and has eccentricity

$$e = \frac{\sqrt{9 - 8.41}}{3} = 0.256\ldots.$$

The second ellipse has as equation

$$\frac{x^2}{9} + \frac{y^2}{0.16} = 1,$$

with eccentricity

$$e = \frac{\sqrt{9 - 0.16}}{3} = 0.991\ldots.$$

9.3.6 The tangent to an ellipse

Again, we will study the intersection of a line with the ellipse. Consider the line with equation

$$y = mx + q, \tag{9.46}$$

and the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1. \tag{9.47}$$

The x-coordinate of an intersection point between the line in (9.46) and the ellipse in (9.47) can be found by substituting y in (9.46) into (9.47):

$$\frac{x^2}{a^2} + \frac{(mx+q)^2}{b^2} = 1 \Leftrightarrow \left(b^2 + a^2m^2\right)x^2 + 2mqa^2x + a^2q^2 - a^2b^2 = 0. \tag{9.48}$$

The discriminant of the quadratic equation[7] in (9.48) (in x) is given by

$$D = 4m^2q^2a^4 - 4\left(a^2q^2 - a^2b^2\right)\left(b^2 + m^2a^2\right) = -4a^2b^2\left(q^2 - b^2 - a^2m^2\right).$$

Again, $D > 0$ implies that there are two possible points of intersection, and $D < 0$ implies that there are no points of intersection. To obtain a tangent to the ellipse (at a point of intersection) we must have $D = 0$, i.e.

$$q^2 = b^2 + a^2m^2. \tag{9.49}$$

We want to write the gradient of the tangent m as a function of known quantities. Assume the point of intersection is $P(x_1, y_1)$ with $y_1 \neq 0$. Then via (9.46),

$$y_1 - mx_1 = q. \tag{9.50}$$

Squaring both sides of (9.50) and substituting q^2 in (9.49), yields

$$a^2m^2 + b^2 - m^2x_1^2 + 2mx_1y_1 - y_1^2 = 0$$

$$\Longleftrightarrow \quad \left(a^2 - x_1^2\right)m^2 + 2x_1y_1m + b^2 - y_1^2 = 0. \tag{9.51}$$

[7] Note that the discriminant always exists since $b^2 + a^2m^2 > 0$.

The quadratic equation in (9.51) (in m) has only one solution since its discriminant is given by

$$\begin{aligned} 4x_1^2y_1^2 - 4\left(b^2 - y_1^2\right)\left(a^2 - x_1^2\right) &= 4b^2x_1^2 + 4a^2y_1^2 - 4a^2b^2 \\ &= 4a^2b^2\left(\frac{x_1^2}{a^2} + \frac{y_1^2}{b^2} - 1\right) \\ &= 0 \quad \text{(via (9.47))}, \end{aligned}$$

since $P(x_1, y_1)$ lies on the ellipse. Therefore the gradient of the tangent at $P(x_1, y_1)$ is given by

$$m = \frac{-2x_1y_1}{2\left(a^2 - x_1^2\right)} = \frac{-x_1y_1b^2}{a^2y_1^2} = -\frac{b^2x_1}{a^2y_1}. \tag{9.52}$$

The equation of the tangent to the ellipse at (x_1, y_1), via (9.52) and (9.46) is

$$\begin{aligned} & y - y_1 = -\frac{b^2x_1}{a^2y_1}(x - x_1) \\ \iff \quad & \frac{y_1}{b^2}(y - y_1) = -\frac{x_1}{a^2}(x - x_1) \\ \iff \quad & \frac{y_1y}{b^2} + \frac{x_1x}{a^2} = \frac{y_1^2}{b^2} + \frac{x_1^2}{a^2} \\ \iff \quad & \frac{x_1x}{a^2} + \frac{y_1y}{b^2} = 1 \quad \text{(via (9.47))}. \end{aligned} \tag{9.53}$$

Therefore, the tangent line to the ellipse in (9.47) at $P(x_1, y_1)$ on the ellipse is given by (9.53), where in fact, $y_1 = 0$ is now allowed i.e. if $(x_1, y_1) = (\pm a, 0)$ then the tangent to the ellipse is $x = \pm a$.

Example 139: Consider the point $P(3/2, \sqrt{3})$ on the ellipse with equation

$$\frac{x^2}{9} + \frac{y^2}{4} = 1.$$

The tangent to the ellipse at P is then given by

$$\frac{(3/2)x}{9} + \frac{\sqrt{3}y}{4} = 1,$$

via (9.53), or equivalently,

$$2x + 3\sqrt{3}y - 12 = 0.$$

This is illustrated in Figure 9.17.

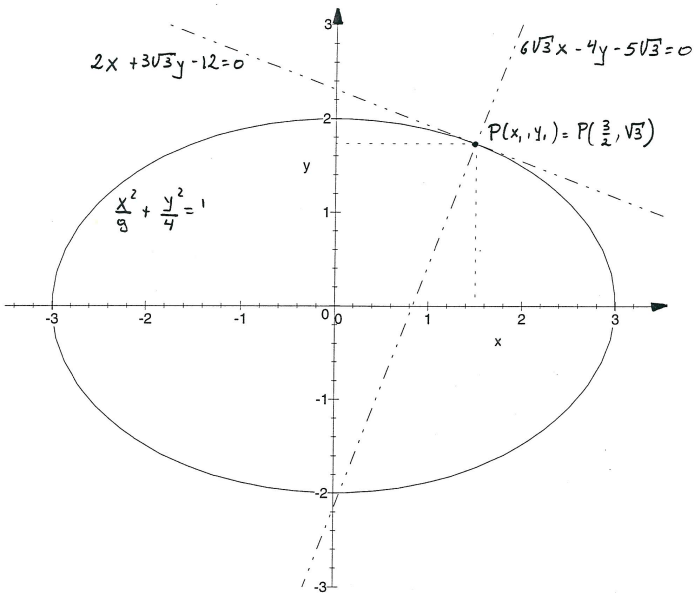


Figure 9.17: The tangent and normal to the ellipse $\frac{x^2}{9} + \frac{y^2}{4} = 1$ at $P(3/2, \sqrt{3})$.

The normal to the ellipse in (9.47) at $P(x_1, y_1)$ can now be determined easily from (9.53). The gradient of the normal to the ellipse at a point $P(x_1, y_1)$ is given by

$$m = \frac{a^2 y_1}{b^2 x_1},$$

for $x_1 \neq 0$, so that the equation of the normal to the ellipse at P (with $x_1 \neq 0$) is given by

$$y - y_1 = \frac{a^2 y_1}{b^2 x_1}(x - x_1). \tag{9.54}$$

Example 140: For the ellipse in Example 139, the normal at $P(3/2, \sqrt{3})$ is given by

$$y - \sqrt{3} = \frac{9\sqrt{3}}{4(3/2)}(x - 3/2),$$

via (9.54), or equivalently,

$$6\sqrt{3}x - 4y - 5\sqrt{3} = 0.$$

This is also illustrated in Figure 9.17.

9.3.7 Reflection property

Theorem 9.18. *The tangent and the normal to an ellipse at a point P are the bisectors of the angles formed by the lines that connect P with the foci of the ellipse.*

This property is illustrated in Figure 9.18.

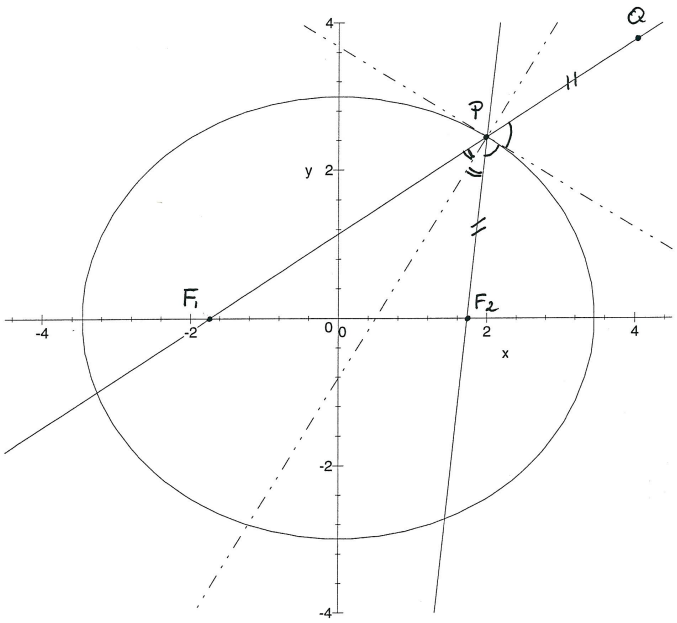


Figure 9.18: The reflection property of an ellipse.

Proof: To prove this property, we construct the bisector of the lines F_1P and F_2P and show the bisector must be normal or tangent (depends which bisector we consider) to the ellipse at P. First, consider the point Q on the line F_1P (outside the ellipse) such that

$$|\vec{PQ}| = |\vec{PF_2}|. \tag{9.55}$$

Each point E on the bisector of the angle $\widehat{F_2PQ}$ defined by the lines F_2P and PQ must then be equally far from F_2 as from Q:

$$|\vec{EF_2}| = |\vec{EQ}|. \tag{9.56}$$

When $E \neq P$ it follows that

$$\begin{aligned} |\vec{EF_2}| + |\vec{EF_1}| &= |\vec{EQ}| + |\vec{EF_1}| && \text{(via (9.56)} \\ &> |\vec{F_1Q}| && \text{(via triangle inequality)} \\ &= |\vec{F_1P}| + |\vec{PQ}| \\ &= |\vec{F_1P}| + |\vec{PF_2}| \\ &= 2a. \end{aligned} \tag{9.57}$$

Therefore, via (9.57), E cannot be on the ellipse. So we have shown that each point on the bisector of the angle $\widehat{F_2PQ}$ not equal to P cannot be on the ellipse, or equivalently,

that the bisector of the angle $\widehat{F_2PQ}$ has only the point P in common with the ellipse. This means that the bisector of the angle $\widehat{F_2PQ}$ is the tangent to the ellipse at P. It then also follows that the normal to the ellipse at P is the bisector of the angle $\widehat{F_1PF_2}$, as required.

