DEPARTMENT OF MATHEMATICS & STATISTICS

MATH1012 MATHEMATICAL THEORY AND METHODS

Acknowledgements: The following members of the Department of Mathematics and Statistics, UWA, have contributed in one way or another for the production of these Lecture Notes: A. Bassom, E. Cripps, A. Devillers, M. Giudici, L. Jennings, K. Judd, D. Hill, J. Hopwood, M. Matthews, A. Niemeyer, G. Royle, T. Stemler, L. Stoyanov.

Contents

1
Systems of linear equations

This chapter covers the systematic solution of *systems of linear equations* using Gaussian elimination and back-substitution and the description, both algebraic and geometric, of their *solution space*.

Before commencing this chapter, students should be able to:

- Plot linear equations in 2 variables, and
- Add and multiply matrices.

After completing this chapter, students will be able to:

- Systematically solve systems of linear equations with many variables, and
- Identify when a system of linear equations has 0, 1 or infinitely many solutions, and
- Give the solution set of a system of linear equations in parametric form.

1.1 Systems of linear equations

A *linear equation* is an equation of the form

$$x + 2y = 4$$

where each term in the equation is either a number[1] (i.e. "6") or a *numerical multiple* of a variable (i.e., "$2x$", "$4y$"). If an equation involves powers or products of variables (x^2, xy, etc.) or any other functions ($\sin x$, e^x, etc.) then it is not linear.

[1] You will often see the word *scalar* used to refer to a number, and *scalar multiple* to describe a numerical multiple of a variable.

A *system* of linear equations is a set of one or more linear equations considered together, such as

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

which is a system of two equations in the two variables[2] x and y.

[2] Often the variables are called "unknowns" emphasizing that *solving* a system of linear equations is a process of *finding* the unknowns.

A *solution* to a system of linear equations is an assignment of values to the variables such that *all* of the equations in the system

are satisfied. For example, there is a *unique solution* to the system given above which is

$$x = 2 \quad , \quad y = 1.$$

Particularly when there are more variables, it will often be useful to give the solutions as *vectors* like

$$(x, y) = (2, 1).$$

In this case we could also just give the solution as $(2, 1)$ where, by convention the first component of the vector is the x-coordinate. Usually the variables will have names like x, y, z or x_1, x_2, ..., x_n and so we can just specify the vector alone and it will be clear which component corresponds to which variable.

If the system of linear equations involves just two variables, then we can visualise the system *geometrically* by plotting the solutions to each equation separately on the xy-plane, as illustrated in Figure 1.1. The solution to the system of linear equations is the point where the two plots intersect, which in this case is the point $(2, 1)$.

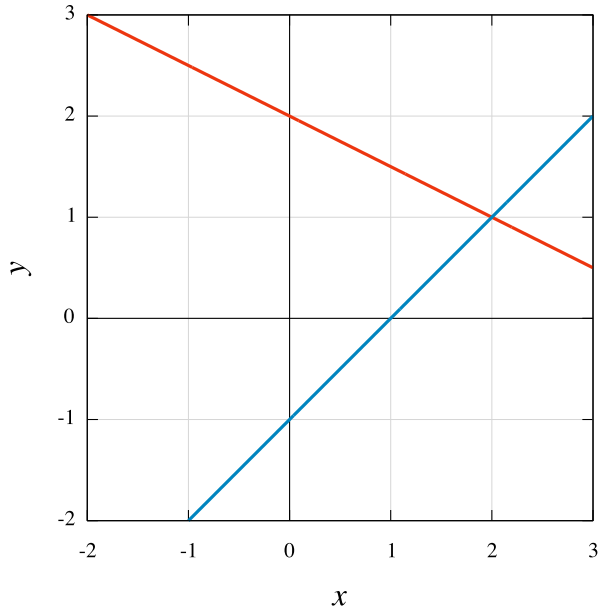


Figure 1.1: The two linear equations $x + 2y = 4$ and $x - y = 1$ plotted as intersecting lines in the xy-plane.

It is easy to visualise systems of linear equations in two variables, but it is more difficult in three variables, where we need 3-dimensional plots. In three dimensions the solutions to a single linear equation such as

$$x + 2y - z = 4$$

form a *plane* in 3-dimensional space. While computer algebra systems can produce somewhat reasonable plots of surfaces in three dimensions, it is hard to interpret plots showing two or more intersecting surfaces.

Recall from MATH1011 that this particular equation describes the plane with normal vector $(1, 2, -1)$ containing the point $(4, 0, 0)$.

With four or more variables any sort of visualisation is essentially impossible and so to reason about systems of linear equations with many variables, we need to develop *algebraic* tools rather than *geometric* ones.

It is still very useful to use *geometric intuition* to think about systems of linear equations with many variables provided you are careful about where it no longer applies.

1.1.1 *Solutions to systems of linear equations*

The system of linear equations shown in Figure 1.1 has a *unique* solution. In other words there is just one (x, y) pair that satisfies

both equations, and this is represented by the unique point of intersection of the two lines. Some systems of linear equations have *no solutions* at all. For example, there are no possible values for (x, y) that satisfy *both* of the following equations

$$\begin{aligned} x + 2y &= 4 \\ 2x + 4y &= 3. \end{aligned}$$

If a system of linear equations has at least one solution, then it is called *consistent*, and otherwise it is called *inconsistent*.

Geometrically, the two equations determine *parallel* but *different* lines, and so they do not meet. This is illustrated in Figure 1.2.

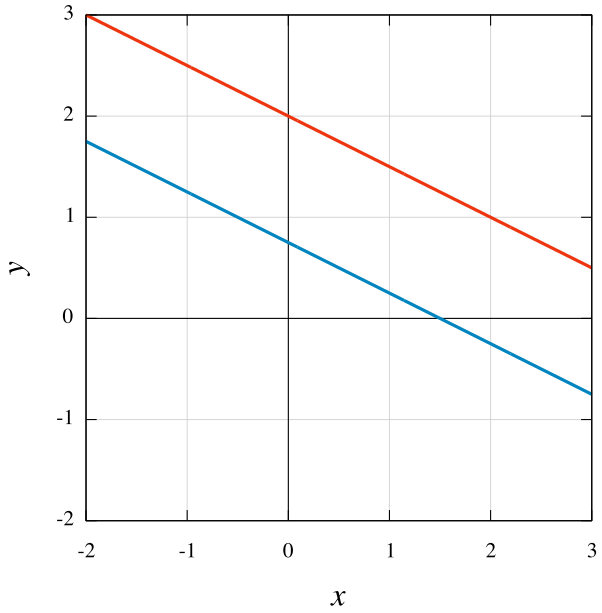


Figure 1.2: The inconsistent system of equations $x + 2y = 4$ and $2x + 4y = 2$ plotted as lines in the xy-plane.

There is another possibility for the number of solutions to a system of linear equations, which is that a system may have *infinitely many* solutions. For example, consider the system

$$\begin{aligned} x + 2y + z &= 4 \\ y + z &= 1. \end{aligned} \tag{1.1}$$

Each of the two equations determines a plane in three dimensions, and as the two planes are not parallel[3], they meet in a *line* and so *every point on the line* is a solution to this system of linear equations.

[3] The two planes are not parallel because the normal vectors to the two planes, that is $\boldsymbol{n}_1 = (1, 2, 1)$ and $\boldsymbol{n}_2 = (0, 1, 1)$ are not parallel.

Remark **1.1.** *How can we describe the solution set to a system of linear equations with infinitely many solutions?*

One way of describing an infinite solution set is in terms of *free parameters* where one (or more) of the variables is left unspecified with the values assigned to the other variables being expressed as *formulas* that depend on the free variables.

Let's see how this works with the system of linear equations given by Equation (1.1): here we can choose z to be the "free variable" but then to satisfy the second equation it will be necessary

to have $y = 1 - z$. Then the first equation can only be satisfied by taking

$$\begin{aligned} x &= 4 - 2y - z \\ &= 4 - 2(1 - z) - z \qquad \text{(using } y = 1 - z\text{)} \\ &= 2 + z. \end{aligned}$$

Thus the *complete* solution set S of system (1.1) is

$$S = \{(2 + z, 1 - z, z) \mid z \in \mathbb{R}\}.$$

To find a *particular* solution to the linear system, you can pick any desired value for z and then the values for x and y are determined. For example, if we take $z = 1$ then we get $(3, 0, 1)$ as a solution, and if we take $z = 0$ then we get $(2, 1, 0)$, and so on. For this particular system of linear equations it would have been possible to choose one of the other variables to be the "free variable" and we would then get a different expression for the same solution set.

EXAMPLE **1.2.** *(Different free variable) To rewrite the solution set* $S = \{(2 + z, 1 - z, z) \mid z \in \mathbb{R}\}$ *so that the y-coordinate is the free variable, just notice that as* $y = 1 - z$*, this implies that* $z = 1 - y$ *and so the solution set becomes* $S = \{(3 - y, y, 1 - y) \mid y \in \mathbb{R}\}$.

A system of linear equations can also be expressed as a single *matrix equation* involving the product of a matrix and a *vector* of variables. So the system of linear equations

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

can equally well be expressed as

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & -1 \\ 2 & 3 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 5 \\ -1 \\ 3 \end{bmatrix}$$

just using the usual rules for multiplying matrices.[4] In general, a system of linear equations with m equations in n variables has the form

$$A\boldsymbol{x} = \boldsymbol{b}$$

where A is an $m \times n$ *coefficient matrix*, $\boldsymbol{x}$ is an $n \times 1$ vector of variables, and $\boldsymbol{b}$ is an $m \times 1$ vector of scalars.

[4] Matrix algebra is discussed in detail in Chapter 3 but for this representation as a system of linear equations, just the definition of the product of two matrices is needed.

1.2 *Solving linear equations*

In this section, we consider a systematic method of solving systems of linear equations. The method consists of two steps, first using *Gaussian elimination* to reduce the system to a simpler system of linear equations, and then *back-substitution* to find the solutions to the simpler system.

In high school, systems of linear equations are often solved with *ad hoc* methods that are quite suitable for small systems, but which are not sufficiently systematic to tackle larger systems. It is *very important* to thoroughly learn the systematic method, as solving systems of linear equations is a fundamental part of many of the questions that arise in linear algebra. In fact, almost every question in linear algebra ultimately depends on setting up and solving a suitable system of linear equations!

1.2.1 *Elementary row operations*

An *elementary row operation* is an operation that transforms a system of linear equations into a *different*, but *equivalent* system of linear equations, where "equivalent" means that the two systems have *identical* solutions. The answer to the obvious question — *" Why bother transforming one system of linear equations into another?"* — is that the new system might be *simpler to solve* than the original system. In fact there are some systems of linear equations that are extremely simple to solve, and it turns out that by systematically applying a *sequence* of elementary row operations, we can transform any system of linear equations into an equivalent system whose solutions are very simple to find.

DEFINITION **1.3.** *(Elementary row operations)*
An elementary row operation *is one of the following three types of transformation applied to a system of linear equations:*

Type 1 *Interchanging two equations.*

Type 2 *Multiplying an equation by a non-zero scalar.*

Type 3 *Adding a multiple of one equation to another equation.*

In a system of linear equations, we let R_i denote the i-th equation, and so we can express an elementary row operation symbolically as follows:

$R_i \leftrightarrow R_j$	Exchange equations R_i and R_j
$R_i \leftarrow \alpha R_i$	Multiply equation R_i by α
$R_i \leftarrow R_i + \alpha R_j$	Add α times R_j to R_i

We will illustrate elementary row operations on a simple system of linear equations:

$$\begin{aligned} x + 2y + z &= 5 \\ y - z &= -1 \\ 2x + 3y - z &= 3 \end{aligned} \tag{1.2}$$

EXAMPLE **1.4.** *(Type 1 Elementary Row Operation) Applying the Type 1 elementary row operation $R_1 \leftrightarrow R_2$ (in words, "interchange equations 1 and 2") to the original system Equation (1.2) yields the system of linear equations:*

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

It is obvious that this new system of linear equations has exactly the same solutions as the original system, because each individual equation is unchanged and listing them in a different order does not alter which vectors satisfy them all. □

EXAMPLE **1.5.** *(Type 2 Elementary Row Operation) Applying the Type 2 elementary row operation $R_2 \leftarrow 3R_2$ (in words, "multiply the second equation by 3") to the original system Equation* (1.2) *gives a new system of linear equations:*

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

Again it is obvious that this system of linear equations has exactly the same solutions as the original system. While the second equation is changed, the solutions *to this individual equation are not changed.*[5] □

[5] This relies on the equation being multiplied by a *non-zero* scalar.

EXAMPLE **1.6.** *(Type 3 Elementary Row Operation) Applying the Type 3 elementary row operation $R_3 \leftarrow R_3 - 2R_1$ (in words,"add -2 times the first equation to the third equation") to the original system Equation* (1.2) *gives a new system of linear equations:*

$$\begin{aligned} x + 2y + z &= 5 \\ y - z &= -1 \\ - y - 3z &= -7 \end{aligned}$$

In this case, it is not obvious that the system of linear equations has the same solutions as the original. In fact, the system is actually different *from the original, but it happens to have the* exact same set of solutions. *This is so important that it needs to be proved.*

As foreshadowed in the last example, in order to use elementary row operations with confidence, we must be sure that the *set of solutions* to a system of linear equations is *not changed* when the system is altered by an elementary row operation. To convince ourselves of this, we need to prove[6] that applying an elementary row operation to a system of linear equations neither destroys existing solutions nor creates new ones.

[6] A *proof* in mathematics is a careful explanation of why some mathematical fact is true. A proof normally consists of a sequence of statements, each following logically from the previous statements, where each individual logical step is sufficiently simple that it can easily be checked. The word "proof" often alarms students, but really it is nothing more than a very simple line-by-line explanation of a mathematical statement. Creating proofs of interesting or useful new facts is the *raison d'être* of a professional mathematician.

THEOREM **1.7.** *Suppose that S is a system of linear equations, and that T is the system of linear equations that results by applying an elementary row operation to S. Then the set of solutions to S is equal to the set of solutions to T.*

Proof. As discussed in the examples, this is obvious for Type 1 and Type 2 elementary row operations. So suppose that T arises from S by performing the Type 3 elementary row operation $R_i \leftarrow R_i + \alpha R_j$. Then S consists of m equations

$$S = \{R_1, R_2, \ldots, R_m\}$$

while T only differs in the i-th equation

$$T = \{R_1, R_2, \ldots, R_{i-1}, R_i + \alpha R_j, R_{i+1}, \ldots, R_m\}.$$

It is easy to check that if a vector satisfies two equations R_i and R_j, then it also satisfies $R_i + \alpha R_j$ and so any solution to S is also a solution to T. What remains to be checked is that any solution to T

is a solution to S. However if a vector satisfies all the equations in T, then it satisfies $R_i + \alpha R_j$ and R_j, and so it satisfies the equation

$$(R_i + \alpha R_j) + (-\alpha R_j)$$

which is just R_i. Thus any solution to T also satisfies S. □

Now let's consider applying an entire sequence of elementary row operations to our example system of linear equations Equation (1.2) to reduce it to a much simpler form.

So, starting with

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

apply the Type 3 elementary row operation $R_3 \leftarrow R_3 - 2R_1$ to get

$$\begin{aligned} x + 2y + z &= 5 \\ y - z &= -1 \\ -y - 3z &= -7 \end{aligned}$$

followed by the Type 3 elementary row operation $R_3 \leftarrow R_3 + R_2$ obtaining

$$\begin{aligned} x + 2y + z &= 5 \\ y - z &= -1 \\ -4z &= -8 \end{aligned}$$

Now notice that the third equation only involves the variable z, and so it can now be solved, obtaining $z = 2$. The second equation involves just y, z and as z is now known, it really only involves y, and we get $y = 1$. Finally, with both y and z known, the first equation only involves x and by substituting the values that we know into this equation we discover that $x = 1$. Therefore, this system of linear equations has the unique solution $(x, y, z) = (1, 1, 2)$.

Notice that the final system was essentially trivial to solve, and so the elementary row operations converted the original system into one whose solution was trivial.

1.2.2 *The augmented matrix*

The names of the variables in a system of linear equations are essentially irrelevant — whether we call three variables x, y and z or x_1, x_2 and x_3 makes no fundamental difference to the equations or their solution. Thus writing out each equation in full when writing down a sequence of systems of linear equations related by elementary row operations involves a lot of unnecessary repetition of the variable names. Provided each equation has the variables in the same order, all the information is contained solely in the *coefficients*, and so these are all we need. Therefore we normally represent a system of linear equations by a matrix known as the *augmented matrix* of the system of linear equations; each row of the matrix represents a single equation, with the coefficients of the variables

to the left of the bar, and the constant term to the right of the bar. Each column to the left of the bar contains all of the coefficients for a single variable. For our example system Equation (1.2), we have the following:

$$\begin{array}{rcl} x + 2y + z = 5 \\ y - z = -1 \\ 2x + 3y - z = 3 \end{array} \qquad \left[\begin{array}{ccc|c} 1 & 2 & 1 & 5 \\ 0 & 1 & -1 & -1 \\ 2 & 3 & -1 & 3 \end{array}\right]$$

EXAMPLE **1.8.** *(From matrix to linear system) What system of linear equations has the following augmented matrix?*

$$\left[\begin{array}{ccc|c} 0 & -1 & 2 & 3 \\ 1 & 0 & -2 & 4 \\ 3 & 4 & 1 & 0 \end{array}\right]$$

The form of the matrix tells us that there are three variables, which we can name arbitrarily, say x_1, x_2 and x_3. Then the first row of the matrix corresponds to the equation $0x_1 - 1x_2 + 2x_3 = 3$, and interpreting the other two rows analogously, the entire system is

$$\begin{array}{rcl} - x_2 + 2x_3 = 3 \\ x_1 - 2x_3 = 4 \\ 3x_1 + 4x_2 + x_3 = 0. \end{array}$$

We could also have chosen any other three names for the variables. □

In other words, the augmented matrix for the system of linear equations $A\boldsymbol{x} = \boldsymbol{b}$ is just the matrix $[A \mid \boldsymbol{b}]$.

1.3 Gaussian elimination

When solving a system of linear equations using the augmented matrix, the elementary row operations[7] are performed directly on the augmented matrix.

As explained earlier, the aim of the elementary row operations is to put the matrix into a simple form from which it is easy to "read off" the solutions; to be precise we need to define exactly the simple form that we are trying to achieve.

[7] This is why they are called elementary *row* operations, rather than elementary equation operations, because they are always viewed as operating on the rows of the augmented matrix.

DEFINITION **1.9.** *(Row echelon form)*
A matrix is in row echelon form *if*

1. *Any rows of the matrix consisting entirely of zeros occur as the last rows of the matrix, and*
2. *The first non-zero entry of each row is in a column strictly to the right of the first non-zero entry in any of the earlier rows.*

This definition is slightly awkward to read, but very easy to

grasp by example. Consider the two matrices

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

Neither matrix has any all-zero rows so the first condition is automatically satisfied. To check the second condition we need to identify the first non-zero entry in each row — this is called the *leading entry*:

$$\begin{bmatrix} \boxed{1} & 1 & -1 & 2 & 0 \\ 0 & 0 & \boxed{-2} & 1 & 3 \\ 0 & 0 & 0 & \boxed{1} & 0 \\ 0 & 0 & 0 & 0 & \boxed{-1} \end{bmatrix} \qquad \begin{bmatrix} \boxed{1} & 1 & -1 & 2 & 0 \\ 0 & 0 & \boxed{-2} & 1 & 3 \\ 0 & \boxed{2} & 0 & 1 & 0 \\ 0 & 0 & \boxed{1} & 2 & -1 \end{bmatrix}$$

In the first matrix, the leading entries in rows 1, 2, 3 and 4 occur in columns 1, 3, 4 and 5 respectively and so the leading entry for each row always occurs *strictly further to the right* than the leading entry in any earlier row. So this first matrix is in row-echelon form. However for the second matrix, the leading entry in rows 2 and 3 occur in columns 3 and 2 respectively, and so the leading entry in row 3 actually occurs *to the left* of the leading entry in row 2; hence this matrix is not in row-echelon form.

EXAMPLE **1.10.** *(Row-echelon form) The following matrices are all in row-echelon form:*

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

□

EXAMPLE **1.11.** *(Not row-echelon form) None of the following matrices are in row-echelon form:*

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

□

Gaussian Elimination (sometimes called *row-reduction*) is a systematic method for applying elementary row operations to a matrix until it is in row-echelon form. We'll see in the next section that a technique called *back substitution*, which involves processing the equations in reverse order, can easily determine the set of solutions to a system of linear equations whose augmented matrix is in row-echelon form.

Without further ado, here is the algorithm for Gaussian Elimination, first informally in words, and then more formally in symbols. The algorithm is defined for *any matrices*, not just the augmented

matrices arising from a system of linear equations, because it has many applications.

DEFINITION **1.12.** *(Gaussian elimination — in words)*
Let A be an $m \times n$ matrix. At each stage in the algorithm, a particular position in the matrix, called the pivot position*, is being processed. Initially the pivot position is at the top-left of the matrix. What happens at each stage depends on whether the* pivot entry *(that is, the number in the pivot position) is zero or not.*

1. *If the pivot entry is zero then, if possible, interchange the pivot row with one of the rows below it, in order to ensure that the pivot entry is non-zero. This will be possible* unless *the pivot entry and every entry below it are zero, in which case simply move the pivot position one column to the right.*
2. *If the pivot entry is non-zero then, by adding a suitable multiple of the pivot row to every row below the pivot row, ensure that every entry* below *the pivot entry is zero. Then move the pivot position one column to the right and one row down.*

When the pivot position is moved off the matrix, then the process finishes and the matrix will be in row-echelon form.

The process of "adding a multiple of the pivot row to every row below it in order to zero out the column below the pivot entry" is called *pivoting* on the pivot entry for short.

EXAMPLE **1.13.** *(Gaussian elimination) Consider the following matrix, with the initial pivot position marked:*

$$\begin{bmatrix} \boxed{2} & 1 & 2 & 4 \\ 2 & 1 & 1 & 0 \\ 4 & 3 & 2 & 4 \end{bmatrix}$$

The initial pivot position is the $(1,1)$ position in the matrix, and the pivot entry is therefore 2. Pivoting on the $(1,1)$-entry is accomplished by performing the two elementary operations $R_2 \leftarrow R_2 - R_1$ and $R_3 \leftarrow R_3 - 2R_1$, leaving the matrix:

$$\begin{bmatrix} 2 & 1 & 2 & 4 \\ 0 & \boxed{0} & -1 & -4 \\ 0 & 1 & -2 & -4 \end{bmatrix} \begin{matrix} \\ R_2 \leftarrow R_2 - R_1 \\ R_3 \leftarrow R_3 - 2R_1 \end{matrix}$$

(The elementary row operations used are noted down next to the relevant rows to indicate how the row reduction is proceeding.) The new pivot entry is 0, but as the entry immediately under the pivot position is non-zero, interchanging the two rows moves a non-zero to the pivot position.

$$\begin{bmatrix} 2 & 1 & 2 & 4 \\ 0 & \boxed{1} & -2 & -4 \\ 0 & 0 & -1 & -4 \end{bmatrix} \begin{matrix} \\ R_2 \leftrightarrow R_3 \\ R_3 \leftrightarrow R_2 \end{matrix}$$

The next step is to pivot on this entry in order to zero out all the entries below it and then move the pivot position. As the only entry below the pivot is already zero, no elementary row operations need be performed, and the only action required is to move the pivot:

$$\begin{bmatrix} 2 & 1 & 2 & 4 \\ 0 & 1 & -2 & -4 \\ 0 & 0 & \boxed{-1} & -4 \end{bmatrix}.$$

Once the pivot position reaches the bottom row, there are no further operations to be performed (regardless of whether the pivot entry is zero or not) and so the process terminates, leaving the matrix in row-echelon form

$$\begin{bmatrix} 2 & 1 & 2 & 4 \\ 0 & 1 & -2 & -4 \\ 0 & 0 & -1 & -4 \end{bmatrix}$$

as required. □

For completeness, and to provide a description more suitable for implementing Gaussian elimination on a computer, we give the same algorithm more formally – in a sort of pseudo-code.[8]

[8] Pseudo-code is a way of expressing a computer program precisely, but without using the syntax of any particular programming language. In pseudo-code, assignments, loops, conditionals and other features that vary from language-to-language are expressed in natural language.

DEFINITION **1.14.** *(Gaussian elimination — in symbols)*
Let $A = (a_{ij})$ be an $m \times n$ matrix and set two variables $r \leftarrow 1$, $c \leftarrow 1$. (Here r stands for "row" and c for "column" and they store the pivot position.) Then repeatedly perform whichever one of the following operations is possible (only one will be possible at each stage) until either $r > m$ or $c > n$, at which point the algorithm terminates.

1. *If $a_{rc} = 0$ and there exists $x > r$ such that $a_{xc} \neq 0$ then perform the elementary row operation $R_r \leftrightarrow R_x$.*
2. *If $a_{rc} = 0$ and $a_{xc} = 0$ for all $x > r$, then set $c \leftarrow c + 1$.*
3. *If $a_{rc} \neq 0$ then, for* each *$x > r$, perform the elementary row operation*

$$R_x \leftarrow R_x - (a_{xc}/a_{rc})R_r,$$

and then set $r \leftarrow r + 1$ and $c \leftarrow c + 1$.

When this algorithm terminates, the matrix will be in row-echelon form.

1.4 Back substitution

Recall that the whole point of elementary row operations is to transform a system of linear equations into a *simpler system* with the *same solutions*; in other words to change the problem to an easier problem with the same answer. So after reducing the augmented matrix of a system of linear equations to row-echelon form, we now need a way to read off the solution set.

The first step is to determine whether the system is consistent or otherwise, and this involves identifying the *leading entries* in each row of the augmented matrix in row-echelon form — these are the *first non-zero entries* in each row. In other words, run a finger along each row of the matrix stopping at the first non-zero entry and noting which *column* it is in.

EXAMPLE **1.15.** *(Leading entries) The following two augmented matrices in row-echelon form have their* leading entries *highlighted.*

$$\left[\begin{array}{cccc|c} \boxed{1} & 0 & -1 & 2 & 3 \\ 0 & 0 & \boxed{2} & 1 & 0 \\ 0 & 0 & 0 & 0 & \boxed{2} \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] \qquad \left[\begin{array}{cccc|c} \boxed{1} & 2 & -1 & 2 & 3 \\ 0 & 0 & \boxed{2} & 1 & 0 \\ 0 & 0 & 0 & \boxed{-1} & 2 \end{array}\right]$$

□

The left-hand matrix of Example 1.15 has the property that one of the leading entries is on the *right-hand side* of the augmenting bar. If we "unpack" what this means for the system of linear equations, then we see that the third row corresponds to the linear equation

$$0x_1 + 0x_2 + 0x_3 + 0x_4 = 2,$$

which can *never be satisfied*. Therefore this system of linear equations has no solutions, or in other words, is *inconsistent*. This is in fact a *defining feature* of an inconsistent system of linear equations, a fact that is important enough to warrant stating separately.

You may wonder why we keep saying "to the right of the augmenting bar" rather than "in the last column". The answer is that if we have more than one linear equation with the same coefficient matrix, say $A\boldsymbol{x} = \boldsymbol{b_1}$, $A\boldsymbol{x} = \boldsymbol{b_2}$, then we can form a "super-augmented" matrix $[A \mid \boldsymbol{b_1}\ \boldsymbol{b_2}]$ and solve both systems with one application of Gaussian elimination. So there may be more than one column to the right of the augmenting bar.

THEOREM **1.16.** *A system of linear equations is inconsistent if and only if one of the leading entries in the row-echelon form of the augmented matrix is to the right of the augmenting bar.*

Proof. Left to the reader. □

The right-hand matrix of Example 1.15 has no such problem, and so we immediately conclude that the system is consistent — it has at least one solution. Every column to the left of the augmenting bar corresponds to one of the variables in the system of linear equations

$$\begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array}\left[\begin{array}{cccc|c} 1 & 2 & -1 & 2 & 3 \\ 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & -1 & 2 \end{array}\right] \tag{1.3}$$

and so the leading entries identify *some of the variables*. In this case, the leading entries are in columns 1, 3 and 4 and so the identified variables are x_1, x_3 and x_4. The variables identified in this fashion are called the *basic variables* (also known as *leading variables*) of the system of linear equations. The following remark is the key to understanding solving systems of linear equations by back substitution:

REMARK **1.17.** *Every* non-basic variable *of a system of linear equations is a* free variable *or* free parameter *of the system of linear equations, while every* basic variable *can be expressed uniquely as a combination of the free parameters and/or constants.*

The process of *back-substitution* refers to examining the equations in reverse order, and for each equation finding the unique expression for the basic variable corresponding to the leading entry of that row. Let's continue our examination of the right-hand matrix of Example 1.15, also shown with the columns identified in Equation (1.3).

The third row of the matrix, when written out as an equation, says that $-1x_4 = 2$, and so $x_4 = -2$, which is an expression for the basic variable x_4 as a constant. The second row of the matrix corresponds to the equation $2x_3 + x_4 = 0$, but as we know now that $x_4 = -2$, this can be substituted in to give $2x_3 - 2 = 0$ or $x_3 = 1$. The *first row* of this matrix corresponds to the equation

$$x_1 + 2x_2 - x_3 + 2x_4 = 3$$

and after substituting in $x_3 = 1$ and $x_4 = -2$ this reduces to

$$x_1 + 2x_2 = 8. \tag{1.4}$$

This equation involves one basic variable (that is, x_1) together with a non-basic variable (that is, x_2) and a constant (that is, 8). The rules of back-substitution say that this should be manipulated to give an expression *for the basic variable* in terms of the other things. So we get

$$x_1 = 8 - 2x_2$$

and the entire solution set for this system of linear equations is given by

$$S = \{(8 - 2x_2, x_2, 1, -2) \mid x_2 \in \mathbb{R}\}.$$

Therefore we conclude that this system of linear equations has *infinitely many solutions* that can be described by *one free parameter*.

The astute reader will notice that (1.4) could equally well be written $x_2 = 4 - x_1/2$ and so we could use x_1 as the free parameter, rather than x_2 — so why does back-substitution need to specify which variable should be chosen as the free parameter? The answer is that there is *always* an expression for the solution set that uses the non-basic variables as the free parameters. In other words, the process as described will *always work*.

EXAMPLE **1.18.** *(Back substitution) Find the solutions to the system of linear equations whose augmented matrix in row-echelon form is*

$$\left[\begin{array}{cccccc|c} 0 & 2 & -1 & 0 & 2 & 3 & 1 \\ 0 & 0 & 1 & 3 & -1 & 0 & 2 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 5 \end{array}\right]$$

First identify the leading entries in the matrix, and therefore the basic and non-basic variables. The leading entries in each row are highlighted below

$$\left[\begin{array}{cccccc|c} 0 & \boxed{2} & -1 & 0 & 2 & 3 & 1 \\ 0 & 0 & \boxed{1} & 3 & -1 & 0 & 2 \\ 0 & 0 & 0 & 0 & \boxed{1} & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \boxed{1} & 5 \end{array}\right]$$

Therefore the basic variables *are* $\{x_2, x_3, x_5, x_6\}$ *while the* free variables *are* $\{x_1, x_4\}$ *and so this system has infinitely many solutions that can be described with two free parameters. Now back-substitute starting from the last equation. The fourth equation is simply that* $x_6 = 5$*, while the third equation gives* $x_5 + x_6 = 0$ *which after substituting the known value for* x_6 *gives us* $x_5 = -5$*. The second equation is*

$$x_3 + 3x_4 - x_5 = 2$$

and so it involves the basic variable x_3 *along with the free variable* x_4 *and the already-determined variable* x_5*. Substituting the known value for* x_5 *and rearranging to give an expression for* x_3*, we get*

$$x_3 = -3 - 3x_4.$$

Finally the first equation is

$$2x_2 - x_3 + 2x_5 + 3x_6 = 1$$

and so substituting all that we have already determined we get

$$2x_2 - (-3 - 3x_4) + 2(-5) + 3(5) = 1$$

which simplifies to

$$x_2 = \frac{-7 - 3x_4}{2}.$$

What about x_1*? It is a variable in the system of linear equations, but it did not actually occur in any of the equations. So if it does not appear in any of the equations, then there are no restrictions on its values and so it can take any value — therefore it is a free variable. Fortunately, the rules for back-substitution have already identified it as a non-basic variable as it should be. Therefore the final solution set for this system of linear equations is*

$$S = \left\{ \left(x_1, \frac{1}{2}(-7 - 3x_4), -3 - 3x_4, x_4, -5, 5 \right) \,\middle|\, x_1, x_4 \in \mathbb{R} \right\}$$

and therefore we have found an expression with two free parameters as expected. □

KEY CONCEPT **1.19.** *(Solving systems of linear equations)*
To solve a system of linear equations of the form $A\boldsymbol{x} = \boldsymbol{b}$*, perform the following steps:*

1. *Form the augmented matrix* $[A \mid \boldsymbol{b}]$.

