Algebra MA183 : Lecture Notes

Dr Rachel Quinlan
School of Mathematics, Statistics and Applied Mathematics, NUI Galway

November 7, 2013

Contents

Preface

> *"Don't just read it, fight it! Ask your own questions, look for your own examples, discover your own proofs. Is the hypothesis necessary? Is the converse true? What happens in the classical special case? What about the degenerate cases? Where does the proof use the hypothesis? "*
>
> *– P. Halmos*

These lecture notes are for Semester 1 of the Algebra section of the first year honours Mathematics course at NUI Galway. The word "algebra" is often used in elementary contexts to refer to mathematical situations in which "symbols" such as letters are used to represent numerical quantities which may be fixed or variable. This is the context in which many people first encounter the word "algebra" but it is very far away from being a description of what algebra is about. Algebra is the study of *algebraic structures* which are very generally defined as systems (of numbers, functions, polynomials, matrices or other objects) with some scheme for combining pairs of these objects to produce new objects of the same type. For example *addition* is a scheme which allows us to combine pairs of integers to produce new integers. Matrix multiplication is a scheme which allows us to take a pair of 2×2 matrices and combine them to produce a new 2×2 matrix.

Algebra is a vast subject with many specializations, and it is the subject of extensive and vigourous research worldwide. NUI Galway has a long and continuing tradition of research in algebra. Areas of current activity here include group theory, linear algebra, representation theory, combinatorics and various aspects of computational algebra.

In this course we will look at two aspects of algebra. Chapter 1 of these notes consists of an introduction to linear algebra. Linear algebra is the theory of *vector spaces* and functions between them. It is closely related to matrix arithmetic and also to Euclidean geometry in 2,3 or more dimensions.

Chapter 2 of these notes consists of an introduction to number theory. Number Theory could be described as the mathematics of the integers or whole numbers. It is concerned with such things as the factorization of integers as products of primes and the distribution of prime numbers.

These lecture notes constitute the "text" for the first semester algebra course. You should study them carefully and attentively. This course is fundamentally about discussing mathematical concepts and reasoning about them, rather than being about learning how to perform particular types of calculations, implement procedures or "work out examples".

Of course, the ability to perform calculations and implement procedures of certain types is extremely important, but it is only useful when it is accompanied by a sound conceptual understanding of the meaning of the calculation and the rationale for carrying it out. For example, one of the procedures that will be encountered in this course will be the multiplication of matrices. Begin able to calculate the product of two matrices is an important procedural skill that you will be expected to master under your own direction, using an example from the lecture notes (or examples from books) for guidance. This takes a bit of practice as matrix multiplication is defined in a non-obvious way. We will be interested in understanding *why* this is the case; this understanding will add meaning to the procedural skill of carrying out the multiplication. In summary this course will be about *why* at least as much as about *how*.

In order to discuss the meaning of mathematical concepts, we need a precise language that is entirely unambiguous and not open to misinterpretation. In written mathematics, *every mark on the page has meaning*, and you must say *exactly what you mean* and not assume that the reader will know what you are talking about if you are vague or sloppy. When you are studying these lecture notes and other mathematical texts, it is very important to be attentive to detail. Every symbol means something, and different symbols have different meanings. For example, the written expressions

$$1, 2 \qquad \{1, 2\} \qquad (1, 2) \qquad [1, 2] \qquad [1, 2) \qquad (1, 2]$$

all mean different things (some of these have more than one meaning just to make things interesting). The symbols $\mathbb{R}$ and $\mathbb{R}^2$, which will be prominent in this course, mean different things and are not interchangeable.

The following statements might look vaguely alike but they have completely different meanings

- For every integer n there exists an even integer m such that $n < m$.
- There exists an even integer m such that $m < n$ for every integer n.

Reading mathematical texts is unlike reading other passages of prose - it is not enough to just read the words. You have to tease out the meaning line by line and satisfy yourself that you are understanding every single bit. This process is painstaking and can be time consuming, especially at first, but if you don't get into the habit of reading mathematics in a critical and questioning maner, you will not make progress. When reading mathematical texts, do not be tempted to skip the passages of prose and only study the "calculations" or the examples. If you do this, you will not realize your potential.

These notes are intended for independent study. Everything in the notes is on the syllabus for the course unless explicitly stated otherwise. I hope that you will find these notes helpful in supporting your activities in this course, which will include attendance at lectures, participation in workshops and tutorials, working on homework assignments and independent study. I will be grateful for any feedback you can offer on the lecture notes and on other aspects of the course.

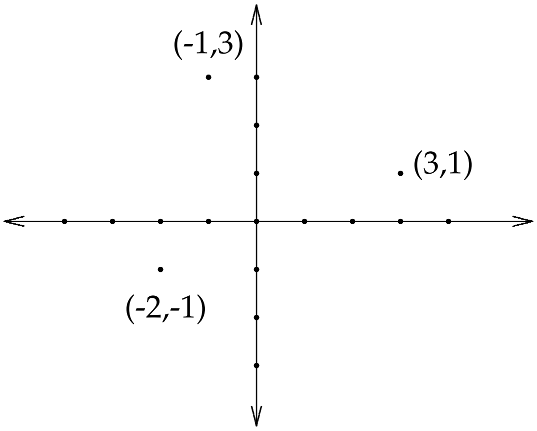
Rachel Quinlan.
School of Mathematics, Statistics and Applied Mathematics
NUI Galway

Chapter 1

Matrices and Linear Transformations

1.1 The Euclidean Plane $\mathbb{R}^2$

The 2-dimensional plane is described by a pair of axes labelled X and Y. Each point in the plane corresponds to the *ordered pair* of real numbers, consisting firstly of its X-coordinate and secondly of its Y-coordinate.



NOTE: The set of real numbers is denoted by $\mathbb{R}$ and can be considered just as a set or as the real number line.

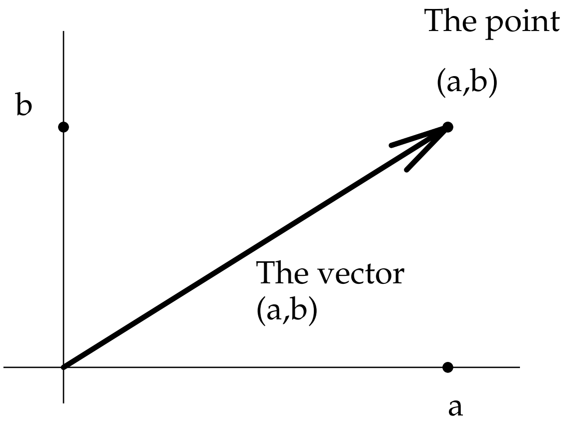
The set of *ordered pairs* of real numbers is denoted by $\mathbb{R}^2$:

$$\mathbb{R}^2 = \{(a, b) : a \in \mathbb{R}, b \in \mathbb{R}\}.$$

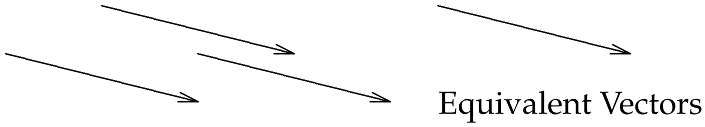
We can think of $\mathbb{R}^2$ either as this set (abstract, set-theoretic description) or as the set of points in the Cartesian plane above (more concrete, geometric description).

QUESTION: What do we understand by the object (a, b) (e.g. $(1, -3), (3/5, 100), (-\pi, \sqrt{3})$ etc.)? It will be useful to understand this object in (at least) three different ways :

1. Just as an ordered pair of numbers.
2. As the *point* in the plane with X-coordinate a and Y-coordinate b.
3. As the *vector* $\vec{v}$ in the plane directed from the origin O $(0, 0)$ to the point (a, b). Vectors in $\mathbb{R}^2$ are line segments with a direction. If $\vec{v} = (a, b)$ is considered as a vector, the numbers a and b are referred to as the X-component and Y-component of $\vec{v}$.



NOTE ON VECTORS: The vector $\vec{v} = (a, b)$ is the line segment directed from $(0, 0)$ to (a, b). Any directed line segment in which the terminal point (end point) can be reached from the initial point (start point) by travelling a units along the X-axis (right or left according as a is positive or negative) and b units along the Y-axis (up if b is positive, down if negative) is said to be *equivalent* to $\vec{v}$. Equivalent vectors are considered to be the same. This means that given a vector $\vec{v}$, we can move it around in $\mathbb{R}^2$ as long as we do not change its length or direction.



ADDITION IN $\mathbb{R}^2$

Let (a_1, b_1) and (a_2, b_2) be elements of $\mathbb{R}^2$. We define their *sum* in $\mathbb{R}^2$ by

$$\boxed{(a_1, b_1) + (a_2, b_2) = (a_1 + a_2, b_1 + b_2).}$$

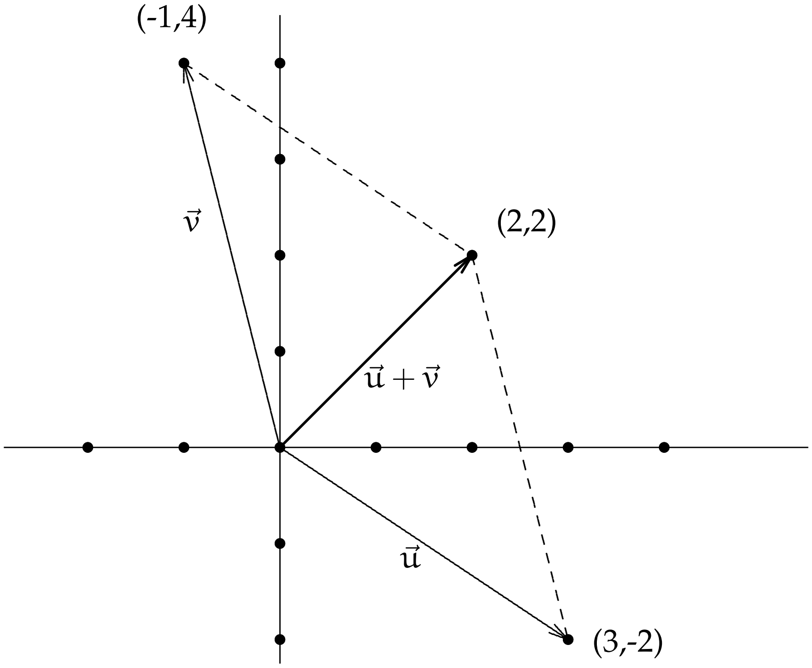
EXAMPLES

1. $(-3, 5) + (2, 4) = (-3 + 2, 5 + 4) = (-1, 9)$.

2. $(\sqrt{2}, 1) + (-\sqrt{3}, \sqrt{5}) = (\sqrt{2} - \sqrt{3}, 1 + \sqrt{5})$.

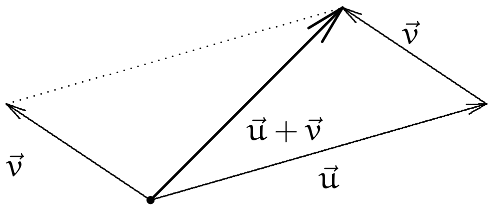
Geometric Interpretation of Addition:

1. **In terms of *points***: The point with coordinates $(a_1 + a_2, b_1 + b_2)$ is the fourth vertex of the parallelogram that has $(a_1, b_1), (0, 0)$ and (b_1, b_2) as three consecutive vertices around its perimeter.



2. **In terms of *vectors***: If $\vec{u}$ and $\vec{v}$ are interpreted as vectors originating at the origin, their sum $\vec{u} + \vec{v}$ is a diagonal of the parallelogram having $\vec{u}$ and $\vec{v}$ as two sides.

   Alternatively : if $\vec{u}$ and $\vec{v}$ are vectors, their sum $\vec{u} + \vec{v}$ can be defined as follows - position $\vec{v}$ with its initial point at the terminal point of $\vec{u}$. The arrow directed from the initial point of $\vec{u}$ to the terminal point of $\vec{v}$ then is the vector $\vec{u} + \vec{v}$.



Scalar Multiplication

In the context of vectors in $\mathbb{R}^2$, real numbers are often referred to as *scalars*. We can multiply an element of $\mathbb{R}^2$ by a scalar as follows : let $v \in \mathbb{R}^2, v = (a, b)$. If $k \in \mathbb{R}$ then $kv \in \mathbb{R}^2$ is defined by

$$\boxed{kv = k(a,b) = (ka,kb).}$$

GEOMETRIC INTERPRETATION OF SCALAR MULTIPLICATION: Consider $\vec{v}$ to be a vector with initial point at $O(0,0)$. Then kv is a vector whose length is $|k|\times$(length of v), and whose direction is

- the same as that of v if $k > 0$ (if k is positive)

- opposite to that of v if $k < 0$ (if k is negative)

NOTE: For a real number k, $|k|$ denotes the *absolute value* of k.
This is equal to k if $k \geqslant 0$, and equal to $-k$ (a positive number) if k is negative.
For example the absolute value of -2 is 2; $|-2| = 2$. The absolute value of a non-zero real number is always positive.

Once equipped with these operations of addition and scalar multiplication, $\mathbb{R}^2$ is no longer just a set - it has *algebraic structure*. (In fact $\mathbb{R}^2$ is an example of a *vector space*).

1.2 Linear Transformations of $\mathbb{R}^2$

Amongst the most fundamental objects in mathematics are *sets*, which are collections of objects known as *elements*. Almost as fundamental as sets are *functions* which are vehicles for travelling from the elements of one set to elements of another. A function from the set A to the set B is an association of some element of B to every element of A.

Sometimes sets are not just amorphous collections of objects, but have extra features, like order, or some arithmetic or algebraic structure. In these cases we might not be interested in *all* functions between such sets, but maybe in functions which have some kind of "good behaviour" with respect to the properties of the set. For example, when doing calculus we are usually not interested in *all* functions from $\mathbb{R}$ to $\mathbb{R}$, but maybe in those that are continuous. Roughly, continuous functions cannot move points that are close together to points that are far apart; they must respect the proximity of points in some way.

Similary in algebra, when looking at functions between sets with algebraic structure, it is common practice to focus on those functions that behave well with respect to that structure.

Example 1.2.1 *(a) Consider the function* $T : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ *defined for all* $(x, y) \in \mathbb{R}^2$ *by*

$$T(x, y) = (3x - 2y, -x).$$

So $T(2, 4) = (3(2) - 2(4), -2) = (-2, -2)$, etc.
Suppose that $u = (a_1, b_1)$ and $v = (a_2, b_2)$ are elements of $\mathbb{R}^2$. Then we can form $u + v = (a_1 + a_2, b_1 + b_2) \in \mathbb{R}^2$.

QUESTION: Will adding u and v and then applying T to the result give the same outcome as applying T separately to u and v and then adding their images?
To check :

$$\begin{aligned}
T(u+v) &= T((a_1, b_1) + (a_2, b_2)) \\
&= T(a_1 + a_2, b_1 + b_2) \\
&= (3(a_1 + a_2) - 2(b_1 + b_2), -(a_1 + a_2)) \\
&= (3a_1 + 3a_2 - 2b_1 - 2b_2, -a_1 - a_2). \\
T(u) + T(v) &= T(a_1, b_1) + T(a_2, b_2) \\
&= (3a_1 - 2b_1, -a_1) + (3a_2 - 2b_2, -a_2) \\
&= (3a_1 - 2b_1 + 3a_2 - 2b_2, -a_1 - a_2) \\
&= (3a_1 + 3a_2 - 2b_1 - 2b_2, -a_1 - a_2) \\
&= T(u+v).
\end{aligned}$$

So $T(u+v) = T(u) + T(v)$ for all $u, v \in \mathbb{R}^2$. We say that T is *additive* or that T respects addition.

(b) Another Question: Let $u = (a, b)$ in $\mathbb{R}^2$ and suppose $k \in \mathbb{R}$. Is multiplying u by k and then applying T the same as applying T to u and then multiplying the result by k?
To check :

$$\begin{aligned}
T(ku) &= T(ka, kb) \\
&= (3ka - 2kb, -ka). \\
kT(u) &= kT(a, b) \\
&= k(3a - 2b, -a) \\
&= (3ka - 2kb, -ka) \\
&= T(ku).
\end{aligned}$$

So $T(ku) = kT(u)$ for all $u \in \mathbb{R}^2$ and $k \in \mathbb{R}$. We say that T respects scalar multiplication.

(c) Suppose a function $S : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is defined by

$$S(x, y) = (xy, -y)$$

for all $(x, y) \in \mathbb{R}^2$. Then S is not additive since for example

$$\begin{aligned} S((1,0) + (0,2)) &= S(1,2) = (2,-2). \\ S(1,0) + S(0,2) &= (0,0) + (0,-2) = (0,-2) \neq (2,-2). \end{aligned}$$

Nor does S respect scalar multiplication since for example

$$\begin{aligned} S(2(1,1)) &= S(2,2) = (4,-2) \\ \text{but } 2S(1,1) &= 2(1,-1) = (2,-2) \neq (4,-2). \end{aligned}$$

NOTE: In (c) above, to show that S is not additive it is enough to show that $S((1,0) + (0,2)) \neq S(1,0) + S(0,2)$, i.e. it is enough to show that the additivity fails for one particular pair of points in $\mathbb{R}^2$. To show that a function *is* additive, it is not enough to just use one particular pair of points. Think about why this is.