A consequence of the reflective property property is that a ray of light which passes through a focus of an elliptical mirror is reflected into the other focus.

9.3.8 Additional equations and properties

Definition 9.19. *For an ellipse with foci at $F_1(c,0)$ and $F_2(-c,0)$ and standard equation*

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1,$$

the lines with equations

$$x=\pm\frac{a}{e}$$

*are known as the **directrices** of the ellipse.*

Theorem 9.20. *The following statements are equivalent:*

(a) P is a point on the ellipse.

(b) the ratio of the distance from P to a focus F_i and the distance from the point P to then corresponding directrix D_i is equal to the eccentricity e, i.e.

$$|PF_i|=e|PD_i|.$$

Proof: Consider the focus F_1 at $(c,0)$. The corresponding directrix has equation

$$D_1:\ x=\frac{a}{e}.$$

Also from Definition 9.17, $c=ae$. Using the standard equation for an ellipse, and

$$b^2=a^2-c^2=a^2(1-e^2), \tag{9.58}$$

for a point $P(x,y)$ on the ellipse,

$$y^2=b^2-\frac{b^2x^2}{a^2}=a^2(1-e^2)-\frac{a^2(1-e^2)x^2}{a^2}=(a^2-x^2)(1-e^2). \tag{9.59}$$

Hence,

$$|\vec{PF_1}| \ = \ \sqrt{(x-ae)^2+y^2}$$

$$\begin{aligned} &= \sqrt{x^2 - 2aex + a^2e^2 + (a^2 - x^2)(1 - e^2)} \quad \text{(via (9.59))} \\ &= \sqrt{x^2e^2 - 2aex + a^2} \\ &= \sqrt{(xe - a)^2} \\ &= e\left|x - \frac{a}{e}\right| \\ &= e|P\vec{D}_1|. \end{aligned} \tag{9.60}$$

Therefore, from (9.60) if (a) is true, then (b) is true.

Now, conversely, if a point $P(x, y)$ satisfies $|P\vec{F}_1| = e|P\vec{D}_1|$, with

$$|P\vec{F}_1| = \sqrt{(x - ae)^2 + y^2}, \quad \text{and} \quad |P\vec{D}_1| = \left|x - \frac{a}{e}\right|, \tag{9.61}$$

then

$$|P\vec{F}_1|^2 = e^2|P\vec{D}_1|^2$$

so that via (9.61),

$$(x - ae)^2 + y^2 = e^2\left(x - \frac{a}{e}\right)^2.$$

Hence,

$$(1 - e^2)x^2 - a^2(1 - e^2) + y^2 = 0,$$

so via (9.58),

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1. \tag{9.62}$$

Equation (9.62) states that P is on the ellipse and hence, if (b) is true then (a) is true, so (a) $\iff$ (b), as required.

Theorem 9.21. ***Parametric equations*** *for an ellipse with foci at $F_1(c, 0)$ and $F_2(-c, 0)$, with $c > 0$, as given by Theorem 9.14, are*

$$\begin{cases} x &= a\cos(t), \\ y &= b\sin(t), \end{cases} \qquad t \in [0, 2\pi).$$

Proof: This can easily be verified by substituting the expressions above into the standard equation of an ellipse.

When using polar coordinates, the following result is useful.

Theorem 9.22. *The* ***polar equation*** *of an ellipse with a focus at $O(0, 0)$ and directrix $x = d$, with $d > 0$ and eccentricity $e \in [0, 1)$, is given by*

$$r(\theta) = \frac{ed}{1 + e\cos(\theta)}.$$

Proof: Consider this as an exercise.

Example 141: The special case $a = b = \rho$ yields the **standard equation** of a circle with centre at the origin and radius ρ:

$$x^2 + y^2 = \rho^2.$$

Notice that a circle, is formally and ellipse with eccentricity $e = 0$, with both foci at the same point, and with directrices 'at infinity'.

The polar equation of a circle is

$$r(\theta) = \rho,$$

with parametric equations given by

$$\begin{cases} x &= \rho\cos(t), \\ y &= \rho\sin(t). \end{cases}$$

9.4 Hyperbola

Definition 9.23. *A hyperbola is the set of points in a plane whose distances from two distinct fixed points F_1 and F_2 in the plane have a constant difference $2a$ with $0 < 2a < |\vec{F_1F_2}|$.*

- The two fixed points are the *foci* of the hyperbola.
- The midpoint between the two foci is called the *centre* of the hyperbola.
- The line through the foci of a hyperbola is the *focal axis* (or *major axis*). The line perpendicular to the focal axis through the centre is the *minor axis*.
- The points where the hyperbola crosses the focal axis are the *vertices*.

If we are given foci at F_1 and F_2 with constant $a > 0$, then these define a hyperbola. A point P on this hyperbola (for instance) satisfies $|\vec{PF_1}| - |\vec{PF_2}| = 2a$ or $|\vec{PF_2}| - |\vec{PF_1}| = 2a$.

We will derive the standard equation of a hyperbola and study how it changes under translation or rotation of the coordinate system. We will derive an equation for the tangent and normal at a point on the hyperbola and briefly discuss its reflective properties.

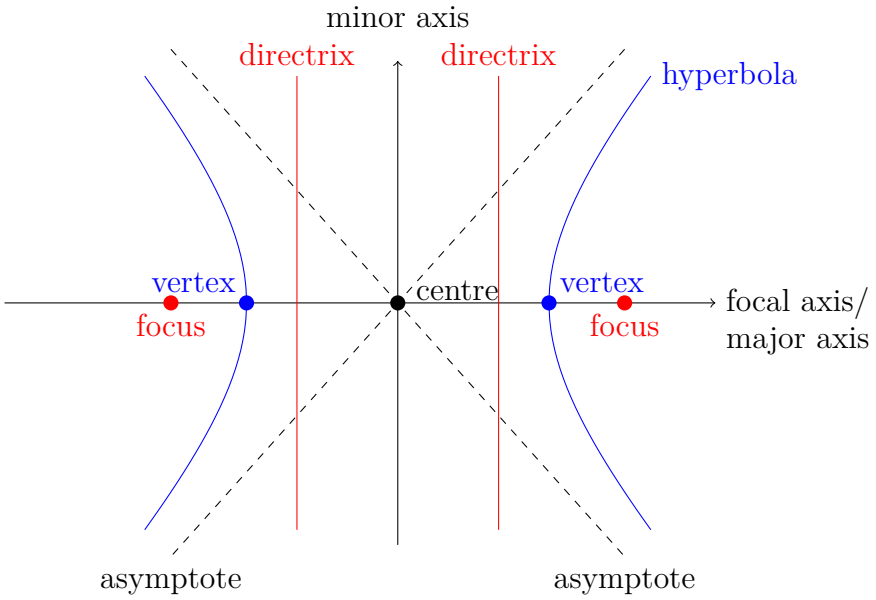


Figure 9.19: Sketch of main terms related to a hyperbola.

9.4.1 Standard equation

Theorem 9.24. *The **standard equation** of a hyperbola with foci at $F_1(c,0)$ and $F_2(-c,0)$, where $c>0$, is given by*

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}=1,$$

where $2a$ is the difference of the distances of a point P on the hyperbola to the two foci, and $b^2=c^2-a^2>0$.

This means that the centre of the hyperbola in standard form is at the origin. This configuration is illustrated in Fig. 9.20.

Proof: Since $2a<|\vec{F_1F_2}|=2c$, it follows that $a<c$ and $b^2=c^2-a^2>0$. A point $P(x,y)$ is on the hyperbola when

$$|\vec{PF_1}|-|\vec{PF_2}|=\pm 2a \tag{9.63}$$

$$\iff \quad \sqrt{(x+c)^2+y^2}-\sqrt{(x-c)^2+y^2}=\pm 2a$$

$$\iff \quad (x+c)^2+y^2-2\sqrt{(x+c)^2+y^2}\sqrt{(x-c)^2+y^2}+(x-c)^2+y^2 = 4a^2$$

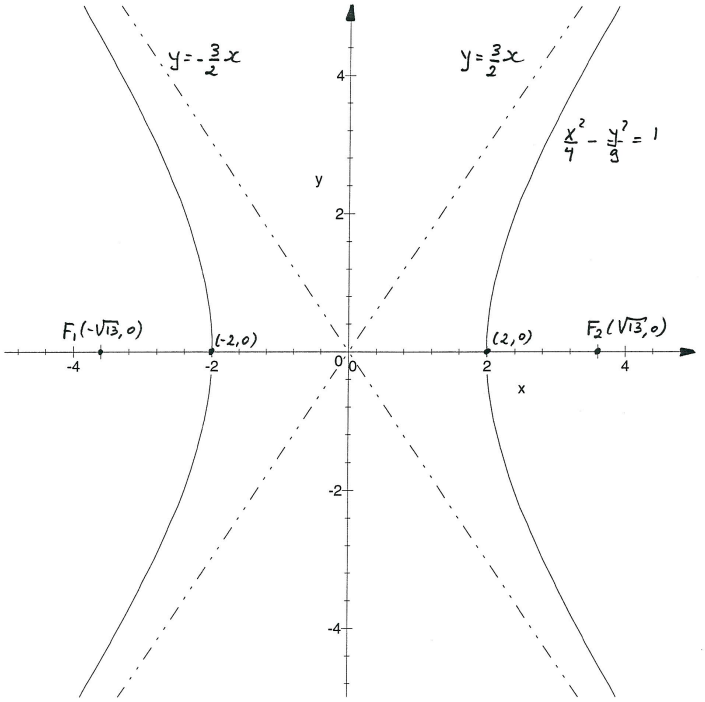


Figure 9.20: The hyperbola with equation $x^2/4 - y^2/9 = 1$ with $c = \sqrt{13}$.