2. *Use Gaussian elimination to put the augmented matrix into row-echelon form.*
3. *Use back-substitution to express each of the basic variables as a combination of the free variables and constants.*

1.5 A more advanced method: Gauss-Jordan elimination

We now explain a method that allows us to do both Gaussian elimination and back-substitution **at the same time, both in matrix form**.

In Gaussian elimination, when we pivot on an entry in the matrix, we use the pivot row in order to zero-out the rest of the column *below* the pivot entry. However there is nothing stopping us from zeroing out the rest of the column *above* the pivot entry as well. We will now do *more* elementary row operations in order to make the system of linear equations *even simpler* than before.

Let's do this on the system with the following augmented matrix and see how useful it is:

$$\left[\begin{array}{ccc|c} -1 & 0 & 1 & 1 \\ 0 & 1 & -1 & 0 \\ 2 & 0 & -1 & 0 \end{array}\right].$$

After pivoting on the $(1,1)$-entry we get

$$\left[\begin{array}{ccc|c} -1 & 0 & 1 & 1 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 2 \end{array}\right] \quad R_3 \leftarrow R_3 + 2R_1$$

which is now in row-echelon form. We can now use the last pivot to zero-out the rest of the third column:

$$\left[\begin{array}{ccc|c} -1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 2 \end{array}\right] \begin{array}{l} R_1 \leftarrow R_1 - R_3 \\ R_2 \leftarrow R_2 + R_3 \end{array}$$

One final elementary row operation puts the augmented matrix into an especially nice form.

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 2 \end{array}\right] \quad R_1 \leftarrow (-1)R_1$$

In this form not even any back substitution is needed to find the solution; the system of equations has solution $x_1 = 1$, $x_2 = 2$ and $x_3 = 2$.

In this example, we've jumped ahead without using the formal terminology or precisely defining the "especially nice form" of the final matrix. We remedy this immediately.

DEFINITION **1.20.** *(Reduced row-echelon form)*
A matrix is in reduced row-echelon form *if it is in row echelon form,* and

1. *The leading entry of each row is equal to one, and*
2. *The leading entry of each row is the only non-zero entry in its column.*

EXAMPLE **1.21.** *(Reduced row-echelon form) The following matrices are both in reduced row echelon form:*

$$\begin{bmatrix} 1 & 0 & 0 & 2 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 3 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \textit{and} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

□

EXAMPLE **1.22.** *(Not in reduced row-echelon form) The following matrices are NOT in reduced row echelon form:*

$$\begin{bmatrix} 1 & 0 & 0 & 2 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 3 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix} \quad \textit{and} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

□

A simple modification to the algorithm for Gaussian elimination yields an algorithm for reducing a matrix to reduced row echelon form; this algorithm is known as *Gauss-Jordan elimination*. It is presented below, with the differences between Gaussian elimination and Gauss-Jordan elimination highlighted in boldface.

DEFINITION **1.23.** *(Gauss-Jordan elimination — in words)*
Let A be an $m \times n$ matrix. At each stage in the algorithm, a particular position in the matrix, called the pivot position, is being processed. Initially the pivot position is at the top-left of the matrix. What happens at each stage depends on whether the pivot entry (that is, the number in the pivot position) is zero or not.

1. *If the pivot entry is zero then, if possible, interchange the pivot row with one of the rows below it, in order to ensure that the pivot entry is non-zero. This will be possible* unless *the pivot entry and every entry below it are zero, in which case simply move the pivot position one column to the right.*
2. *If the pivot entry is non-zero,* **multiply the pivot row to ensure that the pivot entry is 1** *and then, by adding a suitable multiple of the pivot row to every row* **above and** *below the pivot row, ensure that every entry* **above and** *below the pivot entry is zero. Then move the pivot position one column to the right and one row down.*

When the pivot position is moved off the matrix, then the process finishes and the matrix will be in **reduced** *row echelon form.*

This method has the advantage that now the solutions can simply be read off the augmented matrix.

KEY CONCEPT **1.24.** *(Solving systems of linear equations, advanced method)*
To solve a system of linear equations of the form $A\boldsymbol{x} = \boldsymbol{b}$, perform the following steps:

1. *Form the augmented matrix $[A \mid \boldsymbol{b}]$.*
2. *Use Gauss-Jordan elimination to put the augmented matrix into reduced row-echelon form.*
3. *Identify the leading entries (which are all equal to 1) to identify the basic variables; the other variables will be free parameters.*
4. *Read from each row of the reduced row-echelon form matrix what each basic variable is equal to as a combination of the free variables and constants.*

We will now apply this method to an example to illustrate the differences in the method and show how easy it is to get the solution from the reduced row-echelon form. Compare with Example 1.13.

EXAMPLE **1.25.** *(Gauss-Jordan elimination) Consider the system corresponding to the following augmented matrix, with the initial pivot position marked:*

$$\left[\begin{array}{cccc|c} \boxed{2} & 1 & 2 & 4 & -2 \\ 2 & 1 & 1 & 0 & 1 \\ 4 & 3 & 2 & 4 & 3 \end{array}\right]$$

The initial pivot position is the $(1,1)$ position in the matrix, and the pivot entry is therefore 2. Our first step is to multiply the first row by $1/2$ so that the pivot entry is 1.

$$\left[\begin{array}{cccc|c} \boxed{1} & 1/2 & 1 & 2 & -1 \\ 2 & 1 & 1 & 0 & 1 \\ 4 & 3 & 2 & 4 & 3 \end{array}\right] \quad R_1 \leftarrow \frac{1}{2}R_1$$

Pivoting on the $(1,1)$-entry is then accomplished by performing the two elementary operations $R_2 \leftarrow R_2 - 2R_1$ and $R_3 \leftarrow R_3 - 4R_1$, leaving the matrix:

$$\left[\begin{array}{cccc|c} 1 & 1/2 & 1 & 2 & -1 \\ 0 & \boxed{0} & -1 & -4 & 3 \\ 0 & 1 & -2 & -4 & 7 \end{array}\right] \quad \begin{array}{l} \\ R_2 \leftarrow R_2 - 2R_1 \\ R_3 \leftarrow R_3 - 4R_1 \end{array}$$

The new pivot entry is 0*, but as the entry immediately under the pivot position is non-zero, interchanging the two rows moves a non-zero to the pivot position.*

$$\left[\begin{array}{cccc|c} 1 & 1/2 & 1 & 2 & -1 \\ 0 & \boxed{1} & -2 & -4 & 7 \\ 0 & 0 & -1 & -4 & 3 \end{array}\right] \begin{array}{l} \\ R_2 \leftrightarrow R_3 \\ R_3 \leftrightarrow R_2 \end{array}$$

This pivot is already equal to 1 *so the next step is to pivot on this entry in order to zero out all the entries* **above and** *below it and then move the pivot position.*

$$\left[\begin{array}{cccc|c} 1 & 0 & 2 & 4 & -9/2 \\ 0 & 1 & -2 & -4 & 7 \\ 0 & 0 & \boxed{-1} & -4 & 3 \end{array}\right] \quad R_1 \leftarrow R_1 - \frac{1}{2}R_2$$

Now me multiply Row 3 *by* -1 *to make the pivot equal to* 1*.*

$$\left[\begin{array}{cccc|c} 1 & 0 & 2 & 4 & -9/2 \\ 0 & 1 & -2 & -4 & 7 \\ 0 & 0 & \boxed{1} & 4 & -3 \end{array}\right] \quad R_3 \leftarrow -R_3$$

Finally we pivot off that entry to get zeros in all other entries in that column.

$$\left[\begin{array}{cccc|c} 1 & 0 & 0 & -4 & 3/2 \\ 0 & 1 & 0 & 4 & 1 \\ 0 & 0 & 1 & 4 & -3 \end{array}\right] \begin{array}{l} R_1 \leftarrow R_1 - 2R_3 \\ R_2 \leftarrow R_2 + 2R_3 \end{array}$$

This matrix is now in reduced row-echelon form. The leading entries are exactly the positions we used as pivots:

$$\left[\begin{array}{cccc|c} \boxed{1} & 0 & 0 & -4 & 3/2 \\ 0 & \boxed{1} & 0 & 4 & 1 \\ 0 & 0 & \boxed{1} & 4 & -3 \end{array}\right]$$

since no leading entry is to the right of the augmenting bar, this system is consistent. Moreover, we see that the basic variables are x_1*,* x_2*, and* x_3*, and there is one free parameter:* x_4*.*

The first row, written as an equation, is $x_1 - 4x_4 = 3/2$*, thus we immediately get* $x_1 = 4x_4 + 3/2$*. From the second row and third row, we immediately get* $x_2 = -4x_4 + 1$ *and* $x_3 = -4x_4 - 3$*, respectively. Therefore the final solution set for this system of linear equations is*

$$S = \{(4x_4 + 3/2, -4x_4 + 1, -4x_4 - 3, x_4) \,|\, x_4 \in \mathbb{R}\}$$

□

As you can see, there are more steps with matrices, but then no back-substitution is required at all.

REMARK **1.26.** *As you saw in the very first step of the example, having a pivot entry not equal to* 1 *or* -1 *introduces fractions, which can be annoying. If there is an entry in that column which is a* 1 *or* -1*, interchanging the two rows before applying Gauss-Jordan elimination allows us to avoid introducing fractions, and so makes calculations easier.*

1.6 Reasoning about systems of linear equations

Understanding the *process* of Gaussian elimination and back-substitution and Gauss-Jordan elimination also allows us to *reason about* systems of linear equations, even if they are not explicitly defined, and make general statements about the number of solutions to systems of linear equations. One of the most important is the following result, which says that any consistent system of linear equations with more unknowns than equations has infinitely many solutions.

THEOREM **1.27.** *Suppose that $A\boldsymbol{x} = \boldsymbol{b}$ is a system of m linear equations in n variables. If $m < n$, then the system is either inconsistent or has infinitely many solutions.*

Proof. Consider the row-echelon form of the augmented matrix $[A \mid \boldsymbol{b}]$. If the last column (on the right of the augmenting bar) contains the leading entry of some row, then the system is inconsistent. Otherwise, each leading entry is in a column corresponding to a variable, and so there are exactly m basic variables. As there are n variables altogether, this leaves $n - m > 0$ free parameters in the solution set and so there are infinitely many solutions. □

A *homogeneous* system of linear equations is one of the form $A\boldsymbol{x} = \boldsymbol{0}$ and these systems are *always consistent*.[9] Thus Theorem 1.27 has the important corollary that *"every homogeneous system of linear equations with more unknowns than equations has infinitely many solutions"*.

[9] Why is this true?

A second example of reasoning *about* a system of linear equations rather than just solving an explicit system is when the system is not fully determined. For example, suppose that a and b are unknown values. What can be said about the number of solutions of the following system of linear equations?

$$\begin{bmatrix} 1 & 2 & a \\ 0 & 1 & 2 \\ 1 & 3 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -3 \\ b \\ 0 \end{bmatrix}$$

In particular, for which values of a and b will this system have 0, 1 or infinitely many solutions?

To answer this, start performing Gaussian elimination as usual, treating a and b symbolically as their values are not known.[10] The row reduction proceeds in the following steps: the initial augmented matrix is

$$\left[\begin{array}{ccc|c} 1 & 2 & a & -3 \\ 0 & 1 & 2 & b \\ 1 & 3 & 3 & 0 \end{array}\right]$$

and so after pivoting on the top-left position we get

$$\left[\begin{array}{ccc|c} 1 & 2 & a & -3 \\ 0 & 1 & 2 & b \\ 0 & 1 & 3-a & 3 \end{array}\right] \quad R_3 \leftarrow R_3 - R_1$$

[10] Of course, it is necessary to make sure that you never compute anything that *might* be undefined, such as $1/a$. If you need to use $1/a$ during the Gaussian elimination, then you need to separate out the cases $a = 0$ and $a \neq 0$ and do them separately.

and then

$$\left[\begin{array}{ccc|c} 1 & 2 & a & -3 \\ 0 & 1 & 2 & b \\ 0 & 0 & 1-a & 3-b \end{array}\right] \quad R_3 \leftarrow R_3 - R_2$$

From this matrix, we can immediately see that if $a \neq 1$ then $1 - a \neq 0$ and every variable is basic, which means that the system has a unique solution (regardless of the value of b). On the other hand, if $a = 1$ then *either* $b \neq 3$ in which case the system is inconsistent, or $b = 3$ in which case there are infinitely many solutions. We can summarise this outcome:

$a \neq 1$	Unique solution
$a = 1$ and $b \neq 3$	No solution
$a = 1$ and $b = 3$	Infinitely many solutions

2
Vector spaces and subspaces

THIS CHAPTER takes the first steps away from the *geometric* interpretation of vectors in familiar 2− or 3−dimensional space by introducing n−dimensional vectors and the vector space $\mathbb{R}^n$, which must necessarily be described and manipulated *algebraically*.

BEFORE COMMENCING this chapter, students should be able to:

- Solve systems of linear equations.

AFTER COMPLETING this chapter, students will be able to:

- Determine when a set of vectors is a subspace, and
- Determine when a set of vectors is linearly independent, and
- Find a basis for a subspace, and hence determine its dimension.

2.1 The vector space $\mathbb{R}^n$

The *vector space* $\mathbb{R}^n$ consists of all the n−tuples of real numbers, which henceforth we call *vectors*; formally we say that

$$\mathbb{R}^n = \{(x_1, x_2, \ldots, x_n) \mid x_1, x_2, \ldots, x_n \in \mathbb{R}\}.$$

Thus $\mathbb{R}^2$ is just the familiar collection of *pairs* of real numbers that we usually visualise by identifying each pair (x, y) with the point (x, y) on the Cartesian plane, and $\mathbb{R}^3$ the collection of *triples* of real numbers that we usually identify with 3−space.

A vector $\boldsymbol{u} = (u_1, \ldots, u_n) \in \mathbb{R}^n$ may have different meanings:

- when $n = 2$ or 3 it could represent a geometric vector in $\mathbb{R}^n$ that has both a magnitude and a direction;
- when $n = 2$ or 3 it could represent the coordinates of a point in the Cartesian plane or in 3−space;
- it could represent certain quantities, eg u_1 apples, u_2 pears, u_3 oranges, u_4 bananas, ...
- it may simply represent a string of real numbers.

The vector space $\mathbb{R}^n$ also has two *operations* that can be performed on vectors, namely *vector addition* and *scalar multiplication*. Although their definitions are intuitively obvious, we give them anyway:

DEFINITION **2.1.** *(Vector addition)*
If $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)$ *and* $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)$ *are vectors in* $\mathbb{R}^n$ *then their* sum $\boldsymbol{u} + \boldsymbol{v}$ *is defined by*

$$\boldsymbol{u} + \boldsymbol{v} = (u_1 + v_1, u_2 + v_2, \ldots, u_n + v_n).$$

In other words, two vectors are added coordinate-by-coordinate.

DEFINITION **2.2.** *(Scalar multiplication)*
If $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)$ *and* $\alpha \in \mathbb{R}$ *then the* product $\alpha\boldsymbol{v}$ *is defined by*

$$\alpha\boldsymbol{v} = (\alpha v_1, \alpha v_2, \ldots, \alpha v_n)\,.$$

In other words, each coordinate of the vector is multiplied by the scalar.

EXAMPLE **2.3.** *Here are some vector operations:*

$$\begin{aligned}(1,2,-1,3) + (4,0,1,2) &= (5,2,0,5)\\ (3,1,2) + (6,-1,-4) &= (9,0,-2)\\ 5(1,0,-1,2) &= (5,0,-5,10)\end{aligned}$$

□

Row or column vectors?

A vector in $\mathbb{R}^n$ is simply an ordered $n-$tuple of real numbers and for many purposes all that matters is that we write it down in such a way that it is clear which is the first coordinate, the second coordinate and so on.

However a vector can also be viewed as a *matrix*, which is very useful when we use *matrix algebra* (the subject of Chapter 3) to manipulate equations involving vectors, and then a choice has to be made whether to use a $1 \times n$ matrix, i.e. a matrix with one row and n columns or an $n \times 1$ matrix, i.e. a matrix with n rows and 1 column to represent the vector.

Thus a vector in $\mathbb{R}^4$ can be represented either as a *row vector* such as

$$(1,2,3,4)$$

or as a *column vector* such as

$$\begin{bmatrix}1\\2\\3\\4\end{bmatrix}.$$

For various reasons, mostly to do with the conventional notation we use for functions (that is, we usually write $f(x)$ rather than $(x)f$), it is more convenient mathematically to assume that vectors are represented as *column vectors* most of the time. Unfortunately, in *writing about mathematics*, trying to typeset a row vector such as $[1,2,3,4]$ is much more convenient than typesetting a column vector such as $\begin{bmatrix}1\\2\\3\\4\end{bmatrix}$ which, as you can see, leads to ugly and difficult to read paragraphs.

Some authors try to be very formal and use the notation for a matrix transpose (see Chapter 3) to allow them to elegantly typeset a column vector: so their text would read something like: Let $v = (1,2,3,4)^T$ in which case everyone is clear that the vector v is really a column vector.

In practice however, either the distinction between a row- and column-vector is not important (e..g adding two vectors together) or it is obvious from the context; in either case there is never any actual confusion caused by the difference. So to reduce the notational overload of adding a slew of transpose symbols that are almost never needed, in these notes we've decided to *write* all vectors just as rows, but with the understanding that when it matters (in matrix equations), they are really to be viewed as column vectors. In this latter case, it will always be obvious from the context that the vectors *must* be column vectors anyway!

For instance when we write $A\boldsymbol{x} = \boldsymbol{b}$ for a system of linear equaions in Chapter 1, the vectors $\boldsymbol{x}$ and $\boldsymbol{b}$ are column vectors.

One vector plays a special role in linear algebra; the vector in $\mathbb{R}^n$ with all components equal to zero is called the *zero-vector* and denoted

$$\mathbf{0} = (0,0,\ldots,0).$$

It has the obvious properties that $v + \mathbf{0} = \mathbf{0} + v = v$; this means that it is an *additive identity*[1].

[1] This is just formal mathematical terminology for saying that you can add it to any vector without altering that vector.

2.2 *Subspaces*

In the study of $2-$ and $3-$ dimensional geometry, figures such as *lines* and *planes* play a particularly important role and occur in many different contexts. In higher-dimensional and more general vector spaces, a similar role is played by a *vector subspace* or just *subspace* which is a set of vectors that has three special additional properties.

Definition 2.4. *(Vector subspace)*
Let $S \subseteq \mathbb{R}^n$ be a set of vectors. Then S is called a subspace *of $\mathbb{R}^n$ if*

(S1) $\mathbf{0} \in S$, *and*

(S2) $\boldsymbol{u} + \boldsymbol{v} \in S$ *for all vectors* $\boldsymbol{u}, \boldsymbol{v} \in S$, *and*

(S3) $\alpha \boldsymbol{v} \in S$ *for all scalars* $\alpha \in \mathbb{R}$ *and vectors* $\boldsymbol{v} \in S$.

First we'll go through these three conditions in turn and see what they are saying. The first condition (S1) simply says that a subspace must contain the zero vector $\mathbf{0}$; when this condition does *not* hold, it is an easy way to show that a given set of vectors is *not* a subspace.

EXAMPLE **2.5.** *The set of vectors* $S = \{(x, y) \mid x + y = 1\}$ *is not a subspace of* $\mathbb{R}^2$ *because the vector* $\mathbf{0} = (0, 0)$ *does not belong to* S. □

The second condition[2] (S2) says that in order to be a subspace, a set S of vectors must be *closed under vector addition*. This means that if two vectors *that are both in* S are added together, then their sum must remain in S.

[2] Condition (S2) does not restrict what happens to the sum of two vectors that are *not* in S, or the sum of a vector in S and one not in S. It is *only* concerned with the sum of two vectors that are *both* in S.

EXAMPLE **2.6.** *(Closed under vector addition)*[3] *In* $\mathbb{R}^3$, *the xy-plane is the set of all vectors of the form* $(x, y, 0)$ *(where x and y can be anything). The xy-plane is* closed under vector addition *because if we add* any *two vectors in the xy-plane together, then the resulting vector also lies in the xy-plane.* □

[3] In the next section, we'll see how to present a formal proof that a set of vectors is closed under vector addition, but for these examples, geometric intuition is enough to see that what is being claimed is true.

EXAMPLE **2.7.** *(Not closed under vector addition) In* $\mathbb{R}^2$, *the* unit disk *is the set of vectors*

$$\{(x, y) \mid x^2 + y^2 \leq 1\}.$$

This set of vectors is not *closed under vector addition because if we take* $\boldsymbol{u} = (1, 0)$ *and* $\boldsymbol{v} = (0, 1)$, *then both* $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *are in the unit disk, but their sum* $\boldsymbol{u} + \boldsymbol{v} = (1, 1)$ *is* not *in the unit disk.* □

The third condition (S3) says that in order to qualify as a subspace, a set S of vectors must be *closed under scalar multiplication*, meaning that if a vector is contained in S, then *all of its scalar multiples* must also be contained in S.

EXAMPLE **2.8.** *(Closed under scalar multiplication) In* $\mathbb{R}^2$, *the set of vectors on the two axes, namely*

$$S = \{(x, y) \mid xy = 0\}$$

is closed under scalar multiplication, because it is clear that any multiple of a vector on the x-axis remains on the x-axis, and any multiple of a vector on the y-axis remains on the y-axis. Algebraically: if (x, y) *satisfies* $xy = 0$ *then* $(\alpha x, \alpha y)$ *satisfies* $\alpha x . \alpha y = 0$. □

EXAMPLE **2.9.** *(Not closed under scalar multiplication) In* $\mathbb{R}^2$, *the* unit disk, *which was defined in Example 2.7, is* not closed *under scalar multiplication because if we take* $\boldsymbol{u} = (1, 0)$ *and* $\alpha = 2$, *then* $\alpha \boldsymbol{u} = (2, 0)$ *which is not in the unit disk.* □

One of the fundamental skills needed in linear algebra is the ability to identify whether a given set of vectors in $\mathbb{R}^n$ forms a subspace or not. Usually a *set* of vectors will be described in some way, and you will need to be able to tell whether this set of vectors is a subspace. To prove that a given set of vectors *is* a subspace, it is necessary to show that *all three conditions* (S1), (S2) and (S3) are satisfied, while to show that a set of vectors *is not* a subspace, it is only necessary to show that *one of the three* conditions is not satisfied.

2.2.1 Subspace proofs

In this subsection, we consider in more detail how to show whether or not a given set of vectors is a subspace or not. It is *much easier* to show that a set of vectors is *not* a subspace than to show that a set of vectors *is* a subspace. The reason for this is that condition (S2) apply to *every pair* of vectors in the given set. To show that this condition fails, we only need to give a single example where the condition does not hold, but to show that they are true, we need to find a general argument that applies to every pair of vectors. The same concept applies to condition (S3). This asymmetry is so important that we give it a name[4].

[4] The "black swan" name comes from the famous notion that in order to prove or disprove the logical statement "all swans are white", it would only be necessary to find *one single* black swan to *disprove* it, but it would be necessary to check *every possible* swan in order to *prove* it. Subspaces are the same — if a set of vectors is not a subspace, then it is only necessary to find one "black swan" showing that one of the conditions does not hold, but if it is a subspace then it is necessary to "check every swan" by proving that the conditions hold for every pair of vectors and scalars.

KEY CONCEPT **2.10.** *(The "Black Swan" concept)*

1. *To show that a set of vectors S is* not closed *under vector addition, it is sufficient to find a single* explicit example *of two vectors $\boldsymbol{u}, \boldsymbol{v}$ that are contained in S, but whose sum $\boldsymbol{u}+\boldsymbol{v}$ is not contained in S.*

   However, to show that a set of vectors S is closed *under vector addition, it is necessary to give a* formal symbolic proof *that applies to* every pair *of vectors in S.*

2. *To show that a set of vectors S is* not closed *under scalar multiplication, it is sufficient to find a single* explicit example *of one vector $\boldsymbol{v}$ contained in S and one scalar α such that $\alpha\boldsymbol{v}$ is not contained in S.*

   However, to show that a set of vectors S is closed *under scalar multiplication, it is necessary to give a* formal symbolic proof *that applies to* every pair *of one vector in S and one scalar.*

EXAMPLE **2.11.** *(Not a subspace) The set $S = \{(w,x,y,z) \mid wx = yz\}$ in $\mathbb{R}^4$ is* not a subspace *because if $\boldsymbol{u} = (1,0,2,0)$ and $\boldsymbol{v} = (0,1,0,2)$, then $\boldsymbol{u}, \boldsymbol{v} \in S$ but $\boldsymbol{u}+\boldsymbol{v} = (1,1,2,2) \notin S$ so condition (S2) does not hold.* □

Examples 2.7 and 2.9 also use this black swan concept for (S2) and (S3) respectively.

One of the hardest techniques for first-time students of linear algebra is to understand how to structure a proof that a set of vectors

is a subspace, so we'll go slowly. Let

$$S = \{(x, y, z) \mid x - y = 2z\}$$

be a set of vectors in $\mathbb{R}^3$. We wish to check whether or not it is a subspace. Here is a model proof, interleaved with some discussion about the proof.[5]

[5] When you do your proofs it may help to structure them like this model proof, but don't include the discussion — this is to help you understand why the model proof looks like it does, but it is not part of the proof itself.

(S1) It is obvious that

$$0 - 0 = 2(0)$$

and so $\mathbf{0} \in S$.

Discussion: *To check that* $\mathbf{0}$ *is in* S*, it is necessary to verify that the zero vector satisfies the "defining condition" that determines* S*. In this case, the defining condition is that the difference of the first two coordinates (that is,* $x - y$*) is equal to twice the third coordinate (that is,* $2z$*). For the vector* $\mathbf{0} = (0, 0, 0)$ *we have all coordinates equal to* 0*, and so the condition is true.*

(S2) Let $\boldsymbol{u} = (u_1, u_2, u_3) \in S$ and $\boldsymbol{v} = (v_1, v_2, v_3) \in S$. Then

$$u_1 - u_2 = 2u_3 \tag{2.1}$$
$$v_1 - v_2 = 2v_3. \tag{2.2}$$

Discussion: *To prove that* S *is closed under addition, we need to check every possible pair of vectors, which can only be done symbolically. We give symbolic names* $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *to two vectors in* S *and write down the only facts that we currently know — namely that they satisfy the defining condition for* S*. These equations are given labels — in this case, Equations* (2.1) *and* (2.2)*, because the proof must refer to these equations later.*

Now consider the sum

$$\boldsymbol{u} + \boldsymbol{v} = (u_1 + v_1, u_2 + v_2, u_3 + v_3)$$

and test it for membership in S. As

$$\begin{aligned}
(u_1 + v_1) - (u_2 + v_2) &= u_1 + v_1 - u_2 - v_2 && \text{(rearranging)} \\
&= (u_1 - u_2) + (v_1 - v_2) && \text{(rearranging)} \\
&= 2u_3 + 2v_3 && \text{(by Eqs. (2.1) and (2.2))} \\
&= 2(u_3 + v_3) && \text{(rearranging terms)}
\end{aligned}$$

it follows that $\boldsymbol{u} + \boldsymbol{v} \in S$.

Discussion: *To show that* $\boldsymbol{u} + \boldsymbol{v}$ *is in* S*, we need to show that the difference of its first two coordinates is equal to twice its third coordinate. So the sequence of calculations starts with the difference of the first two coordinates and then carefully manipulates this expression in order to show that it is equal to twice the third coordinate. Every stage of the manipulation is justified either just as a rearrangement of the terms or by reference to some previously known fact. At some stage in the manipulation, the proof must use Equations* (2.1) *and* (2.2)*, because the result must depend on the two original vectors being vectors in* S*.*

(S3) Let $\boldsymbol{u} = (u_1, u_2, u_3) \in S$ and $\alpha \in \mathbb{R}$. Then

$$u_1 - u_2 = 2u_3. \tag{2.3}$$

DISCUSSION: *To prove that S is closed under scalar multiplication, we need to check every vector in S and scalar in $\mathbb{R}$. We give the symbolic name $\boldsymbol{u}$ to the vector in S and α to the scalar, and note down the only fact that we currently know — namely that $\boldsymbol{u}$ satisfies the defining condition for S. We'll need this fact later, and so give it a name, in this case Equation* (2.3).

Now consider the vector

$$\alpha\boldsymbol{u} = (\alpha u_1, \alpha u_2, \alpha u_3)$$

and test it for membership in S. As

$$\begin{aligned} \alpha u_1 - \alpha u_2 &= \alpha(u_1 - u_2) && \text{(rearranging)} \\ &= \alpha(2u_3) && \text{(by Equation (2.3))} \\ &= 2(\alpha u_3) && \text{(rearranging)} \end{aligned}$$

it follows that $\alpha\boldsymbol{u} \in S$.

DISCUSSION: *To show that $\alpha\boldsymbol{u}$ is in S, we need to show that the difference of its first two coordinates is equal to twice its third coordinate. So the sequence of calculations starts with the difference of the first two coordinates and then carefully manipulates it in order to show that it is equal to twice the third coordinate. At some stage in the manipulation, the proof must use the equations Equation* (2.3) *because the result must depend on the original vector being a member of S.*

It will take quite a bit of practice to be able to write this sort of proof correctly, so do not get discouraged if you find it difficult at first. Here are some examples to try out.

EXAMPLE **2.12.** *These sets of vectors* are *subspaces:*

1. *The set of vectors $\{(w, x, y, z) \mid w + x + y + z = 0\}$ in $\mathbb{R}^4$.*
2. *The xy-plane in $\mathbb{R}^3$.*
3. *The line $x = y$ in $\mathbb{R}^2$.*
4. *The set of vectors $\{(x_1, x_2, \ldots, x_n) \mid x_1 + x_2 + \ldots + x_{n-1} = x_n\}$ in $\mathbb{R}^n$.*
5. *The set consisting of the unique vector $\mathbf{0}$.*

These sets of vectors are not *subspaces:*

1. *The set of vectors $\{(w, x, y, z) \mid w + x + y + z = 1\}$ in $\mathbb{R}^4$.*
2. *The plane normal to $\boldsymbol{n} = (1, 2, 1)$ passing through the point $(1, 1, 1)$.*
3. *The line $x = y - 1$ in $\mathbb{R}^2$.*
4. *The set of vectors $\{(x_1, x_2, \ldots, x_n) \mid x_1 + x_2 + \ldots + x_{n-1} \geq x_n\}$ in $\mathbb{R}^n$ for $n \geq 2$.*

□

2.2.2 *Exercises*

1. Show that a line in $\mathbb{R}^2$ is a subspace if and only if it passes through the origin $(0,0)$.
2. Find a set of vectors in $\mathbb{R}^2$ that is closed under vector addition, but not closed under scalar multiplication.
3. Find a set of vectors in $\mathbb{R}^2$ that is closed under scalar multiplication, but not closed under vector addition.

2.3 *Spans and spanning sets*

We start this section by considering a simple question:

> What is the smallest subspace S of $\mathbb{R}^2$ containing the vector $v = (1,2)$?

If S is a subspace, then by condition (S1) of Definition 2.4 it must also contain the zero vector, so it follows that S must contain *at least* the two vectors $\{\mathbf{0}, v\}$. But by condition (S2), the subspace S must also contain the *sum* of any two vectors in S, and so therefore S must also contain all the vectors

$$(2,4), (3,6), (4,8), (5,10), \ldots$$

But then, by condition (S3), it follows that S must also contain all the *multiples* of v, such as

$$(-1,-2), (1/2,1), (1/4,1/2), \ldots$$

Therefore S must contain *at least* the set of vectors[6]

$$\{(\alpha, 2\alpha) \mid \alpha \in \mathbb{R}\}$$

and in fact this set contains enough vectors to satisfy the three conditions (S1), (S2) and (S3), and so it is the smallest subspace containing v.

[6] So if a subspace contains a vector v then it must contain *every scalar multiple* of v. In $\mathbb{R}^2$ and $\mathbb{R}^3$, this means that if a subspace contains a point, then it contains the line containing the origin and that point.