Definition 1.2.2 *Let* $T : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ *be a function.* T *is a* linear transformation *of* $\mathbb{R}^2$ *if it respects both addition and scalar multiplication, i.e. if*

$$\begin{aligned} T(u + v) &= T(u) + T(v) \text{ for all } u, v \in \mathbb{R}^2, \text{ and} \\ T(ku) &= kT(u) \text{ for all } u \in \mathbb{R}^2 \text{ and } k \in \mathbb{R}. \end{aligned}$$

Note: Suppose $u = (a, b) \in \mathbb{R}^2$. Then $a, b \in R$ and $u = (a, 0) + (0, b)$. We have $(a, 0) = a(1, 0)$ and $(0, b) = b(0, 1)$. So

$$u = (a, b) = a(1, 0) + b(0, 1).$$

Thus any element of $\mathbb{R}^2$ can be written as the sum of a scalar multiple of $(1, 0)$ and a scalar multiple of $(0, 1)$.

The set $\{(1, 0), (0, 1)\}$ is called the *standard basis* of $\mathbb{R}^2$.

Claim: If $T : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is a linear transformation and we know $T(1, 0)$ and $T(0, 1)$, we can write down $T(x, y)$ for any $(x, y) \in \mathbb{R}^2$.

Example 1.2.3 *Suppose* $T : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ *is a linear transformation. If* $T(1, 0) = (-1, 2)$ *and* $T(0, 1) = (3, 4)$, *what is*

(i) $T(-1, 5)$? (ii) $T(3, -1/2)$?

Solution : (i)

$$\begin{aligned} T(-1,5) &= T((-1,0) + (0,5)) \\ &= T(-1(1,0) + 5(0,1)) \\ &= T(-1(1,0)) + T(5(0,1)) \\ &= -1T(1,0) + 5T(0,1) \\ &= -1(-1,2) + 5(3,4) \\ &= (1,-2) + (15,20) \\ &= (16,18). \end{aligned}$$

(ii) Answer is $(-9/2, 4)$ - Exercise.

In general we have the following statement.

Theorem 1.2.4 *Suppose that* $\mathsf{T} : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ *is a linear transformation satisfying* $\mathsf{T}(1,0) = (a,b)$ *and* $\mathsf{T}(0,1) = (c,d)$*. Then if* (x,y) *is any element of* $\mathbb{R}^2$*, we have*

$$\mathsf{T}(x,y) = (ax + cy, bx + dy).$$

Proof:

$$\begin{aligned}
\mathsf{T}(x,y) &= \mathsf{T}((x,0) + (0,y)) \\
&= \mathsf{T}(x(1,0) + y(0,1)) \\
&= \mathsf{T}(x(1,0)) + \mathsf{T}(y(0,1)) \\
&= x\mathsf{T}(1,0) + y\mathsf{T}(0,1) \\
&= x(a,b) + y(c,d) \\
&= (ax, bx) + (cy, dy) \\
&= (ax + cy, bx + dy).
\end{aligned}$$

1.3 The Matrix of a Linear Transformation

Let $T : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ be a linear transformation. Theorem 1.2.4 tells us that we can calculate $T(x, y)$ for any $(x, y) \in \mathbb{R}^2$ if we have just four pieces of information : namely the x and y coordinates of $T(1, 0)$ and $T(0, 1)$. Suppose

$$T(1, 0) = (a, b), \quad T(0, 1) = (c, d).$$

We can encode this information by writing

$$M_T = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$$

For example if $T(1, 0) = (3, -1)$ and $T(0, 1) = (2, -4)$ we would write

$$M_T = \begin{pmatrix} 3 & 2 \\ -1 & -4 \end{pmatrix}.$$

Definition 1.3.1 M_T *is called the* matrix *of the linear transformation* T.

- In general a matrix is a rectangular array of numbers.
- M_T above is a 2×2 matrix (2 rows, 2 columns).
  A $m \times n$ ("m by n") matrix has m (horizontal) rows and n (vertical) columns. For example $\begin{pmatrix} 1 & 5 & -6 \\ 2 & 3 & -2 \end{pmatrix}$ is a 2×3 matrix.

Back to T with $T(1, 0) = (a, b)$, $T(0, 1) = (c, d)$.
Suppose $(x, y) \in \mathbb{R}^2$. We can use M_T to calculate $T(x, y)$ as follows.

1. Write (x, y) as the two entries in a column vector (a column vector is a matrix with 1 column) : $\binom{x}{y}$.

2. Form the *matrix product* $M_T \binom{x}{y}$ as follows :

$$M_T \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

   This product is a 2×1 matrix whose entries are formed from those of M_T and $\binom{x}{y}$ according to the following instruction.

   The first entry of the product comes from combining the entries a c of the first row of M_T with the entries x y of the column $\binom{x}{y}$ by taking

   (Product of first entries) + (Product of 2nd entries)

$$(a \times x) + (b \times y) = ax + by.$$

   The second component comes from the same procedure applied to the second row of M_T and the column $\binom{x}{y}$:

$$(b \times x) + (c \times y) = bx + dy.$$

   So we have

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax + cy \\ bx + dy \end{pmatrix}.$$

3. This is a definition of matrix multiplication (for a 2×2 matrix by a 2×1 matrix at least).

4. Recall $M_T = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$ is the matrix of T, where $T(1,0) = (a,b)$ and $T(0,1) = (c,d)$. By Theorem 1.2.4 $T(x,y) = (ax + cy, bx + dy)$ for any $(x,y) \in \mathbb{R}^2$. So the coordinates of $T(x,y)$ are the entries of the column vector $M_T \binom{x}{y}$.

Example 1.3.2 *Suppose* $T : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ *is a linear transformation sending* $(1,0)$ *to* $(3,-2)$ *and* $(0,1)$ *to* $(4,1)$. *Write down the matrix of* T *and use it to calculate*

(i) $T(2,3)$ (ii) $T(-1,2)$

Solution: The matrix of T is $M_T = \begin{pmatrix} 3 & 4 \\ -2 & 1 \end{pmatrix}$.

(i) To calculate $T(2,3)$, form the matrix product $M_T \binom{2}{3}$. This is

$$\begin{pmatrix} 3 & 4 \\ -2 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 3(2) + 4(3) \\ -2(2) + 1(3) \end{pmatrix} = \begin{pmatrix} 18 \\ -1 \end{pmatrix}.$$

Thus $T(2,3) = (18,-1)$.

(ii) $T(-1,2) = (5,4)$ (Exercise).

The Sum of Linear Transformations

Suppose $T_1 : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ and $T_2 : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ are linear transformations. Then we can define a function

$$(T_1 + T_2) : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$$

by declaring $(T_1 + T_2)(u) = T_1(u) + T_2(u)$ for all $u \in \mathbb{R}^2$.

Claim: $T_1 + T_2$ is a linear transformation.

Proof: We need to show that $T_1 + T_2$ respects addition and scalar multiplication, given that T_1 and T_2 do. So let $u, v \in \mathbb{R}^2$. Then

$$\begin{aligned} (T_1 + T_2)(u + v) &= T_1(u + v) + T_2(u + v) \\ &= T_1(u) + T_1(v) + T_2(u) + T_2(v) \\ &= T_1(u) + T_2(u) + T_1(v) + T_2(v) \\ &= (T_1 + T_2)(u) + (T_1 + T_2)(v). \end{aligned}$$

So $T_1 + T_2$ respects addition.

Now suppose $u \in \mathbb{R}^2$ and $k \in \mathbb{R}$. We have

$$\begin{aligned} (T_1 + T_2)(ku) &= T_1(ku) + T_2(ku) \\ &= kT_1(u) + kT_2(u) \\ &= k(T_1(u) + T_2(u)) \\ &= k((T_1 + T_2)(u)). \end{aligned}$$

So $T_1 + T_2$ respects scalar multiplication. □

Question: How (if at all) does the matrix of $T_1 + T_2$ depend on the matrices of T_1 and T_2?

Suppose $M_{T_1} = \begin{pmatrix} a_1 & c_1 \\ b_1 & d_1 \end{pmatrix}$ so $T_1(1,0) = (a_1, b_1)$ and $T_1(0,1) = (c_1, d_1)$.

Suppose $M_{T_2} = \begin{pmatrix} a_2 & c_2 \\ b_2 & d_2 \end{pmatrix}$ so $T_2(1,0) = (a_2, b_2)$ and $T_2(0,1) = (c_2, d_2)$.
Then $(T_1 + T_2)(1,0) = (a_1, b_1) + (a_2, b_2) = (a_1 + a_2, b_1 + b_2)$.
Also $(T_1 + T_2)(0,1) = (c_1, d_1) + (c_2, d_2) = (c_1 + c_2, d_1 + d_2)$.
So the matrix of $T_1 + T_2$ is

$$\begin{pmatrix} a_1 + a_2 & c_1 + c_2 \\ b_1 + b_2 & d_1 + d_2 \end{pmatrix}.$$

This motivates the following definition

Definition 1.3.3 *(Matrix Addition) Suppose* A *and* B *are* 2×2 *matrices; write*

$$A = \begin{pmatrix} a_1 & c_1 \\ b_1 & d_1 \end{pmatrix}, \quad B = \begin{pmatrix} a_2 & c_2 \\ b_2 & d_2 \end{pmatrix}.$$

We define $A + B$ *to be the* 2×2 *matrix*

$$\begin{pmatrix} a_1 + a_2 & c_1 + c_2 \\ b_1 + b_2 & d_1 + d_2 \end{pmatrix}.$$

In general if A and B are $m \times n$ matrices their sum $A + B$ is obtained by adding entries of A and B in corresponding positions. For example

$$\begin{pmatrix} 2 & 4 \\ -1 & 2 \end{pmatrix} + \begin{pmatrix} 1 & 1 \\ -2 & 0 \end{pmatrix} = \begin{pmatrix} 2+1 & 4+1 \\ -1+(-2) & 2+0 \end{pmatrix} = \begin{pmatrix} 3 & 5 \\ -3 & 2 \end{pmatrix},$$

and

$$\begin{pmatrix} 1 & 3 & 2 \\ 0 & -1 & -4 \end{pmatrix} + \begin{pmatrix} 2 & -2 & 5 \\ 6 & 3 & 1 \end{pmatrix} = \begin{pmatrix} 1+2 & 3+(-2) & 2+5 \\ 0+6 & -1+3 & -4+1 \end{pmatrix} = \begin{pmatrix} 3 & 1 & 7 \\ 6 & 2 & -3 \end{pmatrix}.$$

NOTE: If A and B are matrices of different sizes (i.e. not both $m \times n$ for the same m and n) we do not define their sum.

MULTIPLICATION OF MATRICES BY SCALARS

Suppose $T : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is a linear transformation with matrix $M_T = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$. If $k \in \mathbb{R}$ we can define a function

$$(kT) : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$$

by $(kT)(u) = k(T(u))$, for $u \in \mathbb{R}^2$.
It is easily checked that kT is a linear transformation, and we have

$$(kT)(1,0) = k(a,b) = (ka, kb), \quad (kT)(0,1) = k(c,d) = (kc, kd).$$

Hence the matrix of kT is $M_{kT} = \begin{pmatrix} ka & kc \\ kb & kd \end{pmatrix}$ - this is M_T with all the entries multiplied by k.

Definition 1.3.4 *Let* M *be a matrix (of any size) and let* $k \in \mathbb{R}$*. The matrix* kM *by definition is obtained from* M *by multiplying all of the entries of* M *by* k*. It has the same size as* M*.*

EXAMPLES:

$-2\begin{pmatrix} 1 & -1 \\ 3 & -2 \end{pmatrix} = \begin{pmatrix} -2 & 2 \\ -6 & 4 \end{pmatrix}$,

$\frac{1}{2}\begin{pmatrix} 2 & 4 & -1 \\ -6 & 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 2 & -\frac{1}{2} \\ -3 & \frac{3}{2} & 2 \end{pmatrix}$.

COMPOSITION OF LINEAR TRANSFORMATIONS : MATRIX MULTIPLICATION

Suppose that $T : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ and $S : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ are linear transformations. Then one can define a function from $\mathbb{R}^2$ to $\mathbb{R}^2$ that sends $u \in \mathbb{R}^2$ to $T(S(u))$, i.e. the function applies S to u first and then applies T to the result. This function is called the *composition* of T with S. It is denoted by $T \circ S$ (read this as T *after* S; first S, then T).

Theorem 1.3.5 $T \circ S : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ *is a linear transformation.*

Proof: Let $u, v \in \mathbb{R}^2$. Then

$$\begin{aligned} T \circ S(u+v) &= T(S(u+v)) \\ &= T(S(u)+S(v)) \ - \text{ because S is additive} \\ &= T(S(u))+T(S(v)) \ - \text{ because T is additive} \\ &= T \circ S(u) + T \circ S(v). \end{aligned}$$

So $T \circ S$ is additive.

To see that $T \circ S$ respects scalar multiplication, let $u \in \mathbb{R}^2$ and let $k \in \mathbb{R}$. Then

$$\begin{aligned} T \circ S(ku) &= T(S(ku)) \\ &= T(kS(u)) \ - \text{ because S respects scalar multiplication} \\ &= k(T(S(u))) \ - \text{ because T respects scalar multiplication} \\ &= k\,T \circ S(u). \end{aligned}$$

So $T \circ S$ respects scalar multiplication. □

REMARK: By the same reasoning $S \circ T$ is also a linear transformation, though typically it is not the same as $T \circ S$.

Question: How does the matrix of $T \circ S$ depend on the matrices of T and S?

Example 1.3.6 *Suppose* $T : \mathbb{R}^2 \longrightarrow R^2$ *and* $S : \mathbb{R}^2 \longrightarrow R^2$ *are linear transformations with*

$$M_T = \begin{pmatrix} 3 & 1 \\ -2 & 4 \end{pmatrix}, \quad M_S = \begin{pmatrix} 1 & -4 \\ 5 & 3 \end{pmatrix}.$$

To write down the matrix of the composition $T \circ S$, we need to calculate the images under $T \circ S$ of $(1,0)$ and $(0,1)$.
First, $T \circ S(1,0) = T(S(1,0)) = T(1,5)$.
This can be found by calculating the matrix product $M_T\binom{1}{5}$:

$$\begin{pmatrix} 3 & 1 \\ -2 & 4 \end{pmatrix}\begin{pmatrix} 1 \\ 5 \end{pmatrix} = \begin{pmatrix} 8 \\ 18 \end{pmatrix} \Longrightarrow T \circ S(1,0) = (8,18).$$

$T \circ S(0,1) = T(-4,3)$.

$$M_T \begin{pmatrix} -4 \\ 3 \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ -2 & 4 \end{pmatrix} \begin{pmatrix} -4 \\ 3 \end{pmatrix} = \begin{pmatrix} -9 \\ 20 \end{pmatrix} \Longrightarrow T \circ S(0,1) = (-9,20).$$

Thus the matrix of $T \circ S$ is given by

$$M_{T \circ S} = \begin{pmatrix} 8 & -9 \\ 18 & 20 \end{pmatrix}.$$

Look at how this was constructed from the entries of M_T and M_S.

$$\begin{aligned} \underbrace{\begin{pmatrix} 3 & 1 \\ -2 & 4 \end{pmatrix}}_{M_T} \underbrace{\begin{pmatrix} 1 & -4 \\ 5 & 3 \end{pmatrix}}_{M_S} &\to \left(\begin{array}{c|c} \text{1st row } M_T & \text{1st row } M_T \\ \text{1st col } M_S & \text{2nd col } M_S \\ \hline \text{2nd row } M_T & \text{2nd row } M_T \\ \text{1st col } M_S & \text{2nd col } M_S \end{array} \right) \\ &\to \begin{pmatrix} 3(1) + 1(5) & 3(-4) + 1(3) \\ -2(1) + 4(5) & -2(-4) + 4(3) \end{pmatrix} \\ &= \begin{pmatrix} 8 & -9 \\ 18 & 20 \end{pmatrix}. \end{aligned}$$

The matrix $\begin{pmatrix} 8 & -9 \\ 18 & 20 \end{pmatrix}$ obtained from M_T and M_S in this way is called the *matrix product* $M_T M_S$. It is the matrix of the transformation $T \circ S$.

Note: To find $M_{S \circ T}$ calculate the product

$$M_S M_T = \begin{pmatrix} 1 & -4 \\ 5 & 3 \end{pmatrix} \begin{pmatrix} 3 & 1 \\ -2 & 4 \end{pmatrix} = \begin{pmatrix} 1(3) + (-4)(-2) & 1(1) + (-4)(4) \\ 5(3) + 3(-2) & 5(1) + 3(4) \end{pmatrix} = \begin{pmatrix} 11 & -15 \\ 9 & 17 \end{pmatrix}$$

We have defined multiplication of 2 × 2 matrices. Note that this is not *commutative* : for 2 × 2 matrices AB and BA, the products AB and BA are typically not the same.

REMARKS

1. This section is essentially about matrix arithmetic and its relationship to linear transformations. In particular, the last part of it is about how matrix multiplication is defined and what it means. Even if you already know how to multiply matrices, do not assume that you have nothing to learn from this section of the lecture notes. It is more about why matrix multiplication is defined as it is that just being about how to multiply matrices.

2. Knowing how to multiply matrices is important and you should practice it, using a textbook if necessary (any book with a name like "Elementary Linear Algebra" will do). Every year I am surprised by the number of people who are not able to carry out this basic procedure in the final exam. Maybe this year will be an exception.