$$\begin{aligned} \iff \quad & 2x^2 + 2y^2 + 2c^2 - 2\sqrt{\left(x^2 + c^2 + y^2 - 2cx\right)\left(x^2 + c^2 + y^2 + 2cx\right)} \\ & = 4a^2. \end{aligned}$$

$$\iff \quad x^2 + y^2 + c^2 - 2a^2 = \sqrt{\left(x^2 + y^2 + c^2\right)^2 - 4c^2x^2}. \tag{9.64}$$

Equation (9.64) implies, but is not necessarily equivalent to,

$$\left(x^2 + y^2 + c^2 - 2a^2\right)^2 = \left(x^2 + y^2 + c^2\right)^2 - 4c^2x^2$$

$$\begin{aligned} \iff \quad & \left(x^2 + c^2 + y^2\right)^2 - 4a^2\left(x^2 + c^2 + y^2\right) + 4a^4 = \\ & \left(x^2 + y^2 + c^2\right)^2 - 4c^2x^2 \end{aligned}$$

$$\iff \quad \left(c^2 - a^2\right)x^2 + a^2\left(a^2 - c^2\right) - a^2y^2 = 0 \tag{9.65}$$

$$\iff \quad b^2x^2 - a^2y^2 = a^2b^2. \tag{9.66}$$

Since $b^2 > 0$, it follows that (9.66) is equivalent to

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1, \tag{9.67}$$

Again, we have lost the equivalence of (9.67) and (9.63) (from (9.64) to (9.66)) since we were not sure

$$x^2 + y^2 + c^2 - 2a^2 \geq 0.$$

Therefore, now assume a point $Q(x_1, y_1)$ satisfies the equation (9.67) with $b^2 = c^2 - a^2 > 0$. This implies that

$$x_1^2 \geq a^2. \tag{9.68}$$

In addition, a positive number is always larger then a negative one, so

$$y_1^2 \geq -b^2. \tag{9.69}$$

Therefore,

$$\begin{aligned} x_1^2 + y_1^2 + c^2 - 2a^2 &\geq a^2 - b^2 + c^2 - 2a^2 && \text{(via (9.68) and (9.69))} \\ &\geq a^2 - b^2 + b^2 - a^2 && \text{(since } c^2 = a^2 + b^2\text{)} \\ &= 0. \end{aligned} \tag{9.70}$$

So when a point satisfies (9.67) with $b^2 = c^2 - a^2 > 0$, we know from (9.70) that

$$\sqrt{\left(x^2 + y^2 + c^2 - 2a^2\right)^2} = x^2 + y^2 + c^2 - 2a^2,$$

and therefore (9.66) implies (9.64) and hence, (since all other relations are iff) (9.64) implies (9.63), as required.

We can therefore refer to (9.67) as the *standard equation* of a hyperbola.

Example 142: The hyperbola with equation

$$\frac{x^2}{4} - \frac{y^2}{9} = 1,$$

is plotted in Figure 9.20. The vertices are located at $V(\pm a, 0) = (\pm 2, 0)$. The foci are then given by $F(\pm\sqrt{13}, 0)$ since $c^2 = a^2 + b^2 = 13$.

9.4.2 Asymptotes

We can rewrite the standard equation for a hyperbola as

$$y^2 = \frac{b^2 x^2}{a^2} - b^2,$$

where for very large values of x, the right hand side will be dominated (this term is much larger than the others) by the x^2 term,

$$y^2 \sim \frac{b^2x^2}{a^2} \quad \text{as} \quad x \to \infty.$$

This means that for large values of x, the hyperbola is close to the two lines

$$y = \frac{b}{a}x \quad \text{and} \quad y = -\frac{b}{a}x. \tag{9.71}$$

These two lines are called the *asymptotes* of the hyperbola.

Example 143: Figure 9.20 displays the two asymptotes given by (9.71) which, for the hyperbola with equation,

$$\frac{x^2}{4} - \frac{y^2}{9} = 1,$$

have equations $y = (3/2)x$ and $y = -(3/2)x$.

9.4.3 Alternative formulae

Focal axis is y-axis

Theorem 9.25. *The equation of a hyperbola with foci at $F_1(0, c)$ and $F_2(0, -c)$, where $c > 0$, is given by*

$$\frac{y^2}{a^2} - \frac{x^2}{b^2} = 1,$$

with a and b as in Theorem 9.24.

Proof: We can find the equation of a hyperbola with foci on the y-axis and centre at the origin by writing the standard equation for a hyperbola in a coordinate system rotated through an angle $\alpha = \pi/2$ and then use the transformation formula, similar to the approach used for the parabola and ellipse. Set

$$\begin{cases} \tilde{x} &= \cos(\pi/2)x + \sin(\pi/2)y &= y, \\ \tilde{y} &= -\sin(\pi/2)x + \cos(\pi/2)y &= -x. \end{cases}$$

and substituting the new coordinates into

$$\frac{\tilde{x}^2}{a^2} - \frac{\tilde{y}^2}{b^2} = 1$$

gives

$$\frac{y^2}{a^2} - \frac{x^2}{b^2} = 1,$$

in the original coordinates, as required.

The hyperbola with equation

$$\frac{y^2}{4} - \frac{x^2}{5} = 1,$$

is depicted in Figure 9.21.

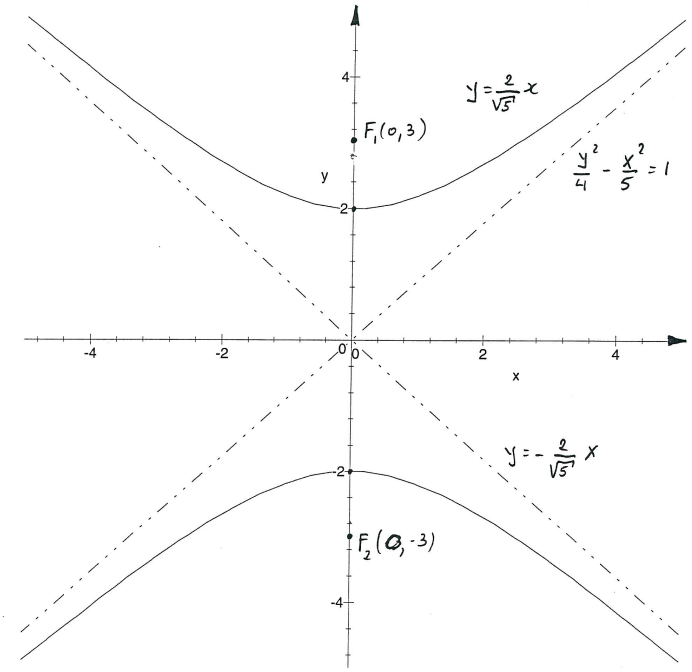


Figure 9.21: Hyperbola with vertical focal axis.

Hyperbola with tilted focal axis

Assume the focal axis of a hyperbola is tilted by an angle α from the positive x-axis in the anti-clockwise direction. Here, we write the standard equation in a coordinate system which is rotated through an angle α in the clockwise direction. Using the transformation formula for a rotation of the coordinate system, we obtain the equation

$$\frac{\left(x\cos(\alpha) + y\sin(\alpha)\right)^2}{a^2} - \frac{\left(-x\sin(\alpha) + y\cos(\alpha)\right)^2}{b^2} = 1,$$

or equivalently,

$$\left(\frac{\cos^2(\alpha)}{a^2} - \frac{\sin^2(\alpha)}{b^2}\right)x^2 +$$

$$\left(\frac{\sin^2(\alpha)}{a^2} - \frac{\cos^2(\alpha)}{b^2}\right) y^2 + 2\sin(\alpha)\cos(\alpha)\left(\frac{1}{a^2} + \frac{1}{b^2}\right) xy = 1. \tag{9.72}$$

Observe that a xy term appears in (9.72), which is not present in the equations of hyperbola in Theorem 9.24 and 9.25.

Example 144: Figure 9.22 depicts a hyperbola with $a = 2$ and $b = \sqrt{5}$ and with the focal axis at an angle $\alpha = \pi/3$ to the positive x-axis. The equation for the hyperbola is given by

$$-\frac{7}{80}x^2 + \frac{11}{80}y^2 + \frac{9\sqrt{3}}{40}xy = 1.$$

Note that there is a non-zero xy term in the equation above.

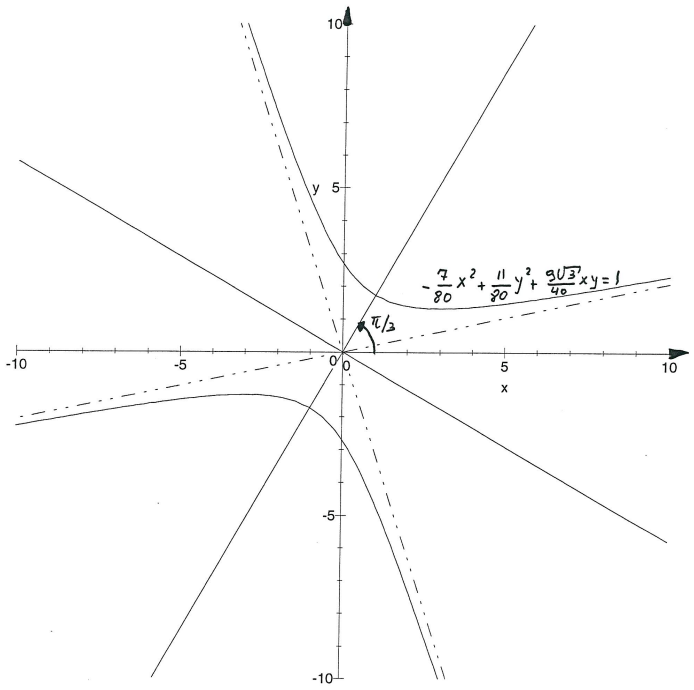


Figure 9.22: Hyperbola with tilted focal axis.

Hyperbola with centre at $P(q,s)$

Theorem 9.26. *The equation of the hyperbola with centre at $P(q,s)$, with centre-to-focus distance $c>0$ and major axis parallel to the x-axis, is given by*

$$\frac{(x-q)^2}{a^2} - \frac{(y-s)^2}{b^2} = 1,$$

with a and b as in Theorem 9.24.