Now let's extend this result by considering the same question but with a bigger starting set of vectors: Suppose that $A = \{v_1, v_2, \ldots, v_k\}$ is a set of vectors in $\mathbb{R}^n$ — what is the smallest *subspace* of $\mathbb{R}^n$ that contains A?

To answer this, we need a couple more definitions:

Definition 2.13. *(Linear combination)*
Let $A = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_k}\}$ be a set of vectors in $\mathbb{R}^n$. Then a linear combination *of the vectors in A is any vector of the form*

$$\boldsymbol{v} = \alpha_1 \boldsymbol{v_1} + \alpha_2 \boldsymbol{v_2} + \cdots + \alpha_k \boldsymbol{v_k}$$

where α_1, α_2, …, $\alpha_k \in \mathbb{R}$ are arbitrary scalars.

By slightly modifying the argument of the last paragraph, it should be clear that if a subspace contains the vectors v_1, v_2, ...,v_k, then it also contains *every linear combination* of those vectors. As we will frequently need to refer to the "set of all possible linear combinations" of a set of vectors, we should give it a name:

Definition 2.14. *(Span)*
The span *of* $A = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_k}\}$ *is the set of* all possible linear combinations *of the vectors in* A*, and is denoted* $span(A)$*. In symbols,*

$$span(A) = \{\alpha_1 \boldsymbol{v_1} + \alpha_2 \boldsymbol{v_2} + \cdots + \alpha_k \boldsymbol{v_k} \mid \alpha_i \in \mathbb{R},\ 1 \leq i \leq k\}.$$

When A is given by a short list of elements, we will sometimes commit a small abuse of notation, writing $\text{span}(v_1, v_2, \ldots, v_k)$ instead of the correct $\text{span}(\{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_k}\})$.

Therefore, if a subspace S contains a *subset* $A \subseteq S$ then it also contains the *span* of A. To answer the original question ("what is the smallest subspace containing A") it is enough to notice that the span of A is always a subspace itself and so no further vectors need to be added. This is sufficiently important to write out formally as a theorem and to give a formal proof.

Theorem 2.15. *(Span of anything is a subspace) Let* $A = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_k}\}$ *be a set of vectors in* $\mathbb{R}^n$*. Then* $span(A)$ *is a subspace of* $\mathbb{R}^n$*, and is the smallest subspace of* $\mathbb{R}^n$ *containing* A*.*

Proof. We must show that the three conditions of Definition 2.4 hold.

(S1) It is clear that
$$\boldsymbol{0} = 0\boldsymbol{v}_1 + 0\boldsymbol{v}_2 + \cdots + 0\boldsymbol{v}_k$$
and so $\boldsymbol{0}$ is a linear combination of the vectors in A and thus $\boldsymbol{0} \in \text{span}(A)$.

(S2) Let $\boldsymbol{u}, \boldsymbol{v} \in \text{span}(A)$. Then there are scalars α_i, β_i $(1 \leq i \leq k)$ such that
$$\boldsymbol{u} = \alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2 + \cdots + \alpha_k \boldsymbol{v}_k \tag{2.4}$$
$$\boldsymbol{v} = \beta_1 \boldsymbol{v}_1 + \beta_2 \boldsymbol{v}_2 + \cdots + \beta_k \boldsymbol{v}_k. \tag{2.5}$$
Now[7] consider the sum $\boldsymbol{u} + \boldsymbol{v}$:
$$\begin{aligned}\boldsymbol{u} + \boldsymbol{v} &= (\alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2 + \cdots + \alpha_k \boldsymbol{v}_k) \\ &\quad + (\beta_1 \boldsymbol{v}_1 + \beta_2 \boldsymbol{v}_2 + \cdots + \beta_k \boldsymbol{v}_k) \\ &\qquad\qquad \text{(by Eqs. (2.4) and (2.5))} \\ &= (\alpha_1 + \beta_1)\,\boldsymbol{v}_1 + (\alpha_2 + \beta_2)\,\boldsymbol{v}_2 + \cdots + (\alpha_k + \beta_k)\,\boldsymbol{v}_k\end{aligned}$$
and so $\boldsymbol{u} + \boldsymbol{v} \in \text{span}(A)$ since $\alpha_i + \beta_i \in \mathbb{R}$ for all i.

[7] This seems like a lot of work just to say something that is almost obvious: if you take two linear combinations of a set of vectors and add them together, then the resulting vector is *also* a linear combination of the original set of vectors!

(S3) Let $\boldsymbol{v} \in \text{span}(A)$ and $\alpha \in \mathbb{R}$. Then there are scalars β_i $(1 \leq i \leq k)$ such that
$$\boldsymbol{v} = \beta_1 \boldsymbol{v}_1 + \beta_2 \boldsymbol{v}_2 + \cdots + \beta_k \boldsymbol{v}_k. \tag{2.6}$$

It is clear that

$$\begin{aligned} \alpha\boldsymbol{v} &= \alpha\left(\beta_1\boldsymbol{v}_1 + \beta_2\boldsymbol{v}_2 + \cdots + \beta_k\boldsymbol{v}_k\right) && \text{(by Equation (2.6))} \\ &= (\alpha\beta_1)\boldsymbol{v}_1 + (\alpha\beta_2)\boldsymbol{v}_2 + \cdots + (\alpha\beta_k)\boldsymbol{v}_k && \text{(rearranging)} \end{aligned}$$

and so $\alpha\boldsymbol{v} \in \text{span}(A)$ since $\alpha\beta_i \in \mathbb{R}$ for all i..

The arguments earlier in this section showed that *any* subspace containing A also contains $\text{span}(A)$ and as $\text{span}(A)$ is a subspace itself, it must be the smallest subspace containing A.

□

REMARK **2.16.** *In $\mathbb{R}^n$, the smallest subspace containing the empty set $\emptyset$ is $\{\mathbf{0}\}$ since any subspace must contain $\mathbf{0}$ and $\{\mathbf{0}\}$ is closed under addition and scalar multiplication. Therefore, by convention, we set*

$$\textit{span}(\emptyset) = \{\mathbf{0}\},$$

so that Theorem 2.15 also holds for A being the empty set.

The span of a set of vectors gives us an easy way to *find* subspaces — start with any old set of vectors, take their span and we get a subspace. If we do have a subspace given in this way, then what can we say about it? Is this a *useful* way to create, or work with, a subspace?

For example, suppose we start with

$$A = \{(1,0,1), (3,2,3)\}$$

as a set of vectors in $\mathbb{R}^3$. Then $\text{span}(A)$ is a subspace of $\mathbb{R}^3$ — what can we say about this subspace? The first thing to notice is that we can *easily check* whether or not a particular vector is contained in $\text{span}(A)$, because it simply involves *solving a system of linear equations*.[8]

[8] This shows that having a subspace of the form $\text{span}(A)$ is a *good* representation of the subspace, because we can easily test membership of the subspace — that is, we "know" which vectors are contained in the subspace.

Continuing our example, to decide whether a vector $\boldsymbol{v}$ is contained in $\text{span}(A)$, we try to *solve* the vector equation

$$\boldsymbol{v} = \lambda_1(1,0,1) + \lambda_2(3,2,3)$$

for the two "unknowns" λ_1, λ_2; this is a system of 3 linear equations in two unknowns and, as discussed in Chapter 1, can easily be solved.

EXAMPLE **2.17.** *(Vector not in span) If $A = \{(1,0,1), (3,2,3)\}$, then $\boldsymbol{v} = (2,4,5)$ is not in $\textit{span}(A)$. This follows because the equation*

$$(2,4,5) = \lambda_1(1,0,1) + \lambda_2(3,2,3)$$

yields the system of linear equations

$$\begin{aligned} \lambda_1 \;+\; 3\lambda_2 &= 2 \\ 2\lambda_2 &= 4 \\ \lambda_1 \;+\; 3\lambda_2 &= 5 \end{aligned}$$

which is obviously inconsistent. *Thus there is* no linear combination *of the vectors in A that is equal to $(2,4,5)$.*

EXAMPLE **2.18.** *(Vector in span) If* $A = \{(1,0,1),(3,2,3)\}$*, then* $\boldsymbol{v} = (5,-2,5)$ is in *span*(A)*. This follows because the equation*

$$(5,-2,5) = \lambda_1(1,0,1) + \lambda_2(3,2,3)$$

yields the system of linear equations

$$\begin{aligned} \lambda_1 + 3\lambda_2 &= 5 \\ 2\lambda_2 &= -2 \\ \lambda_1 + 3\lambda_2 &= 5 \end{aligned}$$

which has the unique solution $\lambda_2 = -1$ *and* $\lambda_1 = 8$*. Therefore* $\boldsymbol{v} \in$ *span*(A) *because we have now found the particular linear combination required.*

Continuing with $A = \{(1,0,1),(3,2,3)\}$, is there another description of the subspace span(A)? It is easy to see that every vector in span(A) must have its first and third coordinates equal, and by trying a few examples, it seems likely that *every* vector with first and third coordinates equal is in span(A). To prove this, we would need to demonstrate that a suitable linear combination of the two vectors can be found for any such vector. In other words, we need to show that the equation

$$(x,y,x) = \lambda_1(1,0,1) + \lambda_2(3,2,3)$$

has a solution *for all values* of x and y. Fortunately this system of linear equations can easily be solved symbolically with the result that the system is always consistent with solution

$$\lambda_1 = x - \frac{3}{2}y \qquad \lambda_2 = \frac{1}{2}y.$$

Therefore we have the fact that

$$\text{span}((1,0,1),(3,2,3)) = \{(x,y,x) \mid x,y \in \mathbb{R}\}$$

2.3.1 *Spanning sets*

So far in this section, we have *started* with a small collection of vectors (that is, the set A), and then *built* a subspace (that is, span(A)) from that set of vectors.

Now we consider the situation where we start with an arbitrary subspace V and try to find a set — hopefully a small set — of vectors A such that $V = \text{span}(A)$. This concept is sufficiently important to warrant a formal definition:

Why do we want to find a spanning set of vectors for a subspace? The answer is that a spanning set is an *effective way* of *describing* a subspace. Every subspace has a spanning set, and so it is also a *universal* way of describing a subspace. Basically, once you know a spanning set for a subspace, you can easily calculate everything about that subspace.

DEFINITION **2.19.** *(Spanning set)*
Let $V \subseteq \mathbb{R}^n$ *be a subspace. Then a set* $A = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_k}\}$ *of vectors, each contained in* V*, is called a* spanning set *for* V *if*

$$V = \textit{span}(A).$$

EXAMPLE **2.20.** *(Spanning set) If* $V = \{(x, y, 0) \mid x, y \in \mathbb{R}\}$*, then* V *is a subspace of* $\mathbb{R}^3$*. The set* $A = \{(1, 0, 0), (0, 1, 0)\}$ *is a* spanning set *for* V*, because every vector in* V *is a linear combination of the vectors in* A*, and every linear combination of the vectors in* A *is in* V*. There are other spanning sets for* V *— for example, the set* $\{(1, 1, 0), (1, -1, 0)\}$ *is another spanning set for* V*.*

EXAMPLE **2.21.** *(Spanning set) If* $V = \{(w, x, y, z) \mid w + x + y + z = 0\}$*, then* V *is a subspace of* $\mathbb{R}^4$*. The set*

$$A = \{(1, -1, 0, 0), (1, 0, -1, 0), (1, 0, 0, -1)\}$$

is a spanning set *for* V *because every vector in* V *is a linear combination of the vectors in* A*, and every linear combination of the vectors in* A *is in* V*. There are many other spanning sets for* V *— for example, the set*

$$\{(2, -1, -1, 0), (1, 0, -1, 0), (1, 0, 0, -1)\}$$

is a different spanning set for the same subspace.

One critical point that often causes difficulty for students beginning linear algebra is understanding the difference between "span" and "spanning set"; the similarity in the phrases seems to cause confusion. To help overcome this, we emphasise the difference.[9]

[9] Another way to think of it is that a "spanning set" is like a list of LEGO® shapes that you can use to build a model while the "span" is the completed model (the subspace). Finding a spanning set for a subspace is like starting with the completed model and asking "What shapes do I need to build this model?".

KEY CONCEPT **2.22.** *(Difference between span and spanning set) To remember the difference between span and spanning set, make sure you understand that:*

- *The* span *of a set* A *of vectors is the* entire subspace *that can be "built" from the vectors in* A *by taking linear combinations in all possible ways.*
- A spanning set *of a subspace* V *is a* set of vectors *that are needed in order to "build"* V*.*

In the previous examples, the spanning sets were just "given" with no explanation of how they were found, and no proof that they were the correct spanning sets. In order to show that a particular set A actually *is* a spanning set for a subspace V, it is necessary to check two things:

1. Check that the vectors in A are actually contained in V — this guarantees that $\text{span}(A) \subseteq V$.
2. Check that *every vector* in V can be made as a linear combination of the vectors in A — this shows that $\text{span}(A) = V$.

The first of these steps is easy and, by now, you will not be surprised to discover that the second step can be accomplished by *solving a system of linear equations*[10].

[10] In fact, almost *everything* in linear algebra ultimately involves nothing more than solving a system of linear equations!

EXAMPLE **2.23.** *(Spanning set with proof) We show that the set* $A = \{(1,1,-1),(2,1,1)\}$ *is a spanning set for the subspace*

$$V = \{(x,y,z) \mid z = 2x - 3y\} \subset \mathbb{R}^3.$$

First notice that both $(1,1,-1)$ *and* $(2,1,1)$ *satisfy the condition that* $z = 2x - 3y$ *and so are actually in* V*. Now we need to show that* every *vector in* V *is a linear combination of these two vectors. Any vector in* V *has the form* $(x,y,2x-3y)$ *and so we need to show that the vector equation*

$$(x,y,2x-3y) = \lambda_1(1,1,-1) + \lambda_2(2,1,1)$$

in the two unknowns λ_1 *and* λ_2 *is consistent, regardless of the values of* x *and* y*. Writing this out as a system of linear equations we get*

$$\begin{aligned} \lambda_1 \;+\; 2\lambda_2 &= x \\ \lambda_1 \;+\; \lambda_2 &= y \\ -\lambda_1 \;+\; \lambda_2 &= 2x - 3y. \end{aligned}$$

Solving this system of linear equations using the techniques of the previous chapter shows that this system always has a unique solution*, namely*

$$\lambda_1 = 2y - x \qquad \lambda_2 = x - y.$$

Hence every vector in V *can be expressed as a linear combination of the two vectors, namely*

$$(x,y,2x-3y) = (2y-x)(1,1,-1) + (x-y)(2,1,1).$$

This shows that these two vectors are a spanning set for V*.*

Actually *finding* a spanning set for a subspace is not so difficult, because it can just be built up vector-by-vector. Suppose that a subspace V is given in some form (perhaps by a formula) and you need to find a spanning set for V. Start by just picking any non-zero vector $\boldsymbol{v_1} \in V$, and examine $\text{span}(\boldsymbol{v_1})$ — if this is equal to V, then you have finished, otherwise there are some vectors in V that cannot yet be built just from $\boldsymbol{v_1}$. Choose one of these "unreachable" vectors, say $\boldsymbol{v_2}$, add it to the set you are creating, and then examine $\text{span}(\boldsymbol{v_1}, \boldsymbol{v_2})$ to see if *this* is equal to V. If it is, then you are finished and otherwise there is another "unreachable" vector, which you call $\boldsymbol{v_3}$ and add to the set, and so on. After some finite number of steps (say k steps), this process will eventually terminate[11] when there are no more unreachable vectors in V in which case

$$V = \text{span}(\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k)$$

and you have found a spanning set for V.

[11] The reason that this process must terminate (i.e. not go on for ever) will become clear over the next few sections.

EXAMPLE **2.24.** *(Finding Spanning Set) Let*

$$V = \{(x,y,z) \mid z = 2x - 3y\}$$

which is a subspace of $\mathbb{R}^3$*. To find a spanning set for* V*, we start by choosing any non-zero vector that lies in* V*, say* $\boldsymbol{v_1} = (1,0,2)$*. It is clear that*

span$(\boldsymbol{v_1})$ *is strictly smaller than* V*, because every vector in span*$(\boldsymbol{v_1})$ *has a zero second coordinate, whereas there are vectors in* V *that do* not *have this property. So we choose any one of these — say,* $\boldsymbol{v_2} = (0, 1, -3)$ *— and now consider span*$(\boldsymbol{v_1}, \boldsymbol{v_2})$*, which we can now prove is actually equal to* V*. Thus, a suitable spanning set for* V *is the set*

$$A = \{(1, 0, 2), (0, 1, -3)\}.$$

2.4 *Linear independence*

In the last section, we learned how a subspace can always be described by giving a *spanning set* for that subspace. There are many spanning sets for any given subspace, and in this section we consider when a spanning set is *efficient* — in the sense that it is as *small* as it can be. For example, here are three different spanning sets for the xy-plane in $\mathbb{R}^3$ (remember that the xy-plane is the set of vectors $\{(x, y, 0) \mid x, y \in \mathbb{R}\}$.

$$A_1 = \{(1, 0, 0), (0, 1, 0)\}$$
$$A_2 = \{(1, 1, 0), (1, -1, 0), (1, 3, 0)\}$$
$$A_3 = \{(2, 2, 0), (1, 2, 0)\}.$$

Which of these is the "best" spanning set to use? There is perhaps nothing much to choose between A_1 and A_3, because each of them contain two vectors[12], but it is clear that A_2 is *unnecessarily big* — if we throw out any of the three vectors in A_2, then the remaining two vectors *still span the same subspace*. On the other hand, both of the spanning sets A_1 and A_3 are *minimal spanning sets* for the xy-plane in that if we discard any of the vectors, then the remaining set no longer spans the whole xy-plane.

[12] But maybe A_1 looks "more natural" because the vectors have such a simple form; later we will see that for many subspaces there is a natural spanning set, although this is not always the case.

The *reason* that A_2 is not a smallest-possible spanning set for the xy-plane is that the third vector is *redundant* — it is *already* a linear combination of the first two vectors: $(1, 3, 0) = 2(1, 1, 0) - (1, -1, 0)$ and therefore any vector in span(A_2) can be produced as a linear combination only of the first two vectors. More precisely any linear combination

$$\alpha_1(1, 1, 0) + \alpha_2(1, -1, 0) + \alpha_3(1, 3, 0)$$

of all three of the vectors in A_2 can be rewritten as

$$\alpha_1(1, 1, 0) + \alpha_2(1, -1, 0) + \alpha_3\left(2(1, 1, 0) - (1, -1, 0)\right)$$

which is equal to

$$(\alpha_1 + 2\alpha_3)(1, 1, 0) + (\alpha_2 - \alpha_3)(1, -1, 0)$$

which is just a linear combination of the first two vectors with altered scalars.

Therefore a spanning set for a subspace is an *efficient* way to represent a subspace if none of the vectors in the spanning set is a linear combination of the other vectors. While this condition is easy

to state, it is hard to work with directly, and so we use a condition that *means* exactly the same thing but is easier to use.

Definition 2.25. *(Linear independence)*
Let $A = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_k}\}$ be a set of vectors in $\mathbb{R}^n$. Then A is called linearly independent *(or just* independent*) if the only solution to the vector equation*

$$\lambda_1 \boldsymbol{v_1} + \lambda_2 \boldsymbol{v_2} + \cdots + \lambda_k \boldsymbol{v_k} = \boldsymbol{0}$$

in the unknowns λ_1, λ_2, …, λ_k is the trivial solution *$\lambda_1 = \lambda_2 = \cdots = \lambda_k = 0$.*

Before seeing why this somewhat strange definition means exactly the same as having no one of the vectors being a linear combination of the others, we'll see a couple of examples. The astute reader will not be surprised to learn that testing a set of vectors for linear independence involves solving a system of linear equations.

Example 2.26. *(Independent set) In order to decide whether the set of vectors $A = \{(1,1,2,2), (1,0,-1,2), (2,1,3,1)\}$ in $\mathbb{R}^4$ is linearly independent, we need to solve the vector equation*

$$\lambda_1(1,1,2,2) + \lambda_2(1,0,-1,2) + \lambda_3(2,1,3,1) = (0,0,0,0).$$

This definitely has at least one solution, namely the trivial solution $\lambda_1 = 0$, $\lambda_2 = 0$, $\lambda_3 = 0$, and so the only question is whether it has more *solutions. The vector equation is equivalent to the system of linear equations*

$$\begin{aligned} \lambda_1 + \lambda_2 + 2\lambda_3 &= 0 \\ \lambda_1 + \lambda_3 &= 0 \\ 2\lambda_1 - \lambda_2 + 3\lambda_3 &= 0 \\ 2\lambda_1 + 2\lambda_2 + \lambda_3 &= 0 \end{aligned}$$

which can easily be shown, by the techniques of Chapter 1 to have a unique solution.

There is an asymmetry here similar to the asymmetry in subspace proofs. To show that a set of vectors is dependent only requires *one* non-trivial linear combination, whereas to show that a set of vectors is independent it is necessary in principle to show that *every* non-trivial linear combination of the vectors is non-zero. Of course in practice this is done by solving the relevant system of linear equations and showing that it has a unique solution, which must therefore be the trivial solution.

A set of vectors that is *not* linearly independent is called *dependent*. To show that a set of vectors is dependent, it is only necessary to find an explicit non-trivial linear combination of the vectors equal to $\boldsymbol{0}$.

Example 2.27. *(Dependent set) Is the set*

$$A = \{(1,3,-1), (2,1,2), (4,7,0)\}$$

in $\mathbb{R}^3$ linearly independent? To decide this, set up the vector equation

$$\lambda_1(1,3,-1) + \lambda_2(2,1,2) + \lambda_3(4,7,0) = (0,0,0)$$

and check how many solutions it has. This is equivalent to the system of linear equations

$$\begin{aligned} \lambda_1 + 2\lambda_2 + 4\lambda_3 &= 0 \\ 3\lambda_1 + \lambda_2 + 7\lambda_3 &= 0 \\ -\lambda_1 + 2\lambda_2 &= 0 \end{aligned}$$

After Gaussian elimination, the augmented matrix for this system of linear equations is

$$\left[\begin{array}{ccc|c} 1 & 2 & 4 & 0 \\ 0 & -5 & -5 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

and so has infinitely many *solutions, because λ_3 is a free parameter. While this is already enough to prove that A is dependent, it is always useful to find an explicit solution which can then be used to double-check the conclusion. As λ_3 is free, we can find a solution by putting $\lambda_3 = 1$, in which case the second row gives $\lambda_2 = -1$ and the first row $\lambda_1 = -2$. And indeed we can check that*

$$-2(1,3,-1) - 1(2,1,2) + 1(4,7,0) = (0,0,0)$$

as required to prove dependence. □

Sometimes a non-trivial linear combination is easy to see, so that there is no need to try to solve a system.

Example **2.28.** *(Dependent set) Is the set*

$$A = \{(1,0,0), (0,1,0), (4,7,0)\}$$

in $\mathbb{R}^3$ linearly independent? We immediately see that

$$-4(1,0,0) - 7(0,1,0) + 1(4,7,0) = (0,0,0)$$

and this proves that A is a dependent set. □

Here is a surprising example:

Example **2.29.** *($\{\mathbf{0}\}$ is dependent) Consider the set $A = \{\mathbf{0}\}$ in $\mathbb{R}^n$. It is a bit counter-intuitive that a single vector can be dependent but the vector equation $\alpha_1\mathbf{0} = \mathbf{0}$ has a non-trivial solution $\alpha_1 = 1$, which proves that A is a dependent set.* □

Now we'll give a rigorous proof of the earlier claim that the definition of linear independence (Definition 2.25) is just a way of saying that none of the vectors is a linear combination of the others.

Theorem **2.30.** *Let $A = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_k}\}$ be a set of vectors in $\mathbb{R}^n$. Then A is linearly independent if and only if none of the vectors in A are a linear combination of the others.*

Proof. We will actually prove the *contrapositive*[13] statement, namely that A is linearly dependent if and only if one of the vectors is a linear combination of the others. First suppose that one of the vectors, say $\boldsymbol{v_i}$, is a linear combination of the others: then there are scalars α_1, α_2, …, α_{i-1}, α_{i+1}, …, α_k such that

$$\boldsymbol{v_i} = \alpha_1\boldsymbol{v_1} + \alpha_2\boldsymbol{v_2} + \cdots + \alpha_{i-1}\boldsymbol{v_{i-1}} + \alpha_{i+1}\boldsymbol{v_{i+1}} + \cdots + \alpha_k\boldsymbol{v_k}$$

and so there is a non-trivial linear combination of the vectors of A equal to $\mathbf{0}$, namely:

$$\alpha_1\boldsymbol{v_1} + \alpha_2\boldsymbol{v_2} + \cdots + \alpha_{i-1}\boldsymbol{v_{i-1}} - 1\boldsymbol{v_i} + \alpha_{i+1}\boldsymbol{v_{i+1}} + \cdots + \alpha_k\boldsymbol{v_k} = \mathbf{0}.$$

[13] Given the statement "A implies B", recall that the contrapositive statement is "not B implies not A". If a statement is true then so is its contrapositive, and vice versa.

(This linear combination is not all-zero because the coefficient of $\boldsymbol{v_i}$ is equal to -1 which is definitely non-zero.)

Next suppose that the set A is linearly dependent. Then there is some non-trivial linear combination of the vectors equal to $\mathbf{0}$:

$$\alpha_1\boldsymbol{v_1} + \alpha_2\boldsymbol{v_2} + \cdots + \alpha_k\boldsymbol{v_k} = \mathbf{0}.$$

Because this linear combination is non-trivial, not *all* of the coefficients are equal to 0, and so we can pick one of them, say α_i, that is non-zero. But then

$$\alpha_i\boldsymbol{v_i} = -\alpha_1\boldsymbol{v_1} - \alpha_2\boldsymbol{v_2} - \cdots - \alpha_{i-1}\boldsymbol{v_{i-1}} - \alpha_{i+1}\boldsymbol{v_{i+1}} - \cdots - \alpha_k\boldsymbol{v_k}$$

and so *because* $\alpha_i \neq 0$ we can divide by α_i and get

$$\boldsymbol{v_i} = -\frac{\alpha_1}{\alpha_i}\boldsymbol{v_1} - \frac{\alpha_2}{\alpha_i}\boldsymbol{v_2} - \cdots - \frac{\alpha_{i-1}}{\alpha_i}\boldsymbol{v_{i-1}} - \frac{\alpha_{i+1}}{\alpha_i}\boldsymbol{v_{i+1}} - \cdots - \frac{\alpha_k}{\alpha_i}\boldsymbol{v_k}$$

so one of the vectors is a linear combination of the others. □

Remark **2.31.** *If A is the empty set $\emptyset$, we can consider that none of the vectors in A are a linear combination of the others (which is a bit of a vacuous condition). For this reason, we will consider that $\emptyset$ is an independent set.*

Remark **2.32.** *The condition*
"none of the vectors in A are a linear combination of the others"
can also be written formally as:
"for all vector $\boldsymbol{v}$ in A, $\boldsymbol{v} \notin span(A \setminus \{\boldsymbol{v}\})$[14]".

Using this condition and remembering that $span(\emptyset) = \{\mathbf{0}\}$, we see that Theorem 2.30 also applies to the set $A = \{\mathbf{0}\}$ (which is dependent, see Example 2.29).

[14] The set notation $A \setminus B$ means the set of vectors that are in A but not in B. So $A \setminus \{\boldsymbol{v}\}$ consists of all the vectors in A except for $\boldsymbol{v}$.

There are two key facts about dependency that are intuitively clear, but useful enough to state formally:

1. If A is a linearly independent set of vectors in $\mathbb{R}^n$, then any *subset* of A is also linearly independent.
2. If B is a linearly *dependent* set of vectors in $\mathbb{R}^n$ then any *superset* of B is also linearly dependent.

In other words, you can remove vectors from an independent set and it remains independent and you can add vectors to a dependent set and it remains dependent. In particular, any set containing the vector $\mathbf{0}$ is dependent.

2.5 Bases

In the last few sections, we have learned that giving a *spanning set* for a subspace is an *effective* way of describing a subspace and that a spanning set is *efficient* if it is linearly independent. Therefore an excellent way to describe or transmit, for example by computer, a subspace is to give a *linearly independent spanning set* for the subspace. This concept is so important that it has a special name:

Definition 2.33. *(Basis)*
Let V be a subspace of $\mathbb{R}^n$. Then a basis *for V is a* linearly independent spanning set *for V. In other words, a basis is a set of vectors $A \subset \mathbb{R}^n$ such that*

- *$V = span(A)$, and*
- *A is linearly independent.*

Example 2.34. *(Basis) The set $A = \{(1,0,0),(0,1,0)\}$ is a* basis *for the xy-plane in $\mathbb{R}^3$, because it is a linearly independent set of vectors and any vector in the xy-plane can be expressed as a linear combination of the vectors of A.*

Example 2.35. *(Basis proof) Let V be the subspace of $\mathbb{R}^3$ defined by $V = \{(x,y,z) \mid x - y + 2z = 0\}$. Then we shall show that $A = \{(2,0,-1),(1,1,0)\}$ is a basis for V. We check three separate things: that the vectors are actually* in *V, that they are linearly independent, and that they are a spanning set for V.*

1. *Check that both vectors are actually in V.*

   This is true because

   $$2 - (0) + 2(-1) = 0$$
   $$1 - 1 + 2(0) = 0.$$

2. *Check that A is linearly independent.*

   Theorem 2.30 shows that two vectors are linearly dependent only if one of them is a multiple of the other. As this is not the case here, we conclude that the set is linearly independent.

3. *Check that A is a spanning set for V*

   Every vector in V can be expressed as a linear combination of the two vectors in A, because all the vectors in V are of the form $\{(x,y,(y-x)/2) \mid x,y \in \mathbb{R}\}$ and using the techniques from Chapter 1 we see that

   $$(x,y,(y-x)/2) = \frac{x-y}{2}(2,0,-1) + y(1,1,0).$$

   Therefore A is a basis for V. □

Note here that we can also write V in the form $\{(y-2z,y,z) \mid y,z \in \mathbb{R}\}$ and then see that $(y-2z,y,z) = -z(2,0,-1) + y(1,1,0)$ to prove that A is a spanning set for V. This makes the computations slightly easier.

A subspace of $\mathbb{R}^n$ can have more than one basis — in fact, a subspace usually has *infinitely many* different *bases*.[15]

[15] The word "bases" is the plural of "basis."

Example 2.36. *(Two different bases) Let V be the subspace of $\mathbb{R}^3$ defined by $V = \{(x,y,z) \mid x - y + 2z = 0\}$. Then*

$$A = \{(2,0,-1),(1,1,0)\}$$
$$B = \{(1,3,1),(3,1,-1)\}$$

are both bases for V. Proving this is left as an exercise. □

The vector space $\mathbb{R}^3$ is itself a subspace and so has a basis. In this case, there is one basis that stands out as being particularly natural. It is called the *standard basis* and contains the three vectors

$$\boldsymbol{e_1} = (1,0,0), \quad \boldsymbol{e_2} = (0,1,0), \quad \boldsymbol{e_3} = (0,0,1)$$

where the *standard basis vectors* are given the special names $\boldsymbol{e_1}$, $\boldsymbol{e_2}$ and $\boldsymbol{e_3}$.[16] More generally, the vector space $\mathbb{R}^n$ has a basis consisting of the n vectors $\{\boldsymbol{e_1}, \boldsymbol{e_2}, \ldots, \boldsymbol{e_n}\}$ where the i-th basis vector $\boldsymbol{e_i}$ is all-zero except for a single 1 in the i-th position.

[16] In Engineering, the standard basis vectors for $\mathbb{R}^3$ are also known as $\boldsymbol{i}$, $\boldsymbol{j}$ and $\boldsymbol{k}$ respectively.

Finding a basis from scratch is straightforward, because the technique described before Example 2.24 (and illustrated in the example) for finding spanning sets by adding vectors one-by-one to an independent set will automatically find a linearly independent spanning set — in other words, a basis. In fact, the same argument shows that you can start with *any* linearly independent set and augment it vector-by-vector to obtain a basis containing the original linearly independent set of vectors[17].

[17] We still have not yet shown that this process will actually terminate, but will do so in the next section.

Another approach to finding a basis of a subspace is to start with a *spanning set* that is linearly dependent and to *remove* vectors from it one-by-one. If the set is linearly dependent then one of the vectors is a linear combination of the others, and so it can be removed from the set without altering the span of the set of vectors. This process can be repeated until the remaining vectors are linearly independent in which case they form a basis.