3. After studying this section you should be able to give a written description of the connection between matrix multiplication and composition of linear transformations, using correct terminology and notation. In particular, make sure that you are using the "$\circ$" notation for composition correctly. This means knowing the difference between $T \circ S$ and $S \circ T$, knowing which is which and knowing which one corresponds to the matrix product $M_T M_S$ and

which corresponds to $M_S M_T$. Confusion and lack of precision over this is another perennial cause of trouble for candidates in examinations in this course. Errors arising in this way are serious, but they can be completely avoided by careful attention to detail when studying the lecture notes and when writing about examples or problems.

1.4 The Inverse of a 2×2 Matrix

QUESTION: Suppose that $T : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is the linear transformation with matrix $\begin{pmatrix} -4 & 9 \\ 2 & -5 \end{pmatrix}$. Is there a linear transformation $S : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ that *reverses the work of* T - i.e. that sends every element of $\mathbb{R}^2$ back to "where it came from" under T?

This would mean that $T \circ S$ and $S \circ T$ would map every element of $\mathbb{R}^2$ to itself. The function from $\mathbb{R}^2$ to $\mathbb{R}^2$ that sends every element to itself is called the identity mapping and denoted by id. It is a linear transformation with

$$\mathrm{id}(1,0) = (1,0), \quad \mathrm{id}(0,1) = (0,1).$$

Thus the matrix of $\mathtt{id}$ is

$$I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Definition 1.4.1 *The matrix* I_2 *is called the* 2×2 *identity matrix. It has the property that* $I_2A = AI_2 = A$ *for all* 2×2 *matrices* A.

In matrix terms our question becomes : does there exist a 2×2 matrix B for which $AB = BA = I_2$, where $A = \begin{pmatrix} -4 & 9 \\ 2 & -5 \end{pmatrix}$?

To answer this question :
Form the matrix

$$\mathrm{adj}(A) = \begin{pmatrix} -5 & -9 \\ -2 & -4 \end{pmatrix}$$

(obtained from A by swapping the entries -5 and -4 on the main diagonal and changing the signs on the other two entries). Observe that

$$A \times \mathrm{adj}(A) = \begin{pmatrix} -4 & 9 \\ 2 & -5 \end{pmatrix}\begin{pmatrix} -5 & -9 \\ -2 & -4 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = 2I_2$$

$$\mathrm{adj}(A) \times A = \begin{pmatrix} -5 & -9 \\ -2 & -4 \end{pmatrix}\begin{pmatrix} -4 & 9 \\ 2 & -5 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = 2I_2$$

NOTE: The number 2 is the *determinant* of the matrix A.

Now form the matrix

$$B = \frac{1}{2}\mathrm{adj}(A) = \frac{1}{2}\begin{pmatrix} -5 & -9 \\ -2 & -4 \end{pmatrix} = \begin{pmatrix} -5/2 & -9/2 \\ -1 & -2 \end{pmatrix}.$$

Then we have $AB = I_2$ and $BA = I_2$.

Definition 1.4.2 *The matrix* B *is called the* inverse *of* A. *In general a pair of* 2×2 *matrices* A *and* B *are called inverses of each other if* $AB = I_2$ *and* $BA = I_2$. *The inverse of* A *is often written* A^{-1}.

ANSWER TO OUR QUESTION: The linear transformation $S : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ whose matrix is

$$\begin{pmatrix} -5/2 & -9/2 \\ -1 & -2 \end{pmatrix}$$

"reverses the work" of T. As a function it is the inverse of T.

QUESTION: Which 2 × 2 matrices have inverses?

Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Define the *adjoint* or *adjugate* of A by

$$\text{adj}(A) = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Note that

$$A \times \text{adj}(A) = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \begin{pmatrix} ad-bc & 0 \\ 0 & ad-bc \end{pmatrix} = (ad-bc)I_2$$

$$\text{adj}(A) \times A = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} ad-bc & 0 \\ 0 & ad-bc \end{pmatrix} = (ad-bc)I_2$$

Definition 1.4.3 *The number* $ad - bc$ *is called the* determinant *of the matrix* $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$*; it is denoted by* $|A|$ *or* $\det(A)$.

If $\det(A) \neq 0$, we can adjust the above equations by multiplying both sides by the real number $\frac{1}{\det(A)}$. We obtain

$$\frac{1}{\det(A)}(A \times \text{adj}(A)) = \frac{1}{\det(A)}(\text{adj}(A) \times A) = I_2.$$

Hence the matrix

$$A^{-1} = \frac{1}{\det(A)}\text{adj}(A)$$

is an inverse for A.

Example 1.4.4 *Find the inverse of the* 2×2 *matrix* $A = \begin{pmatrix} 3 & 1 \\ -2 & 4 \end{pmatrix}$.

SOLUTION: $\det(A) = 3(4) - (1)(-2) = 14$ ($\neq 0$)

$$\text{adj}(A) = \begin{pmatrix} 4 & -1 \\ 2 & 3 \end{pmatrix} \Longrightarrow A^{-1} = \frac{1}{14}\begin{pmatrix} 4 & -1 \\ 2 & 3 \end{pmatrix}.$$

Check that $AA^{-1} = A^{-1}A = I_2$.

What about the case where $\det(A) = 0$?

In this case A does not have an inverse. To see this let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and suppose that $ad - bc = 0$. This means either that $a = c = 0$ or $b = d = 0$ or

$$\begin{pmatrix} b \\ d \end{pmatrix} = k\begin{pmatrix} a \\ c \end{pmatrix}$$

for some real number k. In all of these cases the points (a, c) and (b, d) lie on the same line L through the origin (think about this).

Now let T be the linear transformation whose matrix is A. Since T maps both $(1, 0)$ and $(0, 1)$ on to the line L, it maps every point of $\mathbb{R}^2$ to a point of L. Since the image of T is not all of $\mathbb{R}^2$, there cannot exist a linear transformation $S : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ for which $T \circ S = \text{id}$.

Example 1.4.5 *(Summer 2005) let* $T : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ *be the linear transformation with matrix*

$$A = \begin{pmatrix} -4 & 1 \\ 4 & -4 \end{pmatrix}.$$

Find the line whose image under T *is* $L : x + 2y = 4$.

NOTE: If $T : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is an invertible linear transformation, then the image under T of any line is another line.

In Example 1.4.5 we need to find the image of L under the inverse of T.

Step 1 Write down the matrix of T^{-1}:

$$A^{-1} = \frac{1}{12}\begin{pmatrix} -4 & -1 \\ -4 & -4 \end{pmatrix}.$$

Step 2 Write down the "parametric representation" of points of L. The equation of L is $y = -\frac{1}{2}x + 2$. This means that L consists of all points of the form $(t, -\frac{1}{2}t + 2)$, where $t \in \mathbb{R}$.

Step 3 Calculate the image of a point of this form under T^{-1} :

$$\begin{aligned} A^{-1}\begin{pmatrix} t \\ -\frac{1}{2}t + 2 \end{pmatrix} &= \frac{1}{12}\begin{pmatrix} -4 & -1 \\ -4 & -4 \end{pmatrix}\begin{pmatrix} t \\ -\frac{1}{2}t + 2 \end{pmatrix} \\ &= \frac{1}{12}\begin{pmatrix} -\frac{7}{2}t - 2 \\ \\ -2t - 8 \end{pmatrix} \end{aligned}$$

So $T^{-1}(L)$ consists of those points whose coordinates have the form

$$\left(-\frac{7}{24}t - \frac{1}{6}, -\frac{1}{6}t - \frac{2}{3}\right) \text{ for some } t \in \mathbb{R}.$$

Step 4 Convert this back into the standard form of the equation of a line.

$$x = -\frac{7}{24}t - \frac{1}{6} \Longrightarrow t = -\frac{24}{7}x - \frac{4}{7}.$$

Now

$$\begin{aligned} y &= -\frac{1}{6}t - \frac{2}{3} \\ &= -\frac{1}{6}\left(-\frac{24}{7}x - \frac{4}{7}\right) - \frac{2}{3} \\ &= \frac{4}{7}x - \frac{4}{7} \end{aligned}$$

So the line that is mapped to L by T is $L_1 : y = \frac{4}{7}x - \frac{4}{7}$ or

$$L_1 : 4x - 7y = 4.$$

EXERCISE: Check that $T(L_1) = L$ by checking that points on L_1 have images under T that satisfy the equation of L.

NOTE: The following question appeared on the 2009 Summer paper (for Science students) in this course.

Let $\mathsf{T} : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ be the linear transformations defined for $(x, y) \in \mathbb{R}^2$ by

$$\mathsf{T}(x, y) = (2x + 3y, x - y).$$

Let L be the line with equation $x + y = 4$. Find the equation of the image of L under T.

Quite a number of students answered this question by exactly imitating the technique of Example 1.4.5 above, even though *this example is not the same*. In Example 1.4.5, the question asked for the line *whose image under* T is a given line L - in other words, for the image of L under the inverse of T. In this question from 2009, the problem was to calculate the image (not the inverse image) of a given line L under a given linear transformation T. Which should be used, the matrix of T or its inverse? This experience highlights the importance of understanding the meaning of the techniques that are being used, in order to realize how they need to be adapted for different examples.

1.5 Eigenvalues and Eigenvectors

When a 2×1 vector is multiplied by a 2×2 matrix, the result is another 2×1 vector. So we can think of matrices as objects that move vectors around in the plane. Generally a matrix could move a vector anywhere - we would not normally expect for example that multiplying a vector by a matrix would have the same effect on the vector as multiplying it by some number (scalar). When this does happen, the vector has a special property with respect to that particular matrix.

Example 1.5.1 below gives an indication of what this can mean in geometric terms. Recall from Example 1.4.5 that when a linear transformation is applied to all the points of a line, the result is a new set of points also forming a line. So linear transformations always send lines to lines (this is not the *definition* of a linear transformation, but it is a property that linear transformations possess). It is also true that if a line includes the origin, its image under a linear transformation will always be a line through the origin. Given a linear transformation T then, we can ask whether there are any lines through the origin that are sent to themselves by T. Note that this does not necessarily mean that every point on the line would need to be sent to itself by T, but that every point on the line would be sent by T to a point also belonging to the same line.

Example 1.5.1 *Let* T *be the linear transformation with matrix* $A = \begin{pmatrix} -4 & 2 \\ 3 & 1 \end{pmatrix}$. *Find all lines through the origin in* $\mathbb{R}^2$ *that are fixed (i.e. mapped to themselves) by* T.

Solution: If L is a line through the origin that is fixed by T, let $(x, y) \neq (0, 0)$ be a point of L. Then we must have

$$T(x, y) = \lambda(x, y)$$

for some scalar λ.

Note: "λ" is the Greek letter *lambda*.

This means

$$\begin{pmatrix} -4 & 2 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix},$$

where x and y are not both zero.

How can we solve this for x, y (and λ)?

$$\begin{aligned} A \begin{pmatrix} x \\ y \end{pmatrix} &= \lambda \begin{pmatrix} x \\ y \end{pmatrix} \quad (2 \times 1 \text{ matrices}) \\ \Longrightarrow \lambda \begin{pmatrix} x \\ y \end{pmatrix} - A \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ \Longrightarrow \lambda I_2 \begin{pmatrix} x \\ y \end{pmatrix} - A \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{aligned}$$

Note $I_2 \begin{pmatrix} x \\ y \end{pmatrix}$ is equal to $\begin{pmatrix} x \\ y \end{pmatrix}$ - but I_2 is a 2×2 matrix. Now

$$(\lambda I_2 - A) \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Suppose that the matrix $\lambda I_2 - A$ has an inverse. Then we would have

$$(\lambda I_2 - A)^{-1}(\lambda I_2 - A) \begin{pmatrix} x \\ y \end{pmatrix} = (\lambda I_2 - A)^{-1} \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

that is

$$I_2 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

This would mean $\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ - but we are looking for solutions with $(x, y) \neq (0, 0)$. The above argument says these cannot occur if $\lambda I_2 - A$ is invertible, so we must look at the case where $\lambda I_2 - A$ is not invertible, i.e. $\det(\lambda I_2 - A) = 0$. Now

$$\begin{aligned} \lambda I_2 - A &= \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} -4 & 2 \\ 3 & 1 \end{pmatrix} \\ &= \begin{pmatrix} \lambda + 4 & -2 \\ -3 & \lambda - 1 \end{pmatrix}. \end{aligned}$$

$$\begin{aligned} \det(\lambda I_2 - A) &= (\lambda + 4)(\lambda - 1) - (-2)(-3) \\ &= \lambda^2 + 3\lambda - 4 - 6 \\ &= \lambda^2 + 3\lambda - 10. \end{aligned}$$

$\det(\lambda I_2 - A) = 0$ means $\lambda^2 + 3\lambda - 10 = 0$, i.e. $(\lambda + 5)(\lambda - 2) = 0$ and $\lambda = -5$ or $\lambda = 2$.

1. Suppose $\lambda = -5$. Can we find a solution to

$$\begin{pmatrix} -4 & 2 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = -5 \begin{pmatrix} x \\ y \end{pmatrix}$$

with $(x, y) \neq (0, 0)$? This would mean

$$\begin{aligned} \begin{pmatrix} -4 & 2 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} &= -5 \begin{pmatrix} x \\ y \end{pmatrix} \\ \begin{pmatrix} -4x + 2y \\ 3x + y \end{pmatrix} &= \begin{pmatrix} -5x \\ -5y \end{pmatrix} \end{aligned}$$

$$\begin{array}{rcrcr} -4x & + & 2y & = & -5x \\ 3x & + & y & = & -5y \end{array} \implies \begin{array}{rcrcr} x & + & 2y & = & 0 \\ 3x & + & 6y & = & 0 \end{array}$$

Both of these equations say $x + 2y = 0$ or $y = -\frac{1}{2}x$. We can satisfy this by taking $y = 1, x = -2; y = 3, x = -6;\ y = -1, x = 2$ etc. - we obtain the points $(-2, 1),\ (-6, 3),\ (2, -1)$ and so on.

Conclusion: Every point (x, y) of the line $y = -\frac{1}{2}x$ is mapped by T to $-5(x, y) = (-5x, -5y)$. The line $L_1 : y = -\frac{1}{2}x$ is fixed by T.

2. Suppose $\lambda = 2$. We can solve

$$A \begin{pmatrix} x \\ y \end{pmatrix} = 2 \begin{pmatrix} x \\ y \end{pmatrix}.$$

This means

$$\begin{aligned} \begin{pmatrix} -4 & 2 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} &= 2 \begin{pmatrix} x \\ y \end{pmatrix} \\ \begin{pmatrix} -4x + 2y \\ 3x + y \end{pmatrix} &= \begin{pmatrix} 2x \\ 2y \end{pmatrix} \end{aligned}$$

$$\begin{array}{rcrcr} -4x & + & 2y & = & 2x \\ 3x & + & y & = & 2y \end{array} \implies \begin{array}{rcrcr} -6x & + & 2y & = & 0 \\ 3x & - & y & = & 0 \end{array}$$

Both equations say $y = 3x$. Every point (x, y) of the line $L_2 : y = 3x$ is mapped bu T to $(2x, 2y)$. The line L_2 is fixed by T. Furthermore L_1 and L_2 are the *only* lines through the origin fixed by T.

Definition 1.5.2 1. *Let* A *be a* 2×2 *matrix. An eigenvector of* A *is a column vector* $v \neq \binom{0}{0}$ *for which* $Av = \lambda v$ *for some number* λ.

2. *In the above situation the number* λ *is the eigenvalue of* A *to which the eigenvector* v *corresponds. In Example 1.5.1 we found that* $\begin{pmatrix} -2 \\ 1 \end{pmatrix}$ *is an eigenvector of* $\begin{pmatrix} -4 & 2 \\ 3 & 1 \end{pmatrix}$ *corresponding to the eigenvalue* -5.

3. *The polynomial* $\det(\lambda I_2 - A)$ *is called the characteristic polynomial of* A. *It is quadratic in* λ *(when* A *is* 2×2*).*

4. *The equation* $\det(\lambda I_2 - A) = 0$ *is the characteristic equation of* A. *The solutions of the characteristic equation are the eigenvalues of* A.

Example 1.5.3 *Find the eigenvalues and eigenvectors of the matrix* $A = \begin{pmatrix} -4 & 1 \\ 4 & -4 \end{pmatrix}$.

SOLUTION:

$$\lambda I_2 - A = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} - \begin{pmatrix} -4 & 1 \\ 4 & -4 \end{pmatrix} = \begin{pmatrix} \lambda+4 & -1 \\ -4 & \lambda+4 \end{pmatrix}.$$

$$\det(\lambda I_2 - A) = (\lambda+4)^2 - (-1)(-4) = \lambda^2 + 8\lambda + 12.$$

$$\lambda^2 + 8\lambda + 12 = 0 \Longrightarrow (\lambda+6)(\lambda+2) = 0.$$

Eigenvalues of A : -6, -2.

EIGENVECTORS:

1. If $\lambda = -2$

$$\begin{aligned} \begin{pmatrix} -4 & 1 \\ 4 & -4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} -2x \\ -2y \end{pmatrix} \\ \Longrightarrow \begin{pmatrix} -4x + y \\ 4x - 4y \end{pmatrix} &= \begin{pmatrix} -2x \\ -2y \end{pmatrix} \end{aligned}$$

$$\Longrightarrow \begin{array}{rcrcr} -4x & + & y & = & -2x \\ 4x & - & 4y & = & -2y \end{array} \Longrightarrow \begin{array}{rcrcr} 2x & - & y & = & 0 \\ 4x & - & 2y & = & 0 \end{array}$$

Both equations say $y = 2x$, so any non-zero vector of the form $\binom{x}{y}$ with $y = 2x$ is an eigenvector of A corresponding to the eigenvalue -2; for example $\binom{1}{2}$ or any non-zero scalar multiple thereof.