Proof: We consider the standard equation in a translated coordinate system with origin at $P(q,s)$. The transformation formula (see for example the proof of Theorem 9.16) then leads to the equation

$$\frac{(x-q)^2}{a^2} - \frac{(y-s)^2}{b^2} = 1,$$

as required.

Example 145: The hyperbola with equation

$$\frac{(x-1.3)^2}{4} - \frac{(y-3.1)^2}{5} = 1, \tag{9.73}$$

is depicted in Figure 9.23.

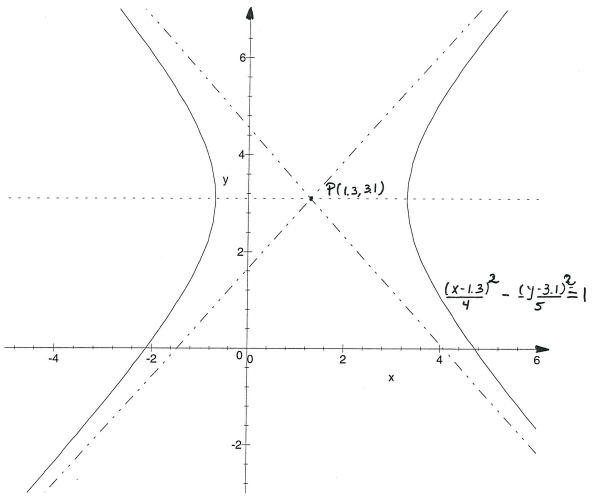


Figure 9.23: Hyperbola with centre in $P(1.3, 3.1)$.

Equation (9.73) can be expanded to yield

$$0.25x^2 - 0.65x - 0.2y^2 + 1.24y = 2.4995.$$

9.4.4 Eccentricity

Definition 9.27. *The eccentricity of a hyperbola (in standard form) is defined as,*

$$e = \frac{c}{a} = \frac{\sqrt{a^2 + b^2}}{a}.$$

Notice that for a hyperbola, $e > 1$ (in contrast to an ellipse).

Example 146: Figure 9.24 depicts two hyperbola with different eccentricities, i.e.

$$\frac{x^2}{2} - \frac{y^2}{0.2} = 1 \quad \text{and} \quad \frac{x^2}{2} - \frac{y^2}{5} = 1.$$

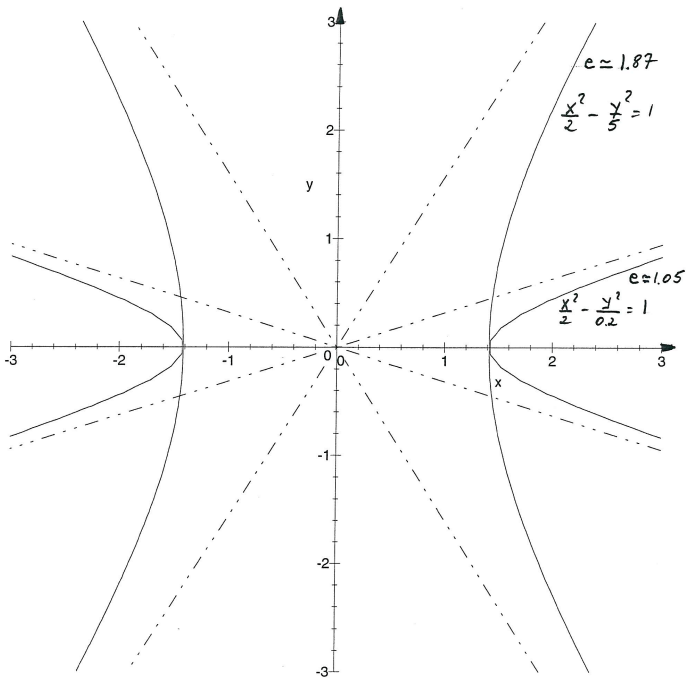


Figure 9.24: Hyperbolae with different eccentricities.

9.4.5 The tangent to a hyperbola

As in the case of the ellipse, the points of intersection of a hyperbola in standard form and the line

$$y = mx + q, \tag{9.74}$$

are determined by substituting (9.74) into the equation in Theorem 9.24, i.e.

$$\frac{x^2}{a^2} - \frac{(mx+q)^2}{b^2} = 1$$

$$\Leftrightarrow \quad \left(b^2 - m^2a^2\right)x^2 - 2mqa^2x - \left(a^2q^2 + a^2b^2\right) = 0.$$

The discriminant of this quadratic equation[8] (in x) is given by

$$D = 4a^2b^2\left(b^2 - m^2a^2 + q^2\right), \tag{9.75}$$

where $D > 0$ corresponds to two point of intersection, $D < 0$ indicates that there are no points of intersection, and, $D = 0$ implies the line is tangent to the hyperbola. We can eliminate q from (9.75) by specifying that the point of intersection $Q(x_1, y_1)$ between the line and the hyperbola is unique, i.e. $q = y_1 - mx_1$. The condition $D = 0$ can then be re-written by substitution of q into (9.75)

$$\left(x_1^2 - a^2\right)m^2 - 2x_1y_1m + b^2 + y_1^2 = 0. \tag{9.76}$$

Since the point (x_1, y_1) is on the hyperbola, it follows that the quadratic equation in (9.76) in m, has zero discriminant, and hence, the gradient of the tangent to the hyperbola at (x_1, y_1) (with $x_1 \neq \pm a$) is equal to

$$m = \frac{x_1y_1}{x_1^2 - a^2} = \frac{x_1b^2}{y_1a^2}. \tag{9.77}$$

The tangent line determined by q, (9.74) and (9.77) can then, after some manipulations, be expressed as,

$$\frac{x_1x}{a^2} - \frac{y_1y}{b^2} = 1. \tag{9.78}$$

Note that (9.78) holds for all points on the hyperbola. Similarly, the equation for the normal to the hyperbola at (x_1, y_1) can be expressed as

$$y - y_1 = -\frac{a^2y_1}{b^2x_1}(x - x_1). \tag{9.79}$$

Example 147: The hyperbola given by

$$\frac{x^2}{4} - \frac{y^2}{9} = 1,$$

has tangent at $P(2\sqrt{2}, 3)$ given by

$$3\sqrt{2}x - 2y - 6 = 0,$$

via (9.78). In addition, the normal to the hyperbola at $P(2\sqrt{2}, 3)$ is given by

$$y = -\frac{\sqrt{2}}{3}x + \frac{13}{3},$$

via (9.79). The tangent and normal to the hyperbola at P are depicted in Figure 9.25.

[8] The discriminant exists iff $b^2 - m^2a^2 \neq 0$ i.e. provided that $m \neq \pm b/a$. If $m = \pm b/a$, then the corresponding lines intersect/cross the hyperbola once at (x_1, y_1) and are parallel to a corresponding asymptote of the hyperbola. Neither of these lines are the tangent line to the hyperbola at (x_1, y_1).

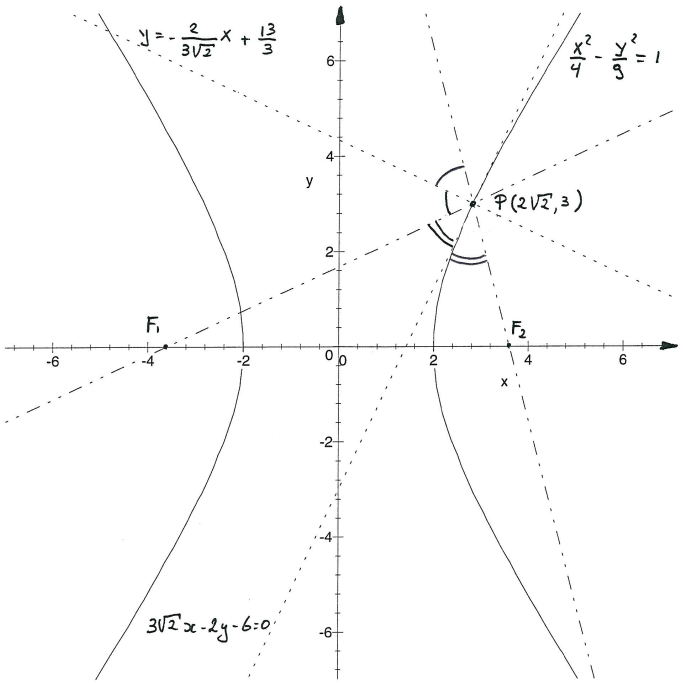


Figure 9.25: Tangent and normal at a hyperbola.

9.4.6 Reflection property

Theorem 9.28. *The tangent and the normal at a point P to a hyperbola are the bisectors of the angles formed by the lines that connect P with the two focal points.*

This property is also illustrated in Figure 9.25.

Proof: To prove this property, we construct the bisector of the lines F_1P and F_2P and show the bisector must be normal or tangent (depends which bisector we consider) to the hyperbola at P. We consider the branch of the hyperbola defined by $|\vec{PF_1}| - |\vec{PF_2}| = 2a$ with the argument for the other branch following the same argument.

Now, consider the point Q on the line F_1P such that

$$|\vec{PQ}| = |\vec{PF_2}|. \tag{9.80}$$

Each point E on the bisector of the angle $\widehat{QPF_2}$ defined by the lines QP and PF_2 must then be equally far from F_2 as from Q:

$$|\vec{EF_2}| = |\vec{EQ}|. \tag{9.81}$$

When $E \neq P$ it follows that

$$\begin{aligned} \|\vec{EF_1}| - |\vec{EF_2}\| &= \|\vec{EF_1}| - |\vec{EQ}\| && \text{(via (9.81)} \\ &< |\vec{F_1Q}| && \text{(via triangle inequality)} \\ &= |\vec{F_1P}| - |\vec{PQ}| \\ &= |\vec{F_1P}| - |\vec{PF_2}| && \text{(via (9.80))} \\ &= 2a. && (9.82) \end{aligned}$$

Therefore, via (9.82),

$$|\vec{EF_1}| - |\vec{EF_2}| \in (-2a, 2a)$$

and hence, E is not on the hyperbola.