EXAMPLE **2.37.** *(Basis from a spanning set) Let*

$$A = \{(1,1,-2), (-2,-2,4), (-1,-2,3), (5,-5,0)\}$$

be a set of vectors in $\mathbb{R}^3$*, and let* $V = span(A)$*. What is a basis for* V *contained in* A*? We start by testing whether* A *is linearly independent by solving the system of linear equations*

$$\lambda_1(1,1,-2) + \lambda_2(-2,-2,4) + \lambda_3(-1,-2,3) + \lambda_4(5,-5,0) = (0,0,0)$$

to see if it has any non-trivial solutions. If so, then one of the vectors can be expressed as a linear combination of the others and discarded. In this case, we discover that $(-2,-2,4) = -2(1,1,-2)$ *and so we can throw out* $(-2,-2,4)$*. Now we are left with*

$$\{(1,1,-2), (-1,-2,3), (5,-5,0)\}$$

and test whether this *set is linearly independent. By solving*

$$\lambda_1(1,1,-2) + \lambda_2(-1,-2,3) + \lambda_3(5,-5,0) = (0,0,0)$$

we discover that $(5,-5,0) = 15(1,1,-2) + 10(-1,-2,3)$ *and so we can discard* $(5,-5,0)$*. Finally the remaining two vectors are linearly independent and so the set*

$$\{(1,1,-2), (-1,-2,3)\}$$

is a basis for V*.* □

2.5.1 Dimension

As previously mentioned, a subspace of $\mathbb{R}^n$ will usually have infinitely many bases. However, these bases will all share one feature — every basis for a subspace V contains *the same number* of vectors. This fact is not at all obvious, and so we will give a proof for it. Actually we will prove a slightly more technical result that has the result about bases, along with a number of other useful results, as simple consequences.[18] While this is a result of fundamental importance, it uses some fiddly notation with lots of subscripts, so do not feel alarmed if you do not understand it first time through.

[18] A result that is a straightforward consequence of a theorem is called a "corollary".

THEOREM **2.38.** *Let $A = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_k}\}$ be a linearly independent set of vectors. Then any set of $\ell > k$ vectors contained in $V = span(A)$ is* dependent.

Proof. Let $\{\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_\ell}\}$ be a set of $\ell > k$ vectors in V. Then each of these can be expressed as a linear combination of the vectors in A, and so there are scalars $\alpha_{ij} \in \mathbb{R}$, where $1 \leq i \leq \ell$ and $1 \leq j \leq k$ such that:

$$\begin{aligned}
\boldsymbol{w}_1 &= \alpha_{11}\boldsymbol{v_1} + \alpha_{12}\boldsymbol{v_2} + \cdots + \alpha_{1k}\boldsymbol{v_k} \\
\boldsymbol{w}_2 &= \alpha_{21}\boldsymbol{v_1} + \alpha_{22}\boldsymbol{v_2} + \cdots + \alpha_{2k}\boldsymbol{v_k} \\
&\vdots \\
\boldsymbol{w}_\ell &= \alpha_{\ell 1}\boldsymbol{v_1} + \alpha_{\ell 2}\boldsymbol{v_2} + \cdots + \alpha_{\ell k}\boldsymbol{v_k}.
\end{aligned}$$

Now consider what happens when we test $\{\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_\ell}\}$ for linear dependence. We try to solve the system of linear equations

$$\beta_1\boldsymbol{w_1} + \beta_2\boldsymbol{w_2} + \cdots + \beta_\ell\boldsymbol{w_\ell} = \boldsymbol{0} \tag{2.7}$$

and determine if there are any non-trivial solutions to this system. By replacing each $\boldsymbol{w_i}$ in Equation (2.7) with the corresponding expression as a linear combination of the vectors in A, we get a huge equation:

$$\begin{aligned}
\boldsymbol{0} = {} & \beta_1(\alpha_{11}\boldsymbol{v_1} + \alpha_{12}\boldsymbol{v_2} + \cdots + \alpha_{1k}\boldsymbol{v_k}) \\
& + \beta_2(\alpha_{21}\boldsymbol{v_1} + \alpha_{22}\boldsymbol{v_2} + \cdots + \alpha_{2k}\boldsymbol{v_k}) \\
& + \cdots \\
& + \beta_\ell(\alpha_{\ell 1}\boldsymbol{v_1} + \alpha_{\ell 2}\boldsymbol{v_2} + \cdots + \alpha_{\ell k}\boldsymbol{v_k}).
\end{aligned} \tag{2.8}$$

However, this is a linear combination of the vectors in A that is equal to the zero vector. Because A is a linearly independent set of vectors, this happens if and only if the coefficients of the vectors $\boldsymbol{v_i}$ in Equation (2.8) are all zero. In other words, the scalars $\{\beta_1, \beta_2, \ldots, \beta_\ell\}$ must satisfy the following system of linear equa-

tions:

$$\begin{aligned}
\alpha_{11}\beta_1 + \alpha_{21}\beta_2 + \cdots + \alpha_{\ell 1}\beta_\ell &= 0 \\
\alpha_{12}\beta_1 + \alpha_{22}\beta_2 + \cdots + \alpha_{\ell 2}\beta_\ell &= 0 \\
\vdots \qquad \vdots \qquad \vdots & \\
\alpha_{1k}\beta_1 + \alpha_{2k}\beta_2 + \cdots + \alpha_{\ell k}\beta_\ell &= 0.
\end{aligned}$$

This is a homogeneous[19] system of linear equations, and so it is consistent. As we discussed in Chapter 1, Theorem 1.27 has the important corollary that every homogeneous system of linear equations with more unknowns than equations has infinitely many solutions. Hence there is at least one non-trivial choice of scalars $\{\beta_1, \beta_2, \ldots, \beta_\ell\}$ satisfying Equation (2.7) (and indeed there are infinitely many such choices), thereby showing that $\{\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_\ell}\}$ is linearly dependent. □

[19] The constant term in each equation is zero.

Corollary **2.39.** *Every basis for a subspace V of $\mathbb{R}^n$ contains the* same number *of vectors.*

Proof. Suppose that $A = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_k}\}$ and $B = \{\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_\ell}\}$ are two bases for V. Then both A and B are linearly independent sets of vectors and both have the same span. By Theorem 2.38, if we had $\ell > k$ then B would be dependent, thus $\ell \leq k$. Now swapping the roles of A and B (and hence swapping ℓ and k), Theorem 2.38 also implies that $k \leq \ell$. Therefore $\ell = k$. □

Definition **2.40.** *(Dimension)*
The dimension *of the subspace V, denoted by* $\dim(V)$*, is the number of vectors in a basis for V. This definition is not ambiguous, thanks to Corollary 2.39.*

Example **2.41.** *(Dimension of $\mathbb{R}^n$) The standard basis for $\mathbb{R}^2$ contains two vectors, the standard basis for $\mathbb{R}^3$ contains three vectors and the standard basis for $\mathbb{R}^n$ contains n vectors, so we conclude that the* dimension *of $\mathbb{R}^n$ is equal to n.*

Example **2.42.** *(Dimension of a line) A* line through the origin *in $\mathbb{R}^n$ is a subspace consisting of all the multiples of a given non-zero vector:*

$$L = \{\lambda \boldsymbol{v} \mid \lambda \in \mathbb{R}\}.$$

The set $\{\boldsymbol{v}\}$ containing the single vector $\boldsymbol{v}$ is a basis for L and so a line is 1-dimensional.

Exercise **2.5.1.** *What is the dimension of the subspace $\{\mathbf{0}\}$ of $\mathbb{R}^n$?*

This shows that the formal algebraic definition of dimension coincides with our intuitive geometric understanding of the word dimension, which is reassuring.[20]

[20] Of course, we would expect this to be the case, because the algebraic notion of "dimension" was developed as an extension of the familiar geometric concept.

Another important corollary of Theorem 2.38 is that we are finally in a position to show that the process of finding a basis by extending [21] a linearly independent set will definitely finish.

[21] "extending" just means "adding vectors to".

COROLLARY **2.43.** *If V is a subspace of $\mathbb{R}^n$, then any linearly independent set A of vectors in V is contained in a basis for V.*

Proof. If $\text{span}(A) \neq V$, then adding a vector in $V \backslash \text{span}(A)$ to A creates a strictly larger independent set of vectors. As no set of $n+1$ vectors in $\mathbb{R}^n$ is linearly independent, this process must terminate in at most n steps, and when it terminates, the set is a basis for V. □

The next result seems intuitively obvious, because it just says that dimension behaves as you would expect, in that a subspace can only contain other subspaces if they have lower dimension.

THEOREM **2.44.** *Suppose that S, T are subspaces of $\mathbb{R}^n$ and that $S \subsetneq T$. Then $\dim(S) < \dim(T)$.*

Proof. Let B_S be a basis for S. Then B_S is a linearly independent set of vectors contained in T, and so it can be extended to a basis for T. As $\text{span}(B_S) \neq T$ the basis for T is strictly larger than the basis for S and so $\dim(S) < \dim(T)$. □

Corollary 2.39 has a number of important consequences. If A is a set of vectors contained in a subspace V, then normally there is no particular relationship between the properties "A is linearly independent" and "A is a spanning set for V" in that A can have none, either or both of these properties. However if A has the *right size* to be a basis, then it must either have none or both of the properties.

COROLLARY **2.45.** *Let V be a k-dimensional subspace of $\mathbb{R}^n$. Then*

1. *Any linearly independent set of k vectors of V is a basis for V.*
2. *Any spanning set of k vectors of V is a basis for V.*

Proof. 1. Let A be a linearly independent set of k vectors in V. By Corollary 2.43, A can be extended to a basis but by Corollary 2.39 this basis contains k vectors and so no vectors can be added to A. Therefore A is already a basis for V.

2. Let B be a spanning set of k vectors. Suppose B is not a basis, that is, B is linearly dependent. As explained above Example 2.37, we can obtain a basis by removing vectors from B. But then this basis would have less than k vectors, contradicting Corollary 2.39. Therefore B is linearly independent.

□

EXAMPLE **2.46.** *(Dimension of a specific plane) The subspace $V = \{(x,y,z) \mid x+y+z = 0\}$ is a plane through the origin in $\mathbb{R}^3$. What is its dimension? We can quickly find two vectors in V, namely $(1,-1,0)$ and $(1,0,-1)$, and as they are not multiples of each other, the set*

$$\{(1,-1,0),(1,0,-1)\}$$

is linearly independent. As $\mathbb{R}^3$ is 3-dimensional, any proper subspace of $\mathbb{R}^3$ has dimension at most 2 *by Theorem 2.44, and so this set of vectors is a basis and V has dimension 2. Similarly any plane through $\mathbf{0}$ in $\mathbb{R}^3$ has dimension 2.* □

2.5.2 Coordinates

The most important property of a basis for a subspace V is that every vector in V can be expressed as a linear combination of the basis vectors *in exactly one way*; we prove this in the next result:

THEOREM **2.47.** *Let $B = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_k}\}$ be a basis for the subspace V. Then for any vector $\boldsymbol{v} \in V$, there is a* unique *choice of scalars $\alpha_1, \alpha_2, \ldots, \alpha_k$ such that*

$$\boldsymbol{v} = \alpha_1 \boldsymbol{v_1} + \alpha_2 \boldsymbol{v_2} + \cdots + \alpha_k \boldsymbol{v_k}.$$

Proof. As B is a spanning set for V, there is *at least* one way of expressing $\boldsymbol{v}$ as a linear combination of the basis vectors. So we just need to show that there cannot be *two different* linear combinations each equal to $\boldsymbol{v}$. So suppose that there are scalars $\alpha_1, \alpha_2, \ldots, \alpha_k$ and $\beta_1, \beta_2, \ldots, \beta_k$ such that

$$\begin{aligned}\boldsymbol{v} &= \alpha_1 \boldsymbol{v_1} + \alpha_2 \boldsymbol{v_2} + \cdots + \alpha_k \boldsymbol{v_k} \\ \boldsymbol{v} &= \beta_1 \boldsymbol{v_1} + \beta_2 \boldsymbol{v_2} + \cdots + \beta_k \boldsymbol{v_k}.\end{aligned}$$

Subtracting these expressions and rearranging, we discover that

$$\mathbf{0} = (\alpha_1 - \beta_1)\boldsymbol{v_1} + (\alpha_2 - \beta_2)\boldsymbol{v_2} + \cdots + (\alpha_k - \beta_k)\boldsymbol{v_k}.$$

As B is a linearly independent set of vectors, the only linear combination equal to $\mathbf{0}$ is the trivial linear combination with all coefficients equal to 0, and so $\alpha_1 - \beta_1 = 0$, $\alpha_2 - \beta_2 = 0$, …, $\alpha_k - \beta_k = 0$ and so $\alpha_i = \beta_i$ for all i. Therefore the two linear combinations for $\boldsymbol{v}$ are actually the same. □

DEFINITION **2.48.** *(Coordinates)*
Let $B = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_k}\}$ be a basis for the subspace V. If

$$\boldsymbol{v} = \alpha_1 \boldsymbol{v_1} + \alpha_2 \boldsymbol{v_2} + \cdots + \alpha_k \boldsymbol{v_k},$$

we call the scalars $\alpha_1, \ldots, \alpha_k$ the coordinates *of $\boldsymbol{v}$ in the basis B, and we write*

$$(\boldsymbol{v})_B = (\alpha_1, \alpha_2, \ldots, \alpha_k).$$

This often applies to V being the full vector space $\mathbb{R}^n$.

REMARK **2.49.** *Note that when we write a vector $\boldsymbol{v}$ as (x, y, z), this just means that x, y, z are the coordinates in the standard basis $S = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$ since*

$$(x, y, z) = x\boldsymbol{e}_1 + y\boldsymbol{e}_2 + z\boldsymbol{e}_3,$$

and similarly in higher dimensions. If we wish to emphasise that fact, we sometimes write $(\boldsymbol{v})_S = (x, y, z)$.

EXAMPLE **2.50.** *Let* $V = \{(x, y, z) \mid x + y + z = 0\}$ *be a plane through the origin in* $\mathbb{R}^3$*. A basis for this subspace is*

$$B = \{(1, -1, 0), (1, 0, -1)\},$$

as seen in Example 2.46

Therefore any vector $\boldsymbol{v}$ *in* V *can be written in a unique way as a linear combination of the basis vectors, for example,*

$$\boldsymbol{v} = (1, -3, 2) = 3(1, -1, 0) - 2(1, 0, -1),$$

hence

$$(\boldsymbol{v})_B = (1, -3, 2)_B = (3, -2).$$

In general, the task of finding the coordinates of a target vector $\boldsymbol{v}$ with respect to a basis $B = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_k}\}$ is that of solving a system of linear equations:

$$\text{Find } \alpha_1, \alpha_2, \cdots, \alpha_n \quad \text{such that} \quad \boldsymbol{v} = \alpha_1 \boldsymbol{v_1} + \alpha_2 \boldsymbol{v_2} + \cdots + \alpha_k \boldsymbol{v_k}.$$

EXAMPLE **2.51.** *(Example 2.50 continued) Express* $\boldsymbol{w} = (1, 4, -5)$ *in terms of the basis B.*

Solution. *We must solve:* $(1, 4, -5) = \alpha_1(1, -1, 0) + \alpha_2(1, 0, -1)$*, which gives us the system*

$$\begin{array}{rrcr} \alpha_1 + & \alpha_2 & = & 1 \\ -\alpha_1 & & = & 4 \\ & -\alpha_2 & = & -5 \end{array}$$

which has the uniques solution $\alpha_1 = -4$ *and* $\alpha_2 = 5$*.*

So we found the coordinates of $\boldsymbol{w}$*:* $(\boldsymbol{w})_B = (1, 4, -5)_B = (-4, 5)$*. The fact that we have found a solution means that* $\boldsymbol{w}$ *lies in the plane* V*. If it did not then the system of equation for* α_1, α_2 *would be inconsistent.*

It is often useful to change the basis that one is working with.

EXAMPLE **2.52.** *(Example 2.50 continued, different basis) Suppose that instead of the basis B in the above example we chose to use another basis, say*

$$C = \{(0, 1, -1), (1, -2, 1)\}.$$

(As was the case for basis B, it is easy to verify that this is a basis for V). After some calculation we can show that

$$(\boldsymbol{v})_C = (-1, 1) \quad , \quad (\boldsymbol{w})_C = (6, 1).$$

In order for us to exploit a judiciously chosen basis we would like to have a simple way of converting from coordinates in one basis to coordinates in another. For example, given $(\boldsymbol{v})_B$ can we find $(\boldsymbol{v})_C$ without having to work out what $\boldsymbol{v}$ itself is? We will answer this question in Chapter 5.

3
Matrices and determinants

This chapter introduces matrix algebra and explains the fundamental relationships between matrices and their properties, and the various subspaces associated with a matrix.

Before commencing this chapter, students should be able to:

- Solve systems of linear equations,
- Confidently identify and manipulate subspaces, including rapidly determining spanning sets and bases for subspaces, and
- Add and multiply matrices.

After completing this chapter, students will be able to:

- Understand the operations of matrix algebra and identify the similarities and differences between matrix algebra and the algebra of real numbers,
- Describe, and find bases for, the row space, column space and null space of a matrix,
- Find the rank and nullity of a matrix and understand how they are related by the rank-nullity theorem, and
- Compute determinants and understand the relationship between determinants, rank and invertibility of matrices.

An $m \times n$ *matrix* is a rectangular array of numbers with m rows and n columns.

3.1 Matrix algebra

In this section we consider the *algebra of matrices* — that is, the system of mathematical operations such as addition, multiplication, inverses and so on, where the operands[1] are matrices, rather than numbers. In isolation, the basic operations are all familiar from high school — in other words, adding two matrices or multiplying two matrices should be familiar to everyone— but matrix algebra is primarily concerned with the relationships between the operations.

[1] This is mathematical terminology for "the objects being operated on".

3.1.1 Basic operations

The basic operations for matrix algebra are *matrix addition*, *matrix multiplication* , *matrix transposition* and *scalar multiplication*. For completeness, we give the formal definitions of these operations:

DEFINITION **3.1.** *(Matrix operations)*
The basic matrix operations are matrix addition, matrix multiplication, matrix transposition *and* scalar multiplication, *which are defined as follows:*
Matrix addition: *Let $A = (a_{ij})$ and $B = (b_{ij})$ be two $m \times n$ matrices. Then their* sum *$C = A + B$ is the $m \times n$ matrix defined by*

$$c_{ij} = a_{ij} + b_{ij}.$$

$A = (a_{ij})$ is a notation which means that the entry in the i-th row and j-th column (the (i, j)-entry, for short) of the matrix A is the real number a_{ij}.

Matrix multiplication: *Let $A = (a_{ij})$ be an $m \times p$ matrix, and $B = (b_{ij})$ be a $p \times n$ matrix. Then their* product *$C = AB$ is the $m \times n$ matrix defined by*

$$c_{ij} = \sum_{k=1}^{k=p} a_{ik} b_{kj}.$$

Matrix transposition: *Let $A = (a_{ij})$ be an $m \times n$ matrix, Then the* transpose *$C = A^T$ of A is the $n \times m$ matrix defined by*

$$c_{ij} = a_{ji}.$$

Scalar multiplication: *Let $A = (a_{ij})$ be an $m \times n$ matrix, and $\alpha \in \mathbb{R}$ be a scalar. Then the* scalar multiple *$C = \alpha A$ is the $m \times n$ matrix defined by*

$$c_{ij} = \alpha a_{ij}.$$

The properties of these operations are mostly obvious but, again for completeness, we list them all and give them their formal names.

THEOREM **3.2.** *If A, B and C are matrices and α, β are scalars then, whenever the relevant operations are defined, the following properties hold:*

1. $A + B = B + A$ *(matrix addition is* commutative*)*
2. $(A + B) + C = A + (B + C)$ *(matrix addition is* associative*)*
3. $\alpha(A + B) = \alpha A + \alpha B$
4. $(\alpha + \beta)A = \alpha A + \beta A$
5. $(\alpha\beta)A = \alpha(\beta A)$
6. $A(BC) = (AB)C$ *(matrix multiplication is associative)*
7. $(\alpha A)B = \alpha(AB)$ *and* $A(\alpha B) = \alpha(AB)$
8. $A(B + C) = AB + AC$ *(multiplication is* left-distributive *over addition)*
9. $(A + B)C = AC + BC$ *(multiplication is* right-distributive *over addition)*

10. $(A^T)^T = A$
11. $(A+B)^T = A^T + B^T$
12. $(AB)^T = B^T A^T$

Proof. All of these can be proved by elementary algebraic manipulation of the expressions for (i,j)-entry of the matrices on both sides of each equation. We omit the proofs because they are slightly tedious and not very illuminating.[2] □

[2] However it may be worth your while working through one of them, say the proof that matrix multiplication is associative, to convince yourself that you can do it.

Almost all of the properties in Theorem 3.2 are unsurprising and essentially mirror the properties of the algebra of real numbers.[3] Probably the only property in the list that is not immediately "obvious" is Property (12) stating that the transpose of a matrix product is the product of the matrix transposes *in reverse order*:

$$(AB)^T = B^T A^T.$$

[3] A 1×1 matrix can be viewed as essentially identical to a real number, and so this is also not surprising.

This property extends to longer products of matrices, for example we can find the transpose of ABC as follows:[4]

$$(ABC)^T = ((AB)C)^T = C^T(AB)^T = C^T(B^T A^T) = C^T B^T A^T.$$

[4] A cautious reader might — correctly — object that ABC is not a legitimate expression in matrix algebra, because we have only defined matrix multiplication to be a product of *two* matrices. To make this a legal expression in matrix algebra, it really needs to be parenthesised, and so we should use either $A(BC)$ or $(AB)C$. However because matrix multiplication is associative, these evaluate to the same matrix and so, by convention, as it does not matter *which* way the product is parenthesised, we omit the parentheses altogether.

It is a nice test of your ability to structure a *proof by induction* to prove formally that

$$(A_1 A_2 \cdots A_{n-1} A_n)^T = A_n^T A_{n-1}^T \cdots A_2^T A_1^T.$$

However, rather than considering the obvious properties that *are* in the list, it is more instructive to consider the most obvious *omission* from the list; in other words, an important property that matrix algebra *does not* share with the algebra of real numbers. This is the property of commutativity of multiplication because, while the multiplication of real numbers is commutative, it is easy to check by example that matrix multiplication is *not commutative*. For example,

$$\begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 3 & -1 \\ 0 & 4 \end{bmatrix} = \begin{bmatrix} 3 & 3 \\ 6 & 2 \end{bmatrix}$$

but

$$\begin{bmatrix} 3 & -1 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 8 & 4 \end{bmatrix}.$$

If A and B are two specific matrices, then it *might* be the case that $AB = BA$, in which case the two matrices are said to *commute*, but usually it will be the case that $AB \neq BA$. This is a key difference between matrix algebra and the algebra of real numbers.

There are some other key differences worth delving into: in real algebra, the numbers 0 and 1 play special roles, being the *additive identity* and *multiplicative identity* respectively. In other words, for any real number $x \in \mathbb{R}$ we have

$$x + 0 = 0 + x = x \quad \text{and } 1.x = x.1 = x$$

and

$$0.x = x.0 = 0. \tag{3.1}$$

In the algebra of *square* matrices (that is, $n \times n$ matrices for some n) we can analogously find an additive identity and a multiplicative identity. The *additive identity* is the matrix O_n with every entry equal to zero, and it is obvious that for any $n \times n$ matrix A,

$$A + O_n = O_n + A = A.$$

The *multiplicative identity* is the matrix I_n where every entry on the *main diagonal*[5] is equal to one, and every entry off the main diagonal is equal to zero. Then for any $n \times n$ matrix A,

$$AI_n = I_n A = A.$$

When the size of the matrices is unspecified, or irrelevant, we will often drop the subscript and just use O and I respectively. As the terms "additive/multiplicative identity" are rather cumbersome, the matrix O is usually called the *zero matrix* and the matrix I is usually called the *identity matrix* or just the *identity*.

The property Equation (3.1) relating multiplication and zero also holds in matrix algebra, because it is clear that

$$AO = OA = O$$

for any square matrix A. However, there are other important properties of real algebra that are *not* shared by matrix algebra. In particular, in real algebra there are no non-zero *zero divisors*, so that if $xy = 0$ then at least one of x and y is equal to zero. However this is not true for matrices — there are products equal to the zero matrix even if neither matrix is zero. For example,

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

[5] The main diagonal consist of the $(1,1)$, $(2,2)$, …, (n,n) positions. As an example,

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

Definition **3.3.** *(A menagerie of square matrices)*

Suppose that $A = (a_{ij})$ is an $n \times n$ matrix. Then

- *A is the* zero matrix *if $a_{ij} = 0$ for all i, j.*
- *A is the* identity matrix *if $a_{ij} = 1$ if $i = j$ and 0 otherwise.*
- *A is a* symmetric matrix *if $A^T = A$.*
- *A is a* skew-symmetric matrix *if $A^T = -A$.*
- *A is a* diagonal matrix *if $a_{ij} = 0$ for all $i \neq j$.*
- *A is an* upper-triangular matrix *if $a_{ij} = 0$ for all $i > j$.*
- *A is a* lower-triangular matrix *if $a_{ij} = 0$ for all $i < j$.*
- *A is an* idempotent matrix *if $A^2 = A$, where $A^2 = AA$ is the product of A with itself.*
- *A is a* nilpotent matrix *if $A^k = O$ for some k, where $A^k = \underbrace{AA \cdots A}_{k \text{ times}}$.*

EXAMPLE **3.4.** *(Matrices of various types) Consider the following* 3×3 *matrices:*

$$A = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 2 & 3 \\ 0 & 0 & 1 \end{bmatrix} \quad B = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad C = \begin{bmatrix} -1 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 3 \end{bmatrix}.$$

Then A *is an* upper-triangular *matrix,* B *is a* diagonal *matrix and* C *is a* symmetric *matrix.* □

EXAMPLE **3.5.** *(Nilpotent matrix) The matrix*

$$A = \begin{bmatrix} 2 & 4 \\ -1 & -2 \end{bmatrix}$$

is nilpotent *because*

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

□

3.2 *Subspaces from matrices*

There are various vector subspaces associated with a matrix, and there are useful relationships between the properties of the matrices and subspaces. The three principal subspaces associated with a matrix are the *row space*, the *column space* and the *null space* of a matrix. The first two of these are defined analogously to each other, while the third is somewhat different. If A is an $m \times n$ matrix, then each of the *rows* of the matrix can be viewed as a vector in $\mathbb{R}^n$, while each of the *columns* of the matrix can be viewed as a vector in $\mathbb{R}^m$.

3.2.1 *The row space and column space*

DEFINITION **3.6.** *(Row and column space)*
Let A *be an* $m \times n$ *matrix. Then the row space and column space are defined as follows:*
Row space *The* row space *of* A *is the subspace of* $\mathbb{R}^n$ *that is* spanned by the rows *of* A*. In other words, the row space is the subspace consisting of all the linear combinations of the rows of* A*. We denote it by* $\text{rowsp}(A)$.
Column space *The* column space *of* A *is the subspace of* $\mathbb{R}^m$ *that is* spanned by the columns *of* A*. In other words, the column space is the subspace consisting of all the linear combinations of the columns of* A*. We denote it by* $\text{colsp}(A)$.

EXAMPLE **3.7.** *(Row space and column space) Let A be the 3×4 matrix*

$$A = \begin{bmatrix} 1 & 0 & 1 & -1 \\ 2 & -1 & 2 & 0 \\ 1 & 0 & 2 & 1 \end{bmatrix}. \tag{3.2}$$

Then the row space *of A is the subspace of $\mathbb{R}^4$ defined by*

$$\text{rowsp}(A) = \mathit{span}(\{(1,0,1,-1),(2,-1,2,0),(1,0,2,1)\}),$$

while the column space *of A is the subspace of $\mathbb{R}^3$ defined by*

$$\text{colsp}(A) = \mathit{span}(\{(1,2,1),(0,-1,0),(1,2,2),(-1,0,1)\}).$$

(Remember the conventions regarding row and column vectors described in Chapter 2.) □

As described in Chapter 2, it is easy to answer any particular question about a subspace if you know a spanning set for that subspace. In particular, it is easy to determine whether a given vector is in the row or column space of a matrix just by setting up the appropriate system of linear equations.

EXAMPLE **3.8.** *(Vector in row space) Is the vector $(2,1,-1,3)$ in the row space of the matrix A shown in Equation (3.2) above? This question is equivalent to asking whether there are scalars λ_1, λ_2 and λ_3 such that*

$$\lambda_1(1,0,1,-1) + \lambda_2(2,-1,2,0) + \lambda_3(1,0,2,1) = (2,1,-1,3).$$

By considering each of the four coordinates in turn, this corresponds to the following system of four linear equations in the three variables:

$$\begin{array}{rcrcrcr} \lambda_1 & + & 2\lambda_2 & + & \lambda_3 & = & 2 \\ & - & \lambda_2 & & & = & 1 \\ \lambda_1 & + & 2\lambda_2 & + & 2\lambda_3 & = & -1 \\ -\lambda_1 & & & + & \lambda_3 & = & 3. \end{array}$$

The augmented matrix for this system is

$$\left[\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & -1 & 0 & 1 \\ 1 & 2 & 2 & -1 \\ -1 & 0 & 1 & 3 \end{array}\right]$$

which, after row-reduction, becomes

$$\left[\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & -3 \\ 0 & 0 & 0 & 13 \end{array}\right]$$

and so the system is inconsistent *and we conclude that $(2,1,-1,3) \notin \text{rowsp}(A)$. Notice that the part on the left of the augmenting bar in the augmented matrix of the system of linear equations is the* transpose *of the original matrix.* □

EXAMPLE **3.9.** *(Vector in column space) Is the vector* $(1,-1,2)$ *in the column space of the matrix* A *shown in Equation* (3.2) *above? This question is equivalent to asking whether there are scalars* λ_1, λ_2, λ_3 *and* λ_4 *such that*

$$\lambda_1(1,2,1)+\lambda_2(0,-1,0)+\lambda_3(1,2,2)+\lambda_4(-1,0,1)=(1,-1,2).$$

By considering each of the three coordinate positions in turn, this corresponds to a system of three equations in the four variables:

$$\begin{array}{rcrcrcl} \lambda_1 & & & + & \lambda_3 & - & \lambda_4 & = & 1 \\ 2\lambda_1 & - & \lambda_2 & + & 2\lambda_3 & & & = & -1 \\ \lambda_1 & & & + & 2\lambda_3 & + & \lambda_4 & = & 2. \end{array}$$

The augmented matrix for this system is

$$\left[\begin{array}{cccc|c} 1 & 0 & 1 & -1 & 1 \\ 2 & -1 & 2 & 0 & -1 \\ 1 & 0 & 2 & 1 & 2 \end{array}\right]$$

which, after row reduction, becomes

$$\left[\begin{array}{cccc|c} 1 & 0 & 1 & -1 & 1 \\ 0 & -1 & 0 & 2 & -3 \\ 0 & 0 & 1 & 2 & 1 \end{array}\right]$$

and so this system of linear equations has three basic variables, one free parameter and therefore infinitely many solutions. So we conclude that $(1,-1,2) \in \text{colsp}(A)$. *We could, if necessary, or just to check, find a particular solution to this system of equations. For example, if we set the free parameter* $\lambda_4 = 1$ *then the corresponding solution is* $\lambda_1 = 3$, $\lambda_2 = 5$, $\lambda_3 = -1$ *and* $\lambda_4 = 1$ *and we can check that*

$$3(1,2,1)+5(0,-1,0)-(1,2,2)+(-1,0,1)=(1,-1,2).$$

Notice that in this case, the part on the left of the augmenting bar in the augmented matrix of the system of linear equations is just the original matrix itself. □

In the previous two examples, the original question led to systems of linear equations whose coefficient matrix had either the original matrix or the transpose of the original matrix on the left of the augmenting bar.

In addition to being able to identify whether particular vectors are in the row space or column space of a matrix, we would also like to be able to find a a *basis* for the subspace and thereby determine its *dimension*. In Chapter 2 we described a technique where any spanning set for a subspace can be *reduced* to a basis by successively throwing out vectors that are linear combinations of the others. While this technique works perfectly well for determining the dimension of the row space or column space of a matrix, there is an alternative approach based on two simple observations:

1. Performing elementary row operations on a matrix *does not change* its row space.[6]

[6] However, elementary row operations *do change* the *column space*!!

2. The non-zero rows of a matrix in row-echelon form[7] are linearly independent, and therefore form a basis for the row space of that matrix.