2. If $\lambda = -6$:
$\binom{1}{-2}$ is an eigenvector for this eigenvalue (Exercise).

Diagonalization

In the above example we found that $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ -2 \end{pmatrix}$ are eigenvectors of the matrix $A = \begin{pmatrix} -4 & 1 \\ 4 & -4 \end{pmatrix}$, with corresponding eigenvalues -2 and -6 respectively. This means

$$\begin{pmatrix} -4 & 1 \\ 4 & -4 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \end{pmatrix} = -2\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \begin{pmatrix} -4 & 1 \\ 4 & -4 \end{pmatrix}\begin{pmatrix} 1 \\ -2 \end{pmatrix} = -6\begin{pmatrix} 1 \\ -2 \end{pmatrix}.$$

Thus

$$\begin{pmatrix} -4 & 1 \\ 4 & -4 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 2 & -2 \end{pmatrix} = \left(-2\begin{pmatrix} 1 \\ -2 \end{pmatrix} \;\; -6\begin{pmatrix} 1 \\ -2 \end{pmatrix}\right) = \begin{pmatrix} -2 & -6 \\ -4 & 12 \end{pmatrix}$$

We have

$$\begin{pmatrix} -4 & 1 \\ 4 & -4 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 2 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 2 & -2 \end{pmatrix}\begin{pmatrix} -2 & 0 \\ 0 & -6 \end{pmatrix}$$

(Think about this). Thus $AE = ED$ where $E = \begin{pmatrix} 1 & 1 \\ 2 & -2 \end{pmatrix}$ has the eigenvectors of A as columns and $D = \begin{pmatrix} -2 & 0 \\ 0 & -6 \end{pmatrix}$ is the *diagonal* matrix having the eigenvalues of A on the main diagonal, in the order in which their corresponding eigenvectors appear as columns of E. (The main diagonal is the diagonal strip from top left to bottom right).

Note on diagonal matrices:

- A 2×2 matrix is *diagonal* if it has the form $\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$ for some real numbers a and b.
- Diagonal matrices behave particularly well with respect to matrix multiplication : if $A = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$ and $B = \begin{pmatrix} c & 0 \\ 0 & d \end{pmatrix}$ are diagonal matrices, then $AB = BA$ and this product is obtained by simply multiplying the entries on the main diagonal. Thus

$$AB = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}\begin{pmatrix} c & 0 \\ 0 & d \end{pmatrix} = \begin{pmatrix} ac & 0 \\ 0 & bd \end{pmatrix} = BA.$$

- In particular for a positive integer n we have $A^n = \begin{pmatrix} a^n & 0 \\ 0 & b^n \end{pmatrix}$.

Back to our Example : We have $AE = ED$. Note that $\det(E) \neq 0$ so E is invertible. Thus

$$\begin{aligned} AE &= ED \\ \Longrightarrow AEE^{-1} &= EDE^{-1} \\ \Longrightarrow A &= EDE^{-1}. \end{aligned}$$

It is convenient to write A in this form if for some reason we need to calculate powers of A. Note for example that

$$\begin{aligned} A^3 &= (EDE^{-1})(EDE^{-1})(EDE^{-1}) \\ &= EDI_2DI_2DE^{-1} \\ &= ED^3E^{-1} \\ &= E\begin{pmatrix} (-2)^3 & 0 \\ 0 & (-6)^3 \end{pmatrix}E^{-1}. \end{aligned}$$

In general $A^n = E\begin{pmatrix} (-2)^n & 0 \\ 0 & (-6)^n \end{pmatrix}E^{-1}$, for any positive integer n. (In fact this is true for negative integers too if we interpret A^{-n} to mean the nth power of the inverse A^{-1} of A).

Example 1.5.4 *(Summer 2005) Solve the recurrence relation*

$$\begin{aligned} x_{n+1} &= -4x_n + 1y_n \\ y_{n+1} &= 4x_n - 4y_n \end{aligned}$$

given that $x_0 = 1,\ y_0 = 1$.

NOTE: this means we have sequences $x_0, x_1, \ldots$ and $y_0, y_1, \ldots$ defined by the above relations. If for some n we know x_n and y_n, the relations tell us how to calculate x_{n+1} and y_{n+1}.

For example

$$\begin{aligned} x_1 &= -4x_0 + y_0 = -4(1) + 1 = -3 \\ y_1 &= 4x_0 - 4y_0 = 4(1) - 4(1) = 0 \\ \\ x_2 &= -4x_1 + y_1 = -4(-3) + 0 = 12 \\ y_2 &= 4x_1 - 4y_1 = 4(-3) - 4(0) = -12. \end{aligned}$$

SOLUTION OF THE PROBLEM:
The relations can be written in matrix form as

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} -4x_n + 1y_n \\ 4x_n - 4y_n \end{pmatrix} = \begin{pmatrix} -4 & 1 \\ 4 & -4 \end{pmatrix}\begin{pmatrix} x_n \\ y_n \end{pmatrix} = A\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix},$$

where A is the matrix $\begin{pmatrix} -4 & 1 \\ 4 & -4 \end{pmatrix}$. Thus

$$\begin{aligned} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} &= A\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = A\begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} &= A\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = A\left(A\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) = A^2\begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ \begin{pmatrix} x_3 \\ y_3 \end{pmatrix} &= A\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = A\left(A^2\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) = A^3\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \text{ etc.} \end{aligned}$$

In general $\begin{pmatrix} x_n \\ y_n \end{pmatrix} = A^n\begin{pmatrix} 1 \\ 1 \end{pmatrix}$.
To obtain general formulae for x_n and y_n we need a general formula for A^n. We have

$$A^n = (EDE^{-1})^n = ED^nE^{-1}$$

where $E = \begin{pmatrix} 1 & 1 \\ 2 & -2 \end{pmatrix}$ and $D = \begin{pmatrix} -2 & 0 \\ 0 & -6 \end{pmatrix}$.
Note

$$E^{-1} = -\frac{1}{4}\begin{pmatrix} -2 & -1 \\ -2 & 1 \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 2 & 1 \\ 2 & -1 \end{pmatrix}.$$

Thus

$$
\begin{aligned}
A^n &= \begin{pmatrix} 1 & 1 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} (-2)^n & 0 \\ 0 & (-6)^n \end{pmatrix} \frac{1}{4} \begin{pmatrix} 2 & 1 \\ 2 & -1 \end{pmatrix} \\
&= \begin{pmatrix} (-2)^n & (-6)^n \\ 2(-2)^n & -2(-6)^n \end{pmatrix} \frac{1}{4} \begin{pmatrix} 2 & 1 \\ 2 & -1 \end{pmatrix} \\
&= \frac{1}{4} \begin{pmatrix} (-2)^n(2) + (-6)^n(2) & (-2)^n - (-6)^n \\ 4(-2)^n - 4(-6)^n & 2(-2)^n + 2(-6)^n \end{pmatrix}
\end{aligned}
$$

and

$$
\begin{aligned}
\begin{pmatrix} x_n \\ y_n \end{pmatrix} = A^n \begin{pmatrix} 1 \\ 1 \end{pmatrix} &= \frac{1}{4} \begin{pmatrix} (-2)^n(2) + (-6)^n(2) & (-2)^n - (-6)^n \\ 4(-2)^n - 4(-6)^n & 2(-2)^n + 2(-6)^n \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
&= \frac{1}{4} \begin{pmatrix} 3(-2)^n + (-6)^n \\ 6(-2)^n - 2(-6)^n \end{pmatrix}
\end{aligned}
$$

We conclude that

$$
\begin{aligned}
x_n &= \frac{3}{4}(-2)^n + \frac{1}{4}(-6)^n \\
y_n &= \frac{3}{2}(-2)^n - \frac{1}{2}(-6)^n
\end{aligned}
$$

for $n \geqslant 0$.
(This is easily verified for small values of n using the recurrence relations). See Problem Sheet 2 for more problems of this type.

This concludes Section 1.5. We recall the main result on diagonalization of 2×2 matrices.

Theorem 1.5.5 *Let A be a 2×2 matrix with eigenvalues λ_1, λ_2. Let v_1 and v_2 be eigenvectors of A corresponding to λ_1 and λ_2 respectively. Then if E denotes the 2×2 matrix having v_1 and v_2 as its columns and E is invertible, we have*

$$E^{-1}AE = D,$$

where D is the diagonal matrix having λ_1 and λ_2 on its main diagonal. It follows that

$$A^n = E \begin{pmatrix} (\lambda_1)^n & 0 \\ 0 & (\lambda_2)^n \end{pmatrix} E^{-1}.$$

1.6 More Matrix Algebra

At this stage we have discussed linear transformations of $\mathbb{R}^2$ and their associated matrices in some detail. The theme of this section is that most of the theory that we have developed applies also to $\mathbb{R}^n$ for positive integers n.

Definition 1.6.1 *Let n be a positive integer. Then $\mathbb{R}^n$ is the set of ordered n-tuples of real numbers, i.e. the set of objects of the form*

$$(a_1, \dots, a_n), \quad a_i \in \mathbb{R} \text{ for } i = 1, \dots, n.$$

Addition and scalar multiplication in $\mathbb{R}^n$ are defined in the obvious way. For example in $\mathbb{R}^4$ we have

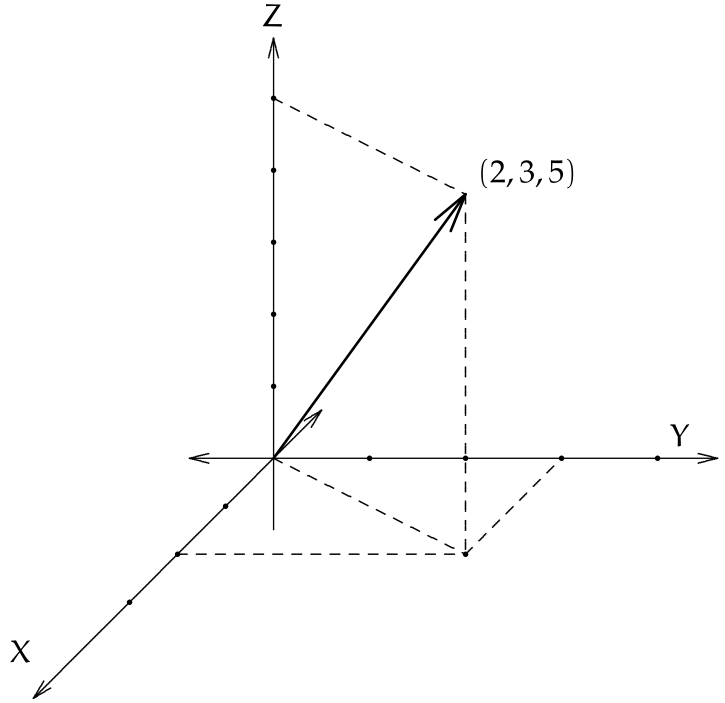
$$(3, 1, -2, 4) + (1, 2, 1, -6) = (3 + 1, 1 + 2, -2 + 1, 4 + (-6) = (4, 3, -1, -2)$$

and

$$3(3, -1, 4, 2) = (9, -3, 12, 6).$$

$\mathbb{R}^n$ is called n-*dimensional Euclidean space*.

$\mathbb{R}^3$ can be considered to be described by three coordinate axes labelled X, Y and Z. Elements of $\mathbb{R}^3$ can be considered as points - the ordered triple $(2, 3, 5)$ for example represents the point with X-coordinate 2, Y-coordinate 3 and Z-coordinate 5.



As in $\mathbb{R}^2$, we consider the triple $(2, 3, 5)$ to represent both the point with these coordinates and the vector directed from the origin to this point.

If $n \geqslant 4$, the pictures are harder to draw and we rely more on algebraic techniques than geometric representation.

Definition 1.6.2 *Let n and m be positive integers. A* linear transformation *from $\mathbb{R}^n$ to $\mathbb{R}^m$ is a function $T : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ satisfying*

1. $T(u+v) = T(u) + T(v)$ for all $u, v \in \mathbb{R}^n$, *and*

2. $T(ku) = kT(u)$ for all $u \in \mathbb{R}^n$ *and* $k \in \mathbb{R}$.

Suppose that T is such a linear transformation. Then T is described by a $m \times n$ matrix M_T defined as follows :

1st column of M_T :	coordinates of $T(1, 0, \dots, 0)$
2nd column of M_T :	coordinates of $T(0, 1, 0 \dots, 0)$
$\vdots$	$\vdots$
nth column of M_T :	coordinates of $T(0, \dots, 0, 1)$

Example 1.6.3 *Suppose* $T : \mathbb{R}^4 \longrightarrow \mathbb{R}^2$ *is defined by*

$$T(x, y, z, w) = (x + 2y + z - w, -x - 3y).$$

Then T is a linear transformation (check) and

$$T(1, 0, 0, 0) = (1, -1),\ T(0, 1, 0, 0) = (2, -3),\ T(0, 0, 1, 0) = (1, 0),\ T(0, 0, 0, 1) = (-1, 0).$$

The matrix of T is

$$M_T = \begin{pmatrix} 1 & 2 & 1 & -1 \\ -1 & -3 & 0 & 0 \end{pmatrix}$$

Then $T(3, 1, 4, -1)$ (for example) can be found by calculating the matrix product

$$M_T \begin{pmatrix} 3 \\ 1 \\ 4 \\ -1 \end{pmatrix}.$$

This is given by

$$\begin{pmatrix} 1 & 2 & 1 & -1 \\ -1 & -3 & 0 & 0 \end{pmatrix} \begin{pmatrix} 3 \\ 1 \\ 4 \\ -1 \end{pmatrix} = \begin{pmatrix} 1(3) + 2(1) + 1(4) + (-1)(-1) \\ -1(3) + (-3)(1) = 0(4) + 0(-1) \end{pmatrix} = \begin{pmatrix} 10 \\ -6 \end{pmatrix}.$$

Thus $T(3, 1, 4, -1) = (10, -6)$. (This can be easily checked using the definition of T).

Matrix Multiplication and Composition

Suppose that $T : \mathbb{R}^n \longrightarrow \mathbb{R}^p$ and $S : \mathbb{R}^q \longrightarrow \mathbb{R}^m$ are linear transformations. Then the transformation $S \circ T$ (S *after* T) can be defined only if makes sense to apply S to the image under T of an element of $\mathbb{R}^n$. The transformation T takes elements of $\mathbb{R}^n$ into $\mathbb{R}^p$. The transformation S can be applied to elements of $\mathbb{R}^q$. So it is possible to apply S *after* T only if $\mathbb{R}^p = \mathbb{R}^q$, i.e. only if $q = p$. In case we have a transformation $S \circ T$ mapping $\mathbb{R}^n$ into $\mathbb{R}^m$.

The matrices M_S and M_T have sizes $m \times p$ and $q \times n$ respectively. The product $M_S M_T$ exists only if $q = p$ and in this case $M_S M_T$ describes the transformation $S \circ T : \mathbb{R}^n \longrightarrow \mathbb{R}^m$; it has size $m \times n$.

Definition 1.6.4 *If* A *is a* $m \times p$ *matrix and* B *is a* $q \times n$ *matrix, then the product* AB *is defined if and only if* $p = q$, *i.e. if and only if*

No. of columns in A = *No. of rows in* B.

(alternatively : No. of entries in a row of A = *No. of entries in a column of* B*)*

In this case the size of AB *is* $m \times n$.

In general the following "cancellation law" holds for the size of matrix products:

$$\text{“}(m \times \cancel{p}) \times (\cancel{p} \times n) = m \times n\text{”}.$$

If A *is a* $m \times p$ *matrix and* B *is a* $p \times n$ *matrix, then the product* AB *is a* $m \times n$ *matrix in which the entry in the* i*th row and* j*th column is given by combining the entries of the* i*th row of* A *with those of the* j*th column of* B *as in the* 2×2 *case.*

Example 1.6.5 *Let* $A = \begin{pmatrix} 2 & -1 & 3 \\ 1 & 0 & -1 \end{pmatrix}$ *and let* $B = \begin{pmatrix} 3 & 1 \\ 1 & -1 \\ 0 & 2 \end{pmatrix}$

Find AB *and* BA.

<u>Solution</u> :

1. $A: \ 2 \times 3, \ B: \ 3 \times 2 \Longrightarrow AB$ will be a 2×2 matrix.

$$\begin{aligned} \begin{pmatrix} 2 & -1 & 3 \\ 1 & 0 & -1 \end{pmatrix}\begin{pmatrix} 3 & 1 \\ 1 & -1 \\ 0 & 2 \end{pmatrix} &= \begin{pmatrix} 2(3)+(-1)(1)+3(0) & 2(1)+(-1)(-1)+3(2) \\ 1(3)+0(1)+(-1)(0) & 1(1)+0(-1)+(-1)(2) \end{pmatrix} \\ &= \begin{pmatrix} 5 & 9 \\ 3 & -1 \end{pmatrix} \end{aligned}$$

2. $B: \ 3 \times 2, \ A: \ 2 \times 3 \Longrightarrow BA$ will be a 3×3.

$$BA = \begin{pmatrix} 7 & -3 & 8 \\ 1 & -1 & 4 \\ 2 & 0 & -2 \end{pmatrix}$$

(Exercise)

Square Matrices : Inverses and Determinants

For a positive integer n, the *identity linear transformation* from $\mathbb{R}^n$ to $\mathbb{R}^n$ is the transformation that maps every element of $\mathbb{R}^n$ to itself. Its matrix is I_n; its entries are 1 along the main diagonal and zero everywhere else. For example

$$I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad I_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

The $n \times n$ identity matrix I_n has the following properties :

1. $AI_n = A$ if A is a matrix with n columns.
2. $I_nA = A$ if A is a matrix with n rows.