Now, consider the continuous function $G : \mathbb{R}^2 \to \mathbb{R}$ given by

$$G(P) = \|\vec{PF_1}| - |\vec{PF_2}\| \quad \forall P \in \mathbb{R}^2.$$

It follows that: $G = 2a$ on the hyperbola; $G < 2a$ in the connected region of $\mathbb{R}^2$, denoted by R, which contains the centre of the hyperbola and is bounded by the 2 arcs of the hyperbola; and $G > 2a$ in the 2 disjoint connected regions, denoted by S_1 and S_2, which are bounded by the arcs of the hyperbola, and that contain the foci F_1 and F_2 respectively. To clarify $\bar{R} \cup \bar{S}_1 \cup \bar{S}_2 = \mathbb{R}^2$.

The bisector of $\widehat{QPF_2}$ is contained in $R \cup P$. Moreover, the lines that intersect the hyperbola at P that are parallel to the asymptotes, necessarily intersect S_2[9]. Therefore, the bisector of $\widehat{QPF_2}$ is the tangent to the hyperbola at P.

The argument for the other branch of the hyperbola follows similarly.

9.4.7 Additional equations and properties

Definition 9.29. *For a hyperbola with foci at $F_1(c, 0)$ and $F_2(-c, 0)$, where $c > 0$, and standard equation*

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1,$$

the lines with equations

$$x = \pm\frac{a}{e},$$

*are known as the **directrices** of the hyperbola.*

[9] As an exercise, you should demonstrate this for a hyperbola in standard form.

Theorem 9.30. *P is a point on the hyperbola if and only if the ratio of the distance from P to a focus F_i and the distance from the point P to the corresponding directrix D_i is equal to the eccentricity e, i.e.*

$$|\vec{PF_i}| = e|\vec{PD_i}|.$$

Proof: Consider this as an exercise.

Theorem 9.31. ***Parametric equations*** *for a hyperbola with foci at $F_1(c,0)$ and $F_2(-c,0)$, with $c>0$, and eccentricity e, are given by*

$$\begin{cases} x &= a\sec(t), \\ y &= b\tan(t), \end{cases} \qquad t \in (-\pi,\pi),$$

or

$$\begin{cases} x &= \pm a\cosh(t), \\ y &= b\sinh(t), \end{cases} \qquad t \in \mathbb{R},$$

where $2a$ is the difference between the distance from a point on the hyperbola to its foci, and $b^2 = a^2(e^2-1)$.

Proof: The result can be verified by substituting the equations above into the standard equation for a hyperbola.

Theorem 9.32. *The **polar equation** of the right hand arc of a hyperbola with focus $F_1(0,0)$, corresponding directrix $D_1 : x = -d$ and eccentricity $e>1$ is*

$$r(\theta) = \frac{ed}{1-e\cos\theta}.$$

Proof: Consider this as an exercise.

9.5 General Equation of a conic

9.5.1 Polar equations of conics*

We can summarise the results which describe conics via polar coordinates in the previous sections in the following result.

Theorem 9.33. *The polar equations*

$$r(\theta) = \frac{ed}{1 \pm e\cos(\theta)},$$

and

$$r(\theta) = \frac{ed}{1 \pm e\sin(\theta)},$$

represent a conic with focus at $O(0,0)$, eccentricity e and with d the distance between the focus and the corresponding directrix.

Proof: Follows from Theorems 9.11, 9.12, 9.22 and 9.32.

The conic is an ellipse for $e < 1$, parabola for $e = 1$ and hyperbola for $e > 1$.

The equations

$$r(\theta) = \frac{ed}{1 \pm e\cos(\theta)},$$

represent conics with a vertical directrix (i.e. $x = \pm d$) and axis of symmetry parallel to the x-axis.

The equations

$$r(\theta) = \frac{ed}{1 \pm e\sin(\theta)},$$

represent conics with a horizontal directrix (i.e. $y = \pm d$) and axis of symmetry parallel to the y-axis.

9.5.2 Classification quadratic equations in 2 variables*

Recall from Sections 9.1.2-9.1.3, that given a quadratic equation in 2 variables,

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0, \tag{9.83}$$

we can choose a rotated coordinate system $\tilde{x}, \tilde{y}$ such that (9.83) is equivalent to

$$\tilde{A}\tilde{x}^2 + \tilde{C}\tilde{y}^2 + \tilde{D}\tilde{x} + \tilde{E}\tilde{y} + \tilde{F} = 0. \tag{9.84}$$

If $\tilde{A}$ and $\tilde{C}$ are non-zero, we can translate the coordinate system (complete the square) to $\hat{x}$, $\hat{y}$ so that (9.84) is equivalent to

$$\tilde{A}(\hat{x})^2 + \tilde{C}\hat{y}^2 = -\Delta. \tag{9.85}$$

From (9.83)-(9.85) we can infer the following classification of quadratic equations in 2 variables:

1. $\tilde{A} = \tilde{C} \neq 0$: circle, point or no graph depending on the sign of Δ.
2. $\tilde{A} = \tilde{C} = 0$: straight line, the whole plane, or no graph depending on $\tilde{D}$, $\tilde{E}$ and $\tilde{F}$.
3. $\tilde{A}\tilde{C} > 0$ and $\tilde{A} \neq \tilde{C}$: ellipse, point or no graph depending on sign of Δ.
4. $\tilde{A}\tilde{C} < 0$: hyperbola $\Delta \neq 0$ or pair of intersecting lines ($\Delta = 0$).
5. $\tilde{A} = 0$, $\tilde{C} \neq 0$ and $\tilde{D} \neq 0$: parabola (slightly different translation is required).
6. $\tilde{C} = 0$, $\tilde{A} \neq 0$ and $\tilde{E} \neq 0$: parabola (slightly different translation is required).
7. $\tilde{A} = \tilde{D} = 0$: two parallel lines, two coinciding parallel lines (i.e. one line), or no graph depending on the sign of $\tilde{E}^2 - 4\tilde{C}\tilde{F}$.
8. $\tilde{C} = \tilde{E} = 0$: two parallel lines, two coinciding parallel lines (i.e. one line), or no graph depending on the sign of $\tilde{D}^2 - 4\tilde{A}\tilde{F}$.

One can show that the classification can also be achieved using the discriminant

$$B^2 - 4AC,$$

which uses the coefficients in the equation (9.1). Ignoring the degenerate cases, the classification becomes:

1. $B^2 - 4AC = 0$: parabola.
2. $B^2 - 4AC > 0$: hyperbola.
3. $B^2 - 4AC < 0$: ellipse.

Appendix A

Sets and Notation

▸ **Learning Outcomes** ◂ After the completion of the lectures and support sessions associated with this chapter you should be able to:

- understand what sets are;
- recall the basic properties of sets;
- understand and utilize common notation used to describe sets; and
- use basic operations to manipulate sets.

[5, p.1-22 and p.165-200] contain alternative presentations of material in this chapter that you may find helpful.

A.1 Introduction

Definition A.1. *A **set** is a collection of uniquely identifiable objects which are known as **elements** or **members**.*

A set which contains elements with a property P is denoted $\{x : x \text{ has property } P\}$ where the notation " : " means "such that". In simple cases, we sometimes list the elements of a set, for example, $\{x : x = a \text{ or } x = b\}$ can be written $\{a, b\}$.

Example 148: Consider the sets

$$S = \{1, 11, 2000\}, \ R = \{\clubsuit, \spadesuit, 5\} \text{ and } Q = \{\text{pear}, \text{plum}, \text{banana}, \text{pear}\}.$$

Do all 3 sets contain the same number of elements?

Note that the sets S, R and Q in Example 148 have **three** elements since there is a repeated element (pear) in Q i.e. repetitions are ignored (since they correspond to elements that are already uniquely defined). We usually use $\{\cdot\}$ (curly brackets) to denote a set, and for small sets we can simply list the elements. Sometimes, **if it is clear**, we abbreviate using the following notation for larger sets:

$$\{1, 3, 5, \ldots, 99\}$$

or

$$\{2, 4, \ldots, 1024\}.$$

The rule in this second example is not clear. Are the elements even numbers or powers of 2? Without additional information, one cannot know which elements are in this set.

Example 149: We use notation below to denote standard sets that appear throughout the course[1]:

$$\begin{aligned}
\mathbb{N} &= \{1, 2, 3, 4, 5, \ldots\} \quad \text{(natural numbers)},\\
\mathbb{Z} &= \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\} \quad \text{(integers)},\\
\mathbb{Q} &= \left\{\frac{m}{n} : m, n \in \mathbb{Z}, n > 0\right\} \quad \text{(rational numbers)}.
\end{aligned}$$

Observe that $\frac{2}{3}, 23, \frac{-7}{8}, \frac{-77}{8} \in \mathbb{Q}$ where " $\in$ " means "is an element of". We also have the following special cases:

$$\begin{aligned}
\mathbb{N}_0 &= \{0, 1, 2, 3, 4, 5, \ldots\},\\
\mathbb{Z}^+ &= \{x \in \mathbb{Z} : x > 0\} = \{1, 2, 3, \ldots\} = \mathbb{N},\\
\mathbb{Z}_0^+ &= \{x \in \mathbb{Z} : x \geq 0\} = \{0, 1, 2, 3, 4, 5, \ldots\} = \mathbb{N}_0,\\
\mathbb{Z}^- &= \{x \in \mathbb{Z} : x < 0\} = \{\ldots, -3, -2, -1\},\\
\mathbb{Z}_0^- &= \{x \in \mathbb{Z} : x \leq 0\} = \{\ldots, -3, -1, -1, 0\},\\
\mathbb{Q}^+ &= \{x \in \mathbb{Q} : x > 0\},\\
\mathbb{Q}_0^+ &= \{x \in \mathbb{Q} : x \geq 0\},\\
\mathbb{Q}^- &= \{x \in \mathbb{Q} : x < 0\},\\
\mathbb{Q}_0^- &= \{x \in \mathbb{Q} : x \leq 0\}.
\end{aligned}$$

In Example 149, we have defined a set by a **defining property** which is written as

$$A = \{x : x \text{ satisfies } P\}.$$

[1] These notations may differ slightly in reference texts and other courses.