[7] Reduced row-echelon form has the same property but will yiled a different basis.

The consequence of these two facts is that it is very easy to find a basis for the row space of a matrix — simply put it into row-echelon form using gauss elimination and then write down the non-zero rows that are found. However the basis that is found by this process will not usually be a subset of the *original* rows of the matrix. If it is necessary to find a basis for the row space whose vectors are all original rows of the matrix, then technique described above can be used.

EXAMPLE **3.10.** *(Basis of row space) Consider the problem of finding a basis for the row space of the* 4×5 *matrix*

$$A = \begin{bmatrix} 1 & 2 & -1 & -1 & 4 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 1 \\ 3 & 0 & 1 & -1 & 6 \end{bmatrix}.$$

After performing the elementary row operations $R_3 \leftarrow R_3 - 1R_1$, $R_4 \leftarrow R_4 - 3R_1$ *then* $R_2 \leftrightarrow R_3$ *and finally* $R_4 \leftarrow R_4 - 2R_2$, *we end up with the following matrix, which we denote* A', *which is in row-echelon form.*

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

The key point is that the elementary row operations have not changed the row space of the matrix in any way and so $\text{rowsp}(A) = \text{rowsp}(A')$. *However it is obvious that the two non-zero rows*

$$\{(1, 2, -1, -1, 4), (0, -3, 2, 1, -3)\}$$

are a basis for the row space of A', *and so they are also a basis for the row space of* A. □

To find a basis for the *column space* of the matrix A, we cannot do elementary *row* operations because they alter the column space. However it is clear that $\text{colsp}(A) = \text{rowsp}(A^T)$, and so just transposing the matrix and then performing the same procedure will find a basis for the column space of A.

EXAMPLE **3.11.** *(Basis of column space) What is a basis for the* column space *of the matrix A of Example 3.10? We first transpose the matrix, getting*

$$A^T = \begin{bmatrix} 1 & 0 & 1 & 3 \\ 2 & 0 & -1 & 0 \\ -1 & 0 & 1 & 1 \\ -1 & 0 & 0 & -1 \\ 4 & 0 & 1 & 6 \end{bmatrix}$$

and then perform Gaussian elimination to obtain the row-echelon matrix

$$\begin{bmatrix} 1 & 0 & 1 & 3 \\ 0 & 0 & -3 & -6 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

whose row space has basis $\{(1,0,1,3),(0,0,-3,-6)\}$. *Therefore these two vectors are a basis for the* column space *of A.* □

What can be said about the *dimension* of the row space and column space of a matrix? In the previous two examples, we found that the row space of the matrix A is a 2-dimensional subspace of $\mathbb{R}^5$, and the column space of A is a 2-dimensional subspace of $\mathbb{R}^4$. In particular, even though they are subspaces of different ambient vector spaces, the *dimensions* of the row space and column space turn out to be equal. This is not an accident, and in fact we have the following surprising result:

THEOREM **3.12.** *Let A be an $m \times n$ matrix. Then the dimension of its row space is equal to the dimension of its column space.*

Proof. Suppose that $\{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_k}\}$ is a basis for the column space of A. Then each column of A can be expressed as a linear combination of these vectors; suppose that the j-th column $\boldsymbol{c_j}$ is given by

$$\boldsymbol{c_j} = \gamma_{1j}\boldsymbol{v_1} + \gamma_{2j}\boldsymbol{v_2} + \cdots \gamma_{kj}\boldsymbol{v_k}.$$

Now form two matrices as follows: B is an $m \times k$ matrix whose columns are the basis vectors $\boldsymbol{v_j}$, while $C = (\gamma_{ij})$ is a $k \times n$ matrix whose j-th column contains the coefficients $\gamma_{1j}, \gamma_{2j}, \ldots, \gamma_{kj}$. It then follows[8] that $A = BC$.

However we can also view the product $A = BC$ as expressing the *rows* of A as a linear combination of the rows of C with the i-th row of B giving the coefficients for the linear combination that determines the i-th row of A. Therefore the rows of C are a spanning set for the row space of A, and so the dimension of the row space of A is at most k. We conclude that

$$\dim(\text{rowsp}(A)) \leq \dim(\text{colsp}(A)).$$

Applying the same argument to A^T we also conclude that

$$\dim(\text{colsp}(A)) \leq \dim(\text{rowsp}(A))$$

and hence these values are equal. □

This number — the common dimension of the row space and column space of a matrix — is an important property of a matrix and has a special name:

[8] You may have to try this out with a few small matrices first to see why this is true. It is not difficult when you see a small situation, but it is not immediately obvious either.

Definition 3.13. *(Matrix Rank)*
The dimension of the row space (and column space) of a matrix is called the rank *of the matrix. If an $n \times n$ matrix has rank n, then the matrix is said to be* full rank.

There is a useful characterisation of the row and column spaces of a matrix that is sufficiently important to state separately.

Theorem 3.14. *If A is an $m \times n$ matrix, then the set of vectors*

$$\{A\boldsymbol{x} \mid \boldsymbol{x} \in \mathbb{R}^n\}$$

Here $\boldsymbol{x}$ is a vector seen as a column vector, that is, as an $(n \times 1)$-matrix.

is equal to the column space of A, while the set of vectors

$$\{\boldsymbol{y}A \mid \boldsymbol{y} \in \mathbb{R}^m\}$$

Here $\boldsymbol{y}$ is a vector seen as a row vector, that is, as an $(1 \times m)$-matrix.

is equal to the row space of A.

Proof. Suppose that $\{\boldsymbol{c_1}, \boldsymbol{c_2}, \ldots, \boldsymbol{c_n}\} \in \mathbb{R}^m$ are the n columns of A. Then if $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, it is easy to see that $A\boldsymbol{x} = x_1\boldsymbol{c}_1 + x_2\boldsymbol{c}_2 + \ldots + x_n\boldsymbol{c}_n$. Therefore every vector of the form $A\boldsymbol{x}$ is a linear combination of the columns of A, and every linear combination of the columns of A can be obtained by multiplying A by a suitable vector. A similar argument applies for the row space of A. □

3.2.2 *The null space*

Suppose that A is an $m \times n$ matrix. Obviously, for $\mathbf{0} \in \mathbb{R}^n$, $A\mathbf{0} = \mathbf{0}$[9]. Moreover if two vectors $\boldsymbol{v_1}$, $\boldsymbol{v_2}$ of $\mathbb{R}^n$ have the property that $A\boldsymbol{v_1} = \mathbf{0}$ and $A\boldsymbol{v_2} = \mathbf{0}$, then simple manipulation shows that

$$A(\boldsymbol{v_1} + \boldsymbol{v_2}) = A\boldsymbol{v_1} + A\boldsymbol{v_2} = \mathbf{0} + \mathbf{0} = \mathbf{0}$$

and for any $\lambda \in \mathbb{R}$,

$$A(\lambda\boldsymbol{v_1}) = \lambda A\boldsymbol{v_1} = \lambda\mathbf{0} = \mathbf{0}.$$

Therefore the set of vectors $\boldsymbol{v}$ with the property that $A\boldsymbol{v} = \mathbf{0}$ contains the zero vector, is closed under vector addition and scalar multiplication, and therefore it satisfies the requirements to be a *subspace* of $\mathbb{R}^n$.

[9] Note this zero vector belongs to $\mathbb{R}^m$

Definition 3.15. *(Null space)*
Let A be an $m \times n$ matrix. The set of vectors

$$\{\boldsymbol{v} \in \mathbb{R}^n \mid A\boldsymbol{v} = \mathbf{0}\}$$

is a subspace of $\mathbb{R}^n$ called the null space *of A and denoted by* $\text{nullsp}(A)$.

EXAMPLE **3.16.** *(Null space) Is the vector* $\boldsymbol{v} = (0, 1, -1, 2)$ *in the null space of the matrix*

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

All that is needed is to check $A\boldsymbol{v}$ *and see what arises. As*

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

it follows that $\boldsymbol{v} \in \text{nullsp}(A)$.

This shows that testing membership of the null space of a matrix is a very easy task. What about finding a *basis* for the null space of a matrix? This turns out[10] to be intimately related to the techniques we used Chapter 1 to solve systems of linear equations.

[10] No surprise here!

So, suppose we wish to find a basis for the nullspace of the matrix

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

from Example 3.16. The matrix equation $A\boldsymbol{x} = \boldsymbol{0}$ yields the following system of linear equations

$$\begin{aligned} x_1 + 2x_2 + 2x_3 &= 0 \\ 3x_1 + 2x_3 + x_4 &= 0 \end{aligned}$$

which has augmented matrix

$$\left[\begin{array}{cccc|c} 1 & 2 & 2 & 0 & 0 \\ 3 & 0 & 2 & 1 & 0 \end{array}\right].$$

Applying the Gauss-Jordan algorithm, we perform the elementary row operations $R_2 \leftarrow R_2 - 3R_1$, $R_2 \leftarrow -\frac{1}{6}R_2$, $R_1 \leftarrow R_1 - 2R_2$:

$$\left[\begin{array}{cccc|c} 1 & 2 & 2 & 0 & 0 \\ 0 & -6 & -4 & 1 & 0 \end{array}\right], \quad \left[\begin{array}{cccc|c} 1 & 2 & 2 & 0 & 0 \\ 0 & 1 & 2/3 & -1/6 & 0 \end{array}\right], \quad \left[\begin{array}{cccc|c} 1 & 0 & 2/3 & 1/3 & 0 \\ 0 & 1 & 2/3 & -1/6 & 0 \end{array}\right].$$

The last matrix is in reduced row echelon form.

Therefore x_3 and x_4 are *free parameters* and we directly get $x_1 = -\frac{1}{3}(2x_3 + x_4)$ and $x_2 = \frac{1}{6}(x_4 - 4x_3)$. Thus, following the techniques of Chapter 1 we can describe the solution set as

$$S = \left\{ \left(-\frac{1}{3}(2x_3 + x_4),\ \frac{1}{6}(x_4 - 4x_3),\ x_3,\ x_4 \right) \middle| \, x_3, x_4 \in \mathbb{R} \right\}.$$

In order to find a *basis* for S notice that we can rewrite the solution as a linear combination of vectors by *separating out the terms* involving x_3 from the terms involving x_4

$$\begin{aligned} &\left(-\frac{1}{3}(2x_3 + x_4),\ \frac{1}{6}(x_4 - 4x_3),\ x_3,\ x_4 \right) \\ &= \left(-\frac{2}{3}x_3,\ -\frac{4}{6}x_3,\ x_3,\ 0 \right) + \left(-\frac{1}{3}x_4,\ \frac{1}{6}x_4,\ 0,\ x_4 \right) \\ &= x_3 \left(-\frac{2}{3},\ -\frac{2}{3},\ 1,\ 0 \right) + x_4 \left(-\frac{1}{3},\ \frac{1}{6},\ 0,\ 1 \right). \end{aligned}$$

Therefore we can express the solution set S as follows:

$$S = \left\{ x_3 \left(-\frac{2}{3},\ -\frac{2}{3},\ 1,\ 0 \right) + x_4 \left(-\frac{1}{3},\ \frac{1}{6},\ 0,\ 1 \right) \middle|\ x_3, x_4 \in \mathbb{R} \right\}.$$

However this immediately tells us that S just consists of *all the linear combinations* of the two vectors $\left(-\frac{2}{3},\ -\frac{2}{3},\ 1,\ 0 \right)$ and $\left(-\frac{1}{3},\ \frac{1}{6},\ 0,\ 1 \right)$ and therefore we have found, almost by accident, a *spanning set* for the subspace S. It is immediate that these two vectors are linearly independent and therefore they form a basis for the null space of A.

Remark **3.17.** *The astute student will notice that after just one elementary row operation, the matrix would be in reduced row echelon form if the fourth column was the second column. Therefore we can stop calculations there and take x_1 and x_4 as the basic parameters. We immediately get*

$$S = \{(-2x_2 - 2x_3,\ x_2, x_3, 6x_2 + 4x_3) |\ x_2, x_3 \in \mathbb{R}\}.$$

Therefore we get that $(-2, 1, 0, 6)$ and $(-2, 0, 1, 4)$ form also a (simpler) basis for the null space of A. The lesson to take from this is that we can sometimes find the nullspace quicker by not following the Gauss-Jordan algorithm blindly, and remembering that we can take other variables as the basic ones than the one given by the leading entries in the (reduced) row echelon form [11].

[11] In this example, you could also perform the operation $R_1 \leftarrow \frac{1}{2} R_1$ then take x_2, x_4 as the basic parameters.

In general, this process will *always* find a basis for the null space of a matrix. If the set of solutions to the system of linear equations has s free parameters, then it can be expressed as a linear combination of s vectors. These s vectors will *always* be linearly independent because in each of the s coordinate positions corresponding to the free parameters, just one of the s vectors will have a non-zero entry.

Definition **3.18.** *(Nullity)*
The dimension of the null space of a matrix A is called the nullity *of A.*

We close this section with one of the most important results in elementary linear algebra, which is universally called the *Rank-Nullity Theorem*, which has a surprisingly simple proof.

Theorem **3.19.** *Suppose that A is an $m \times n$ matrix. Then*

$$\text{rank}(A) + \text{nullity}(A) = n.$$

Proof. Consider the system of linear equations $A\boldsymbol{x} = \mathbf{0}$, which is a system of m equations in n unknowns. This system is solved by applying Gaussian elimination to the augmented matrix $[A \mid \mathbf{0}]$, thereby obtaining the matrix $[A' \mid \mathbf{0}]$ in row-echelon form. The rank of A is equal to the number of non-zero rows of A', which is equal to the number of *basic variables* in the system of linear equations. The nullity of A is the number of free parameters in the solution

set to the system of linear equations and so it is equal to the number of *non-basic variables*. So the rank of A plus the nullity of A is equal to the number of basic variables plus the number of non-basic variables. As each of the n variables is either basic or non-basic, the result follows. □

Given a matrix, it is important to be able to put all the techniques together and to determine the rank, the nullity, a basis for the null space and a basis for the row space of a given matrix. This is demonstrated in the next example.

EXAMPLE **3.20.** *(Rank and nullity) Find the rank, nullity and bases for the row space and null space for the following* 4×4 *matrix:*

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

All of the questions can be answered once the matrix is in reduced row-echelon form, and so the first task is to apply Gauss-Jordan elimination, which will result in the following matrix:

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

The matrix has 3 non-zero rows and so the rank *of A is equal to 3. These non-zero rows form a basis for the rowspace of A and so a basis for the rowspace of A is*

$$\{(1,0,2,0),\ (0,1,-3,0),\ (0,0,0,1)\}.$$

By the Rank-Nullity theorem, we immediately know that the nullity of A is the difference between the number of columns (4) and the rank (3) so is equal to 1. We now determine a basis for the null space.

The null space is the set of solutions to the matrix equation $A\boldsymbol{x} = \mathbf{0}$, and solving this equation by performing Gauss-Jordan elimination on the augmented matrix $[A \mid \mathbf{0}]$ would yield the augmented matrix $[A' \mid \mathbf{0}]$. [12] *So given the augmented matrix*

$$A' = \left[\begin{array}{cccc|c} 1 & 0 & 2 & 0 & 0 \\ 0 & 1 & -3 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

we see that x_1, x_2 and x_4 are basic variables, while the solution space has x_3 as its only free parameter. Expressing the basic variables in terms of the free parameter, we determine that the solution set is

$$S = \{(-2x_3, 3x_3, x_3, 0) \mid x_3 \in \mathbb{R}\}$$

and it is clear that a basis for this is $\{(-2,3,1,0)\}$, which confirms that the nullity of A is equal to 1. □

[12] In other words, the Gauss-Jordan elimination part only needs to be done once, because everything depends only on the form of the matrix A'. However it is important to remember that although it is the same matrix, we are using it in two quite distinct ways. This distinction is often missed by students studying linear algebra for the first time.

3.3 Solving systems of linear equations

We have seen that the set of all solutions to the system of linear equations $A\boldsymbol{x} = \boldsymbol{0}$ is the nullspace of A. What can we say about the set of solutions of

$$A\boldsymbol{x} = \boldsymbol{b} \tag{3.3}$$

when $\boldsymbol{b} \neq \boldsymbol{0}$?

Suppose that we know one solution $\boldsymbol{x}_1$ and that $\boldsymbol{v}$ lies in the nullspace of A. Then

$$A(\boldsymbol{x}_1 + \boldsymbol{v}) = A\boldsymbol{x}_1 + A\boldsymbol{v} = \boldsymbol{b} + \boldsymbol{0} = \boldsymbol{b}.$$

Hence if we are given one solution we can create many more by simply adding elements of the nullspace of A.

Moreover, given any two solutions $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ of Equation (3.3) we have that

$$A(\boldsymbol{x}_2 - \boldsymbol{x}_1) = A\boldsymbol{x}_2 - A\boldsymbol{x}_1 = \boldsymbol{b} - \boldsymbol{b} = \boldsymbol{0}$$

and so $\boldsymbol{x}_2 - \boldsymbol{x}_1$ lies in the nullspace of A. In particular, every solution of $A\boldsymbol{x} = \boldsymbol{b}$ is of the form $\boldsymbol{x}_1 + \boldsymbol{v}$ for some $\boldsymbol{v} \in \text{nullsp}(A)$. This is so important that we state it as a theorem.

THEOREM **3.21.** *Let $A\boldsymbol{x} = \boldsymbol{b}$ be a system of linear equations and let $\boldsymbol{x}_1$ be one solution. Then the set of all solutions is*

$$S = \{\boldsymbol{x}_1 + \boldsymbol{v} \mid \boldsymbol{v} \in \text{nullsp}(A)\}.$$

With a slight abuse of notation, we can write the solution set S as $\boldsymbol{x}_1 + \text{nullsp}(A)$. If we know a basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ for the null space of A, we can even write $S = \boldsymbol{x}_1 + \text{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$.

Consider for instance the solution set for Example 1.25:

$$S = \{(4x_4 + 3/2, -4x_4 + 1, -4x_4 - 3, x_4) | \, x_4 \in \mathbb{R}\}.$$

Here our solution $\boldsymbol{x}_1$ is $(3/2, 1, -3, 0)$ and the null space is

$$\{(4x_4, -4x_4, -4x_4, x_4) \, | x_4 \in \mathbb{R}\},$$

so we can write $S = (3/2, 1, -3, 0) + \text{span}(\{(4, -4, -4, 1)\})$. This is a convenient and efficient way to write the solution set to a system.

This corollary follows from the theorem.

COROLLARY **3.22.** *The number of free parameters required for the set of solutions of $A\boldsymbol{x} = \boldsymbol{b}$ is the nullity of A.*

3.4 Matrix inversion

In this section we consider the *inverse* of a matrix. The theory of inverses in matrix algebra is far more subtle and interesting than that of inverses in real algebra, and it is intimately related to the ideas of independence and rank that we have explored so far.

In real algebra, every non-zero element has a *multiplicative inverse*; in other words, for every $x \neq 0$ we can find another number, x' such that

$$xx' = x'x = 1.$$

Rather than calling it x', we use the notation x^{-1} to mean "the number that you need to multiply x by in order to get 1", thus getting the familiar

$$2^{-1} = \frac{1}{2}.$$

In matrix algebra, the concept of a multiplicative identity and hence an inverse only makes sense for square matrices. However, even then, there are some complicating factors. In particular, because multiplication is not commutative, it is *conceivable* that a product of two square matrices might be equal to the identity only if they are multiplied in a particular order. Fortunately, this does not actually happen:

THEOREM **3.23.** *Suppose that A and B are square $n \times n$ matrices such that $AB = I_n$. Then $BA = I_n$.*

Proof. First we observe that B has rank equal to n. Indeed, suppose that $B\boldsymbol{v} = \mathbf{0}$ for any vector $\boldsymbol{v}$, then $AB\boldsymbol{v} = \mathbf{0}$. Since $AB = I_n$, we get $I_n\boldsymbol{v} = \boldsymbol{v} = \mathbf{0}$. So the null space of B contains only the zero vector, so B has nullity 0 and therefore, by the Rank-Nullity theorem, B has rank n.

Secondly we do some simple manipulation using the properties of matrix algebra that were outlined in Theorem 3.2:

$$\begin{aligned} O_n &= AB - I_n && \text{(because } AB = I_n\text{)} \\ &= B(AB - I_n) && \text{(because } BO_n = O_n\text{)} \\ &= BAB - B && \text{(distributivity)} \\ &= (BA - I_n)B. && \text{(distributivity)} \end{aligned}$$

This manipulation shows that the matrix $(BA - I_n)B = O_n$. However because the rank of B is n, it follows that the column space of B is the whole of $\mathbb{R}^n$, and so *any vector* $\boldsymbol{v} \in \mathbb{R}^n$ can be expressed in the form $B\boldsymbol{x}$ for some $\boldsymbol{x}$ by Theorem 3.14. Therefore

$$\begin{aligned} (BA - I_n)\boldsymbol{v} &= (BA - I_n)B\boldsymbol{x} && \text{(because } B \text{ has rank } n\text{)} \\ &= O_n\boldsymbol{x} && \text{(because } (BA - I_n)B = O_n\text{)} \\ &= \mathbf{0}. && \text{(properties of zero matrix)} \end{aligned}$$

Applying Theorem 3.14 to the matrix $BA - I_n$, this means the column space of $BA - I_n$ is just $\{\mathbf{0}\}$, so $BA - I_n = O_n$ or $BA = I_n$ as required. □

This theorem shows that when defining the inverse of a matrix, we don't need to worry about the *order* in which the multiplication occurs.[13]

Another property of inverses in the algebra of real numbers is that a non-zero real number has a *unique* inverse. Fortunately, this property also holds for matrices:

[13] In some text-books, the authors introduce the idea that B is the "left-inverse" of A if $BA = I$ and the "right-inverse" of A if $AB = I$, and then immediately prove Theorem 3.23 showing that a left-inverse is a right-inverse and vice versa.

THEOREM **3.24.** *If A, B and C are square matrices such that*

$$AB = I_n \quad \text{and} \quad AC = I_n$$

then $B = C$.

Proof. The proof just proceeds by manipulation using the properties of matrix algebra outlined in Theorem 3.2.

$$\begin{aligned} B &= BI_n && \text{(identity matrix property)} \\ &= B(AC) && \text{(hypothesis of theorem)} \\ &= (BA)C && \text{(associativity)} \\ &= I_nC && \text{(by Theorem 3.23)} \\ &= C. && \text{(identity matrix property)} \end{aligned}$$

Therefore a matrix has *at most* one inverse. □

DEFINITION **3.25.** *(Matrix inverse)*
Let A be an $n \times n$ matrix. If there is a matrix B such that

$$AB = I_n$$

then B is called the inverse *of A, and is denoted A^{-1}. From Theorems 3.23 and 3.24, it follows that B is uniquely determined, that $BA = I_n$, and that $B^{-1} = A$.*

A matrix is called invertible *if it has an inverse, and* non-invertible *otherwise.*

EXAMPLE **3.26.** *(Matrix inverse) Suppose that*

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

Then if we take

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

then it is easy to check that

$$AB = I_3.$$

Therefore we conclude that A^{-1} exists and is equal to B and, naturally, B^{-1} exists and is equal to A. □

In real algebra, *every* non-zero number has an inverse, but this is *not* the case for matrices:

EXAMPLE **3.27.** *(Non-zero matrix with no inverse) Suppose that*

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}.$$

Then there is no possible matrix B *such that* $AB = I_2$. *Why is this? If the matrix* B *existed, then it would necessarily satisfy*

$$\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

In order to satisfy this matrix equation, then $b_{11} + b_{21}$ *must equal* 1, *while* $2b_{11} + 2b_{21} = 2(b_{11} + b_{21})$ *must equal* 0 —– *clearly this is impossible. So the matrix* A *has no inverse.* □

One of the common mistakes made by students of elementary linear algebra is to assume that *every* matrix has an inverse. If an argument or proof about a generic matrix A ever uses A^{-1} as part of the manipulation, then it is necessary to first demonstrate that A is actually invertible. Alternatively, the proof can be broken down into two separate cases, one covering the situation where A is assumed to be invertible and a separate one for where it is assumed to be non-invertible.

3.4.1 *Finding inverses*

This last example of the previous section (Example 3.27) essentially shows us how to *find* the inverse of a matrix because, as usual, it all boils down to solving systems of linear equations. If A is an $n \times n$ matrix then finding its inverse, if it exists, is just a matter of finding a matrix B such that $AB = I_n$. To find the *first column* of B, it is sufficient to solve the equation $A\boldsymbol{x} = \boldsymbol{e_1}$, then the second column is the solution to $A\boldsymbol{x} = \boldsymbol{e_2}$, and so on. If any of these equations has no solutions then A does not have an inverse.[14] Therefore, finding the inverse of an $n \times n$ matrix involves solving n separate systems of linear equations. However because each of the n systems has the same coefficient matrix on the left of the augmenting bar (that is, A), there are shortcuts that make this procedure easier.

[14] In fact, if any of them have *infinitely many* solutions, then it also follows that A has no inverse because if the inverse exists it must be unique. Thus if one of the equations has infinitely many solutions, then one of the *other* equations must have no solutions.

To illustrate this, we do a full example for a 3×3 matrix, although the principle is the same for any matrix. Suppose we want to find the inverse of the matrix

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

The results above show that we just need to find a matrix $B = (b_{ij})$ such that

$$\begin{bmatrix} -1 & 0 & 1 \\ 0 & 1 & -1 \\ 2 & 0 & -1 \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

This can be done by solving *three* separate systems of linear equations, one to determine each column of B:

$$A \begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad A \begin{bmatrix} b_{12} \\ b_{22} \\ b_{32} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad A \begin{bmatrix} b_{13} \\ b_{23} \\ b_{33} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Then the matrix A has an inverse if and only if all three of these systems of linear equations have a solution, and in fact, each of them must have a *unique* solution. If any one of the three equations is inconsistent, then A is one of the matrices that just doesn't have an inverse.

Consider how solving these systems of linear equations will proceed: for the first column, we get the augmented matrix

$$\left[\begin{array}{ccc|c} -1 & 0 & 1 & 1 \\ 0 & 1 & -1 & 0 \\ 2 & 0 & -1 & 0 \end{array}\right]$$

which we solved in Section 1.5 by doing the elementary row operations $R_3 \leftarrow R_3 + 2R_1, R_1 \leftarrow R_1 - R_3, R_2 \leftarrow R_2 + R_3, R_1 \leftarrow -R_1$: it has solution $b_{11} = 1$, $b_{21} = 2$ and $b_{31} = 2$

Now we solve for the *second column* of B; this time the augmented matrix is

$$\left[\begin{array}{ccc|c} -1 & 0 & 1 & 0 \\ 0 & 1 & -1 & 1 \\ 2 & 0 & -1 & 0 \end{array}\right]$$

and after pivoting on the $(1,1)$-entry we get

$$\left[\begin{array}{ccc|c} -1 & 0 & 1 & 0 \\ 0 & 1 & -1 & 1 \\ 0 & 0 & 1 & 0 \end{array}\right] \quad R_3 \leftarrow R_3 + 2R_1$$

We can now use the last pivot to zero-out the rest of the third column:

$$\left[\begin{array}{ccc|c} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{array}\right] \begin{array}{l} R_1 \leftarrow R_1 - R_3 \\ R_2 \leftarrow R_2 + R_3 \end{array}$$

and we finish with

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{array}\right] \quad R_1 \leftarrow -R_1$$

It is immediately apparent that we used the *exact same elementary row operations* as we did for the previous system of linear equations because, naturally enough, the coefficient matrix is the same matrix. And obviously, we'll do the same elementary row operations *again* when we solve the third system of linear equations! So to avoid repeating work unnecessarily, it is better to solve *all three systems* simultaneously. This is done by using a sort of "super-augmented" matrix that has three columns to the right of the augmenting bar, representing the right-hand sides of the three separate equations:

$$\left[\begin{array}{ccc|ccc} -1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 1 & 0 \\ 2 & 0 & -1 & 0 & 0 & 1 \end{array}\right].$$

Then performing an elementary row operation on this bigger matrix has exactly the same effect as doing it on each of the three systems separately. We will apply Gauss-Jordan elimination (see Section 1.5) to this "super-augmented" matrix.

$$\left[\begin{array}{ccc|ccc} \boxed{1} & 0 & -1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 1 & 0 \\ 2 & 0 & -1 & 0 & 0 & 1 \end{array}\right] \quad R_1 \leftarrow -R_1$$

$$\left[\begin{array}{ccc|ccc} 1 & 0 & -1 & -1 & 0 & 0 \\ 0 & \boxed{1} & -1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 2 & 0 & 1 \end{array}\right] \quad R_3 \leftarrow R_3 - 2R_1$$

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 2 & 1 & 1 \\ 0 & 0 & \boxed{1} & 2 & 0 & 1 \end{array}\right] \quad \begin{array}{l} R_1 \leftarrow R_1 + R_3 \\ R_2 \leftarrow R_2 + R_3 \end{array}$$

The first system of equations has solution $b_{11} = 1$, $b_{21} = 2$ and $b_{31} = 2$, while the second has solution $b_{12} = 0$, $b_{22} = 1$ and $b_{32} = 0$, and the final system has solution $b_{13} = 1$, $b_{23} = 1$ and $b_{33} = 1$. Thus the inverse of the matrix A is given by

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

which is just exactly the matrix that was found to the right of the augmenting bar!

Formally, the procedure for finding the inverse of a matrix is as follows. Remember, however, that this is simply a way of *organising* the calculations efficiently, and that there is *nothing more sophisticated* occurring than solving systems of linear equations.

KEY CONCEPT **3.28.** *(Finding the inverse of a matrix) In order to find the inverse of an $n \times n$ matrix A, proceed as follows:*

1. *Form the "super-augmented" matrix*
$$[A \mid I_n].$$
2. *Apply Gauss-Jordan elimination to this matrix to place it into reduced row-echelon form*
3. *If the resulting reduced row echelon matrix has an identity matrix to the left of the augmenting bar, then it must have the form*
$$[I_n \mid A^{-1}]$$
and so A^{-1} will be the matrix on the right *of the augmenting bar.*
4. *If the reduced row echelon matrix* does not *have an identity matrix to the left of the augmenting bar, then the matrix A is not invertible.*

It is interesting to note that, while it is important to understand what a matrix inverse is and *how* to calculate a matrix inverse, it is almost never necessary to actually find an explicit matrix inverse in practice. An explicit problem for which a matrix inverse might be useful can almost always be solved directly (by some form of Gaussian elimination) without actually computing the inverse.

However, as we shall see in the next section, understanding the procedure for calculating an inverse is useful in developing theoretical results.

3.4.2 *Characterising invertible matrices*

In the last two subsections we have defined the inverse of a matrix, demonstrated that some matrices have inverses and others don't and given a procedure that will either find the inverse of a matrix or demonstrate that it does not exist. In this subsection, we consider some of the special properties of *invertible matrices* focussing on what makes them invertible, and what particular properties are enjoyed by invertible matrices.

THEOREM **3.29.** *An $n \times n$ matrix is invertible if and only if it has rank equal to n.*

Proof. This is so important that we give a couple of proofs[15] in slightly different language, though the fundamental concept is the same in both proofs.
PROOF 1: Applying elementary row operations to a matrix does not alter its row space, and hence its rank. If a matrix A is invertible, then Gauss-Jordan elimination applied to A will yield the identity matrix, which has rank n. If A is not invertible, then applying Gauss-Jordan elimination to A yields a matrix with at least one row of zeros, and so it does not have rank n.
PROOF 2: If a matrix A is invertible then there is always a solution to the matrix equation

$$A\boldsymbol{x} = \boldsymbol{v}$$

for every $\boldsymbol{v}$. Indeed we can just take $\boldsymbol{x} = A^{-1}\boldsymbol{v}$. Thus the column space of A, which is $\{A\boldsymbol{x} | \boldsymbol{x} \in \mathbb{R}^n\}$ is equal to the whole of $\mathbb{R}^n$, and so the rank of A is n. Conversely, assume A has full rank. Then the column space of A, which is $\{A\boldsymbol{x} | \boldsymbol{x} \in \mathbb{R}^n\}$, has dimension n so is equal to $\mathbb{R}^n$. Therefore there exist $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$ such that $A\boldsymbol{x_j} = \boldsymbol{e_j}$ for each j. Now construct the matrix B whose j-th column is the vector $\boldsymbol{x_j}$. Then it can be checked that $AB = I$ and so A is invertible. □

[15] Note that in the proof of Theorem 3.23, we already saw that if B is the inverse of A, then B has full rank.