In particular if A is an $n \times n$ matrix then $AI_n = I_nA = A$.

For a positive integer n, we denote the set of $n \times n$ matrices with entries in $\mathbb{R}$ by $M_n(\mathbb{R})$. If we deal only with matrices in $M_n(\mathbb{R})$, then we can add any pair of matrices or multiply any matrix by any other, and we stay within $M_n(\mathbb{R})$. ($M_n(\mathbb{R})$ is an example of the type of algebraic structure known as a *ring*).

In the remainder of this section we will consider the problem of how to calculate determinants and inverses of square matrices in general. This problem is particularly easy in the 2×2 case because of the small size of the matrices involved. Although our examples in this section will be 3×3, the techniques that we will develop are generally applicable.

Definition 1.6.6 *Let A and B be $n \times n$ matrices. Then A and B are called inverses of each other if*

$$AB = BA = I_n.$$

We note that a matrix can have only one inverse, for suppose that B and C are both inverses of some matrix A. Then we have

$$B(AC) = BI_n = B \text{ and also } B(AC) = (BA)C = I_nC = C.$$

Thus $B = C$ and A can have only one inverse.

For every n by n matrix A the adjoint $\text{adj}(S)$ (a $n \times n$ matrix) and the determinant $\det(A)$ (a number) are defined and are related to each other by

$$\text{adj}(A) \times A = A \times \text{adj}(A) = \det(A)I_n.$$

Thus $A^{-1} = \dfrac{1}{\det(A)}\text{adj}(A)$ provided $\det(A) \neq 0$; as in the 2×2 case A does not have an inverse if $\det(A) = 0$.

We now discuss a technique for calculating the determinant and adjoint of a $n \times n$ matrix, using a 3×3 example as a guide.

Example 1.6.7 *Let* $A = \begin{pmatrix} 1 & 3 & 0 \\ 2 & -2 & 1 \\ -4 & 1 & -1 \end{pmatrix}$. *Find* A^{-1}

We begin with some definitions :

- $(A)_{ij}$ (or simply A_{ij}) denotes the entry in the ith row and jth column of A.
- For each entry $(A)_{ij}$ of A, we define the *minor* M_{ij} of $(A)_{ij}$ to be the determinant of the 2×2 matrix which remains when the ith row and jth column (i.e. the row and column containing $(A)_{ij}$) are deleted from A.
  For example
  $$M_{11} = \det\begin{pmatrix} -2 & 1 \\ 1 & -1 \end{pmatrix} = -2(-1) - (1)(1) = 1$$
  $$M_{12} = \det\begin{pmatrix} 2 & 1 \\ -4 & -1 \end{pmatrix} = 2(-1) - (1)(-4) = 2.$$
- We define the *cofactor* C_{ij} of the entry $(A)_{ij}$ of A as follows:
  $$\begin{array}{lcrll} C_{ij} & = & M_{ij} & \text{if} & i + j \text{ is even} \\ C_{ij} & = & -M_{ij} & \text{if} & i + j \text{ is odd} \end{array}$$

In the 3×3 case we have the following pattern of signs : in the positions marked "$-$", $C_{ij} = -M_{ij}$, and in the positions marked "$+$", $C_{ij} = M_{ij}$:

$$\left(\begin{array}{ccc} + & - & + \\ - & + & - \\ + & - & + \end{array}\right)$$

In our example
$C_{11} = M_{11} = 1$, since the $(1,1)$ position (top left) is marked with "$+$" in the pattern of signs.
$C_{12} = -M_{12} = -2$, since the $(1,2)$ position (1st row, 2nd column) is marked with "$-$" in the pattern of signs.

The *determinant* of a $n \times n$ matrix A can be calculated as follows.

- Choose a row or column of A. (Any row or column will do, but as we will see it is a good idea to choose the one with the largest possible number of entries equal to zero).
- Calculate the cofactor of each entry in the chosen row or column.
- Multiply each entry in the chosen row or column by *its own cofactor*. The sum of these products is the determinant of A.

This method of calculating a determinant is called *cofactor expansion* along a row or column.

Back to our example : $A = \left(\begin{array}{rrr} 1 & 3 & 0 \\ 2 & -2 & 1 \\ -4 & 1 & -1 \end{array}\right)$ To calculate $\det(A)$ by cofactor expansion along the first row :
We already know $C_{11} = 1,\ C_{12} = -2$. We have

$$\det(A) = A_{11}C_{11} + A_{12}C_{12} + A_{13}C_{13} = 1(1) + 3(-2) + 0(C_{13}) = -5.$$

We should get the same result if we apply cofactor expansion along the first column :

$$\begin{array}{rcl} C_{11} & = & 1 \\ C_{21} & = & -M_{21} = -\det\left(\begin{array}{rr} 3 & 0 \\ 1 & -1 \end{array}\right) = 3(-1) - (0)(1) = 3 \\ C_{31} & = & M_{31} = \det\left(\begin{array}{rr} 3 & 0 \\ -2 & 1 \end{array}\right) = 3(1) - (0)(-2) = 3 \end{array}$$

Then

$$\det(A) = A_{11}C_{11} + A_{21}C_{21} + A_{31}C_{31} = 1(1) + 2(3) + (-4)3 = -5,$$

as expected.

This method can be used to calculate the determinant of a square matrix of any size. Of course the cofactors of a 4×4 matrix are 3×3 determinants, etc.

To calculate $\mathrm{adj}(A)$ for our 3×3 matrix A we proceed as follows.

Step 1 *The Matrix of Minors* We calculate the 9 minors :

$M_{11}:\ M_{11} = 1$

$M_{12}:\ M_{12} = 2$

$M_{13}:\ M_{13} = \det\left(\begin{array}{rr} 2 & -2 \\ -4 & 1 \end{array}\right) = 2(1) - (-2)(-4) = -6$

$M_{21}: \ M_{21} = -3$

$M_{22}: \ M_{22} = \det\begin{pmatrix} 1 & 0 \\ -4 & -1 \end{pmatrix} = 1(-1) - (0)(-4) = -1$

$M_{23}: \ M_{23} = \det\begin{pmatrix} 1 & 3 \\ -4 & 1 \end{pmatrix} = 1(1) - (3)(-4) = 13$

$M_{31}: \ M_{31} = 3$

$M_{32}: \ M_{32} = \det\begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} = 1(1) - (0)(2) = 1$

$M_{33}: \ M_{11} = \det\begin{pmatrix} 1 & 3 \\ 2 & -2 \end{pmatrix} = 1(-2) - (3)(2) = -8$

We now write the *matrix of minors* M of A defined by

$$(M)_{ij} = \text{the minor of } (A)_{ij}.$$

$$M = \begin{pmatrix} 1 & 2 & -6 \\ -3 & -1 & 13 \\ 3 & 1 & -8 \end{pmatrix}$$

Step 2 *The Matrix of Cofactors*

We now write down C, the *matrix of cofactors* of A. The matrix C differs from M by the pattern of signs mentioned earlier. Its entry in the ith row and jth column is the cofactor of $(A)_{ij}$. The matrix of cofactors differs from the matrix of minors in the signs of the entries in the positions marked "−" in the pattern of signs.

$$C = \begin{pmatrix} +(1) & -(2) & +(-6) \\ -(-3) & +(-1) & -(13) \\ +(3) & -(1) & +(-8) \end{pmatrix} = \begin{pmatrix} 1 & -2 & -6 \\ 3 & -1 & -13 \\ 3 & -1 & -8 \end{pmatrix}$$

Step 3 *The Adjoint*

The *adjoint* of A is C^{tr}, the *transpose* of the matrix of cofactors. The transpose B^T of a matrix B is the matrix having the entries of the first row of B in its first column, having the entries of the second row of B in its second column, etc. The transpose of a square matrix is square of the same size and in general the transpose of a $m \times n$ matrix is $n \times m$.

$$\text{adj}(A) = \begin{pmatrix} 1 & 3 & 3 \\ -2 & -1 & -1 \\ -6 & -13 & -8 \end{pmatrix}$$

We conclude that

$$A^{-1} = -\frac{1}{5} \times \text{adj}(A) = -\frac{1}{5}\begin{pmatrix} 1 & 3 & 3 \\ -2 & -1 & -1 \\ -6 & -13 & -8 \end{pmatrix}$$

We can confirm now that $A \times A^{-1} = I_3$.

$$A \times A^{-1} = \begin{pmatrix} 1 & 3 & 0 \\ 2 & -2 & 1 \\ -4 & 1 & -1 \end{pmatrix}\left(-\frac{1}{5}\right)\begin{pmatrix} 1 & 3 & 3 \\ -2 & -1 & -1 \\ -6 & -13 & -8 \end{pmatrix} = -\frac{1}{5}\begin{pmatrix} -5 & 0 & 0 \\ 0 & -5 & 0 \\ 0 & 0 & -5 \end{pmatrix} = I_3$$

Also $A^{-1} \times A = I_3$ (Check).

We conclude this section with an application to the problem of solving a system of linear equation with a square coefficient matrix.

Example 1.6.8 *Solve the following system of linear equations.*

$$\begin{array}{rcrcrcr} x_1 & + & 3x_2 & & & = & 3 \\ 2x_1 & - & 2x_2 & + & x_3 & = & -8 \\ -4x_1 & + & x_2 & - & x_3 & = & 12 \end{array}$$

<u>Solution</u>: The system can be written in matrix form as follows :

$$\underbrace{\begin{pmatrix} 1 & 3 & 0 \\ 2 & -2 & 1 \\ -4 & 1 & -1 \end{pmatrix}}_{\text{coefficient matrix of the system}} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 3 \\ -8 \\ 12 \end{pmatrix}$$

The left-hand side here is the matrix A multiplied by the column vector with entries x_1, x_2 and x_3. This product is a column vector whose three entries are precisely the left-hand sides of the three equations in the system.

So, instead of the original 3 equations we have the *single* matrix equation

$$A \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 3 \\ -8 \\ 12 \end{pmatrix},$$

which we need to solve for x_1, x_2 and x_3. Note that the coefficient matrix A is the matrix of Example 1.6.7 above.

We know A has an inverse - if we multiply the above equation on the left by A^{-1} we obtain :

$$\begin{aligned} A^{-1}A \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} &= A^{-1} \times \begin{pmatrix} 3 \\ -8 \\ 12 \end{pmatrix} \\ \Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} &= A^{-1} \begin{pmatrix} 3 \\ -8 \\ 12 \end{pmatrix} \end{aligned}$$

The right-hand side above will be a column with three entries : these will be the values of x_1, x_2 and x_3 in the (unique) solution of the system.

$$\begin{aligned} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} &= -\frac{1}{5} \begin{pmatrix} 1 & 3 & 3 \\ -2 & -1 & -1 \\ -6 & -13 & -8 \end{pmatrix} \begin{pmatrix} 3 \\ -8 \\ 12 \end{pmatrix} \\ &= -\frac{1}{5} \begin{pmatrix} 1(3) + 3(-8) + 3(12) \\ -2(3) - 1(-8) - 1(12) \\ -6(3) - 13(-8) - 8(12) \end{pmatrix} = -\frac{1}{5} \begin{pmatrix} 15 \\ -10 \\ -10 \end{pmatrix} \end{aligned}$$

Solution: $x_1 = -3,\ x_2 = 2,\ x_3 = 2$

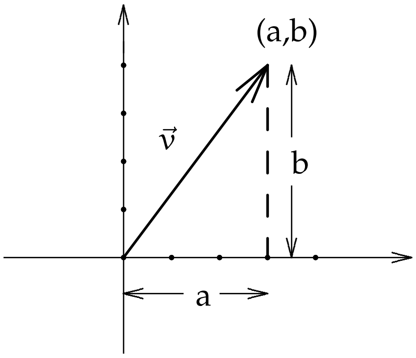
1.7 Some Vector Geometry

Recall that in $\mathbb{R}^2$ (and $\mathbb{R}^3$), the expression (a, b) (or (a, b, c) in $\mathbb{R}^3$) denotes both the point with these coordinates and the vector directed from the origin to that point.

1. The Length of a Vector

Let $\vec{v} = (a, b)$ be a vector in $\mathbb{R}^2$. The *length* of $\vec{v}$ is the length of a line segment representing $\vec{v}$. It is denoted by $||\vec{v}||$. This is $\sqrt{a^2 + b^2}$; we write

$$||(a, b)|| = \sqrt{a^2 + b^2}.$$



For any vector $\vec{v}$, $||\vec{v}||$ is a non-negative real number. Also $||\vec{v}||$ can be equal to zero only if $\vec{v} = (0, 0)$.

Example

(i) $||(5, 12)|| = \sqrt{(5)^2 + (12)^2} = \sqrt{25 + 144} = \sqrt{169} = 13$

(ii) $||(-6, 8)|| = \sqrt{(-6)^2 + (8)^2} = \sqrt{36 + 64} = \sqrt{100} = 10$

(iii) $||(-2, -4)|| = \sqrt{(-2)^2 + (-4)^2} = \sqrt{4 + 16} = \sqrt{20} = 2\sqrt{5}$

2. The Scalar Product

Let $\vec{u} = (a_1, b_1)$ and $\vec{v} = (a_2, b_2)$ be (non-zero) vectors in $\mathbb{R}^2$. We define their *scalar product* (or *dot* product) by

$$\vec{u}.\vec{v} = a_1a_2 + b_1b_2.$$

So $\vec{u}.\vec{v}$ is a *number* (scalar).

In general We define the scalar product for vectors $\vec{u} = (u_1, u_2, \dots, u_n)$ and $\vec{v} = (v_1, \dots, v_n)$ in $\mathbb{R}^n$ by

$$\vec{u}.\vec{v} = u_1v_1 + u_2v_2 + \cdots + u_nv_n.$$

Examples

1. In $\mathbb{R}^2$, $(2, -3).(4, 1) = 2(4) + 3(1) = 5$.
2. In $\mathbb{R}^3$, $(1, -1, 2).(2, 1, 2) = 1(2) + (-1)(1) + 2(2) = 5$.
3. In $\mathbb{R}^4$, $(0, 1, -2, 1).(5, 6, -1, -1) = 0(5) + 1(6) + (-2)(-1) + 1(-1) = 7$.

Note: For matrices A and B, when we calculate the product AB we are basically taking scalar products of the rows of A with the columns of B.

Geometric Significance of the Scalar Product

Theorem 1.7.1 *For any vectors $\vec{u}$ and $\vec{v}$ in $\mathbb{R}^2$*

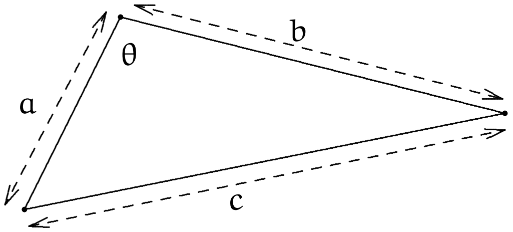
$$\vec{u}.\vec{v} = ||\vec{u}||\,||\vec{v}||\cos\theta$$

where θ is the angle between $\vec{u}$ and $\vec{v}$.

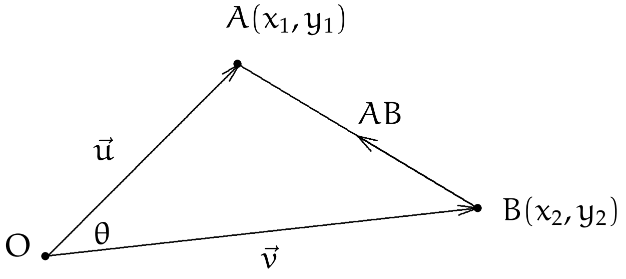
The proof of Theorem 1.7.1 uses the Cosine Rule, which we recall here:

If a triangle has sides of lengths a, b and c, and θ is the angle opposite the side of length c, then

$$c^2 = a^2 + b^2 - 2ab\cos\theta.$$



Proof of Theorem 1.7.1: Let $\vec{u} = (x_1, y_1)$, $\vec{v} = (x_2, y_2)$ and let A and B respectively denote the points (x_1, y_1) and (x_2, y_2).Form the triangle with $\vec{u}$, $\vec{v}$ and the line segment AB as sides.



Cosine Rule : $|AB|^2 = ||\vec{u}||^2 + ||\vec{v}||^2 - 2||\vec{u}||\,||\vec{v}||\cos\theta$
Then

$$\begin{aligned}
(x_2 - x_1)^2 + (y_2 - y_1)^2 &= ||(x_1, y_1)||^2 + ||(x_2, y_2)||^2 - 2||\vec{u}||\,||\vec{v}||\cos\theta \\
\Longrightarrow (x_1 - x_2)^2 + (y_1 - y_2)^2 &= (x_1^2 + y_1^2) + (x_2^2 + y_2^2) - 2||\vec{u}||\,||\vec{v}||\cos\theta \\
\Longrightarrow x_2^2 + x_1^2 - 2x_1x_2 + y_2^2 + y_1^2 - 2y_1y_2 &= x_1^2 + y_1^2 + x_2^2 + y_2^2 - 2||\vec{u}||\,||\vec{v}||\cos\theta \\
\Longrightarrow -2x_1x_2 - 2y_1y_2 &= -2||\vec{u}||\,||\vec{v}||\cos\theta \\
\Longrightarrow x_1x_2 + y_1y_2 &= ||\vec{u}||\,||\vec{v}||\cos\theta \\
\Longrightarrow \vec{u}.\vec{v} &= ||\vec{u}||\,||\vec{v}||\cos\theta
\end{aligned}$$

We say $\vec{u}$ and $\vec{v}$ are *orthogonal* ($\vec{u} \perp \vec{v}$) if the angle between $\vec{u}$ and $\vec{v}$ is $\frac{\pi}{2}$ (90°).