Sometimes ' : ' is replaced by ' | ' so that the set A may be written

$$A = \{x \mid x \text{ satisfies } P\}.$$

Example 150: Let A be the set of all even natural numbers,

$$\begin{aligned} A &= \{x : x \in \mathbb{N} \text{ and is divisibe by } 2\} \\ &= \{2, 4, 6, 8, 10, 12, \ldots\}. \end{aligned}$$

The set A can also be expressed as $2\mathbb{N}$ (the set with elements of the form $2n$ where n is a natural number).

To 'complete' the rational numbers (to express everything on the number-line), we require irrational numbers (for instance $\sqrt{2}$, $\sqrt{3}$, π, $\ln 2$, e, ...). The set containing all rational and irrational numbers is referred to as the set of real numbers[2] and denoted by

$$\mathbb{R} = \text{the set of all real numbers.}$$

$\mathbb{R}$ can be thought of as being inside $\mathbb{C}$ (the set of complex numbers), which will be discussed in Chapter 7.

Some sets are **finite**, for example $A = \{1, 2, 3\}$, and others are **infinite**, for example $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{N}$ etc. For finite sets we use $|A|$ to denote the number of elements in the set or equivalently, the size of the set. $|A|$ is known as the **order** or **cardinality** of the set A. So for $A = \{1, 2, 3\}$, the cardinality is given by $|A| = 3$.

A.2 Interval Notation

Consider two real numbers, a and b, with $a \leq b$. Intervals of finite length are represented as follows:

Closed Interval	$[a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$ with $a \leq b$ $[a, a] = \{a\}$
Open Interval	$(a, b) = \{x \in \mathbb{R} : a < x < b\}$
Half Open	$(a, b] = \{x \in \mathbb{R} : a < x \leq b\}$ $[a, b) = \{x \in \mathbb{R} : a \leq x < b\}$

[2]For a detailed construction of the real numbers, see [14, p.101-110], but note that this is not within the scope of this course.

Moreover, intervals of infinite length can be represented as follows:

$$[a, \infty) = \{x \in \mathbb{R} \ : x \geq a\},$$

$$(-\infty, b) = \{x \in \mathbb{R} \ : x < b\}.$$

⋆⋆ **WARNING** '∞' is not a number: it is used here as a form of shorthand ⋆⋆

Example 151: We can express $\mathbb{R}^+$, $\mathbb{R}_0^+$, $\mathbb{R}^-$ and $\mathbb{R}_0^-$ in interval notation:

$$\begin{aligned}
\mathbb{R}^+ &= \{x \in \mathbb{R} : x > 0\} = (0, \infty), \\
\mathbb{R}_0^+ &= \{x \in \mathbb{R} : x \geq 0\} = [0, \infty), \\
\mathbb{R}^- &= \{x \in \mathbb{R} : x < 0\} = (-\infty, 0), \\
\mathbb{R}_0^- &= \{x \in \mathbb{R} : x \leq 0\} = (-\infty, 0].
\end{aligned}$$

A.3 Inclusion Among Sets

Definition A.2. *We write $A \subseteq B$ when every element of A is also an element of B. Equivalently, $A \subseteq B$ if either of the following equivalent statements hold:*

- *for all x, if $x \in A$ then $x \in B$; and*
- $\forall x : \ x \in A \Longrightarrow x \in B$.

Note that the notation " $\forall$ " means "for all/for each" or "for any", and " $\Longrightarrow$ " means "implies" or "then". Additionally the " : " in Definition A.2 is just a colon, not "such that". If $A \subseteq B$, we say that "A is a subset of B".

Example 152:

$$2\mathbb{N} \subseteq \mathbb{N},$$

$$\{1, 2\} \subseteq \{1, 2, 3\},$$

$$\{1, 2, 3\} \subseteq \{1, 2, 3\},$$

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}.$$

For **strict inclusion** we write

$$A \subset B,$$

which means

$$A \subseteq B \text{ but } A \neq B.$$

When $A \subset B$ we say "A is a **strict subset** of B."

Definition A.3. *Two sets A and B are **equal**, denoted $A = B$, if and only if,*

$$A \subseteq B \textbf{ and } B \subseteq A.$$

As for Definition A.2, Definition A.3 can be equivalently given as:

$$A = B \text{ if and only if, for all } x : (x \in A \Longrightarrow x \in B) \text{ and } (x \in B \Longrightarrow x \in A),$$

or

$$A = B \text{ iff } \forall x : x \in A \Longleftrightarrow x \in B.$$

Note that in order to prove that two sets are equal, you must prove both inclusions. Moreover, the words "if and only if" are often shortened to "iff". The notation " $A \Longleftrightarrow B$ " means "$A \Longrightarrow B$ and $B \Longrightarrow A$ ".

Example 153: Which set inclusions hold for $A = \{1, 2, 3\}$ and $B = \{1, 2, 3, \{3\}\}$?

Answer: Since $1, 2, 3 \in B$ and $A = \{1, 2, 3\}$, it follows that for all x, if $x \in A$ then $x \in B$. Therefore $A \subseteq B$. However, since $\{3\} \in B$ but $\{3\} \notin A$, it follows that $A \neq B$ and hence $A \subset B$. We can also conclude that $B \nsubseteq A$.

A.4 Intersection, Union and Difference

Definition A.4. *For sets A and B, we define $A \cap B$, the **intersection** of A and B as,*

$$A \cap B = \{x : (x \in A) \textit{ and } (x \in B)\}.$$

Example 154:

$$\{1, 2, 3\} \cap \{2, 4, 6\} = \{2\},$$

$$\mathbb{N} \cap 2\mathbb{N} = 2\mathbb{N}.$$

Definition A.5. *For sets A and B, we define $A \cup B$, the **union** of A and B, as,*

$$A \cup B = \{x : (x \in A) \textit{ or } (x \in B)\}.$$

In this case the "or" is an **inclusive**[3] "or" that embraces $x \in A$ or $x \in B$, or, both $x \in A$ and $x \in B$.

[3]Note that an "exclusive or" here would mean ($x \in A$ and $x \notin B$) or ($x \in B$ and $x \notin A$)

Example 155:

$$\{1,2,3\} \cup \{2,4,6\} = \{1,2,3,4,6\},$$

$$\mathbb{N} \cup 2\mathbb{N} = \mathbb{N}.$$

Definition A.6. *For sets A and B, we define $A \setminus B$, the **difference** of A and B, as,*

$$A \setminus B = \{x : (x \in A) \text{ and } (x \notin B)\}.$$

Example 156:

$$\{1,2,3\} \setminus \{2,3,6\} = \{1\},$$

$$\mathbb{N} \setminus 2\mathbb{N} = \{1,3,5,7,9,\ldots\}.$$

Some authors use alternative notation for the difference of sets A and B, for example, $A-B$. It is often useful to use number-lines or Venn diagrams to determine the intersection, union or difference of sets.

A.5 The Empty Set

We can define the empty set in the following way:

Definition A.7. *The **empty set** is a set that contains no elements and is denoted by*

$$\varnothing.$$

Equivalently,

$$\varnothing = \{x : x \neq x\}.$$

It follows that:

$$|\varnothing| = 0,$$

$$|\{\varnothing\}| = 1,$$

$$|\{\varnothing, \{\varnothing\}\}| = 2.$$

Also note that for any set A, $A \cup \varnothing = A$ and $A \cap \varnothing = \varnothing$.

A.6 Operations With Sets

Before considering operations with sets, recall these properties which describe familiar properties of numbers (for example, numbers in $\mathbb{R}$):

Remark A.8. *Let $a, b, c \in \mathbb{R}$.*

(i) Addition is Commutative.

That is $a + b = b + a$, i.e. order of the addition operation does not matter;

(ii) Addition is Associative.

That is $a + (b + c) = (a + b) + c$, i.e. re-bracketing does not effect addition; and

(iii) Addition satisfies the Distributive Law.

That is $a(b + c) = ab + ac$, i.e. expanding brackets is allowed.

We can look for similar properties in set operations, specifically:

$$A \cap B = B \cap A \text{ and } A \cup B = B \cup A,$$

so intersect and union set operations are commutative. However, clearly the difference set operation is not commutative[4], since in general,

$$A \setminus B \neq B \setminus A.$$

Theorem A.9. *For any three sets A, B and C:*

(i) $A \cap (B \cap C) = (A \cap B) \cap C$,

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

Proof: To establish (i), we must show that for arbitrary x: $x \in A \cap (B \cap C)$ implies $x \in (A \cap B) \cap C$, and; $x \in (A \cap B) \cap C$ implies $x \in A \cap (B \cap C)$. Since,

$$\begin{aligned} x \in (A \cap B) \cap C &\iff x \in (A \cap B) \text{ and } x \in C \\ &\iff x \in A \text{ and } x \in B \text{ and } x \in C \\ &\iff x \in A \text{ and } x \in (B \cap C) \\ &\iff x \in A \cap (B \cap C), \end{aligned}$$

[4]To make this clear, we should provide a suitable counter-example.

we conclude that

$$(A \cap B) \cap C \subseteq A \cap (B \cap C)$$

and

$$A \cap (B \cap C) \subseteq (A \cap B) \cap C.$$

Hence, $(A \cap B) \cap C = A \cap (B \cap C)$, as required.

To prove (ii), a similar argument can be used, and you should provide the details (see practice questions).

Theorem A.9 states that the set operations union and intersection, are associative (or satisfy the associative law).

Example 157: Is $A \cap (B \cup C) = (A \cap B) \cup C$?