There are some other characterisations of invertible matrices that may be useful, but they are all really just elementary restatements of Theorem 3.29.

THEOREM **3.30.** *Let A be an $n \times n$ matrix. Then*

1. *A is invertible if and only if its rows are linearly independent.*

2. *A is invertible if and only if its columns are linearly independent.*
3. *A is invertible if and only if its row space is* $\mathbb{R}^n$.
4. *A is invertible if and only if its column space is* $\mathbb{R}^n$.

Proof. These are all ways of saying "the rank of A is n". □

EXAMPLE **3.31.** *(Non-invertible matrix) The matrix*

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

is not invertible *because*

$$(0,1,2) + (1,2,-1) = (1,3,1)$$

is a dependency among the rows, and so the rows are not linearly independent. □

Now let's consider some of the *properties* of invertible matrices.

THEOREM **3.32.** *Suppose that A and B are invertible $n \times n$ matrices, and k is a positive integer. Then*

1. *The matrix AB is invertible, and*

$$(AB)^{-1} = B^{-1}A^{-1}.$$

2. *The matrix A^k is invertible, and*

$$\left(A^k\right)^{-1} = \left(A^{-1}\right)^k.$$

Recall that $A^k = \underbrace{AA\cdots A}_{k \text{ times}}$

3. *The matrix A^T is invertible, and*

$$\left(A^T\right)^{-1} = \left(A^{-1}\right)^T.$$

Proof. To show that a matrix is invertible, it is sufficient to demonstrate the existence of *some matrix* whose product with the given matrix is the identity. Thus to show that AB is invertible, we must find something that we can multiply AB by in order to end up with the identity.

$$\begin{aligned}
(AB)(B^{-1}A^{-1}) &= A(BB^{-1})A^{-1} && \text{(associativity)} \\
&= AI_nA^{-1} && \text{(properties of inverses)} \\
&= AA^{-1} && \text{(properties of identity)} \\
&= I_n. && \text{(properties of inverses)}
\end{aligned}$$

This shows that AB is invertible, and that its inverse is $B^{-1}A^{-1}$ as required. The remaining two statements are straightforward to prove using matrix properties (and induction for the second property). □

This theorem shows that the collection of invertible $n \times n$ matrices is *closed* under matrix multiplication. In addition, there is a multiplicative identity (the matrix I_n) and every matrix has an inverse (obviously!). These turn out to be the conditions that define an algebraic structure called a *group*. The group of invertible $n \times n$ matrices plays a fundamental role in the mathematical subject of *group theory* which is an important topic in higher-level Pure Mathematics.

3.5 Determinants

From high-school we are all familiar with the formula for the inverse of a 2×2 matrix:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \quad \text{if } ad - bc \neq 0$$

where the inverse does not exist if $ad - bc = 0$. In other words, a 2×2 matrix has an inverse if and only if $ad - bc \neq 0$. This number is called the *determinant* of the matrix, and it is either denoted $\det(A)$ or just $|A|$.

EXAMPLE **3.33.** *(Determinant notation) If*

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

then we say either

$$det(A) = 2 \quad or \quad \begin{vmatrix} 3 & 5 \\ 2 & 4 \end{vmatrix} = 2$$

because $3 \cdot 4 - 2 \cdot 5 = 2$. □

In this section, we'll extend the concept of determinant to $n \times n$ matrices and show that it characterises invertible matrices in the same way — a matrix is invertible if and only if its determinant is non-zero.

The *determinant* of a square matrix is a scalar value (i.e. a number) associated with that matrix that can be *recursively defined* as follows:

DEFINITION **3.34.** *(Determinant)*
If $A = (a_{ij})$ is an $n \times n$ matrix, then the determinant *of A is a real number, denoted $det(A)$ or $|A|$, that is defined as follows:*

1. *If $n = 1$, then $|A| = a_{11}$.*
2. *If $n > 1$, then*

$$|A| = \sum_{j=1}^{j=n} (-1)^{1+j} a_{1j} |A[1, j]| \tag{3.4}$$

where $A[i, j]$ is the $(n-1) \times (n-1)$ matrix obtained from A by deleting the i-th row and the j-th column

Notice that when $n > 1$, this expresses an $n \times n$ determinant as an alternating sum of n terms, each of which is a real number multiplied by an $(n-1) \times (n-1)$ determinant.

EXERCISE **3.5.1.** *Check that this method yields the formula you know for 2×2 matrices.*

EXAMPLE **3.35.** *(A 3×3 determinant) What is the determinant of the matrix*

$$A = \begin{bmatrix} 2 & 5 & 3 \\ 4 & 3 & 6 \\ 1 & 0 & 2 \end{bmatrix}?$$

First let's identify the matrices $A[1,1]$, $A[1,2]$ and $A[1,3]$; recall these are obtained by deleting one row and column from A. For example, $A[1,2]$ is obtained by deleting the first row and second column from A, thus

$$A[1,2] = \begin{bmatrix} 2 & 5 & 3 \\ 4 & 3 & 6 \\ 1 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 4 & 6 \\ 1 & 2 \end{bmatrix}.$$

The term $(-1)^{1+j}$ simply alternates between $+1$ and -1 and so the first term is added because $(-1)^2 = 1$, the second subtracted because $(-1)^3 = -1$, the third added, and so on. Using the formula we get

$$\begin{aligned} |A| &= 2 \cdot \begin{vmatrix} 3 & 6 \\ 0 & 2 \end{vmatrix} - 5 \cdot \begin{vmatrix} 4 & 6 \\ 1 & 2 \end{vmatrix} + 3 \cdot \begin{vmatrix} 4 & 3 \\ 1 & 0 \end{vmatrix} \\ &= 2 \cdot 6 - 5 \cdot 2 + 3 \cdot (-3) \\ &= -7 \end{aligned}$$

where the three 2×2 determinants have just been calculated using the usual rule. □

This procedure for calculating the determinant is called *expanding along the first row*, because each of the terms $a_{1j}A[1,j]$ is associated with an entry in the first row. However it turns out, although we shall not prove it[16], that it is possible to do the expansion along *any row* or indeed, *any column*. So in fact we have the following result:

[16] Proving this is not difficult but it involves a lot of manipulation of subscripts and nested sums, which is probably not the best use of your time.

THEOREM **3.36.** *Let $A = (a_{ij})$ be an $n \times n$ matrix. Then for any* **fixed row index** *i we have*

$$|A| = \sum_{j=1}^{j=n} (-1)^{i+j} a_{ij} |A[i,j]|$$

and for any **fixed column index** *j, we have*

$$|A| = \sum_{i=1}^{i=n} (-1)^{i+j} a_{ij} |A[i,j]|.$$

(Notice that the first of these two sums involves terms obtained from the i-th row of the matrix, while the second involves terms from the j-th column of the matrix.)

EXAMPLE **3.37.** *(Expanding down the second column) Determine the determinant of*

$$A = \begin{bmatrix} 2 & 5 & 3 \\ 4 & 3 & 6 \\ 1 & 0 & 2 \end{bmatrix}$$

by expanding down the second column. Notice that because we are using the second *column, the signs given by the* $(-1)^{i+j}$ *terms alternate* -1, $+1$, -1 *starting with a negative, not a positive. So the calculation gives*

$$\begin{aligned} |A| &= (-1)\cdot 5\cdot \begin{vmatrix} 4 & 6 \\ 1 & 2 \end{vmatrix} + 3\cdot \begin{vmatrix} 2 & 3 \\ 1 & 2 \end{vmatrix} + (-1)\cdot 0\cdot \textit{(don't care)} \\ &= -5\cdot 2 + 3\cdot 1 + 0 \\ &= -7. \end{aligned}$$

Also notice that because $a_{32} = 0$, *the term* $(-1)^{3+2}\, a_{32}\, |A[3,2]|$ *is forced to be zero, and so there is no need to actually calculate* $|A[3,2]|$.

In general, you should choose the row or column of the matrix that has lots of zeros in it, in order to make the calculation as easy as possible!

EXAMPLE **3.38.** *(Easy if you choose right) To determine the determinant of*

$$A = \begin{bmatrix} 2 & 5 & 0 & 3 \\ 4 & 3 & 0 & 6 \\ 1 & 0 & 0 & 2 \\ 1 & 1 & 3 & 2 \end{bmatrix}$$

use the third column *which has only one non-zero entry, and get*

$$\begin{aligned} |A| &= (+1)\cdot 0 + (-1)\cdot 0 + (+1)\cdot 0 + (-1)\cdot 3\cdot \begin{vmatrix} 2 & 5 & 3 \\ 4 & 3 & 6 \\ 1 & 0 & 2 \end{vmatrix} \\ &= (-3).(-7) = 21 \end{aligned}$$

rather than getting an expression with three or four 3×3 *determinants to evaluate!* □

From this we can immediately deduce some theoretical results:

THEOREM **3.39.** *Let* A *be an* $n \times n$ *matrix. Then*

1. $|A^T| = |A|$,
2. $|\alpha A| = \alpha^n |A|$, *and*
3. *If* A *has a row of zeros, then* $|A| = 0$.
4. *If* A *is an upper (or lower) triangular matrix, then* $|A|$ *is the product of the entries on the diagonal of the matrix.*

Proof. To prove the first statement, we use induction on n. Certainly the statement is true for 1×1 matrices. So now suppose that it is true for all matrices of size $n-1$. Notice that expanding along the

first row of A gives a sum with the same coefficients as expanding down the first *column* of A^T, the signs of each term are the same because $(-1)^{i+j} = (-1)^{j+i}$, and all the $(n-1) \times (n-1)$ determinants in the first sum are just the transposes of those in the second sum, and so are equal by the inductive hypothesis.

For the second statement we again use induction. Certainly the statement is true for 1×1 matrices. So now suppose that it is true for all matrices of size up to $n-1$. Then

$$\begin{aligned}
|\alpha A| &= \sum_{j=1}^{j=n} (-1)^{i+j} \, (\alpha a_{ij}) \, |\alpha A[i,j]| \\
&= \sum_{j=1}^{j=n} (-1)^{i+j} \, (\alpha a_{ij}) \, \alpha^{n-1} |A[i,j]| \qquad \text{(inductive hypothesis)} \\
&= \alpha \alpha^{n-1} \sum_{j=1}^{j=n} (-1)^{i+j} \, a_{ij} \, |A[i,j]| \qquad \text{(rearranging)} \\
&= \alpha^n |A|.
\end{aligned}$$

The third statement is immediate because if we expand along the row of zeros, then every term in the sum is zero.

The fourth statement again follows from an easy induction argument. Intuitively, for an upper triangular matrix, then keep expanding along the first column at each step. □

EXAMPLE **3.40.** *(Determinant of matrix in row echelon form) A matrix in row echelon form is necessarily upper triangular, and so its determinant can easily be calculated. For example, the matrix*

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

which is in row echelon form has determinant equal to $2 \cdot 1 \cdot (-3) \cdot 1 = -6$ *because this is the product of the diagonal entries. We can verify this easily from the formula by repeatedly expanding down the first column. So*

$$\begin{vmatrix} 2 & 0 & 1 & -1 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & -3 & 2 \\ 0 & 0 & 0 & 1 \end{vmatrix} = 2 \cdot \begin{vmatrix} 1 & 2 & 1 \\ 0 & -3 & 2 \\ 0 & 0 & 1 \end{vmatrix} = 2 \cdot 1 \cdot \begin{vmatrix} -3 & 2 \\ 0 & 1 \end{vmatrix}.$$

□

3.5.1 Calculating determinants

The recursive definition of a determinant expresses an $n \times n$ determinant as a linear combination of n terms each involving an $(n-1) \times (n-1)$ determinant. So to find a 10×10 determinant like this involves computing ten 9×9 determinants, each of which involves nine 8×8 determinants, each of which involves eight 7×7

determinants, each of which involves seven 6×6 determinants, each of which involves six 5×5 determinants, each of which involves five 4×4 determinants, each of which involves four 3×3 determinants, each of which involves three 2×2 determinants. While this is possible (by computer) for a 10×10 matrix, even the fastest supercomputer would not complete a 100×100 matrix in the lifetime of the universe.

However, in practice, a computer can easily find a 100×100 determinant, so there must be another more efficient way. Once again, this way is based on elementary row operations.

THEOREM **3.41.** *Suppose A is an $n \times n$ matrix, and that A' is obtained from A by performing a single elementary row operation.*

1. *If the elementary row operation is of Type 1 ($R_i \leftrightarrow R_j$), then*

$$|A'| = -|A|.$$

   In other words, a Type 1 elementary row operation multiplies the determinant by -1.

2. *If the elementary row operation is of Type 2, say $R_i \leftarrow \alpha R_i$ then*

$$|A'| = \alpha |A|.$$

   In other words, multiplying a row by the scalar α multiplies the determinant by α.

3. *If the elementary row operation is of Type 3 ($R_i \leftarrow R_i + \alpha R_k$), then*

$$|A| = |A'|.$$

   In other words, adding a multiple of one row to another does not change the determinant.

Proof. 1. (This proof is very technical so we will only give a sketch.) Consider the elementary row operation of Type 1 $R_i \leftrightarrow R_j$ where $i < j$. If you expand along row i then row j (which has now become row $j-1$ after row i was deleted) for A, and expand along row j then row i for A'. You will get two linear combinations of $n(n-1)$ terms involving $(n-2) \times (n-2)$ determinants, and you will notice the two combinations are exactly the opposite of each other. Therefore $|A'| = -|A|$.

2. Consider the elementary row operation of Type 2 $R_i \leftarrow \alpha R_i$. Expand both A and A' along row i and notice that, if we remove row i, A and A' are the same matrix so $|A[i,j]| = |A'[i,j]|$ for each j. Therefore

Note this is an elementary row operation only if $\alpha \neq 0$ but this result holds even if $\alpha = 0$.

$$\begin{aligned} |A'| &= \sum_{j=1}^{j=n} (-1)^{i+j} a'_{ij} |A'[i,j]| = \sum_{j=1}^{j=n} (-1)^{i+j} \alpha a_{ij} |A[i,j]| \\ &= \alpha \sum_{j=1}^{j=n} (-1)^{i+j} a_{ij} |A[i,j]| = \alpha |A|. \end{aligned}$$

3. Consider the elementary row operation of Type 3 $R_i \leftarrow R_i + \alpha R_k$. Expand A' along row i and notice that $|A[i,j]| = |A'[i,j]|$ for each j.

$$|A'| = \sum_{j=1}^{j=n} (-1)^{i+j} a'_{ij} |A'[i,j]| = \sum_{j=1}^{j=n} (-1)^{i+j} (a_{ij} + \alpha a_{kj}) |A[i,j]|$$
$$= \sum_{j=1}^{j=n} (-1)^{i+j} a_{ij} |A[i,j]| + \alpha \sum_{j=1}^{j=n} (-1)^{i+j} a_{kj} |A[i,j]| = |A| + \alpha |B|,$$

where B is obtained from A by replacing row i by row k. Therefore B has twice the same row: rows i and k are the same. Now apply the row operation $R_i \leftrightarrow R_k$ to B: it did not change the matrix! Thus, by part 1, $|B| = -|B|$, and so $|B| = 0$.

□

Another common mistake for beginning students of linear algebra is to assume that row-reduction preserves *every* interesting property of a matrix. Instead, row-reduction preserves some properties, alters others in a controlled fashion, and destroys others. It is important to always know *why* the row-reduction is being done.

Previously we have used elementary row operations to find solutions to systems of linear equations, and to find the basis and dimension of the row space of a matrix. In both these applications, the elementary row operations did not change the answer that was being sought. For finding determinants however, elementary row operations *do change* the determinant of the matrix, but they change it *in a controlled fashion* and so the process is still useful.

EXAMPLE **3.42.** *(Finding determinant by row-reduction) We return to an earlier example, of finding the determinant of*

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

Suppose that this unknown value is denoted d. Then after doing the Type I elementary row operation $R_1 \leftrightarrow R_3$ we get the matrix

$$\begin{bmatrix} 1 & 0 & 2 \\ 4 & 3 & 6 \\ 2 & 5 & 3 \end{bmatrix}$$

which has determinant $-d$, because Type 1 elementary row operations multiply the determinant by -1. If we now perform the Type 3 elementary row operations, $R_2 \leftarrow R_2 - 4R_1$ and $R_3 \leftarrow R_3 - 2R_1$, then the resulting matrix

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

still has determinant $-d$ because Type 3 elementary row operations do not alter the determinant. Finally, the elementary row operation $R_3 \leftarrow R_3 - (5/3)R_2$ yields the matrix

$$\begin{bmatrix} 1 & 0 & 2 \\ 0 & 3 & -2 \\ 0 & 0 & 7/3 \end{bmatrix}$$

which still *has determinant $-d$. Using property 4 of Theorem 3.39, we get that the determinant of this final matrix is 7, and so $-d = 7$, which immediately tells us that $d = -7$ confirming the results of Examples 3.35 and 3.37* □

Thinking about this process in another way shows us that if a matrix A has determinant $|A|$ and the matrix A' is the row-echelon matrix obtained by performing Gaussian elimination on A, then

$$|A'| = \beta|A| \quad \text{for some } \beta \neq 0.$$

Combining this with the fourth property of Theorem 3.39 allows us to state the single most important property of determinants:

THEOREM **3.43.** *A matrix A is invertible if and only if its determinant is non-zero.*

Proof. Consider the row echelon matrix A' obtained by applying Gaussian elimination to A. If A is invertible, then A' has no zero rows and so every diagonal entry is non-zero and thus $|A'| \neq 0$, while if A is not invertible, A' has at least one zero row and thus $|A'| = 0$. As the determinant of A is a non-zero multiple of the determinant of A', it follows that A has non-zero determinant if and only if it is invertible. □

3.5.2 *Properties of the determinant*

We finish this chapter with some of the properties of determinants, most of which follow immediately from the following theorem, which shows that the determinant function is *multiplicative*.

THEOREM **3.44.** *If A and B are two $n \times n$ matrices, then*

$$|AB| = |A| \cdot |B|.$$

Proof. There are several proofs of this result, none of which are very nice. We give a sketch outline[17] of the most illuminating proof. First note that if either A or B (or both) is not invertible, then AB is not invertible and so the result is true if any of the determinants is zero.

Then proceed in the following steps:

1. Define an *elementary matrix* to be a matrix obtained by performing a single elementary row operation on the identity matrix.
2. Note that premultiplying a matrix A by an elementary matrix E, thereby forming the matrix EA, is exactly the same as performing the same elementary row operation on A.
3. Show that elementary matrices of Type 1, 2 and 3 have determinant -1, α and 1 respectively (where α is the non-zero scalar associated with an elementary row operation of Type 2).
4. Conclude that the result is true if A is an elementary matrix or a product of elementary matrices.

[17] This is a very brief outline of the proof so do not worry if you cannot follow it without some guidance on how to fill in the gaps.

5. Finish by proving that a matrix is invertible if and only if it is the product of elementary matrices, because Gauss-Jordan elimination will reduce any invertible matrix to the identity matrix.

The other proofs of this result use a different description of the determinant as the weighted sum of $n!$ products of matrix entries, together with extensive algebraic manipulation. □

Just for fun, we'll demonstrate this result for 2×2 matrices purely algebraically, in order to give a flavour of the alternative proofs. Suppose that

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad B = \begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix}.$$

Then

$$AB = \begin{bmatrix} aa' + bc' & ab' + bd' \\ a'c + c'd & b'c + dd' \end{bmatrix}.$$

Therefore

$$\begin{aligned} |AB| &= (aa' + bc')(b'c + dd') - (ab' + bd')(a'c + c'd) \\ &= aa'b'c + aa'dd' + bc'b'c + bc'dd' - ab'a'c - ab'c'd - bd'a'c - bd'c'd \\ &= (aa'b'c - ab'a'c) + aa'dd' + bc'b'c + (bc'dd' - bd'c'd) - ab'c'd - bd'a'c \\ &= 0 + (ad)(a'd') + (bc)(b'c') + 0 - (ad)(b'c') - (a'd')(bc) \\ &= (ad - bc)(a'd' - b'c') \\ &= |A||B|. \end{aligned}$$

The multiplicativity of the determinant immediately gives the main properties.

THEOREM **3.45.** *Suppose that A and B are $n \times n$ matrices and k is a positive integer. Then*

1. $|AB| = |BA|$
2. $|A^k| = |A|^k$
3. *If A is invertible, then $|A^{-1}| = 1/|A|$*

Proof. The first two are immediate, and the third follows from the fact that $AA^{-1} = I_n$ and so $|A||A^{-1}| = 1$. □

4
Linear transformations

4.1 Introduction

First we will give the general definition of a function.

Definition 4.1. *(Function)*
Given two sets A and B, a function $f : A \to B$ *is a rule that assigns to each element of A a* unique *element of B. We often write*

$$\begin{array}{rcl} f : A & \longrightarrow & B \\ a & \longmapsto & f(a) \end{array}$$

where $f(a)$ is the element of B assigned to a, called the image *of a under f. The set A is called the* domain *of f and is sometimes denoted* $\mathrm{dom}(f)$*. The set B is called the* codomain *of f. The* range *of f, (sometimes denoted* $\mathrm{range}(f)$*) is the set of all elements of B that are the image of some element of A.*

Often $f(a)$ is defined by some equation involving a (or whatever variable is being used to represent elements of A), for example $f(a) = a^2$. However, sometimes you may see f defined by listing $f(a)$ for each $a \in A$. For example, if $A = \{1, 2, 3\}$ we could define f by

$$\begin{array}{rcl} f : A & \longrightarrow & \mathbb{R} \\ 1 & \longmapsto & 10 \\ 2 & \longmapsto & 10 \\ 3 & \longmapsto & 102. \end{array}$$

If $f(x)$ is defined by some rule and the domain of f is not explicitly given then we assume that the domain of f is the set of all values on which $f(x)$ is defined.

Note that the range of f need not be all of the codomain. For example, if $f : \mathbb{R} \to \mathbb{R}$ is the function defined by $f(x) = x^2$ then the codomain of f is $\mathbb{R}$ while the range of f is the set $\{x \in \mathbb{R} \mid x \geq 0\}$.

A *linear transformation* is a function from one vector space to another preserving the structure of vector spaces, that is, it preserves vector addition and scalar multiplication.

More precisely:

Definition 4.2. *(Linear transformation)*
A function f from $\mathbb{R}^n$ to $\mathbb{R}^m$ is a linear transformation if:

1. $f(\boldsymbol{u}+\boldsymbol{v}) = f(\boldsymbol{u}) + f(\boldsymbol{v})$ *for all* $\boldsymbol{u}, \boldsymbol{v}$ *in* $\mathbb{R}^n$;
2. $f(\alpha\boldsymbol{v}) = \alpha f(\boldsymbol{v})$ *for all* $\boldsymbol{v}$ *in* $\mathbb{R}^n$ *and all* α *in* $\mathbb{R}$.

An interesting case is when $n = m$, in which case the domain and codomain are the same vector space.

Example 4.3. *(Linear transformation) In $\mathbb{R}^3$, the* orthogonal projection *to the xy-plane is a linear transformation. This maps the vector (x, y, z) to $(x, y, 0)$.* □

Check the two conditions to convince yourself

Example 4.4. *(Not a linear transformation) The function from $\mathbb{R}^3$ to $\mathbb{R}$ given by $f(x, y, z) = x^2 + y^2 + z^2$ is* not *a linear transformation. Indeed for $\boldsymbol{v} = (1, 0, 0)$ and $\alpha = 2$, $f(\alpha\boldsymbol{v}) = f(2, 0, 0) = 4$ while $\alpha f(\boldsymbol{v}) = 2.1 = 2$.* □

Example 4.5. *(Not a linear transformation) Let $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = ax + b$. Note that $f(1) = a + b$ while $f(2) = 2a + b \neq 2(a + b)$ when $b \neq 0$. Thus when $b \neq 0$, the function f is not a linear transformation of the vector space $\mathbb{R}$. We call f an* affine *function.* □

These two examples illustrate again the black swan concept: we only need to find one concrete counter-example to prove that a function is **not** a linear transformation.

Example 4.6. *Let A be an $m \times n$ matrix. Then the function f from $\mathbb{R}^n$ to $\mathbb{R}^m$ such that $f(\boldsymbol{x}) = A\boldsymbol{x}$ is a linear transformation (where we see $\boldsymbol{x}$ as an $n \times 1$ column vector, as described in Chapter 2). Indeed*

$$f(\boldsymbol{u}+\boldsymbol{v}) = A(\boldsymbol{u}+\boldsymbol{v}) = A\boldsymbol{u} + A\boldsymbol{v} = f(\boldsymbol{u}) + f(\boldsymbol{v})$$

(using Property (8) of Theorem 3.2), and

$$f(\alpha\boldsymbol{v}) = A(\alpha\boldsymbol{v}) = \alpha A\boldsymbol{v} = \alpha f(\boldsymbol{v})$$

(using Property (7) of Theorem 3.2). □

Theorem 4.7. *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be a linear transformation.*

(i) $f(\mathbf{0}) = \mathbf{0}$.

(ii) The range of f is a subspace of $\mathbb{R}^m$.

Proof. (i) Since $\mathbf{0}$ is an additive identity, we have that $\mathbf{0} = \mathbf{0} + \mathbf{0}$. Applying f to this identity yields: $f(\mathbf{0}) = f(\mathbf{0} + \mathbf{0}) = f(\mathbf{0}) + f(\mathbf{0})$. The result follows from subtracting $f(\mathbf{0})$ on each side.

(ii) We need to prove the three subspace conditions for range(f) (see Definition 2.4).

- (S1) $\mathbf{0} \in$ range(f) since $\mathbf{0} = f(\mathbf{0})$ by Part (i).
- (S2) Let $\boldsymbol{u}, \boldsymbol{v} \in$ range(f), that is, $\boldsymbol{u} = f(\boldsymbol{u'})$, $\boldsymbol{v} = f(\boldsymbol{v'})$ for some $\boldsymbol{u'}, \boldsymbol{v'} \in \mathbb{R}^n$. Then $\boldsymbol{u} + \boldsymbol{v} = f(\boldsymbol{u'}) + f(\boldsymbol{v'}) = f(\boldsymbol{u'} + \boldsymbol{v'})$ and so $\boldsymbol{u} + \boldsymbol{v} \in$ range(f).
- (S3) Let $\alpha \in \mathbb{R}$ and $\boldsymbol{v} \in$ range(f), that is, $\boldsymbol{v} = f(\boldsymbol{v'})$ for some $\boldsymbol{v'} \in \mathbb{R}^n$. Then $\alpha\boldsymbol{v} = \alpha f(\boldsymbol{v'}) = f(\alpha\boldsymbol{v'})$ and so $\alpha\boldsymbol{v} \in$ range(f).

□

4.2 Linear transformations and bases

A linear transformation can be given by a formula, but there are other ways to describe it. In fact, if we know the images under a linear transformation of each of the vectors in a basis, then the *rest* of the linear transformation is completely determined.

THEOREM **4.8.** *Let* $\{\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_n}\}$ *be a basis for* $\mathbb{R}^n$ *and let* $\boldsymbol{t_1}$, $\boldsymbol{t_2}$, …, $\boldsymbol{t_n}$ *be n vectors of* $\mathbb{R}^m$. *Then there exists a* unique *linear transformation f from* $\mathbb{R}^n$ *to* $\mathbb{R}^m$ *such that* $f(\boldsymbol{u_1}) = \boldsymbol{t_1}$, $f(\boldsymbol{u_2}) = \boldsymbol{t_2}$, …, $f(\boldsymbol{u_n}) = \boldsymbol{t_n}$.

For instance the basis of $\mathbb{R}^n$ can be the standard basis.

Proof. We know by Theorem 2.47 that any vector $\boldsymbol{v} \in \mathbb{R}^n$ can be written in a unique way as $\boldsymbol{v} = \alpha_1\boldsymbol{u_1} + \alpha_2\boldsymbol{u_2} + \ldots + \alpha_n\boldsymbol{u_n}$ (where the α_i's are real numbers). Define $f(\boldsymbol{v}) = \alpha_1\boldsymbol{t_1} + \alpha_2\boldsymbol{t_2} + \ldots + \alpha_n\boldsymbol{t_n}$. Then f satisfies $f(\boldsymbol{u_i}) = \boldsymbol{t_i}$ for all i between 1 and n and we can easily check that f is linear. So we have that f exists.

Now suppose g is also a linear transformation satisfying $g(\boldsymbol{u_i}) = \boldsymbol{t_i}$ for all i between 1 and n. Then

$$\begin{aligned} g(\boldsymbol{v}) &= g(\alpha_1\boldsymbol{u_1} + \alpha_2\boldsymbol{u_2} + \cdots + \alpha_n\boldsymbol{u_n}) \\ &= g(\alpha_1\boldsymbol{u_1}) + g(\alpha_2\boldsymbol{u_2}) + \cdots + g(\alpha_n\boldsymbol{u_n}) \\ &\qquad \text{(by the first condition for a linear function)} \\ &= \alpha_1 g(\boldsymbol{u_1}) + \alpha_2 g(\boldsymbol{u_2}) + \cdots + \alpha_n g(\boldsymbol{u_n}) \\ &\qquad \text{(by the second condition for a linear function)} \\ &= \alpha_1\boldsymbol{t_1} + \alpha_2\boldsymbol{t_2} + \cdots + \alpha_n\boldsymbol{t_n} \\ &= f(\boldsymbol{v}). \end{aligned}$$

Thus $g(\boldsymbol{v}) = f(\boldsymbol{v})$ for all $\boldsymbol{v} \in \mathbb{R}^n$ so they are the same linear transformation, that is, f is unique. □

EXERCISE **4.2.1.** *Let f be a linear transformation from* $\mathbb{R}^2$ *to* $\mathbb{R}^3$ *with* $f(1,0) = (1,2,3)$ *and* $f(0,1) = (0,-1,2)$. *Determine* $f(x,y)$.

4.3 Linear transformations and matrices

We have seen that it is useful to choose a basis of the domain, say $B = \{\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_n}\}$. Now we will also take a basis for the codomain: $C = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_m}\}$. For now we can think of both the bases as the standard ones, but later we will need the general case.

By Theorem 2.47, each vector $f(\boldsymbol{u_j})$ $(1 \leq j \leq n)$ has unique coordinates in the basis C of $\mathbb{R}^m$. More precisely:

$$f(\boldsymbol{u_j}) = a_{1j}\boldsymbol{v_1} + a_{2j}\boldsymbol{v_2} + \cdots + a_{mj}\boldsymbol{v_m}, \text{ that is } (f(\boldsymbol{u_j}))_C = (a_{1j}, a_{2j}, \ldots, a_{mj}).$$

For short we write

$$f(\boldsymbol{u_j}) = \sum_{i=1}^{m} a_{ij}\boldsymbol{v_i}.$$

Now we can determine the image of any vector $\boldsymbol{x}$ in $\mathbb{R}^n$. If $\boldsymbol{x} = x_1\boldsymbol{u_1} + x_2\boldsymbol{u_2} + \cdots + x_n\boldsymbol{u_n}$ (so $(\boldsymbol{x})_B = (x_1, x_2, \ldots, x_n)$), then we have:

$$\begin{aligned} f(\boldsymbol{x}) &= f(x_1\boldsymbol{u_1} + x_2\boldsymbol{u_2} + \cdots + x_n\boldsymbol{u_n}) \\ &= f(\sum_{j=1}^{n} x_j\boldsymbol{u_j}) \\ &= \sum_{j=1}^{n} x_j f(\boldsymbol{u_j}) && \text{(by linearity)} \\ &= \sum_{j=1}^{n} x_j \left(\sum_{i=1}^{m} a_{ij}\boldsymbol{v_i}\right) \\ &= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij}x_j\right)\boldsymbol{v_i}. \end{aligned}$$

Notice that $\sum_{j=1}^{n} a_{ij}x_j$ is exactly the i-th element of the $m \times 1$ matrix $A(\boldsymbol{x})_B$, where $A = (a_{ij})$ is the $m \times n$ matrix defined by $f(\boldsymbol{u_j}) = \sum_{i=1}^{m} a_{ij}\boldsymbol{v_i}$. This is saying that the coordinates with respect to basis C of $f(\boldsymbol{x})$ are just $A(\boldsymbol{x})_B$.

This gives us a very convenient way to express a linear transformation (as the matrix A) and to calculate the image of any vector.

DEFINITION **4.9.** *(Matrix of a linear transformation)*
The matrix of a linear transformation f, with respect to the basis B of the domain and the basis C of the codomain, is the matrix A whose j-th column contains the coordinates in the basis C of the image under f of the j-th basis vector of B.