Corollary 1.7.2 *Let $\vec{u}$ and $\vec{v}$ be non-zero vectors in $\mathbb{R}^2$ or $\mathbb{R}^3$. Then $\vec{u} \perp \vec{v}$ if and only if $\vec{u}.\vec{v} = 0$.*

Proof: $\vec{u}.\vec{v} = ||\vec{u}||\,||\vec{v}|| \cos\theta$ by Theorem 1.7.1. Since neither $||\vec{u}||$ nor $||\vec{v}||$ is zero, we have $\vec{u}.\vec{v} = 0$ if and only if $\cos\theta = 0$. This happens precisely if $\vec{u} \perp \vec{v}$.

Example 1.7.3 *Let $\vec{u} = (2,1)$, $\vec{v} = (-4,2)$. Find $\cos\theta$ if θ is the angle between $\vec{u}$ and $\vec{v}$.*

SOLUTION:

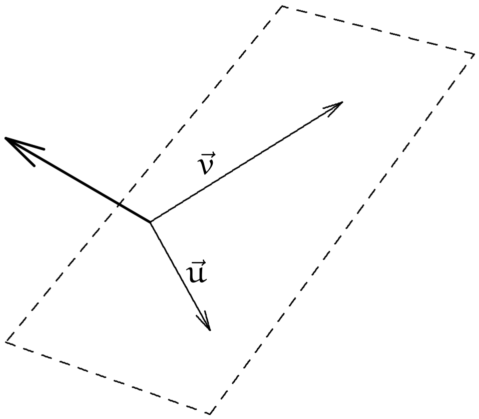
$$\begin{aligned}
\vec{u}.\vec{v} &= ||\vec{u}||\,||\vec{v}|| \cos\theta \\
\vec{u}.\vec{v} &= 2(-4) + 1(2) = -8 + 2 = -6 \\
||\vec{u}|| &= \sqrt{(2)^2 + (1)^2} = \sqrt{5} \\
||\vec{v}|| &= \sqrt{(-4)^2 + (2)^2} = \sqrt{20} \\
\Longrightarrow -6 &= \sqrt{5}\sqrt{20}\cos\theta = \sqrt{100}\cos\theta = 10\cos\theta \\
\cos\theta &= -\frac{6}{10} = -\frac{3}{5}
\end{aligned}$$

Example 1.7.4 *The vectors $\vec{u} = (-1,2,4)$ and $\vec{v} = (2,-1,1)$ are orthogonal in $\mathbb{R}^3$, since*

$$\vec{u}.\vec{v} = (-1,2,4).(2,-1,1) = -2 - 2 + 4 = 0.$$

3. The Cross Product in $\mathbb{R}^3$

Let $\vec{u} = (u_1, u_2, u_3)$ be $\vec{v} = (v_1, v_2, v_3)$ be vectors in $\mathbb{R}^3$. Suppose that $\vec{u}$ and $\vec{v}$ do not point in the same (or opposite) directions, so that they point along different lines. Then there is a unique direction in $\mathbb{R}^3$ that is orthogonal to both $\vec{u}$ and $\vec{v}$.



Definition 1.7.5 *Let $\vec{u} = (u_1, u_2, u_3)$ and $\vec{v} = (v_1, v_2, v_3)$. Then the vector given by*

$$(u_2v_3 - u_3v_2, u_3v_1 - u_1v_3, u_1v_2 - u_2v_1)$$

is called the cross product *(or* vector product*) of $\vec{u}$ and $\vec{v}$ and denoted $\vec{u} \times \vec{v}$.*

Example 1.7.6 *Suppose $\vec{u} = (1,2,-1)$ and $\vec{v} = (2,1,0)$. Find $\vec{u} \times \vec{v}$ and show it is orthogonal to both $\vec{u}$ and $\vec{v}$.*

SOLUTION:

$$\begin{aligned}
\vec{u} \times \vec{v} &= (u_2v_3 - u_3v_2, u_3v_1 - u_1v_3, u_1v_2 - u_2v_1) \\
&= (2(0) - (-1)(1),\ -1(2) - (1)(0),\ 1(1) - 2(2)) \\
&= (1, -2, -3)
\end{aligned}$$

$\vec{u}.(\vec{u} \times \vec{v}) = (1, 2, -1).(1, -2, -3) = 1 - 4 + 3 = 0$
$\vec{v}.(\vec{u} \times \vec{v}) = (2, 1, 0).(1, -2, -3) = 2 - 2 + 0 = 0$
Thus $\vec{u} \perp \vec{u} \times \vec{v}$ and $\vec{v} \perp \vec{u} \times \vec{v}$.

Claim: In general $\vec{u} \times \vec{v}$ is orthogonal to both $\vec{u}$ and $\vec{v}$. To see this calculate $\vec{u}.(\vec{u} \times \vec{v})$:

$$\begin{aligned} \vec{u}.(\vec{u} \times \vec{v}) &= (u_1, u_2, u_3).(u_2v_3 - u_3v_2, u_3v_1 - u_1v_3, u_1v_2 - u_2v_1) \\ &= u_1(u_2v_3 - u_3v_2) + u_2(u_3v_1 - u_1v_3) + u_3(u_1v_2 - u_2v_1) \\ &= u_1u_2v_3 - u_1u_3v_2 + u_2u_3v_1 - u_2u_1v_3 + u_3u_1v_2 - u_3u_2v_1 \\ &= 0 \end{aligned}$$

Similarly one can show that $\vec{v}.(\vec{u} \times \vec{v}) = 0$.

REMARKS

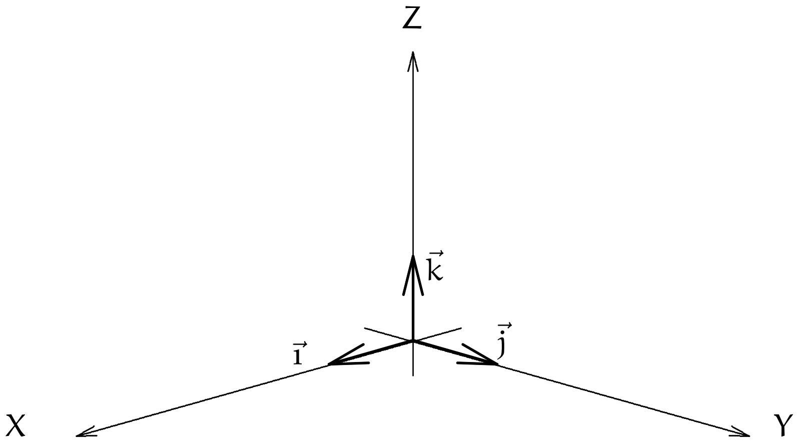
1. $\vec{u} \times \vec{v} = (0, 0, 0)$ (the zero vector) if either $\vec{u}$ or $\vec{v}$ is the zero vector or if $\vec{u}$ is a scalar multiple of $\vec{v}$, i.e. if $\vec{u}$ and $\vec{v}$ have the same (or opposite) directions.
2. The cross product is *not* commutative. In fact $\vec{v} \times \vec{u} = -(\vec{u} \times \vec{v})$. So $\vec{u} \times \vec{v}$ and $\vec{v} \times \vec{u}$ have *opposite* directions.

THE CROSS PRODUCT AS A DETERMINANT

It is conventional in $\mathbb{R}^3$ to write

$$\vec{\imath} = (1, 0, 0), \quad \vec{\jmath} = (0, 1, 0), \quad \vec{k} = (0, 0, 1).$$

Thus $\vec{\imath}$, $\vec{\jmath}$ and $\vec{k}$ are vectors of length 1 pointing along the positive X, Y and Z axes respectively.



Then for example $(2, 3, 4) = 2\vec{\imath} + 3\vec{\jmath} + 4\vec{k}$ and in general the vector (a, b, c) may also be written $a\vec{\imath} + b\vec{\jmath} + c\vec{k}$. The "$\vec{\imath}$, $\vec{\jmath}$, $\vec{k}$" notation will be convenient for computing cross products.

Let $\vec{u} = (u_1, u_2, u_3)$ and $\vec{v} = (v_1, v_2, v_3)$. Then cofactor expansion along the first row confirms that $\vec{u} \times \vec{v}$ is the determinant of the matrix

$$\begin{pmatrix} \vec{\imath} & \vec{\jmath} & \vec{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{pmatrix}.$$

To calculate $(3,1,-2) \times (-2,2,4)$:

$$\begin{aligned}
\det\begin{pmatrix} \vec{\imath} & \vec{\jmath} & \vec{k} \\ 3 & 1 & -2 \\ -2 & 2 & 4 \end{pmatrix} &= \vec{\imath}\det\begin{pmatrix} 1 & -2 \\ 2 & 4 \end{pmatrix} + \vec{\jmath}\left(-\det\begin{pmatrix} 3 & -2 \\ -2 & 4 \end{pmatrix}\right) + \vec{k}\det\begin{pmatrix} 3 & 1 \\ -2 & 2 \end{pmatrix} \\
&= \vec{\imath}(8) + \vec{\jmath}(-8) + \vec{k}(8) \\
\Longrightarrow (3,1,-2) \times (-2,2,4) &= (8,-8,8).
\end{aligned}$$

It is easily checked that $(8,-8,8)$ is orthogonal to both $(3,1,-2)$ and $(-2,2,4)$.

Length of the Cross Product

So far our discussion of the cross product has focussed on its direction. The *length* of $\vec{u} \times \vec{v}$ also has significance, relating to the angle θ between $\vec{u}$ and $\vec{v}$.

Fact 1.7.7 *(Lagrange's Identity) For any vectors $\vec{u}$ and $\vec{v}$ in $\mathbb{R}^3$*

$$||\vec{u} \times \vec{v}||^2 = ||\vec{u}||^2||\vec{v}||^2 - (\vec{u}.\vec{v})^2$$

This can be proved by writing each term in terms of components of $\vec{u}$ and $\vec{v}$.

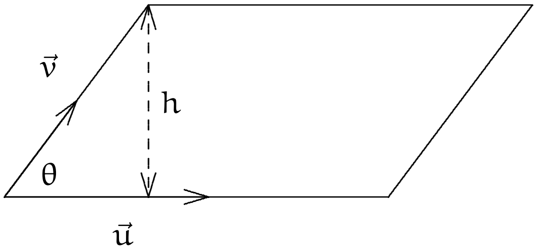
Replacing $\vec{u}.\vec{v}$ by $||\vec{u}||\,||\vec{v}||\cos\theta$, Lagrange's Identity becomes :

$$\begin{aligned}
||\vec{u} \times \vec{v}||^2 &= ||\vec{u}||^2||\vec{v}||^2 - (||\vec{u}||\,||\vec{v}||\cos\theta)^2 \\
&= ||\vec{u}||^2||\vec{v}||^2 - ||\vec{u}||^2||\vec{v}||^2\cos^2\theta \\
&= ||\vec{u}||^2||\vec{v}||^2(1-\cos^2\theta) \\
&= ||\vec{u}||^2||\vec{v}||^2(\sin^2\theta) \\
\Longrightarrow ||\vec{u} \times \vec{v}|| &= ||\vec{u}||\,||\vec{v}||\sin\theta
\end{aligned}$$

(Note that $\sin\theta \geqslant 0$ since θ is between 0 and π (180°)).

Application : Area of a Parallelogram

Suppose $\vec{u}$ and $\vec{v}$ are vectors representing adjacent sides of a parallelogram P. The area of P is $||\vec{u}|| \times h$, where $\vec{u}$ is regarded as the base, and h denotes the perpendicular height of P above $\vec{u}$.



Then

$$\sin\theta = \frac{h}{||\vec{v}||} \Longrightarrow h = ||\vec{v}||\sin\theta$$

Area of $P = ||\vec{u}||h = ||\vec{u}||\,||\vec{v}||\sin\theta = ||\vec{u} \times \vec{v}||$

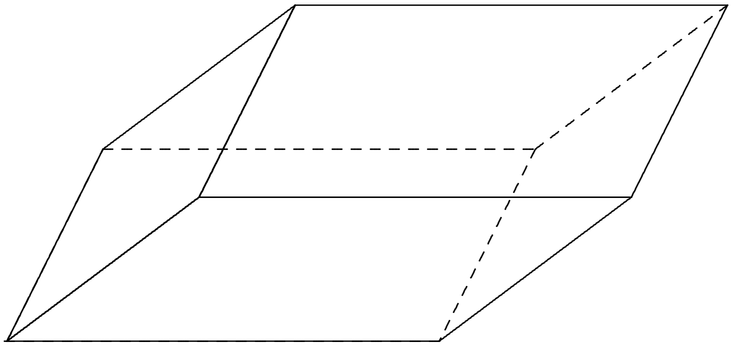
Example 1.7.8 *Find the area of the parallelogram in $\mathbb{R}^3$ having the vectors $(1,2,3)$ and $(7,6,-7)$ as adjacent sides.*

SOLUTION: Area is $||(1,2,3) \times (7,6,-7)||$.
$(1,2,3) \times (7,6,-7) = (-32,28,-8)$ (Check)
Area of parallelogram $= ||(-32,28,-8)|| = 4||(-8,7,-2)|| = 4\sqrt{64+49+4} = 4\sqrt{117} = 12\sqrt{13}$.

Note that the area of triangle with adjacent sides $\vec{u}$ and $\vec{v}$ is given by $\frac{1}{2}||\vec{u} \times \vec{v}||$.

ANOTHER APPLICATION : VOLUME OF A PARALLELEPIPED

A *parallelepiped* in $\mathbb{R}^3$ is a six-faced object, in which pairs of opposite faces consist of similar parallelograms.

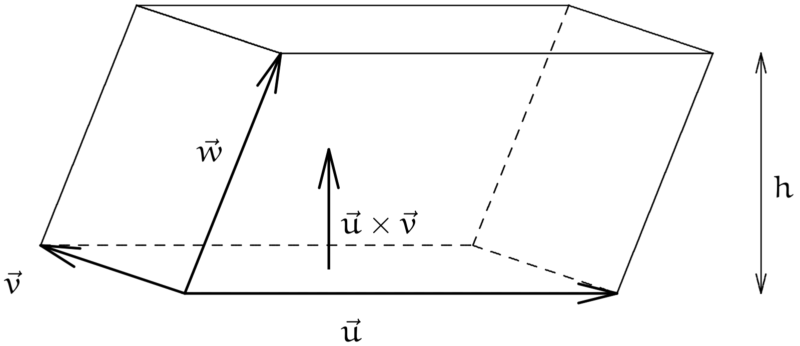


A Parallelepiped

If $\vec{u}$, $\vec{v}$ and $\vec{w}$ are vectors in $\mathbb{R}^3$ having different directions and initial points at O, they form three adjacent sides of a unique parallelepiped.

Example 1.7.9 *Find the volume of the parallelepiped* P *having* $\vec{u} = (1,2,3)$, $\vec{v} = (7,6,-7)$ *and* $\vec{w} = (4,5,-3)$ *as adjacent sides.*

Solution: Suppose the parallelogram with $\vec{u}$ and $\vec{v}$ as sides forms the "base" of P.



Then
$$V = \text{Volume of P} = A \times h$$
where A is the area of the base and h is the (perpendicular) height of P above this base.
$$A = ||\vec{u} \times \vec{v}||$$
The vector $\vec{u} \times \vec{v}$ is perpendicular to the base of P and from the diagram we see that
$$h = ||\vec{w}||\,|\cos\theta| = \frac{|w.(\vec{u} \times \vec{v})|}{||\vec{u} \times \vec{v}||},$$

where θ is the angle between $\vec{w}$ and $\vec{u} \times v$. Thus

$$h = \frac{|\vec{w}.(\vec{u} \times \vec{v})|}{||\vec{u} \times \vec{v}||}$$

Then

$$V = A \times h = ||\vec{u} \times \vec{v}|| \frac{|\vec{w}.(\vec{u} \times \vec{v})|}{||\vec{u} \times \vec{v}||}$$

$$\boxed{\text{Volume of } P \ = \ |\vec{w}.(\vec{u} \times \vec{v})|}$$

Now if $\vec{u} = (u_1, u_2, u_3),\ \vec{v} = (v_1, v_2, v_3)$ and $\vec{w} = (w_1, w_2, w_3)$, we have

$$\vec{w}.(\vec{u} \times \vec{v}) = w_1(u_2v_3 - u_3v_2) + w_2(u_3v_1 - u_1v_3) + w_3(u_1v_2 - u_2v_1).$$

This is exactly the determinant of the matrix

$$\begin{pmatrix} w_1 & w_2 & w_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{pmatrix}.$$

So in our example, the volume of P is the absolute value of the determinant of the matrix

$$\begin{pmatrix} 4 & 5 & -3 \\ 1 & 2 & 3 \\ 7 & 6 & -7 \end{pmatrix}.$$

This is 36.