Answer: In general, the answer is no. For example let $A = \varnothing$, $B = \mathbb{Z}$ and $C = \mathbb{Q}$. Then since $\varnothing \subset \mathbb{Z} \subset \mathbb{Q}$,

$$A \cap (B \cup C) = A \cap \mathbb{Q} = \varnothing \cap \mathbb{Q} = \varnothing$$

and

$$(A \cap B) \cup C = (\varnothing \cap \mathbb{Z}) \cup \mathbb{Q} = \mathbb{Q}.$$

Therefore,

$$A \cap (B \cup C) \neq (A \cap B) \cup C.$$

Example 158: Consider the sets A, B and C given by,

$$A = \{\text{apple}, \text{passion fruit}\},$$

$$B = \{\text{apple}, \text{pear}\},$$

$$C = \{\text{tomato}, \text{pear}\}.$$

Show that

$$A \cap (B \cup C) \neq (A \cap B) \cup C.$$

Answer: Since

$$A \cap (B \cup C) = \{\text{apple}, \text{passion fruit}\} \cap \{\text{apple}, \text{pear}, \text{tomato}\} = \{\text{apple}\}$$

and

$$(A \cap B) \cup C = \{\text{apple}\} \cup \{\text{tomato}, \text{pear}\} = \{\text{apple}, \text{tomato}, \text{pear}\},$$

it follows that

$$A \cap (B \cup C) \neq (A \cap B) \cup C.$$

Note that in the proof of Theorem A.9 and Examples 157 and 158, we see an important feature of mathematics: a **true** statement (a theorem, proposition, lemma etc) **requires proof**, whereas a **false** statement requires a **counterexample**.

A.7 The Universal Set

The sets we are interested in are often contained in a larger **universal set**, and hence, we can define the **complement** of a set A contained within this universal set.

Definition A.10. *Let U be a set, namely, the universal set. For $A \subseteq U$, we define the **complement** of A (with respect to U) as A', with*

$$A' = \{x \in U : x \notin A\}.$$

The complement of $A \subseteq U$ as in Definition A.10 can be expressed as $A' = U \setminus A$.

Example 159: If $U = \mathbb{R}$ and $A = \mathbb{Z}$ then

$$\begin{aligned} A' &= \{x \in \mathbb{R} \ : x \notin \mathbb{Z} \ \} \\ &= \{x \in \mathbb{R} \ : x \neq \ldots, -2, -1, 0, 1, 2, \ldots\}. \end{aligned}$$

The set in Example 159 can also be written using interval notation, specifically:

$$\begin{aligned} A' &= \cdots \cup (-2,-1) \cup (-1,0) \cup (0,1) \cup (1,2) \cup \cdots \\ &= \bigcup_{n \in \mathbb{Z}} (n, n+1). \end{aligned}$$

Example 160: What are the complements of the universal set and the empty set?

Answer: $U' = \varnothing$ and $\varnothing' = U$.

A.8 de Morgan's Laws

Theorem A.11. ***(de Morgan's Laws)*** *For sets A and B contained in a universal set U:*

(i) $(A \cup B)' = A' \cap B'$,

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

Proof: To establish (i), we must show that for arbitrary x: $x \in (A \cup B)'$ implies $x \in A' \cap B'$, and; $x \in A' \cap B'$ implies $x \in (A \cup B)'$. Since for $x \in U$,

$$\begin{aligned} x \in (A \cup B)' &\iff x \notin (A \cup B) \\ &\iff x \notin A \text{ and } x \notin B \end{aligned}$$

$$\begin{aligned} &\iff && x \in A' \text{ and } x \in B' \\ &\iff && x \in A' \cap B', \end{aligned}$$

it follows (as in the proof of Theorem A.9 since all implications above are if and only if) that $(A \cup B)' = A' \cap B'$, as required.

To prove (ii), a similar argument can be used, and ... again, you should provide it (see practice questions).

Corollary A.12. *For any sets A, B and C contained in a universal set U:*

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

Proof: Let $A_1 = A$ and $A_2 = B \cup C$. Then, via two applications of Theorem A.11, we have,

$$\begin{aligned} (A \cup B \cup C)' &= (A \cup (B \cup C))' \\ &= (A_1 \cup A_2)' \\ &= A_1' \cap A_2' \\ &= A_1' \cap (B \cup C)' \\ &= A_1' \cap (B' \cap C') \\ &= A' \cap B' \cap C', \end{aligned}$$

as required.

Theorem A.13. ***Distributive law for intersection and union:*** *For any sets A, B and C:*

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

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

Proof: To establish (i), since

$$\begin{aligned} x \in A \cap (B \cup C) &\iff && (x \in A) \text{ and } x \in (B \cup C) \\ &\iff && (x \in A) \text{ and } (x \in B \text{ or } x \in C) \\ &\iff && (x \in A \text{ and } x \in B) \text{ or } (x \in A \text{ and } x \in C) \\ &\iff && (x \in A \cap B) \text{ or } (x \in A \cap C) \\ &\iff && x \in (A \cap B) \cup (A \cap C), \end{aligned}$$

it follows that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, as required.

To establish (ii), a similar argument can be employed, and ... once again, you should provide it (see practice questions).

Results similar to Corollary A.12 and Theorem A.13 which are for operations on n sets (for $n \in \mathbb{N}$ with $n \geq 3$) can be established by mathematical induction[5], which we will consider in Chapter 2.

A.9 Cartesian Product

Definition A.14. *The Cartesian product of sets A and B, denoted $A \times B$, is defined as,*

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

Example 161: Let $A = \{1, 2, 3\}$ and $B = \{3, 5\}$. Then,

$$A \times B = \{(1, 3), (1, 5), (2, 3), (2, 5), (3, 3), (3, 5)\}.$$

Note that $|A| = 3$, $|B| = 2$ and $|A \times B| = 6$.

Example 162: Let $A = \{\heartsuit, \clubsuit\}$ and $B = \{x, y, z\}$. Observe that

$$A \times B = \{(\heartsuit, x), (\heartsuit, y), (\heartsuit, z), (\clubsuit, x), (\clubsuit, y), (\clubsuit, z)\},$$
$$B \times A = \{(x, \heartsuit), (y, \heartsuit), (z, \heartsuit), (x, \clubsuit), (y, \clubsuit), (z, \clubsuit)\},$$

and hence, $A \times B \neq B \times A$.

From Definition A.14 we can define functions from a relational approach[6].

Definition A.15. *Consider the sets X, Y and $R \subseteq X \times Y$. The set R is referred to as a binary relation over X and Y. Suppose that:*

- *for each $x \in X$, there exists (x, y) in R; and*
- *if $(x, y_1), (x, y_2) \in R$ for some $x \in X$ and $y_1, y_2 \in Y$, then $y_1 = y_2$.*

Then R defines a function $f : X \to Y$ via the rule $f(x) = y$ for all $(x, y) \in R$.

For example, $R = X \times Y$ in Example 161 does not define a function (since $(x, y_1), (x, y_2) \in R$ does not imply that $y_1 = y_2$). However, $R = \{(1, 3), (2, 5), (3, 5)\} \subset X \times Y$ defines a function $f : X \to Y$ given by: $f(1) = 3$, $f(2) = 5$ and $f(3) = 5$.

[5] We note here that "De Morgan's Laws" are named after British Mathematician Augustus De Morgan (1806-1871) who also gave the first rigorous treatment of mathematical induction [9].

[6] You will consider *relations* in more detail in J1AC [7].

Appendix B

Mathematical Induction

▸ **Learning Outcomes** ◂ After the completion of the lectures and support sessions associated with this chapter you should be able to:

- recall the statement of common forms of mathematical induction and understand how these are equivalent in simple cases; and
- produce correctly structured arguments using the principle of mathematical induction.

[2, p.109-115] and [5, p.201-207] contain alternative presentations of material in this chapter that you may find helpful.

Mathematical Induction is a method used to prove statements which hold for all $n \in \mathbb{N}$. We denote such statements by $P(n)$ for each $n \in \mathbb{N}$.

Example 163: Consider the statement $P(n)$ for $n \in \mathbb{N}$ given by,

$$\text{the sum of the first } n \text{ natural numbers is equal to } \frac{1}{2}n(n+1).$$

Equivalently, the statement $P(n)$ for $n \in \mathbb{N}$ is given by,

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

Theorem B.1. ***(Principle of mathematical induction)*** *Let $P(n)$ be a statement for each $n \in \mathbb{N}$. Additionally, suppose that both of the following statements are satisfied:*

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

(ii) for each $k \in \mathbb{N}$, we have

$$P(k) \text{ is true} \Longrightarrow P(k+1) \text{ is true.}$$

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

Note that we refer to $P(n)$, condition (i) and condition (ii) in Theorem B.1 as the **induction hypothesis**, the **base step** and the **induction step**, respectively. Additionally, note that we will prove Theorem B.1 using **proof by contradiction**, a widely used method of proof.

Proof: Suppose that for some $n \in \mathbb{N}$, the statement $P(n)$ is false. Therefore, there exists $\bar{n} \in \mathbb{N}$, for which, the statement $P(\bar{n})$ is false, and via (i), $P(k)$ is true for all $k \in \mathbb{N}$ such that $1 \leq k \leq \bar{n} - 1$.

Since $P(\bar{n} - 1)$ is true, (ii) implies that $P(\bar{n})$ is true, which contradicts the statement $P(\bar{n})$ is false. Therefore, the supposition is invalid, and we conclude that $P(n)$ is true for all $n \in \mathbb{N}$, as required.

Great care is needed to write **good** (and hence useful) proofs using the PMI (Principle of Mathematical Induction). The following example (with footnotes) provides a good model which you should use as a template in your proofs which use the PMI.

Example 164: Show, using the principle of mathematical induction, that for all $n \in \mathbb{N}$,

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

Answer: Let $P(n)$ for $n \in \mathbb{N}$ be the statement[1]

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

To establish condition (i) in Theorem B.1, observe that when $n = 1$, the LHS (left hand side) and RHS (right hand side) of (B.1) are given respectively by

$$\sum_{i=1}^{1} i = 1 \text{ and } \frac{1}{2}(1)(1+1) = 1, \tag{B.2}$$

[1] You should first state the *induction hypothesis* $P(n)$.

so

$$P(1) \text{ is true,} \tag{B.3}$$

i.e. the base step of Theorem B.1 is satisfied[2].