If we want to emphasise the choice of bases we write $A = A_{CB}$.

When both B and C are the standard bases then we refer to the matrix A as the standard matrix *of f.*

Whenever $m = n$ we usually take $B = C$.

The argument above and this definition yield the following theorem.

THEOREM **4.10.** *Let f be a linear transformation, B a basis of its domain, C a basis of its codomain, and A_{CB} as above. Then*

$$(f(\boldsymbol{x}))_C = A_{CB}(\boldsymbol{x})_B.$$

In the case where $B = C = S$ (where S is the standard basis) then we can just write $f(\boldsymbol{x}) = A\boldsymbol{x}$, where $A = A_{SS}$ is the standard matrix for f.

Together with Example 4.6, this tells us that linear transformations are essentially the same as matrices (after you have chosen a basis of the domain and a basis of the codomain).

KEY CONCEPT **4.11.** *Let A be the matrix of a linear transformation f.*

- *The number of* **rows** *of A is the dimension of the* **codomain** *of f.*

- *The number of* **columns** *of A is the dimension of the* **domain** *of f.*

In other words, if $f : \mathbb{R}^n \to \mathbb{R}^m$ then A is an $m \times n$ matrix, whatever bases we choose for the domain and codomain.

EXAMPLE **4.12.** *(identity) The* identity matrix I_n *corresponds to the linear transformation that fixes every basis vector, and hence fixes every vector in $\mathbb{R}^n$.* □

EXAMPLE **4.13.** *(dilation) The linear transformation with matrix* $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ *(with respect to the standard basis, for both the domain and codomain) maps $\boldsymbol{e_1}$ to $2\boldsymbol{e_1}$ and $\boldsymbol{e_2}$ to $2\boldsymbol{e_2}$: it is a dilation of ratio 2 in $\mathbb{R}^2$.* □

A *dilation* is a function that maps every vector to a fixed multiple of itself: $\boldsymbol{x} \mapsto \lambda \boldsymbol{x}$, where λ is called the *ratio* of the dilation.

EXAMPLE **4.14.** *(rotation) The linear transformation with matrix* $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ *(with respect to the standard basis) maps $\boldsymbol{e_1}$ to $\boldsymbol{e_2}$ and $\boldsymbol{e_2}$ to $-\boldsymbol{e_1}$: it is the anticlockwise rotation of the plane by an angle of 90 degrees (or $\pi/2$) around the origin.* □

EXERCISE **4.3.1.** *In $\mathbb{R}^2$, an anticlockwise rotation of angle θ around the origin is a linear transformation. What is its matrix with respect to the standard basis?*

4.4 Rank-nullity theorem revisited

Remember the Rank-Nullity Theorem (Theorem 3.19): $\text{rank}(A) + \text{nullity}(A) = n$ for an $m \times n$ matrix A. We now know that A represents a linear transformation $f : \boldsymbol{x} \to A\boldsymbol{x}$, so we are going to interpret what the rank and the nullity are in terms of f.

The rank of A is the dimension of the column space of A. We have seen that the columns represent the images $f(\boldsymbol{u_j})$ for each basis vector $\boldsymbol{u_j}$ of $\mathbb{R}^n$.

THEOREM **4.15.** *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be a linear transformation and let $\{\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_n}\}$ be a basis for $\mathbb{R}^n$. Then*

$$\text{range}(f) = \textit{span}(\{f(\boldsymbol{u_1}), f(\boldsymbol{u_2}), \ldots, f(\boldsymbol{u_n})\}).$$

Proof. We first show that any element in the range of f is in the span, that is can be written as a linear combination of the vectors. Let $\boldsymbol{y} \in \mathbb{R}^m$ be in the range of f, that is $\boldsymbol{y} = f(\boldsymbol{x})$ for some x in $\mathbb{R}^n$. By Theorem 2.47, $\boldsymbol{x}$ can be written in a unique way as $\boldsymbol{x} = \alpha_1\boldsymbol{u_1} + \alpha_2\boldsymbol{u_2} + \cdots + \alpha_n\boldsymbol{u_n}$ (where the α_i's are real numbers). Using the linearity of f, it immediately follows that $f(\boldsymbol{x}) = \alpha_1 f(\boldsymbol{u_1}) + \alpha_2 f(\boldsymbol{u_2}) + \cdots + \alpha_n f(\boldsymbol{u_n})$. Hence $\boldsymbol{y} \in \text{span}(\{f(\boldsymbol{u_1}), f(\boldsymbol{u_2}), \ldots, f(\boldsymbol{u_n})\})$.

Now we need to prove the converse: that every element in the the span must be in the range of f. Let $\boldsymbol{v}$ be in $\text{span}(\{f(\boldsymbol{u_1}), f(\boldsymbol{u_2}), \ldots, f(\boldsymbol{u_n})\})$. By definition $\boldsymbol{v} = \alpha_1 f(\boldsymbol{u_1}) + \alpha_2 f(\boldsymbol{u_2}) + \cdots + \alpha_n f(\boldsymbol{u_n})$ for some

scalars $\alpha_1, \alpha_2, \ldots, \alpha_n$. Using the linearity of f, it immediately follows that $\boldsymbol{v} = f(\alpha_1 \boldsymbol{u_1} + \alpha_2 \boldsymbol{u_2} + \cdots + \alpha_n \boldsymbol{u_n})$, and so $\boldsymbol{v}$ is in the range of f since it is the image of some vector $\alpha_1 \boldsymbol{u_1} + \alpha_2 \boldsymbol{u_2} + \cdots + \alpha_n \boldsymbol{u_n}$ of $\mathbb{R}^n$. □

Therefore the column space corresponds exactly to the range of f. By Theorem 4.7, we know that the range of f is a subspace, and so has a dimension: the rank of A corresponds to the dimension of the range of f.

The nullity of A is the dimension of the null space of A. Recall that the null space of A is the set of vectors $\boldsymbol{x}$ of $\mathbb{R}^n$ such that $A\boldsymbol{x} = \boldsymbol{0}$. In terms of f, it corresponds to the vectors $\boldsymbol{x}$ of $\mathbb{R}^n$ such that $f(\boldsymbol{x}) = \boldsymbol{0}$. This set is called the *kernel* of f.

Definition 4.16. *(Kernel)*
The kernel of a linear transformation $f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is the set

$$\text{Ker}(f) = \{\boldsymbol{x} \in \mathbb{R}^n | f(\boldsymbol{x}) = \boldsymbol{0}\}.$$

The kernel of f is a subspace[1] of $\mathbb{R}^n$ and so has a dimension: the nullity of A corresponds to the dimension of the kernel of f.

[1] Try proving it!

We can now rewrite the Rank-Nullity Theorem as follows:

Theorem 4.17. *Let f be a linear transformation. Then*

$$\dim(\text{range}(f)) + \dim(\text{Ker}(f)) = \dim(\text{dom}(f)).$$

We immediately get:

Corollary 4.18. *The dimension of the range of a linear transformation is at most the dimension of its domain.*

4.5 Composition

Whenever you have two linear transformations such that the codomain of the first one is the same vector space as the domain of the second one, we can apply the first one followed by the second one.

Definition 4.19. *(Composition)*
Let $f : A \to B$ and $g : B \to C$ be functions. Then the function $g \circ f : A \to C$ defined by

$$(g \circ f)(a) = g(f(a)) \text{ for all } a \in A$$

is the composition of f by g.

Notice we read composition from right to left.

Theorem 4.20. *If $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^p$ are linear transformations, then $g \circ f$ is also a linear transformation, from $\mathbb{R}^n$ to $\mathbb{R}^p$.*

Proof. We need to prove the two conditions for a linear transformation.
For all $\boldsymbol{u}, \boldsymbol{v}$ in $\mathbb{R}^n$:

$$\begin{aligned}
(g \circ f)(\boldsymbol{u}+\boldsymbol{v}) &= g(f(\boldsymbol{u}+\boldsymbol{v})) && \text{(definition of composition)}\\
&= g(f(\boldsymbol{u})+f(\boldsymbol{v})) && (f \text{ is a linear transformation})\\
&= g(f(\boldsymbol{u}))+g(f(\boldsymbol{v}))) && (g \text{ is a linear transformation})\\
&= (g \circ f)(\boldsymbol{u})+(g \circ f)(\boldsymbol{v}). && \text{(composition definition)}
\end{aligned}$$

For all $\boldsymbol{v}$ in $\mathbb{R}^n$ and all α in $\mathbb{R}$:

$$\begin{aligned}
(g \circ f)(\alpha \boldsymbol{v}) &= g(f(\alpha \boldsymbol{v})) && \text{(definition of composition)}\\
&= g(\alpha f(\boldsymbol{v})) && (f \text{ is a linear transformation})\\
&= \alpha g(f(\boldsymbol{v})) && (g \text{ is a linear transformation})\\
&= \alpha (g \circ f)(\boldsymbol{v}). && \text{(definition of composition)}
\end{aligned}$$

□

Let $B = \{\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_n}\}$, $C = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_m}\}$, $D = \{\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_p}\}$ be bases of $\mathbb{R}^n$, $\mathbb{R}^m$, and $\mathbb{R}^p$ respectively. Let $F = F_{CB} = (f_{ij})$ be the matrix corresponding to f with respect to the bases B of the domain and C of the codomain. Let $G = G_{DC} = (g_{ij})$ be the matrix corresponding to g with respect to the bases C of the domain and D of the codomain. So F is an $m \times n$ matrix and G is a $p \times m$ matrix. Let us look at the image of $\boldsymbol{u_1}$ under $g \circ f$.

We first apply f, so the image $f(\boldsymbol{u_1})$ corresponds to the first column of A: $f(\boldsymbol{u_1}) = f_{11}\boldsymbol{v_1} + f_{21}\boldsymbol{v_2} + \cdots + f_{m1}\boldsymbol{v_m} = \sum_{i=1}^{m} f_{i1}\boldsymbol{v_i}$. Then we apply g to $f(\boldsymbol{u_1})$:

$$\begin{aligned}
(g \circ f)(\boldsymbol{u_1}) &= g\left(\sum_{i=1}^{m} f_{i1}\boldsymbol{v_i}\right)\\
&= \sum_{i=1}^{m} f_{i1} g(\boldsymbol{v_i}) && (g \text{ is a linear transformation})\\
&= \sum_{i=1}^{m} f_{i1}\left(\sum_{j=1}^{p} g_{ji}\boldsymbol{w_j}\right)\\
&= \sum_{j=1}^{p}\left(\sum_{i=1}^{m} f_{i1} g_{ji}\right)\boldsymbol{w_j} && \text{(rearranging terms)}\\
&= \sum_{j=1}^{p}\left(\sum_{i=1}^{m} g_{ji} f_{i1}\right)\boldsymbol{w_j}\\
&= \sum_{j=1}^{p} (GF)_{j1}\boldsymbol{w_j}.
\end{aligned}$$

This says that the first column of the matrix GF yields the coordinates of $(g \circ f)(\boldsymbol{u_1})$ with respect to the basis D. We can do the same calculation with any $\boldsymbol{u_j}$ $(1 \leq j \leq n)$ to see that the image $(g \circ f)(\boldsymbol{u_j})$ corresponds exactly to the j-th column of the matrix GF. Hence the matrix corresponding to $g \circ f$ with respect to the basis B of the domain and the basis D of the codomain is $GF = G_{DC}F_{CB}$.

KEY CONCEPT **4.21.** *Composition of linear transformations is the same thing as multiplication of the corresponding matrices, where we order the matrices from right to left, just as composition.*

You may have thought that matrix multiplication was defined in a strange way: it was defined precisely so that it corresponds with composition of linear transformations.

4.6 Inverses

An *inverse function* is a function that "undoes" another function: if $f(x) = y$, the inverse function g maps y to x.

DEFINITION **4.22.** *(Inverse function)*
Let $f : A \longrightarrow B$ be a function. We say that f is invertible *if there exists a function $g : B \longrightarrow A$ such that $g(f(x)) = x$, meaning that $g \circ f$ is the identity function. The inverse function g is then uniquely determined by f and is denoted by f^{-1}. We have*

$$(f^{-1} \circ f)(a) = a \text{ for all } a \in A \text{ and } (f \circ f^{-1})(b) = b \text{ for all } b \in B$$

It is routine to show that f is invertible if and only if it is *bijective*.

DEFINITION **4.23.** *(Bijective function)*
Let $f : A \longrightarrow B$ be a function.

- *We say that f is* one-to-one *if no two elements of A have the same image. More formally:*

$$f(a_1) = f(a_2) \Rightarrow a_1 = a_2$$

- *We say that f is* onto *if the range of f is equal to the codomain B.*
- *We say that f is* bijective *if f is both one-to-one and onto.*

We now look at the particular case where f is a linear transformation.

THEOREM **4.24.** *Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be a linear transformation.*

1. *If f is invertible then f^{-1} is a linear transformation.*
2. *f is invertible if and only if $n = m$ and* range$(f) = \mathbb{R}^n$.

Proof. 1. Suppose first that f is invertible. We need to show that $f^{-1} : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ satisfies the two properties of a linear transformation. Take $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^m$. Then:

$$\begin{aligned} f\left(f^{-1}(\boldsymbol{u}+\boldsymbol{v})\right) &= \left(f \circ f^{-1}\right)(\boldsymbol{u}+\boldsymbol{v}) \\ &= \boldsymbol{u}+\boldsymbol{v} = \left(f \circ f^{-1}\right)(\boldsymbol{u}) + \left(f \circ f^{-1}\right)(\boldsymbol{v}) \\ &= f\left(f^{-1}(\boldsymbol{u}) + f^{-1}(\boldsymbol{v})\right). \end{aligned}$$

Since f is one-to-one, it follows that $f^{-1}(\boldsymbol{u}+\boldsymbol{v}) = f^{-1}(\boldsymbol{u}) + f^{-1}(\boldsymbol{v})$. The other property is proved in a similar fashion.

2. Suppose first that f is invertible. Since f is onto, $\text{range}(f) = \mathbb{R}^m$, which has dimension m. By Corollary 4.18, we get that $m \leq n$. Now the inverse function f^{-1} is also onto, so that $\text{range}(f^{-1}) = \mathbb{R}^n$, which has dimension n. Since f^{-1} is a linear transformation, we can apply Corollary 4.18 to f^{-1}, and so $n \leq m$. Therefore $m = n$ and $\text{range}(f) = \mathbb{R}^m = \mathbb{R}^n$.

   Conversely, suppose that $n = m$ and $\text{range}(f) = \mathbb{R}^n$. It follows immediately that f is onto so we only need to prove that f is one-to-one. Suppose $f(\boldsymbol{u}) = f(\boldsymbol{v})$ for some $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$, we want to show that $\boldsymbol{u} = \boldsymbol{v}$. By linearity $f(\boldsymbol{u} - \boldsymbol{v}) = \boldsymbol{0}$, that is $\boldsymbol{u} - \boldsymbol{v} \in \text{Ker}(f)$. By Theorem 4.17, $\dim(\text{Ker}(f)) = \dim(\text{dom}(f)) - \dim(\text{range}(f)) = n - n = 0$. The only subspace of dimension 0 is the trivial subspace $\{\boldsymbol{0}\}$, so $\text{Ker}(f) = \{\boldsymbol{0}\}$. It follows that $\boldsymbol{u} - \boldsymbol{v} = \boldsymbol{0}$, and so $\boldsymbol{u} = \boldsymbol{v}$. □

Consider an invertible linear transformation f. By Theorem 4.24, the domain and codomain of f are the same, say $\mathbb{R}^n$, and f^{-1} is also a linear transformation with domain and codomain $\mathbb{R}^n$. Let A be the matrix corresponding to f and B be the matrix corresponding to f^{-1} (all with respect to the standard basis, say); both are $n \times n$ matrices. We have seen in Example 4.12 that the matrix corresponding to the identity function is the identity matrix I_n. By the Key Concept 4.21, we have that $BA = I_n$. In other words, B is the inverse matrix of A.

Hence we have:

THEOREM **4.25.** *The matrix corresponding to the inverse of an invertible linear transformation f is the inverse of the matrix corresponding to f (with respect to a chosen basis).*

5
Change of basis

In Chapter 2, we saw that choosing a basis for a subspace (for instance the whole vector space $\mathbb{R}^n$) determines coordinates. In Chapter 4, we saw how a choice of bases for the domain and for the codomain allows us to write a linear transformation as a matrix. In this Chapter, we will study ways to change bases in these two cases.

5.1 Change of basis for vectors

A change of coordinates from one basis to another can be achieved by multiplication of the given coordinate vector by a so-called *change of coordinates matrix*. Consider a subspace V of dimension n of $\mathbb{R}^m$ and two different bases for V:

$$B = \{\boldsymbol{u_1}, \boldsymbol{u_2}, \cdots, \boldsymbol{u_n}\} \quad \text{and} \quad C = \{\mathbf{w_1}, \mathbf{w_2}, \cdots, \mathbf{w_n}\}.$$

A given vector $\boldsymbol{v} \in V$ will have cordinates in each of these bases:

$$(\boldsymbol{v})_B = (\alpha_1, \alpha_2, \ldots, \alpha_n) \quad , \quad (\boldsymbol{v})_C = (\beta_1, \beta_2, \ldots, \beta_n),$$

that is, $\boldsymbol{v} = \sum_{k=1}^{n} \alpha_k \boldsymbol{u_k} = \sum_{i=1}^{n} \beta_i \boldsymbol{w_i}$. Our task is to find an *invertible* $n \times n$ matrix P_{CB} for which

$$(\boldsymbol{v})_C = P_{CB}\,(\boldsymbol{v})_B \quad \text{and} \quad (\boldsymbol{v})_B = P_{BC}\,(\boldsymbol{v})_C$$

where $P_{BC} = P_{CB}^{-1}$. That is, pre-multiplication by P_{CB} will convert coordinates in basis B to coordinates in basis C and pre-multiplication by P_{BC} will convert those in C to those in B.

Let $P_{CB} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}$.

We compute $P_{CB}\,(\boldsymbol{v})_B$ by computing the $(i,1)$-entry for each i (recall the formula for matrix multiplication in Definition 3.1 and that $P_{CB}\,(\boldsymbol{v})_B$is an $n \times 1$ matrix):

$$(P_{CB}\,(\boldsymbol{v})_B)_{i1} = \sum_{k=1}^{n} p_{ik}\alpha_k$$

Therefore $\beta_i = \sum_{k=1}^{n} p_{ik}\alpha_k$ for each i.

It follows that

$$\begin{aligned} \boldsymbol{v} &= \sum_{i=1}^{n} \beta_i \boldsymbol{w_i} \\ &= \sum_{i=1}^{n} \left(\sum_{k=1}^{n} p_{ik}\alpha_k \right) \boldsymbol{w_i} \\ &= \sum_{k=1}^{n} \alpha_k \left(\sum_{i=1}^{n} p_{ik} \boldsymbol{w_i} \right) \end{aligned}$$

On the other hand,

$$\boldsymbol{v} = \sum_{k=1}^{n} \alpha_k \boldsymbol{u_k},$$

so

$$\boldsymbol{u_k} = \sum_{i=1}^{n} p_{ik} \boldsymbol{w_i}.$$

In other words, the coefficients of the matrix P_{CB} satisfy:

$$\begin{aligned} (\boldsymbol{u_1})_C &= (p_{11}, p_{21}, \ldots, p_{n1}) \\ (\boldsymbol{u_2})_C &= (p_{12}, p_{22}, \ldots, p_{n2}) \\ &\vdots \\ (\boldsymbol{u_n})_C &= (p_{1n}, p_{2n}, \ldots, p_{nn}). \end{aligned}$$

That is each column corresponds to the coordinates of the vectors in the basis B with respect to basis C. For this reason, we might prefer to write the vectors $(\boldsymbol{u_i})_C$ as column vectors in this case:

$$(\boldsymbol{u_1})_C = \begin{bmatrix} p_{11} \\ \vdots \\ p_{n1} \end{bmatrix} \quad , \quad (\boldsymbol{u_2})_C = \begin{bmatrix} p_{12} \\ \vdots \\ p_{n2} \end{bmatrix} \quad , \quad \cdots \quad (\boldsymbol{u_n})_C = \begin{bmatrix} p_{1n} \\ \vdots \\ p_{nn} \end{bmatrix}.$$

KEY CONCEPT **5.1.** *To convert coordinates of the vector $\boldsymbol{v}$ from basis B to basis C we perform the matrix multiplication*

$$(\boldsymbol{v})_C = P_{CB} (\boldsymbol{v})_B ,$$

where P_{CB} is the matrix whose i-th column is the coordinates with respect to basis C of the i-th basis vector in B.

Moreover $P_{BC} = P_{CB}^{-1}$.

The matrix P_{CB} will be *invertible* because the elements of basis B are *linearly independent* and the elements of basis C are *linearly independent*.

EXAMPLE **5.2.** *(Example 2.50 revisited again) Recall we considered $B = \{(1,-1,0),(1,0,-1)\}$ and $C = \{(0,1,-1),(1,-2,1)\}$, two bases for the vector space V. To find P_{CB} we need to determine the coordinates of the basis elements of B with respect to basis C.*

The required coordinates are

$$(1,-1,0)_C = \begin{bmatrix}1\\1\end{bmatrix} \quad , \quad (1,0,-1)_C = \begin{bmatrix}2\\1\end{bmatrix}$$

and hence

$$P_{CB} = \begin{bmatrix}1 & 2\\1 & 1\end{bmatrix} \quad \text{and} \quad P_{BC} = P_{CB}^{-1} = \begin{bmatrix}-1 & 2\\1 & -1\end{bmatrix}.$$

We can verify these. Recall that $(\boldsymbol{v})_B = (3,-2)$ *and* $(\boldsymbol{w})_C = (6,1)$*. Then*

$$(\boldsymbol{v})_C = P_{CB}\,(\boldsymbol{v})_B = \begin{bmatrix}1 & 2\\1 & 1\end{bmatrix}\begin{bmatrix}3\\-2\end{bmatrix} = \begin{bmatrix}-1\\1\end{bmatrix}$$

which agrees with what we got before, and

$$(\boldsymbol{w})_B = P_{BC}\,(\boldsymbol{w})_C = \begin{bmatrix}-1 & 2\\1 & -1\end{bmatrix}\begin{bmatrix}6\\1\end{bmatrix} = \begin{bmatrix}-4\\5\end{bmatrix}$$

which also agrees with what we got before. □

5.2 *Change of bases for linear transformations*

Recall that linear transfomations f can be represented using matrices:

$$f(\boldsymbol{x}) = A_{SS}\boldsymbol{x}.$$

For example, counterclockwise rotation about the origin through an angle θ in $\mathbb{R}^2$ using the standard basis for the original and transformed vectors has transformation matrix

$$A_{SS} = \begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}.$$

We have used the subscript SS to indicate that the standard basis is being used to represent the original vectors (domain of the linear transformation) and also the rotated vectors (codomain of the linear transformation). For example, the vector $\boldsymbol{e}_2 = [0,1]$ under a rotation of $\dfrac{\pi}{2}$ becomes

$$\begin{bmatrix}0 & -1\\1 & 0\end{bmatrix}\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}-1\\0\end{bmatrix} \;=\; -\boldsymbol{e}_1, \text{ as expected.}$$

If we desire to use different bases to represent the coordinates of the vectors, say basis B for the domain and basis C for codomain, then recall from Definition 4.9 that we label the transformation matrix A_{CB} and the linear transformation will be

$$(f(\boldsymbol{x}))_C = A_{CB}\,(\boldsymbol{x})_B\,. \tag{5.1}$$

We need a way of deducing A_{CB}. This can be achieved by employing the change of basis matrices P_{BS} and P_{CS} where

$$(\boldsymbol{x})_B = P_{BS}\,(\boldsymbol{x})_S \quad \text{and} \quad (f(\boldsymbol{x}))_C = P_{CS}\,(f(\boldsymbol{x}))_S\,.$$

Substitution of these formulae in Equation (5.1) gives

$$P_{CS}\,(f(\boldsymbol{x}))_S = A_{CB}P_{BS}\,(\boldsymbol{x})_S \quad \Rightarrow \quad (f(\boldsymbol{x}))_S = P_{SC}A_{CB}P_{BS}\,(\boldsymbol{x})_S$$

(recalling that $P_{CS} = P_{SC}^{-1}$). Using the standard basis for the transformation would be $(f(\boldsymbol{x}))_S = A_{SS}\,(\boldsymbol{x})_S$ and hence we must have $A_{SS} = P_{SC}A_{CB}P_{BS}$, which we can rearrange to get the linear transformation change of basis formula

$$A_{CB} = P_{CS}A_{SS}P_{SB}.$$

Note that if we use the same basis B for both the domain and codomain then we have

$$A_{BB} = P_{BS}A_{SS}P_{SB}. \tag{5.2}$$

Two matrices M and N are *similar* if there exists an invertible matrix Q such that $N = Q^{-1}MQ$. Equation (5.2) tells us that all linear transformation matrices in which the same basis is used for the domain and codomain are similar.

EXAMPLE **5.3.** *We determine the change of basis matrix from a basis $B = \{\boldsymbol{u_1}, \boldsymbol{u_2}, \cdots, \boldsymbol{u_n}\}$ to the standard basis S. We note that $(\boldsymbol{u_i})_S = \boldsymbol{u_i}$ for all $i = 1, \cdots, n$ and hence we can immediately write*

$$P_{SB} = [\boldsymbol{u_1}\ \boldsymbol{u_2}\ \cdots\ \boldsymbol{u_n}]$$

where the i-th column is the vector $\boldsymbol{u_i}$ (written as usual as an n-tuple).

□

EXAMPLE **5.4.** *We determine the transformation matrix for counterclockwise rotation through an angle θ in $\mathbb{R}^2$ using the basis*

$$B = \{(1,-1),(1,1)\}$$

to represent both the original vectors and the transformed vectors.

We need to calculate P_{SB} and P_{BS}. We can write down immediately that

$$P_{SB} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \quad \text{and that} \quad P_{BS} = P_{SB}^{-1} = \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix}$$

and so the desired transformation matrix is

$$\begin{aligned} A_{BB} = P_{BS}A_{SS}P_{SB} &= \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} \cos\theta + \sin\theta & \cos\theta - \sin\theta \\ \sin\theta - \cos\theta & \sin\theta + \cos\theta \end{bmatrix} \\ &= \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \end{aligned}$$

That is, $A_{BB} = A_{SS}$, but if we think about the geometry then this makes sense. □

EXAMPLE **5.5.** *We determine the transformation matrix for counterclockwise rotation through an angle θ in $\mathbb{R}^2$ using the basis*

$$C = \{(1,0),(1,1)\}$$

to represent both the original vectors and the transformed vectors.

We need to calculate P_{SC} and P_{CS}. We can write down immediately that

$$P_{SC} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad \text{and that} \quad P_{CS} = P_{SC}^{-1} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$$

and so the desired transformation matrix is

$$\begin{aligned} A_{CC} = P_{CS}A_{SS}P_{SC} &= \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & \cos\theta - \sin\theta \\ \sin\theta & \sin\theta + \cos\theta \end{bmatrix} \\ &= \begin{bmatrix} \cos\theta - \sin\theta & -2\sin\theta \\ \sin\theta & \sin\theta + \cos\theta \end{bmatrix}. \end{aligned}$$

□

EXAMPLE **5.6.** *The matrix of a particular linear transfomation in $\mathbb{R}^3$ in which the standard basis has been used for both the domain and codomain is*

$$A_{SS} = \begin{bmatrix} 1 & 3 & 4 \\ 2 & -1 & 1 \\ -3 & 5 & 1 \end{bmatrix}.$$

Determine the matrix of the linear transfomation if basis B is used for the domain and basis C is used for the codomain, where

$$B = \{(0,0,1),(0,1,0),(1,0,0)\} \quad , \quad C = \{(1,1,1),(0,1,1),(0,0,1)\}.$$

Solution. *First we need to calculate P_{CS} and P_{SB}. We can immediately write down P_{SB} and P_{SC}, and use the fact that $P_{CS} = P_{SC}^{-1}$. We have*

$$P_{SB} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad P_{SC} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \quad \Rightarrow \quad P_{CS} = P_{SC}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

and hence

$$A_{CB} = P_{CS}A_{SS}P_{SB} = \begin{bmatrix} 4 & 3 & 1 \\ -3 & -4 & 1 \\ 0 & 6 & -5 \end{bmatrix}.$$

□

Of course we would like to make the matrix of the linear transformation *simpler* by choosing an appropriate basis B (most often, if the matrix is square, we chose the same basis for the domain and codomain).

EXAMPLE **5.7.** *Consider the linear transformation with matrix*

$$A_{SS} = \begin{bmatrix} 2 & 6 \\ 3 & 5 \end{bmatrix}.$$

Find A_{BB} where $B = \{(1,1),(-2,1)\}$.

Solution. *We need to calculate P_{SB} and P_{BS}. We can write down immediately that*

$$P_{SB} = \begin{bmatrix} 1 & -2 \\ 1 & 1 \end{bmatrix} \quad \text{and that} \quad P_{BS} = P_{SB}^{-1} = \begin{bmatrix} 1/3 & 2/3 \\ -1/3 & 1/3 \end{bmatrix}$$

and so the desired transformation matrix is

$$\begin{aligned} A_{BB} = P_{BS}A_{SS}P_{SB} &= \begin{bmatrix} 1/3 & 2/3 \\ -1/3 & 1/3 \end{bmatrix} \begin{bmatrix} 2 & 6 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1/3 & 2/3 \\ -1/3 & 1/3 \end{bmatrix} \begin{bmatrix} 8 & 2 \\ 8 & -1 \end{bmatrix} \\ &= \begin{bmatrix} 8 & 0 \\ 0 & -1 \end{bmatrix}. \end{aligned}$$

The matrix A_{BB} has a very simple form, which is nice for calculations. It also makes it easy to visualise. The first vector in the basis B is stretched 8 times, and the second vector is mapped onto its opposite. □

In the next chapter, we will learn how to determine a nice basis B for a given linear transformation and its corresponding matrix.

6
Eigenvalues and eigenvectors

6.1 Introduction

Matrix multiplication usually results in a change of direction, for example,

$$\begin{bmatrix} 2 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \end{bmatrix} = \begin{bmatrix} 2 \\ 13 \end{bmatrix} \quad \text{which is not parallel to} \quad \begin{bmatrix} 1 \\ 4 \end{bmatrix}.$$

The *eigenvectors* of a given (square) matrix A are those special non-zero vectors $\boldsymbol{v}$ that map to multiples of themselves under multiplication by the matrix A, and the *eigenvalues* of A are the corresponding scale factors.

DEFINITION **6.1.** *(Eigenvectors and eigenvalues)*
Let A be a square matrix. An eigenvector *of A is a vector $\boldsymbol{v} \neq \mathbf{0}$ such that*

$$A\boldsymbol{v} = \lambda\boldsymbol{v} \text{ for some scalar } \lambda.$$

An eigenvalue *of A is a scalar λ such that*

$$A\boldsymbol{v} = \lambda\boldsymbol{v} \text{ for some vector } \boldsymbol{v} \neq \mathbf{0}.$$

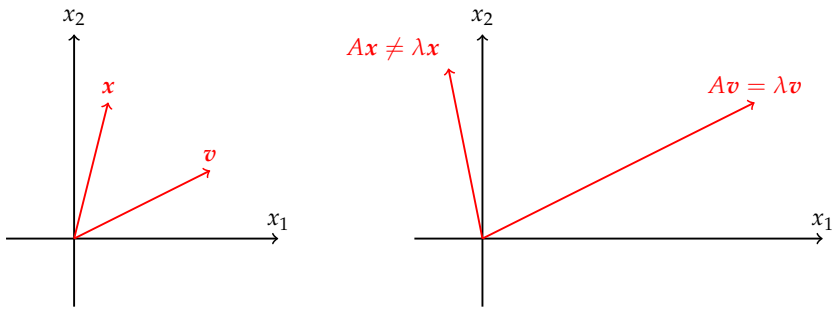
Geometrically, the eigenvectors of a matrix A are stretched (or shrunk) on multiplication by A, whereas any other vector is rotated as well as being stretched or shrunk.

Note that by definition $\mathbf{0}$ *is not an eigenvector* but we do allow 0 to be an eigenvalue.