REMARKS:

1. If $\vec{w},\ \vec{u}$ and $\vec{v}$ are adjacent sides of a parallelepiped P, the volume of P is given by $|\vec{w}.(\vec{u} \times \vec{v})|$.

2. $\vec{w}.(\vec{u} \times \vec{v})$ is called the *scalar triple product* of $\vec{w},\ \vec{u}$ and $\vec{v}$. Although this definition looks non-symmetric, it turns out that $|\vec{w}.(\vec{u} \times \vec{v})|$ does not depend on the order in which $\vec{u},\ \vec{v}$ and $\vec{w}$ are written; i.e.

   $$|\vec{w}.(\vec{u} \times \vec{v})| = |\vec{u}.(\vec{v} \times \vec{w})| = |\vec{v}.(\vec{w} \times \vec{u})| = |\vec{w}.(\vec{v} \times \vec{u})| = |\vec{v}.(\vec{u} \times \vec{w})| = |\vec{u}.(\vec{w} \times \vec{v})|,$$

   for any vectors $\vec{u},\ \vec{v},\ \vec{w}$ in $\mathbb{R}^3$. (Typically three of the six expressions inside the absolute value signs above will be negative and three positive, but all will have the same absolute value).

   In Example 1.7.9 there was no particular reason to choose the parallelogram defined by $\vec{u}$ and $\vec{v}$ as the base : choosing a different face, for example the one defined by $\vec{u}$ and $\vec{w}$ would have resulted in $|\vec{v}.(\vec{u} \times \vec{w})|$ as the volume formula.

3. Suppose $T : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ is a linear transformation and let M_T be its matrix. Then the columns of M_T have as their entries the components of the vectors $T(1,0,0),\ T(0,1,0)$ and $T(0,0,1)$. The parallelepiped P having $(1,0,0),\ (0,1,0)$ and $(0,0,1)$ as adjacent edges (which has volume 1) is transformed by T to the parallelepiped P' having $T(1,0,0),\ T(0,1,0)$ and $T(0,0,1)$ as adjacent edges. The absolute value of the determinant of the matrix of T is the volume of this parallelepiped P'. If $\det(M_T) = 0$, this means that P is transformed not to another parallelepiped but to a parallelogram, line segment, or a single point.

   It is also true that if T is a linear transformation from $\mathbb{R}^2$ to $\mathbb{R}^2$, the absolute value of the determinant of the matrix of T is the area of the parallelogram to which the square with the vectors $(1,0)$ and $(0,1)$ is transformed by T (see Problem Sheet 3).

Chapter 2

Introduction to Number Theory

2.1 The Well-Ordering Axiom for $\mathbb{Z}$

INTEGERS AND NATURAL NUMBERS
The set $\mathbb{Z}$ of integers includes all the "whole numbers" :

$$\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, 3, \dots\}.$$

The set $\mathbb{N}$ of natural numbers or "counting numbers" is given by

$$\mathbb{N} = \{1, 2, 3, \dots\}.$$

NOTE: Some authors include 0 in the set of natural numbers, there is a lack of consensus about this. We will write $\mathbb{N}_0$ for the set of *non-negative* integers :

$$\mathbb{N}_0 = \{0, 1, 2, 3, \dots\}.$$

We also have the set $\mathbb{Q}$ of rational numbers and the set $\mathbb{R}$ of real numbers, and among these sets we have the following inclusions :

$$\mathbb{N} \subset \mathbb{N}_0 \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}.$$

All of the number systems mentioned here are *ordered*. This means : choose a and b in $\mathbb{Z}$ (or $\mathbb{Q}$ or $\mathbb{R}$). Then either $a \leqslant b$ (a is less than or equal to B) or $b \leqslant a$ (b is less than or equal to a), and these occur simultaneously if and only if the elements a and b are equal.

However, the order on the set $\mathbb{Z}$ of integers differs substantially in its properties from the order on the set $\mathbb{Q}$ of rational numbers.

Definition 2.1.1 *Let S be a non-empty subset of $\mathbb{Z}$. An integer b is a* lower bound *for S if $b \leqslant s$ for every element s of S.*

EXAMPLES

1. $\mathbb{N}$ has 1 as a lower bound, since $1 \leqslant n$ for every natural number n. Any integer less than 1 is also a lower bound for $\mathbb{N}$.
2. If $S = \{4, -3, 5, 56\}$ then -5 is a lower bound for S. So are $-10, -6$ and any integer less than or equal to -3.
3. The set $2\mathbb{Z} = \{\dots, -4, -2, 0, 2, 4, 6, \dots\}$ of even integers does not have a lower bound. Given any integer b, there exists an even integer c for which c is less than b.

The definition of lower bound above can be applied to $\mathbb{Q}$ or $\mathbb{R}$ as well as to $\mathbb{Z}$. In $\mathbb{Q}$ for example, 0 is a lower bound for the set of positive rational numbers.

Definition 2.1.2 *Let S be a non-empty subset of $\mathbb{Z}$ (or $\mathbb{Q}$ or $\mathbb{R}$). An element b of $\mathbb{Z}$ (or $\mathbb{Q}$ or $\mathbb{R}$) is the* least element *of S if*

- b *is a lower bound for S, and*
- $b \in S$ *(b is an element of S).*

Note: A subset of $\mathbb{Z}$ (or $\mathbb{Q}$ or $\mathbb{R}$) can have at most one least element, for suppose b and c are both least elements of a non-empty subset S of $\mathbb{Z}$ (or $\mathbb{Q}$ or $\mathbb{R}$). Then $b \leqslant c$ and $c \leqslant b$, which means $b = c$.

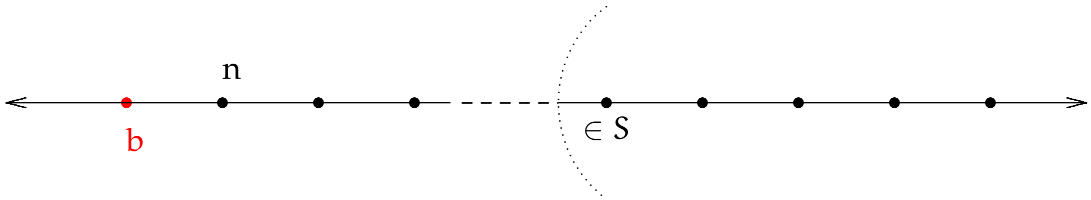
EXAMPLES

1. $\mathbb{Z}$ has no least element (but it has no lower bound).
2. The least element of $\mathbb{N}$ is 1.
3. The least element of the subset $\{-3, -30, -1, 16\}$ of $\mathbb{Z}$ is -30.
4. The set of positive rational numbers has no least element, although it does have lower bounds (for example 0).

The Well-Ordering Axiom for $\mathbb{Z}$ states the following :
Let S be a non-empty subset of $\mathbb{Z}$ and suppose that S has a lower bound. Then S has a least element.

NOTES:

1. The Well-Ordering Axiom does not hold for $\mathbb{Q}$, since the set $\mathbb{Q}^+$ of positive rational numbers has no least element. To see this let q be any positive rational number. Then $\frac{1}{2}q$ is a positive rational number less than q, so q is not a lower bound for $\mathbb{Q}^+$. Thus $\mathbb{Q}^+$ has no least element, although it does have lower bounds.
2. To understand why the Well-Ordering Axiom makes sense in $\mathbb{Z}$, suppose that S is a non-empty subset of $\mathbb{Z}$ and let b be a lower bound for S.



Then on the number line, every element of S is to the right of b. The integers are regularly spaced along the number line, and if $b \notin S$ we can we travel right from b and we will encounter a first integer n (n is the smallest integer for which $b < n$). If $n \notin S$ we can proceed to $n+1$, $n+2$ etc, making progress along the number line as we go. Eventually we will encounter for the first time an element of S.

3. This approach does not work for $\mathbb{Q}$ because, whereas the integers are regularly spaced along the number line, the rational numbers are densely packed into the number line. For example the interval $\left[\frac{1}{4}, \frac{1}{2}\right]$ contains infinitely many rational numbers but no integer. Given a rational number a, there is no "next" rational number after a.

4. Every non-empty subset of $\mathbb{N}$ has a lower bound in $\mathbb{Z}$, for example 0. Thus every non-empty subset of $\mathbb{N}$ has a least element. This is sometimes given as a formulation of the well-ordering axiom.

2.2 The Principle of Mathematical Induction

Suppose we have a statement about every natural number n (or every natural number $\geqslant k$ for some fixed $k \in \mathbb{N}$). For example

- The sum of the first n positive integers is $\dfrac{n(n+1)}{2}$, i.e. for all $n \geqslant 1$
$$1+2+\cdots+n=\frac{n(n+1)}{2}.$$
i.e. $\displaystyle\sum_{i=1}^{n} i=\frac{n(n+1)}{2}$.
- For all $n \geqslant 4$, $n! \geqslant 2^n$.
$n!$ is n *factorial*, the product of the first n positive integers.
NOTE: The statement "for all $n \geqslant 4$, $n! \geqslant 2^n$ encapsulates separate statements for $n = 4, 5, 6, \ldots$.
When $n = 4$ this statement says $4! \geqslant 2^4$.
When $n = 5$ it says $5! \geqslant 2^5$.
When $n = 100$ it says $100! \geqslant 2^{100}$.
- For $n \geqslant 1$,
$$1(2)(4)+2(3)(5)+\cdots+n(n+1)(n+3)=\frac{1}{12}n(n+1)(n+2)(3n+13)$$
(i.e. $\displaystyle\sum_{i=1}^{n} i(i+1)(i+3)=\frac{1}{12}n(n+1)(n+2)(3n+13)$).

The truth of such statements can be checked for individual values of n, but how can we prove that they hold for *every* relevant value of n?

One strategy is to use the *Principle of Mathematical Induction*. This means

1. The Base.
Check directly that the statement holds for the least relevant value of n.
2. The Induction Step.
Prove that if the statement holds when $n = k$ then it also holds when $n = k+1$.
This is the heart of the proof, and finding a deductive argument to convince your reader that the statement about $k+1$ somehow follows from the statement about k is not always an easy task. There is no template or set of instructions for how to establish the induction step, separate arguments have to be developed for separate examples.
3. Suppose the base occurs when $n = n_0$. Then, having checked the base, the step tells us that the statement also holds when $n = n_0 + 1$. Another application of the step tells us that that the statement also holds when $n = n_0 + 2$. For any integer $\geqslant n_0$, a finite number of applications of the induction step tells us that the statement holds when n has this value. This is the Principle of Mathematical Induction.

Example 2.2.1 *(Summer 2005) Prove by induction on* n *that*

$$1(2)(4)+2(3)(5)+\cdots+n(n+1)(n+3)=\frac{1}{12}n(n+1)(n+2)(3n+13)$$

for $n \geqslant 1$.

Proof:

1. The Base.
   When $n = 1$ we have on the left $1(2)(4) = 8$
   and on the right $\frac{1}{12}1(2)(3)(16) = 8$.
   So the statement holds when $n = 1$.

2. The Induction Step.
   Assume that the statement holds when $n = k$ (this is *the Induction Hypothesis*) and try to deduce that it holds when $n = k + 1$, i.e. that

$$\begin{aligned} 1(2)(4) + 2(3)(5)+ \quad \ldots \quad &+k(k+1)(k+3) + (k+1)(k+2)(k+4) \\ = \quad & \frac{1}{12}(k+1)((k+1)+1)((k+1)+2)(3(k+1)+13) \\ = \quad & \frac{1}{12}(k+1)((k+2)((k+3)(3k+16). \end{aligned}$$

   We have

$$1(2)(4)+2(3)(5)+\cdots+k(k+1)(k+3)+(k+1)(k+2)(k+4) = \frac{1}{12}k(k+1)(k+2)(3k+13)+(k+1)(k+2)(k+4)$$

   by the induction hypothesis. Thus

$$\begin{aligned} 1(2)(4) + \cdots + (k+1)(k+2)(k+4) &= \frac{1}{12}(k+1)(k+2)\left(k(3k+13) + 12(k+4)\right) \\ &= \frac{1}{12}(k+1)(k+2)(3k^2 + 25k + 48) \\ &= \frac{1}{12}(k+1)(k+2)(k+3)(3k+16) \\ &= \frac{1}{12}(k+1)((k+1)+1)((k+1)+2)(3(k+1)+13) \end{aligned}$$

   as required. This establishes the induction step.

3. By 1. (the base), 2. (the step) and the Principle of Induction, the proof is complete.

REMARKS

1. The Principle of Induction works because of the Well-Ordering Axiom for $\mathbb{Z}$. It works because it is possible to go from one integer *to the next*. It would not be possible to use the Principle of Induction in the same way to prove a statement about rational or real numbers.

2. In some cases, instead of just assuming the truth of a statement for one value $n = k$ and deducing it for $n = k + 1$, we need to assume it for *all* $n \leqslant k$ and deduce it for $n = k + 1$. This variant is sometimes called the *Strong* Principle of Mathematical Induction.

Example 2.2.2 *Suppose that for each natural number n the integer u_n is defined by*

$$u_1 = 3, \; u_2 = 5, \; u_n = 3u_{n-1} - 2u_{n-2} \text{ for } n \geqslant 3.$$

Prove that $u_n = 2^n + 1$ *for* $n \geqslant 1$.

Proof:

1. The Base.
   When $n = 1$ we have $u_1 = 3 = 2^1 + 1$.
   When $n = 2$ we have $u_2 = 5 = 2^2 + 1$.
   So the statement holds when $n \leqslant 2$.

2. The Induction Step.
   Induction Hypothesis : Assume that $u_n = 2^n + 1$ for all $n \leqslant k$.
   Then $u_{k+1} = 3u_k - 2u_{k-1}$ and by the induction hypothesis

   $$u_{k+1} = 3(2^k + 1) - 2(2^{k-1} + 1) = 3(2^k) + 3 - 2^k - 2 = 2(2^k) + 1 = 2^{k+1} + 1,$$

   as required.

3. By 1. and 2. and the (strong) principle of mathematical induction, $u_n = 2^n + 1$ for every natural number n.

2.3 The Division Algorithm in $\mathbb{Z}$

If we "divide" 34 by 3 in $\mathbb{Z}$ we obtain a quotient of 11 and a remainder of 1. Thus

$$34 = 3(11) + 1.$$

Divide	Quotient	Remainder
50 by 6	8	2
45 by 7	6	3
35 by 5	7	0

In all cases the remainder is non-negative and less than the positive integer by which we are dividing.

Theorem 2.3.1 *(The Division Algorithm in $\mathbb{Z}$) Let a and b be integers, with b positive. Then there exist unique integers q and r for which*

$$a = qb + r \text{ and } 0 \leqslant r < b.$$

(The integers q and r are respectively called the *quotient* and *remainder* on dividing a by b.)
NOTE: The following proof is included in these notes for completeness; this proof will not be the subject of any exam questions.

Proof: Let $S = \{x \in \mathbb{N}_0 : x = a - yb \text{ for some } y \in \mathbb{Z}\}$. So S is the set consisting of those non-negative integers that differ from a by a multiple of b. Then S is not empty : to prove this we need to show that some non-negative integer can be written as $a - yb$ for some integer y.

- If $a \geqslant 0$ we can put $y = 0$ to obtain $a \in S$.
- If $a < 0$, put $y = a$ to get $x = a - ab = a(1 - b)$ - a non-negative integer.

So S is non-empty and therefore S has a least element r, and

$$r = a - qb \text{ for some } q \in \mathbb{Z};\ a = qb + r.$$

Now

$$r - b = a - qb - b = a - (q + 1)b.$$

However $r - b \notin S$ since r is the least element of S. Thus we conclude that $r - b$ is negative, $0 \leqslant r < b$. This establishes the *existence* part of the theorem.

For the uniqueness, suppose that

$$a = bq_1 + r_1 \text{ and } a = bq_2 + r_2,$$

where $q_1, q_2, r_1, r_2 \in \mathbb{Z}$ and $0 \leqslant r_1 < b$, $o \leqslant r_2 < b$. We can assume that $r_1 \geqslant r_2$. Then $0 = (q_1 - q - 2)b + (r_1 - r_2)$ and

$$r_1 - r - 2 = (q_2 - q_1)b.$$

Since $0 \leqslant r_1 - r_2 < b$, this is possible only if $r_1 - r_2 = 0$ and $(q_2 - q_1)b = 0$ which means $q_2 - q - 1 = 0$. Thus $r_1 = r_2$ and $q_1 = q_2$. This proves the uniqueness of q and r in the statement of the theorem. □

Definition 2.3.2 *Let a and b be integers. We say that b divides a in $\mathbb{Z}$ if $a = bc$ for some integer c. We write $b|a$ to indicate that b divides a.*

EXAMPLES

- $3|12$ since $12 = 3 \times 4$.

- $6|(-42)$ since $-42 = 6 \times (-7)$.
- $5 \not| 21$ (5 does not divide 21).

The statement $b|a$ can also be expressed as

- b is a divisor (or factor) of a in $\mathbb{Z}$.
- a is a multiple of b in $\mathbb{Z}$.
- (if $b > 0$) the remainder on dividing a by b in $\mathbb{Z}$ is 0.

NOTE ON NOTATION: A common error is to confuse the symbol "|" for "divides" with a slash as in the fraction 2/5. The symbol for "divides" is a vertical bar not a forward or back slash or a dash. The statement "$b|a$ (in $\mathbb{Z}$)" means "a is an integer multiple of b". This is not related to the notation used in the fractions a/b or b/a.