To establish condition (ii) in Theorem B.1 for $P(n)$, suppose that $P(k)$ is true for some $k \in \mathbb{N}$, or equivalently via (B.1), assume that

$$\sum_{i=1}^{k} i = \frac{1}{2}k(k+1). \tag{B.4}$$

For $n = k+1$, the LHS of (B.1) is given by

$$\begin{aligned}\sum_{i=1}^{k+1} i &= \sum_{i=1}^{k} i + (k+1) \\ &= \left(\frac{1}{2}k(k+1)\right) + (k+1) \quad \text{(via (B.4))} \\ &= (k+1)\left(\frac{1}{2}k+1\right) \\ &= \frac{1}{2}(k+1)(k+2).\end{aligned} \tag{B.5}$$

Since the RHS of (B.5) is the RHS of (B.1) for $n = k+1$, it follows from (B.4)-(B.5) that for $k \in \mathbb{N}$,

$$P(k) \text{ is true} \implies P(k+1) \text{ is true,} \tag{B.6}$$

i.e. $P(n)$ satisfies the induction step[3] in Theorem B.1. Therefore, via (B.3) and (B.6), $P(n)$ given by (B.1) satisfies the conditions of Theorem B.1, and we conclude[4] that $P(n)$ is true for all $n \in \mathbb{N}$, as required.

Note that the statement $P(n)$ is written explicitly in (B.1). Moreover, observe that the equations are numbered in the proof, and these are referred to in the text of the proof (and subsequent remark). When proving statements yourself, you are advised to include numbering next to all calculations (at first) so you can refer to them in the text that justifies your arguments (unused equation numbers can be removed afterwards).

[2]Secondly you should show that the *base step* is satisfied.

[3]Thirdly you should show that $P(n)$ satisfies the *induction step*.

[4]Finally, refer to the principle of mathematical induction (in this case Theorem B.1) to conclude your proof.

Theorem B.2. ***(Principle of Mathematical Induction)*** *Let $P(n)$ be a statement for each $n \in \mathbb{Z}$ with $n \geq m$, for some $m \in \mathbb{Z}$. Additionally, suppose that both of the followings statements are satisfied:*

(i) $P(m)$ is true; and

(ii) for each $k \in \mathbb{Z}$ with $k \geq m$, we have

$$P(k) \text{ is true} \implies P(k+1) \text{ is true.}$$

Then, $P(n)$ is true for all $n \in \mathbb{Z}$ with $n \geq m$.

Proof: The statement $Q(n)$ defined by

$$Q(n) = P(n+m-1) \quad \forall n \in \mathbb{N},$$

satisfies the conditions of Theorem B.1. From Theorem B.1 we conclude that $Q(n)$ is true for all $n \in \mathbb{N}$ and therefore, $P(n)$ is true for all $n \in \mathbb{Z}$ with $n \geq m$, as required.

Example 165: Show, using the principle of mathematical induction, that

$$3^n \geq 10 + 2^n \quad \forall n \in \mathbb{Z} \text{ with } n \geq 3.$$

Answer: Let $P(n)$ for $n \in \mathbb{Z}$ with $n \geq 3$, be the statement

$$3^n \geq 10 + 2^n. \tag{B.7}$$

We now show that $P(n)$ satisfies the conditions of Theorem B.2 with $m = 3$.

To establish that $P(n)$ satisfies condition (i), observe that the LHS and RHS of (B.7) with $n = 3$ are given respectively by,

$$3^3 = 27 \text{ and } 10 + 2^3 = 18. \tag{B.8}$$

Since $27 > 18$ it follows from (B.7) and (B.8) that

$$P(3) \text{ is true.} \tag{B.9}$$

To establish that $P(n)$ satisfies condition (ii) with $m = 3$, suppose that for some $k \in \mathbb{Z}$ with $k \geq 3$, that $P(k)$ is true, i.e.

$$3^k \geq 10 + 2^k. \tag{B.10}$$

Via (B.7), the LHS of $P(k+1)$ is given by

$$\begin{aligned} 3^{k+1} &= 3 \times 3^k \\ &\geq 3 \times (10 + 2^k) \quad \text{(via (B.10))} \end{aligned}$$

$$
\begin{aligned}
&= 30 + (3 \times 2^k) \\
&\geq 10 + 2 \times 2^k \\
&= 10 + 2^{k+1} = \text{ RHS of } P(k+1).
\end{aligned} \tag{B.11}
$$

Thus, via (B.10) and (B.11), for $k \in \mathbb{Z}$ with $k \geq 3$, we conclude that

$$
P(k) \text{ is true} \Longrightarrow P(k+1) \text{ is true.} \tag{B.12}
$$

Therefore, via (B.9) and (B.12) respectively, the base step and induction step in Theorem B.2 are satisfied ($m = 3$) and hence, via Theorem B.2 we conclude that for all $n \in \mathbb{Z}$ with $n \geq 3$, that $P(n)$ given by (B.7) is true, as required.

The hardest situations arise where some initial guess has to be made for the base step, for example:

Example 166: $n!$ (pronounced n factorial) is defined as

$$
n! = \prod_{i=1}^{n} i = n \times (n-1) \times (n-2) \times \ldots 3 \times 2 \times 1.
$$

Using Theorem B.2, find $m \in \mathbb{Z}$ such that $n! > 10^n$ for all $n \in \mathbb{Z}$ with $n \geq m$. For example, when $n = 3$, $3! = 6$ and $10^3 = 1000$ i.e. the statement does not hold for $n = 3$.

Note that to apply the principle of mathematical induction, as stated in Theorems B.1 and B.2, the base step and induction step need to be satisfied by the induction hypothesis.

If one does not establish conditions in these theorems before drawing conclusions from them, some 'invalid' statements might be claimed. To highlight this, see if you can find the mistakes in the following two examples.

Example 167: All natural numbers of the form $F_n = 2^{2^n} + 1$ with $n \in \mathbb{N}_0$ are prime numbers.

Proof: It follows that

$$
\begin{aligned}
&F_0 = 3 \text{ which is a prime number.} \\
&F_1 = 5 \text{ which is a prime number.} \\
&F_2 = 17 \text{ which is a prime number.} \\
&F_3 = 257 \text{ which is a prime number.} \\
&F_4 = 65537 \text{ which is a prime number.} \\
&\vdots
\end{aligned}
$$

One can establish that for $k \in \mathbb{N}_0$, if F_k is a prime number, then F_{k+1} is a prime number. Therefore, the conditions of Theorem B.2 are satisfied (with $m = 0$), and we conclude that $P(n)$ is true for all $n \in \mathbb{N}_0$, as required.

Example 168: Prove that every person in the room has the same hairstyle.

Proof: Consider the statement $P(n)$ for $n \in \mathbb{N}$ given by

For any n people in this room, they all have the same hairstyle.

Now, for any 1 person in the room, they have the same hairstyle as ... themselves, and hence P(1) is true.

Now assume that $P(k)$ is true for some $k \geq 1$. Then for any group of $k+1$ people, remove 1 person, and hence via the assumption, all remaining k people have the same hairstyle. Additionally, consider the same $k+1$ people and remove a different person. Via the assumption, it follows that all remaining k people have the same hairstyle, and hence all $k+1$ people have the same hairstyle. Therefore, the conditions of Theorem B.1 are satisfied, and we conclude that $P(n)$ is true for all $n \in \mathbb{N}$, as required.

Index

Bibliography

[1] 2011. URL http://www.cliffsnotes.com/assets/256137.png. [Accessed 2017-2-14].

[2] R. A. Adams and C. Essex. *Calculus: a Complete course*. Pearson, 8th edition, Toronto, 2013.

[3] S. C. Althoen and R. Mclauglin. Gauss-Jordan reduction: A brief history. *The American mathematical monthly*, 94(2):130–142, 1987.

[4] R. A. Beezer. A first course in linear algebra, version 3.5, 2015. URL http://linear.pugetsound.edu./. [Accessed 2017-17-12].

[5] J. E. Fields. *A gentle introduction to the art of mathematics, Version 3.1*. GNU, 2013. URL http://giam.southernct.edu/GIAM/GIAM.pdf.

[6] The Quality Assurance Agency for Higher Education. Academic credit in higher education in england - an introduction, 2009. URL http://www.qaa.ac.uk/en/Publications/Documents/Academic-credit-in-higher-education-in-England---an-introduction.pdf. [Accessed 2017-03-12].

[7] S. Goodwin. Algebra and combinatorics (lecture notes), 2020.

[8] D. Pixton M. Beck, G. Marchesi and L. Sabalka. A first course in complex analysis, 2012. URL http://math.sfsu.edu/beck/papers/complexorth.pdf. [Accessed 2019-01-19].

[9] J. J. O'Connor and E. F. Robertson. Augustus de morgan, 1996. URL http://www-history.mcs.st-andrews.ac.uk/Biographies/De_Morgan.html. [Accessed 2017-10-12].

[10] J. J. O'Connor and E. F. Robertson. René descartes, 1996. URL http://www-groups.dcs.st-and.ac.uk/history/Biographies/Descartes.html. [Accessed 2018-01-22].

[11] J. J. O'Connor and E. F. Robertson. Leonhard euler, 1996. URL http://www-history.mcs.st-and.ac.uk/Biographies/Euler.html. [Accessed 2018-04-14].

[12] The Norwegian Academy of Science and Letters. Niels henrik abel. URL http://www.abelprize.no/c53672/seksjon/vis.html?tid=53910. [Accessed 2018-04-14].

[13] H. A. Priestley. *Introduction to Complex Analysis*. Oxford University Press, 2nd edition, GB, 2003.

[14] T. Tao. *Analysis I*. Hindustan, 2nd edition, New Delhi, 2009.

[15] Y. Wang and J. Huang. Real analysis and the calculus (lecture notes), 2019.

These lecture notes include contributions from Dr H. C. Wilkie, Prof. J. R. Blake, Dr J. Kyle, Mr B. J. Philp, Dr D. F. M. Hermans, Dr J. C. Meyer and Prof. C. W. Parker. The original mathematical typesetting was undertaken using LaTeX by Mr B. Taylor.