2.4 Greatest Common Divisors and the Euclidean Algorithm

Let a and b be integers. An integer c is a *common divisor* of a and b if $c|a$ and $c|b$. For example 3 is a common divisor of 15 and 30.

Definition 2.4.1 *The integer* d *is the* greatest common divisor *(gcd) of* a *and* b *if*

- $d|a$ *and* $d|b$ - d *is a common divisor of* a *and* b.
- *If* c *is any common divisor of* a *and* b*, then* $c|d$.
- $d \geqslant 1$.

The greatest common divisor of 30 and 45 is 15 - we write $\gcd(45, 30) = 15$ (some authors just write $(45, 30) = 15$).

We will show that every pair (a, b) of (non-zero) integers has a unique gcd. We note that we can assume that both a and b are positive, since $-a$ and a have the same integer divisors, as do $-b$ and b.

Given positive integers a and b with $a > b$, we can calculate $\gcd(a, b)$ as in the following example.

Example 2.4.2 *Calculate* $\gcd(770, 528)$.

Step 1 Write $a = bq_1 + r_1$ where $0 \leqslant r_1 < b$. Then $r_1 = a - bq_1$, so every common divisor of a and b is a divisor of r_1, and hence a common divisor of b and r_1. On the other hand since $a = bq_1 + r_1$, every common divisor of b and r_1 is a divisor of a, and hence a common divisor of a and b. Thus the pairs (a, b) and (b, r_1) have the same sets of common divisors and

$$\gcd(a, b) = \gcd(b, r_1).$$

In our example

$$770 = 528(1) + 242, \ \ r_1 = 242.$$

Step 2 Now write $b = r_1q_2 + r_2$ where $0 \leqslant r_2 < r_1$. By the above reasoning $\gcd(b, r_1) = \gcd(r_1, r_2) = \gcd(a, b)$.

$$528 = 242(2) + 44, \ \ r_2 = 44.$$

Step 3 Now write $r_1 = r_2q_3 + r_3$ with $0 \leqslant r_3 < r_2$. Continuing like this we create a sequence

$$b > r_1 > r_2 > \cdots > 0.$$

This is a strictly decreasing sequence of non-negative integers, so it reaches 0 after a finite number of steps. Each pair of successive terms in the sequence has the same gcd and this is $\gcd(a, b)$. So $\gcd(a, b)$ is the last non-zero term in the sequence. We have

$$242 = 44(5) + 22, \ \ r_3 = 22.$$

$$44 = 22(2) + 0, \ \ r_4 = 0.$$

We conclude that $\gcd(770, 528) = 22$.

Note $22 = \gcd(22, 44) = \gcd(44, 242) = \gcd(242, 528) = \gcd(528, 770)$.

The procedure that we have just used to calculate $\gcd(770, 528)$ is called the *Euclidean Algorithm.*

Example 2.4.3 *Calculate* $\gcd(1704, 1344)$ *using the Euclidean Algorithm.*

SOLUTION:

1. $1704 = 1344(1) + 360, \quad r_1 = 360$
2. $1344 = 360(3) + 264, \quad r_2 = 264$
3. $360 = 264(1) + 96, \quad r_3 = 96$
4. $264 = 96(2) + 72, \quad r_4 = 72$
5. $96 = 72(1) + 24, \quad r_5 = 24$
6. $72 = 24(3) + 0.$

So $\gcd(1704, 1344) = 24$.

The next theorem, which is one of the main themes of this chapter, captures an important property of the greatest common divisor.

Theorem 2.4.4 *Let a and b be integers and let $d = \gcd(a, b)$. Then there exist integers m and n for which*

$$d = ma + nb.$$

REMARKS

1. What this theorem says about Example 2.4.3 is that there exist integers m and n for which

   $$24 = 1704m + 1344n.$$

   To understand what the theorem is about, think about what integers you might expect to get by adding a multiple (positive or negative) of 1704 to a multiple (positive or negative) of 1344. Convince yourself that any number that could possibly arise that way would have to be a multiple of 24. What is not obvious yet is that 24 itself arises this way - that is the content of the theorem.

2. More generally, it is easy enough to see that any integer that can be written in the form $ma + nb$ for integers m and n must be divisible by all common divisors of a and b, and hence by $\gcd(a, b)$. It is perhaps less obvious that $\gcd(a, b)$ can be written in this form.

3. Theorem 2.4.4 can be proved by going backwards through the steps involved in calculating $\gcd(a, b)$ using the Euclidean algorithm. Rather than giving a formal proof of this theorem we will demonstrate how it works by writing 24 in the form $1704m + 1344n$ for integers m and n.

Step 1 Look at Step 5 in the calculation of $\gcd(1704, 1344)$. This says

$$24 = 96 + 72(-1).$$

Step 2 Now use Step 4 to replace 72 with a combination of 96 and 264.

$$24 = 96 + 72(-1) = 96 + (264 + 96(-2))(-1) = 96(3) + 264(-1).$$

Step 3 Use Step 3 to replace 96 with a combination of 360 and 264.

$$24 = 96(3) + 264(-1) = (360 + 264(-1))(3) + 264(-1) = 360(3) + 264(-4).$$

Step 4 Use Step 2 to write 264 as a combination of 1344 and 360. Then

$$24 = 360(3) + 264(-4) = 360(3) + (1344 + 360(-3))(-4) = 360(15) + 1344(-4).$$

Step 5 Finally use Step 1 to write 360 as a combination of 1344 and 1704. Then

$$24 = 360(15) + 1344(-4) = (1704 + 1344(-1))(15) + 1344(-4) = 1704(15) + 1344(-19).$$

So we have succeeded in writing 24 in the form $1704m+1344n$, where $m = 15$ and $n = -19$.

Definition 2.4.5 *Let a and b be non-zero integers. The a and b are* coprime *or* relatively prime *(to each other) if* $\gcd(a, b) = 1$.

Equivalently a and b are relatively prime if they have no common divisors except 1 and -1.

From Theorem 2.4.4 we can say that a and b are relatively prime if and only if there exist integers m and n for which

$$1 = ma + nb.$$

Example 2.4.6 *Find integers m and n for which*

$$1 = 98m + 85n.$$

SOLUTION: First apply the Euclidean algorithm to 98 and 85.

1. $98 = 85(1) + 13$
2. $85 = 13(6) + 7$
3. $13 = 7(1) + 6$
4. $7 = 6(1) + 1$
5. $6 = 1(6) + 0$

Now reverse the steps :

4. $1 = 7 + 6(-1)$
3. $1 = 7 + (13 + 7(-1))(-1) = 7(2) + 13(-1)$
2. $1 = (85 + 13(-6))(2) + 13(-1) = 85(2) + 13(-13)$
1. $1 = 85(2) + (98 + 85(-1))(-13) = 85(15) + 98(-13)$

So we have $1 = 85(15) + 98(-13);\ m = -13,\ n = 15$.

REMARKS:

1. The integers m and n in these problems are not unique. For example in the above problem we could obtain another solution as follows :

   $$1 = 85(15) + 98(-13) = 85(15) + 85(-98) + 98(85) + 98(-13) = 85(-83) + 98(72).$$

   So $m = 72, n = -83$ would be another solution.
   Exercise - think about how all possible solutions are related.

2. Let a and b be non-zero integers. An integer d can be written as $ma+nb$ for integers m and n if and only if $\gcd(a, b)$ divides d.

2.5 Factorization of Integers

Definition 2.5.1 *A positive integer* p *is said to be* prime *if* $p \geqslant 2$ *and the only positive integers that divide* p *are 1 and* p.

The list of primes begins as follows :

$$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61 \ldots$$

Theorem 2.5.2 *Every integer* $\geqslant 2$ *can be written as the product of a finite number of prime factors.*

Examples

$$\begin{array}{rcccl} 100 & = & 2 \times 2 \times 5 \times 5 & = & 2^2 \times 5^2 \\ 891 & = & 11 \times 3 \times 3 \times 3 \times 3 & = & 3^4 \times 11 \\ 794 & = & 2 \times 397 & & \end{array}$$

Proof of Theorem 2.5.2: Suppose the theorem is false. Then the set of integers $\geqslant 2$ that cannot be written as a product of primes is non–empty, and by the Well-Ordering Axiom it has a least element m. Then m is not prime (otherwise it would be the product of one prime) so $m = m_1 m_2$, where m_1 and m_2 are positive integers both strictly less than m.

Now since $m_1 < m$ we have $m_1 = p_1 \times p_2 \times \cdots \times p_k$ and since $m_2 < m$ we have $m_2 = q_1 \times q_2 \times \cdots \times q_l$, where $p_1, p_2, \ldots, p_k$ and $q_1, q_2, \ldots, q_l$ are primes. Then

$$m = p_1 \times p_2 \times \cdots \times p_k \times q_1 \times q_2 \times \cdots \times q_l,$$

so m is after all the product of a finite number of primes. This contradiction proves the theorem. □

The uniqueness of the expression as a product of primes of a given positive integer is a consequence of the following important lemma.

Lemma 2.5.3 *Let* a *and* b *be positive integers and let* p *be a prime that divides the product* ab. *Then* $p|a$ *or* $p|b$ *(or both).*

Note: This statement becomes false if p is replaced with a composite (i.e. non-prime) integer. For example 6 divides $72 = 8 \times 9$, but 6 divides neither 8 nor 9. The lemma says that for example 5 cannot divide the product of two integers neither of which is a multiple of 5.

Proof of Lemma 2.5.3: Since $p|ab$ we can write $ab = pk$ for some integer k. Suppose that p does not divide a. Then $\gcd(a, p) = 1$ since p is prime, and so there exist integers m and n for which $1 = pm + an$. Then

$$b = pmb + abn = pmb + pkn = p(mb + kn).$$

Thus b is a multiple of p, i.e. $p|b$.

We have shown : if $p \not| a$, then $p|b$. Thus p divides at least one of a and b. □

Corollary 2.5.4 *Let* p *be a prime and suppose that* p *divides the product* $a_1 a_2 \ldots a_n$, *where* $a_1, \ldots, a_n$ *are positive integers. Then* p *divides at least one of the* a_i.

Proof: By induction on n.
Base: The case $n = 1$ is clear.
The case $n = 2$ is exactly Lemma 2.5.3.

Induction Hypothesis: Assume that whenever p divides the product of k positive integers, it divides at least one of the factors.

Induction Step: Suppose that p divides $a_1 a_2 \dots a_{k+1}$ where $a_1, \dots, a_{k+1}$ are positive integers. Then, by Lemma 2.5.3, either p divides $a_1 a_2 \dots a_k$ or p divides a_{k+1}. In the first case p divides at least one of $a_1, \dots, a_k$ by the induction hypothesis. So in all cases, p divides at least one of $a_1, \dots, a_{k+1}$. This proves the corollary, by the principle of induction. □

Theorem 2.5.5 *(The Fundamental Theorem of Arithmetic) The expression for an integer $\geqslant 2$ as a product of primes is unique.*

Thus if for some integer $n \geqslant 2$ we have

$$n = p_1 \times p_2 \times \cdots \times p_r \text{ and } n = q_1 \times q_2 \times \cdots \times q_s$$

for primes $p_1, \dots, p_r$ and $q_1, \dots, q_s$, then $s = r$ and $p_1, \dots, p_r$ are exactly $q_1, \dots, q_r$ in some order.

Examples

Integer	Factorization
230	$2 \times 5 \times 23$
576	$2^6 \times 3^2$
1017	$3^2 \times 113$

Proof of Theorem 2.5.5: Suppose for some integer $n \geqslant 2$ we have

$$n = p_1 \times p_2 \times \cdots \times p_r \text{ and } n = q_1 \times q_2 \times \cdots \times q_s$$

for primes $p_1, \dots, p_r$ and $q_1, \dots, q_s$. We can assume $s \geqslant r$. Then by Corollary 2.5.4

$$p_1 | q_1 q_2 \dots q_s \Longrightarrow p_1 | q_k$$

for some $k \in \{1, \dots, s\}$. Thus $p_1 = q_k$ since p_1 and q_k are both primes. After reordering the q_j we can assume $p_1 = q_1$ and

$$p_1 p_2 \dots p_r = p_1 q_2 \dots q_s \Longrightarrow p_2 \dots p_r = q_2 \dots q_s.$$

Repeating this step, and reordering the q_j when necessary, we obtain

$$p_1 = q_1,\ p_2 = q_2, \dots, p_r = q_r,$$

and after r steps we have

$$1 = q_{r+1} \dots q_s.$$

It follows that $s = r$ and that $p_1, \dots, p_r$ are the original $q_1, \dots, q_r$ in some order. □

We conclude this section (and this set of lecture notes) with a description of some more propoerties of prime numbers. First we describe a proof due to Euclid of the well-known statement that the number of primes is infinite.

Theorem 2.5.6 *There are infinitely many primes.*

Proof (Euclid) : Suppose that the set of primes is finite, and suppose that $p_1, p_2, \dots, p_k$ is the full list of primes. Define a positive integer P by

$$P = p_1 \times p_2 \times \cdots \times p_k + 1.$$

Then by Theorem 2.5.2, P is a product of prime numbers. However none of $p_1, \dots, p_k$ can divide P, since we obtain a remainder of 1 upon division of P by any of these. Thus there exist primes outside the set $\{p_1, \dots, p_k\}$, and the full set of primes is infinite. □

An algorithm for determining all the primes less than N for a fixed N was developed by Eratosthenes of Cyrene in the 3rd century BC. This technique is known as the *Sieve of Eratosthenes* and it uses the following fact :

Suppose that a positive integer $m \leqslant N$ is composite. Then if $m = ab$ for integers a and b strictly less than m, at least one of a and b is less than (or equal to) $\sqrt{N}$. So every composite integer less than N is a multiple of some integer that is at most equal to $\sqrt{N}$.

To implement the Sieve of Eratosthenes :

1. Write out the integers from 2 to N.
2. Strike out all the multiples of 2 that are greater than 2.
3. Move to the next remaining number and strike out all of its multiples (greater than itself).
4. Repeat Step 3 until the next remaining number exceeds $\sqrt{N}$.
5. The remaining numbers are the primes in the range 1 to N.

Example 2.5.7 *Use the Sieve of Eratosthenes to find all the primes in the range 1 to 30.*

Step 1		2	3	4	5	6	7	8	9	10
	11	12	13	14	15	16	17	18	19	20
	21	22	23	24	25	26	27	28	29	30
Step 2		2	3		5		7		9	
	11		13		15		17		19	
	21		23		25		27		29	
Step 3		2	3		5		7			
	11		13				17		19	
			23		25				29	
Step 4		2	3		5		7			
	11		13				17		19	
			23						29	

We can stop now since the next remaining number, 7, exceeds $\sqrt{30}$. So the primes in the range 1 to 30 are 2,3,5,7,11,13,17,19,23 and 29.

The set of prime numbers has historically been, and continues to be, a subject of intense study. While much is known about prime numbers, many famous and interesting questions remain unanswered. We conclude now with a selection of important known theorems about primes, and a selection of (as yet) unsolved problems.

Some Facts and Open Problems about Primes

1. If $n \geqslant 2$ is an integer, then $\pi(n)$ denotes the number of primes in the range from 1 to n. (So $\pi(6) = 3$ for example). The *Prime Number Theorem* states that

$$\lim_{n\to\infty} \frac{\pi(n)}{n/\log n} = 1.$$

This was first proved (independently) by Hadamard and de la Vallée Poussin in 1896.

2. An efficient algorithm for determining if a given integer is prime was discovered in 2002 by Agarwal, Kayal and Saxena. No efficient algorithm is known for finding factors of very large integers, and many modern cryptographic systems rely on this fact.

3. *The Goldbach Conjecture (Goldbach 1742)*. Every even integer greater than 2 is the sum of two primes.
   At present it is known that every positive even integer is the sum of six or fewer primes (Ramaré 1995). The Goldbach conjecture has been verified for all integers up to ~ 10^{17}).
   (See the novel *Uncle Petros and Goldbach's Conjecture* by Apostolos Doxiadis).

4. Is every positive even integer the difference of two primes?

5. *The Twin Prime Conjecture*. A pair of *twin primes* is a pair of primes whose difference is 2, for example 5 and 7, 17 and 19, etc. The twin prime conjecture says that there are infinitely many pairs of twin primes.

   A more general conjecture says that for any positive even integer $2n$, there are infinitely many pairs of consecutive primes whose difference is $2n$. But the related question 4 above is still open.

6. *Fermat Primes*. The *Fermat Number* F_n is defined for $n \geqslant 0$ by $F_n = 2^{2^n} + 1$:

   $$F_0 = 3,\ F_1 = 5,\ F_2 = 17,\ F_3 = 257,\ F_4 = 65537,$$

   all of which are prime. Fermat conjectured that F_n is prime for all n but in fact to date only the above five Fermat primes have been discovered, and F_n is known to be composite for $5 \leqslant n \leqslant 32$. It is now conjectured that the number of Fermat primes is finite.

7. *Mersenne Primes*. For a prime p, the *Mersenne Number* M_p is defined by $M_p = 2^p - 1$. For many values of p we find that M_p is a prime number, called a *Mersenne prime*:

   $$M_2 = 3,\ M_3 = 7,\ M_5 = 31,\ M_7 = 127, \ldots$$

   However M_p is composite for some values of p, for example

   $$M_{11} = 2^{11} - 1 = 2047 = 23 \times 89.$$

   It is not known whether the number of Mersenne primes is finite, or whether the number of composite Mersenne numbers is finite. It is onjectured that both are infinite.