EXAMPLE **6.2.** *Recall Example 5.7, and we compute $A\boldsymbol{v}$ for each $\boldsymbol{v}$ in the basis $B = \{(1,1), (-2,1)\}$.*

$$\begin{bmatrix} 2 & 6 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 8 \\ 8 \end{bmatrix} = 8 \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

$$\begin{bmatrix} 2 & 6 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} -2 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \end{bmatrix} = -1 \begin{bmatrix} -2 \\ 1 \end{bmatrix}.$$



Hence 8 *and* -1 *are eigenvalues of the matrix* $\begin{bmatrix} 2 & 6 \\ 3 & 5 \end{bmatrix}$ *with* $\lambda = 8$ *having corresponding eigenvector* $(1,1)$ *and* $\lambda = -1$ *having corresponding eigenvector* $(-2,1)$. □

Example 5.7 illustrates that if we change the basis to eigenvectors we get a diagonal matrix. In other words, solving the eigenvalue-eigenvector problem is equivalent to finding a basis in which the linear transformation has a particularly simple (in this case, diagonal) matrix representation.

Definition 6.3. *(Eigenspace)*
Let λ *be an eigenvalue for* A. *Then the* eigenspace *corresponding to the eigenvalue* λ *is the set*

$$E_\lambda = \{v \mid Av = \lambda v\},$$

that is, the set of eigenvectors corresponding to λ *together with the zero vector.*

Theorem 6.4. *Let* λ *be an eigenvalue for the* $n \times n$ *matrix* A. *Then the eigenspace* E_λ *is a subspace of* $\mathbb{R}^n$ *of dimension at least* 1.

Proof. We need to show the three subspace conditions.

(S1) $A\mathbf{0} = \lambda\mathbf{0}$ so $\mathbf{0} \in E_\lambda$.

(S2) Let $u, v \in E_\lambda$. Then $Au = \lambda u$ and $Av = \lambda v$. We want to show that $u + v \in E_\lambda$ so we test the membership condition:

$$A(u+v) = Au + Av = \lambda u + \lambda v = \lambda(u+v).$$

(S3) Let $\alpha \in \mathbb{R}$ and $v \in E_\lambda$. Then $Av = \lambda v$. We want to show that $\alpha v \in E_\lambda$ so we test the membership condition:

$$A(\alpha v) = \alpha(Av) = \alpha(\lambda v) = \lambda(\alpha v).$$

By definition of eigenvalue, there exists a non-zero vector v such that $Av = \lambda v$, so we can construct a basis for E_λ containing at least the vector v. Thus the dimension of E_λ is at least 1. □

Geometrically, for $n = 3$, eigenspaces are lines or planes through the origin in $\mathbb{R}^3$ (or sometimes even the whole of $\mathbb{R}^3$).

In Example 6.2, we have at least two eigenvalues 8 and -1 and each eigenspace contains at least all the scalar multiples of the eigenvectors we found. In this case, there are only two eigenspaces, both of dimension 1. The eigenspaces for Example 6.2 are shown in Figure 6.1.

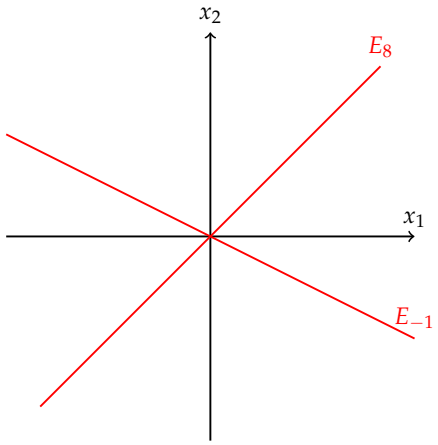


Figure 6.1: Eigenspaces for Example 6.2.

THEOREM **6.5.** *Consider a square matrix A and its eigenvalues/eigenspaces.*

1. *If 0 is an eigenvalue, then the eigenspace E_0 is exactly the null space of A.*
2. *0 is an eigenvalue if and only if the null space of A has dimension at least 1.*
3. *For each eigenvalue $\lambda \neq 0$, E_λ is a subspace of the column space of A.*

Proof. Suppose 0 is an eigenvalue. Then $E_0 = \{\boldsymbol{v} \mid A\boldsymbol{v} = \mathbf{0}\}$ is the null space and has dimension at least 1 by Theorem 6.4. If 0 is not an eigenvalue, then the only vector $\boldsymbol{v}$ such that $A\boldsymbol{v} = \mathbf{0}$ is the zero vector, so the null space of A has dimension 0. This proves the first two statements.

Recall that the column space of A is equal to $\{A\boldsymbol{x} | \boldsymbol{x} \in \mathbb{R}^n\}$. For an eigenvalue $\lambda \neq 0$, each eigenvector $\boldsymbol{v} = \frac{1}{\lambda} A\boldsymbol{v} = A(\frac{1}{\lambda}\boldsymbol{v})$, and so $\boldsymbol{v}$ belongs to the column space of A. It follows that E_λ is a subspace of the column space of A. □

6.2 *Finding eigenvalues and eigenvectors*

Let A be a given $n \times n$ matrix. Recall that the algebraic definition of an eigenvalue-eigenvector pair is

$$A\boldsymbol{v} \quad = \quad \lambda\boldsymbol{v}$$

where λ is a scalar and $\boldsymbol{v}$ is a nonzero column vector of length n. We begin by rearranging and regrouping, and noting that $\boldsymbol{v} = I\boldsymbol{v}$

for any vector $\boldsymbol{v}$, where I is the $n \times n$ identity matrix, as follows:

$$A\boldsymbol{v} - \lambda\boldsymbol{v} = \mathbf{0}$$

$$A\boldsymbol{v} - \lambda I\boldsymbol{v} = \mathbf{0}$$

$$(A - \lambda I)\boldsymbol{v} = \mathbf{0}.$$

This is a homogeneous system of linear equations for the components of $\boldsymbol{v}$, with augmented matrix $[A - \lambda I|\mathbf{0}]$. If the matrix $A - \lambda\mathbf{I}$ were invertible then the solution would simply be $\boldsymbol{v} = \mathbf{0}$ but this is not allowed by definition and in any case would be of no practical use in applications. We hence require that $A - \lambda\mathbf{I}$ be not invertible and, by Theorem 3.43, this will be the case if

$$\det(A - \lambda I) = 0. \tag{6.1}$$

When we evaluate this determinant we will have a polynomial equation of degree n in the unknown λ. The solutions of this equation will be the required eigenvalues. Equation (6.1) is called the *characteristic equation* of the matrix A. The polynomial $\det(A - \lambda I)$ is called the *characteristic polynomial* of A.

EXAMPLE **6.6.** *(Example 6.2 revisited) We start by forming*

$$A - \lambda I = \begin{bmatrix} 2 & 6 \\ 3 & 5 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2-\lambda & 6 \\ 3 & 5-\lambda \end{bmatrix}.$$

The determinant of this matrix is

$$\det(A - \lambda I) = (2-\lambda)(5-\lambda) - 18 = \lambda^2 - 7\lambda - 8$$

and hence the characteristic equation is

$$\lambda^2 - 7\lambda - 8 = 0 \quad \Rightarrow \quad (\lambda - 8)(\lambda + 1) = 0 \quad \Rightarrow \quad \lambda = 8, -1.$$

That is, the eigenvalues of the given matrix A are $\lambda = 8$ and $\lambda = -1$.

For each eigenvalue λ we must solve the system

$$(A - \lambda I)\boldsymbol{v} = \mathbf{0}$$

to determine the corresponding eigenspace. In other words we must solve the system with augmented matrix $[A - \lambda I|\mathbf{0}]$, using the techniques learned in Chapter 1. When $\lambda = 8$ we have

$$\left[\begin{array}{cc|c} -6 & 6 & 0 \\ 3 & -3 & 0 \end{array}\right] \quad \Rightarrow \quad E_8 = span((1,1)).$$

When $\lambda = -1$ we have

$$\left[\begin{array}{cc|c} 3 & 6 & 0 \\ 3 & 6 & 0 \end{array}\right] \quad \Rightarrow \quad E_{-1} = span((-2,1)).$$

In solving these systems of equations we end up with the complete eigenspace in each case. For reasons that will become clear shortly, it is useful to determine a basis for each eigenspace. □

It is important to note that the reduced row echelon form of $(A - \lambda I)$ *will always have at least one row of zeros*. The number of zero rows equals the number of free parameters, so is the dimension of the eigenspace.

EXAMPLE **6.7.** *Consider the upper triangular matrix*

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

Since A is upper triangular, so is $A - \lambda I$ and hence its determinant is just the product of the diagonal. Thus the characteristic equation is

$$(1-\lambda)(3-\lambda)(4-\lambda) = 0 \quad \Rightarrow \quad \lambda = 1, 3, 4.$$

Note that these are in fact the diagonal elements of A. *The respective eigenspaces are*

$$E_1 = span((1,0,0)), \quad E_3 = span((1,1,0)), \quad E_4 = span((16,15,3)).$$

□

The eigenvalues, and corresponding eigenvectors, could be complex-valued. If the matrix A is real-valued then the eigenvalues, that is, the roots of the characteristic polynomial, will occur in complex conjugate pairs.

EXAMPLE **6.8.** *Consider*

$$A = \begin{bmatrix} 2 & 1 \\ -5 & 4 \end{bmatrix}.$$

The characteristic polynomial is $\lambda^2 - 6\lambda + 13 = 0$. The eigenvalues are $\lambda = 3 + 2i$ and $\lambda = 3 - 2i$. When we solve $(A - \lambda I)\boldsymbol{v} = \mathbf{0}$ we will get solutions containing complex numbers. Although we can't interpret them as vectors in $\mathbb{R}^2$ there are many applications (particularly in Engineering) in which there is a natural interpretation in terms of the problem under investigation. The corresponding eigenvectors are

$$\begin{bmatrix} 1-2i \\ 5 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1+2i \\ 5 \end{bmatrix}.$$

□

The characteristic polynomial may have repeated roots. If it factors into the form

$$(\lambda_1 - \lambda)^{m_1} \cdots (\lambda_j - \lambda)^{m_j} \cdots (\lambda_p - \lambda)^{m_p}$$

we say that the *algebraic multiplicity* of the eigenvalue λ_j is m_j. For example, if the characteristic polynomial were $(\lambda + 3)(\lambda - 2)^4(\lambda - 5)$ then the algebraic multiplicity of the eigenvalue 2 would be 4.

EXAMPLE **6.9.** *Consider*

$$A = \begin{bmatrix} -2 & 2 & -3 \\ 2 & 1 & -6 \\ -1 & -2 & 0 \end{bmatrix}.$$

The characteristic equation is

$$-\lambda^3 - \lambda^2 + 21\lambda + 45 = 0$$

and so

$$(3+\lambda)^2(5-\lambda) = 0 \quad \Rightarrow \quad \lambda = -3, -3, 5.$$

Repeating the root -3 reflects the fact that $\lambda = -3$ has algebraic multiplicity 2.

To find the eigenvectors corresponding to $\lambda = 5$ we solve

$$(A - 5I)\boldsymbol{v} = \boldsymbol{0} \quad \Rightarrow \quad \left[\begin{array}{ccc|c} -7 & 2 & -3 & 0 \\ 2 & -4 & -6 & 0 \\ -1 & -2 & -5 & 0 \end{array}\right].$$

After some work we arrive at the reduced row echelon form

$$\left[\begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right].$$

Note that we have one row of zeros. The solution will therefore involve one free parameter, namely v_3. We readily get the solution $v_1 = -v_3$ and $v_2 = -2v_3$. Hence the eigenspace corresponding to $\lambda = 5$ is

$$E_5 = \{(-v_3, -2v_3, v_3)\} = span((-1, -2, 1)).$$

Similarly, to find the eigenvectors corresponding to $\lambda = -3$ we solve

$$(A + 3I)\boldsymbol{v} = \boldsymbol{0} \quad \Rightarrow \quad \left[\begin{array}{ccc|c} 1 & 2 & -3 & 0 \\ 2 & 4 & -6 & 0 \\ -1 & -2 & 3 & 0 \end{array}\right].$$

The reduced row echelon form is

$$\left[\begin{array}{ccc|c} 1 & 2 & -3 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right],$$

which has two rows of zeros and hence the solution will involve two free parameters. The eigenspace corresponding to $\lambda = -3$ is

$$E_{-3} = \{(-2v_2 + 3v_3, v_2, v_3)\} = span((-2, 1, 0), (3, 0, 1)).$$

□

The dimension of the eigenspace of an eigenvalue is called its *geometric multiplicity*. In the above example the geometric multiplicity of $\lambda = -3$ is 2 and that of $\lambda = 5$ is 1. Note that the eigenspace corresponding to $\lambda = -3$ is a plane through the origin and the eigenspace corresponding to $\lambda = 5$ is a line through the origin, as displayed in Figure 6.2. Moreover $\{(-1, -2, 1), (-2, 1, 0), (3, 0, 1)\}$ is a basis of $\mathbb{R}^3$ all of whose elements are eigenvectors.

We summarise the two definitions of multiplicity.

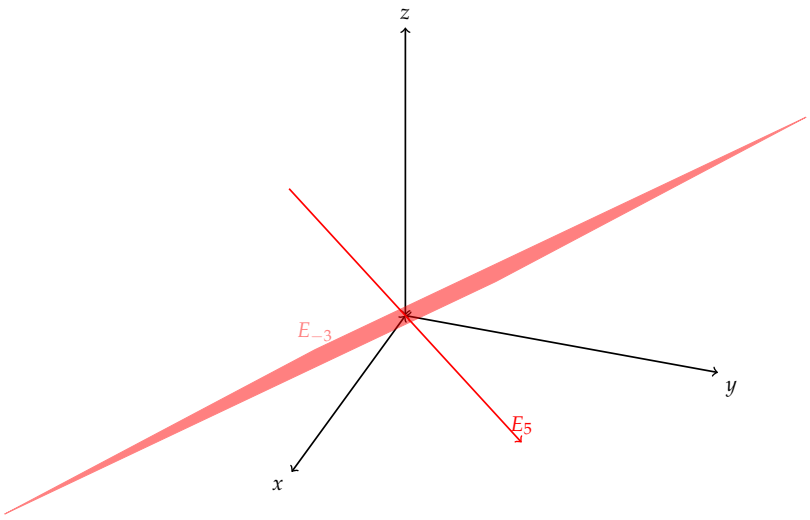


Figure 6.2: The eigenspaces for Example 6.9.

DEFINITION **6.10.** *(Multiplicity of eigenvalue)*
Let λ_i be an eigenvalue of the matrix A. The **geometric** *multiplicity of λ_i is $\dim(E_{\lambda_i})$, while the* **algebraic** *multiplicity of λ_i is the number of factors $(\lambda_i - \lambda)$ in the factorisation of the characteristic polynomial of A.*

It can be proved that the geometric multiplicity of an eigenvalue is always at most its algebraic multiplicity.

6.3 *Some properties of eigenvalues and eigenvectors*

Let A be an $n \times n$ matrix with eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$, where we include all complex and repeated eigenvalues. Then:

- The determinant of the matrix A equals the product of the eigenvalues:
$$\det(A) = \lambda_1 \lambda_2 \ldots \lambda_n.$$
- The *trace* of a square matrix is the sum of its diagonal entries. The trace of the matrix A equals the sum of the eigenvalues:
$$\text{trace}(A) = a_{11} + a_{22} + \cdots + a_{nn} = \lambda_1 + \lambda_2 + \cdots + \lambda_n.$$

Note that in both of these formulae *all n eigenvalues must be counted*.

- The eigenvalues of A^{-1} (if it exists) are
$$\frac{1}{\lambda_1}, \quad \frac{1}{\lambda_2}, \quad \cdots, \quad \frac{1}{\lambda_n}.$$

- The eigenvalues of A^T (that is, the transpose of A) are the same as for the matrix A:

$$\lambda_1, \quad \lambda_2 \quad , \cdots , \quad \lambda_n.$$

- If k is a scalar then the eigenvalues of the matrix kA are

$$k\lambda_1, \quad k\lambda_2 \quad , \cdots , \quad k\lambda_n.$$

- If k is a scalar and I the identity matrix then the eigenvalues of the matrix $A + kI$ are

$$\lambda_1 + k, \quad \lambda_2 + k \quad , \cdots , \quad \lambda_n + k.$$

- If k is a positive integer then the eigenvalues of A^k are

$$\lambda_1^k, \quad \lambda_2^k \quad , \cdots , \quad \lambda_n^k.$$

- Any matrix polynomial in A:

$$A^n + \alpha_{n-1}A^{n-1} + \cdots + \alpha_1 A + \alpha_0 I$$

  has eigenvalues

$$\lambda^n + \alpha_{n-1}\lambda^{n-1} + \cdots + \alpha_1\lambda + \alpha_0 \quad \text{for} \quad \lambda = \lambda_1, \lambda_2, \ldots, \lambda_n.$$

- **The Cayley-Hamilton Theorem:** A matrix A satisfies its own characteristic equation, that is, if the characteristic equation is

$$(-1)^n\lambda^n + c_{n-1}\lambda^{n-1} + \cdots + c_1\lambda + c_0 = 0$$

  where $c_1, c_2, \cdots, c_n$ are constants then

$$(-1)^n A^n + c_{n-1}A^{n-1} + \cdots + c_1 A + c_0 I = 0.$$

- It can be shown that any set of ℓ vectors from ℓ different eigenspaces, that is, corresponding to different eigenvalues, is a *linearly independent* set.

6.4 Diagonalisation

Suppose that the $n \times n$ matrix $A = A_{SS}$ has enough eigenvectors so that we can construct a basis $B = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_n}\}$ only consisting of eigenvectors. Then if we perform a change of basis on the matrix to determine A_{BB}, then that matrix will be diagonal! Indeed the i-th column of A_{BB} represents $A\boldsymbol{v}_i$ with coordinates in terms of the basis B. Since $A\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$, we get all zero coordinates except for the i-th one equal to λ_i.

We know from Chapter 5 that $A_{BB} = P_{BS}A_{SS}P_{SB}$. To reflect that A_{BB} is a diagonal matrix, we write $D = A_{BB}$, and for short we will write $P = P_{SB}$[1]. Then we also have $P_{BS} = P^{-1}$. Hence we can rewrite the diagonalisation formula.

[1] Recall from Example 5.3 that the matrix P_{SB} is particularly easy to determine: just take as columns the vectors of B.

Let A be an $n \times n$ matrix and suppose there exists a basis B of $\mathbb{R}^n$ only consisting of eigenvectors of A, then there exists a diagonal matrix D and an invertible matrix P such that

$$D = P^{-1}AP.$$

More precisely, P has for columns the eigenvectors in B, and D is a diagonal matrix with entries the eigenvalues corresponding to those eigenvectors (in the same order).

Two matrices M and N are called *similar matrices* if there exists an invertible matrix Q such that

$$N = Q^{-1}MQ.$$

Clearly A and the diagonal matrix D constructed from the eigenvalues of A are similar.

EXAMPLE **6.11.** *(Example 6.9 revisited) Consider*

$$A = \begin{bmatrix} -2 & 2 & -3 \\ 2 & 1 & -6 \\ -1 & -2 & 0 \end{bmatrix}.$$

We found a basis of eigenvectors $B = \{(-1,-2,1),(-2,1,0),(3,0,1)\}$. *Thus we take*

$$P = \begin{bmatrix} -1 & -2 & 3 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \quad D = \begin{bmatrix} 5 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & -3 \end{bmatrix}.$$

Note that the columns of D and P must correspond, e.g. $\lambda = 5$ *is in column 1 of D so the corresponding eigenvector must be in column 1 of P. We easily check that the diagonalisation formula holds.* □

EXAMPLE **6.12.** *(An engineering example) In a number of branches of engineering one encounters stress (and strain) tensors. These are in fact matrix representations of (linearised) mechanical considerations. For example, the stress tensor is used to calculate the stress in a given direction at any point of interest (*$\boldsymbol{T}^{(\boldsymbol{n})} = \sigma \boldsymbol{n}$ *in one of the standard notations). The eigenvalues are referred to as* principal stresses *and the eigenvectors as* principal directions*. The well-known (to these branches of engineering) transformation rule for the stress tensor is essentially the diagonalisation formula.* □

Diagonalisation is the process of determining a matrix P such that $P^{-1}AP$ is diagonal. All we need to do is to find the eigenvalues and eigenvectors of A and form P as described above.

Note, however, that not every matrix is diagonalisable.

EXAMPLE **6.13.** *Consider*

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

The eigenvalues of A are $\lambda = -2, 1, 1$ (that is, $\lambda = 1$ has algebraic multiplicity 2).

The eigenspace corresponding to $\lambda = -2$ is

$$E_{-2} = span((1, 3, -9)).$$

The eigenspace corresponding to $\lambda = 1$ is

$$E_1 = span((1, 0, 0)).$$

Both $\lambda = -2$ and $\lambda = 1$ have geometric multipliciy 1. This means that for $\lambda = 1$ the geometric multiplicity is less than the algebraic multiplicity. In order to get the matrix P we need a basis of eigenvectors. Unfortunately, there are not enough linearly independent eigenvectors to enable us to build matrix P. Hence matrix A is not diagonalisable. □

In order for a matrix to be diagonisable, the characteristic polynomial must be factorisable fully into linear factors, and all geometric multiplicities must be equal to the algebraic multiplicities. Otherwise we won't have enough linearly independent eigenvectors.

REMARK **6.14.** *If an $n \times n$ matrix has n distinct eigenvalues then it will be diagonalisable because each eigenvalue will give a representative eigenvector and these will be linearly independent because they correspond to different eigenvalues.*

REMARK **6.15.** *Recall from Definition 3.3 that a matrix A is called a* symmetric matrix *if it equals its transpose, that is*

$$A = A^T.$$

It can be shown that the eigenvalues of a real symmetric *$n \times n$ matrix A are all real and that we can always find enough linearly independent eigenvectors to form matrix P, even if there are less than n distinct eigenvalues. That is, a real symmetric matrix can always be diagonalised.*

7
Improper integrals

Recall that definite integrals were defined in MATH1011 for bounded functions on finite intervals $[a, b]$. In this chapter we describe how to generalise this to unbounded functions and unbounded domains, using limits. These are called *improper integrals*.

7.1 Improper integrals over infinite intervals

Definition 7.1. *(*Type I improper integrals*)*
(a) Let the function f be defined on $[a, \infty)$ for some $a \in \mathbb{R}$ and integrable over $[a, t]$ for any $t > a$. The improper integral of f over $[a, \infty)$ is defined to be

$$\int_a^\infty f(x)\,dx = \lim_{t\to\infty} \int_a^t f(x)\,dx.$$

If the limit exists then the improper integral is called convergent. If the limit does not exist then the improper integral is called divergent.

(b) Similarly

$$\int_{-\infty}^b f(x)\,dx = \lim_{t\to-\infty} \int_t^b f(x)\,dx.$$

(c) Finally, suppose $f(x)$ is defined for all $x \in \mathbb{R}$. Consider an arbitrary $c \in \mathbb{R}$ and define

$$\int_{-\infty}^\infty f(x)\,dx = \int_{-\infty}^c f(x)\,dx + \int_c^\infty f(x)\,dx.$$

The improper integral is convergent if and only if for some $c \in \mathbb{R}$ <u>both</u> integrals on the right-hand-side are convergent.

Remark 7.2. *It can be shown that the choice of c is not important; i.e. if for one particular choice of c the integrals on the right-hand-side are convergent, then the same is true for any other choice of c and the sum of the two integrals is always the same*[1].

[1] Try to prove this.

EXAMPLE **7.3.** *Find the improper integral* $\int_1^\infty \frac{1}{x^3}\,dx$ *if it is convergent or show that it is divergent.*

Solution: $\int_1^\infty \frac{1}{x^3}\,dx = \lim_{t\to\infty}\int_1^t \frac{1}{x^3}\,dx$ *by definition. For any* $t > 1$,

$$\int_1^t \frac{1}{x^3}\,dx = \left[-\frac{1}{2x^2}\right]_1^t = \left(-\frac{1}{2t^2} + \frac{1}{2}\right)$$

which converges to $\frac{1}{2}$ *when* $t \to \infty$. *Hence the improper integral is convergent and its value is* $\frac{1}{2}$ *(cf. Figure 7.1)* □

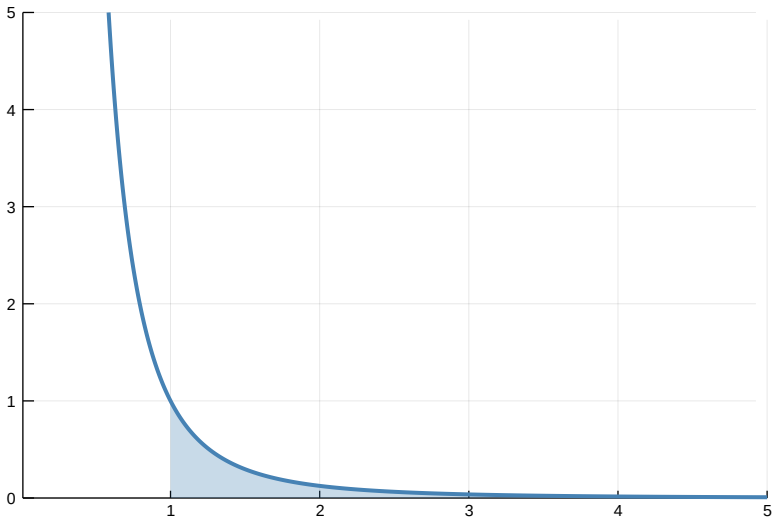


Figure 7.1: The area A under the graph of $f(x) = \frac{1}{x^3}$ and above the interval $[1,\infty)$ is finite even though the 'boundaries' of the area are infinitely long.

EXAMPLE **7.4.** *Find the improper integral* $\int_1^\infty \frac{1}{x}\,dx$ *if it is convergent or show that it is divergent.*

Solution:

$$\begin{aligned}\int_1^\infty \frac{1}{x}\,dx &= \lim_{t\to\infty}\int_1^t \frac{1}{x}\,dx\\ &= \lim_{t\to\infty}[\ln x]_1^t\\ &= \lim_{t\to\infty}\ln t\end{aligned}$$

which does not exist. Hence the integral $\int_1^\infty \frac{1}{x}\,dx$ *is divergent (cf. Figure 7.2).* □

EXAMPLE **7.5.** *Find all constants* $p \in \mathbb{R}$ *such that the improper integral* $\int_1^\infty \frac{1}{x^p}\,dx$ *is convergent.*

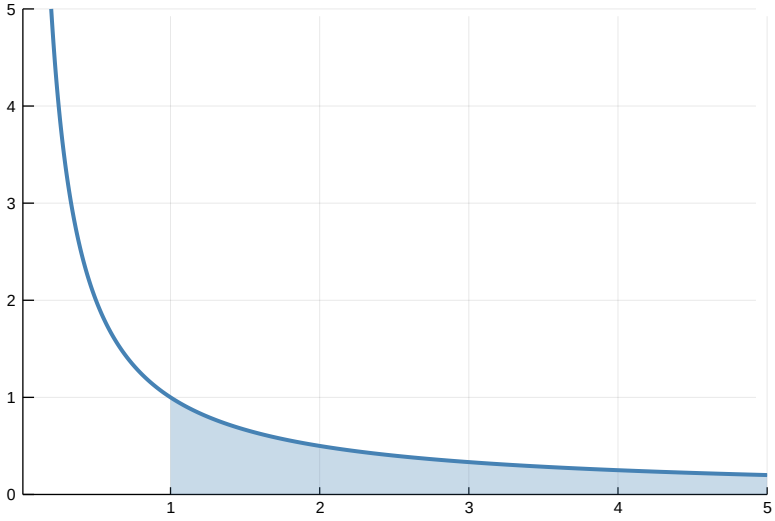


Figure 7.2: The area under the graph of $f(x) = \frac{1}{x}$ and above the interval $[1, \infty)$ is unbounded.

Solution: *For $p = 1$ the previous example shows that the integral is divergent. Suppose $p \neq 1$. Then*

$$\int_1^t \frac{1}{x^p}\,dx = \left[\frac{x^{-p+1}}{-p+1}\right]_1^t = \frac{t^{1-p}-1}{1-p}.$$

When $1-p<0$, $\lim_{t\to\infty} t^{1-p} = 0$ so the integral is convergent (to $\frac{1}{p-1}$).
When $1-p>0$, $\lim_{t\to\infty} t^{1-p} = \infty$ and therefore the integral is divergent.
Hence the integral is divergent for $p \leq 1$ and otherwise convergent. □

7.2 *Improper integrals of unbounded functions over finite intervals*

Sometimes we want to integrate a function over an interval, even though the function is not defined at some points in the interval.

Definition 7.6. (Type II improper integrals)
(a) Assume that for some $a < b$ the function f is defined and continuous on $[a,b)$ however it has some kind of singularity at b, e.g. $f(x) \to \infty$ or $-\infty$ as $x \to b^-$. Define the improper integral of f over $[a,b]$ by

$$\int_a^b f(x)\,dx = \lim_{t\to b^-} \int_a^t f(x)\,dx.$$

If the limit exists, then the improper integral is called convergent, otherwise it is divergent.

(b) In a similar way, if f is continuous on $(a,b]$ however it has some kind of singularity at a, e.g. $f(x) \to \infty$ or $-\infty$ as $x \to a^+$. We define

$$\int_a^b f(x)\,dx = \lim_{t\to a^+} \int_t^b f(x)\,dx.$$

Recall from MATH1011 that the notation $\lim_{t\to b^-}$, $\lim_{t\to a^+}$ represent the left-hand and right-hand limits respectively.

(c) If for some $c \in (a,b)$, f is continuous on each of the intervals $[a,c)$ and $(c,b]$, however it has some kind of singularity at c, define

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx$$
$$= \lim_{t\to c^-} \int_a^t f(x)\,dx + \lim_{t\to c^+} \int_t^b f(x)\,dx.$$

The improper integral on the left-hand side is convergent if and only if both improper integrals on the right-hand-side are convergent.

EXAMPLE **7.7.** *Consider the integral $\int_0^1 \frac{1}{\sqrt{x}}\,dx$. This is an improper integral, since $\frac{1}{\sqrt{x}}$ is not defined at 0 and $\frac{1}{\sqrt{x}} \to \infty$ as $x \to 0$.*

By definition, $\int_0^1 \frac{1}{\sqrt{x}}\,dx = \lim_{t\to 0^+} \int_t^1 \frac{1}{\sqrt{x}}\,dx$.

For $0 < t < 1$, we have

$$\int_t^1 \frac{1}{\sqrt{x}}\,dx = [2\sqrt{x}]_t^1 = 2 - 2\sqrt{t}$$

which converges to 2 as $t \to 0^+$. So the improper integral is convergent and its value is 2. (This example shows that the area A under the graph of $f(x) = \frac{1}{\sqrt{x}}$ and above the interval $(0,1]$ is finite even though the 'boundaries' of the area are infinitely long.) □

EXAMPLE **7.8.** *Consider the integral $\int_0^2 \frac{1}{x-1}\,dx$. It is improper, since $f(x) = 1/(x-1)$ is not defined at $x = 1$ and is unbounded near $x = 1$. However, $f(x)$ is continuous on $[0,1) \cup (1,2]$.*

Therefore

$$\int_0^2 \frac{1}{x-1}\,dx = \int_0^1 \frac{1}{x-1}\,dx + \int_1^2 \frac{1}{x-1}\,dx$$

and the integral on the left-hand-side is convergent if and only if both integrals on the right-hand-side are convergent.

Consider

$$\int_0^1 \frac{1}{x-1}\,dx = \lim_{t\to 1^-} \int_0^t \frac{1}{x-1}\,dx.$$

Given $0 < t < 1$, we have

$$\int_0^t 1/(x-1)\,dx = [\ln|x-1|]_0^t = \ln|t-1| - 0 = \ln(1-t)$$

which does not exist as $t \to 1^-$.

Hence this integral is divergent and therefore the original integral is also divergent[2]*. (The area under the graph of $f(x) = \dfrac{1}{x-1}$ and above the interval $[0,2]$ is unbounded.)* □

[2] Note that if one integral is shown to be divergent we do not need to check if the other one is convergent or not, we already know the answer: the original improper integral is divergent.

EXERCISE **7.2.1.** *For what values of p is $\displaystyle\int_0^1 x^p \, dx$ improper, and in that case for what values of p is the integral divergent? Compare with Example 7.5.*

7.3 More complicated improper integrals

Sometimes you can have a combination of problems: multiple points with singularities, perhaps also infinite intervals. The method in this case is to split the domain of integration to get a sum of improper integrals which can each be solved independently. They all need to be convergent in order for the original integral to be convergent.

KEY CONCEPT **7.9.** *(Splitting the domain of integration for improper integrals)*

1. *Identify all the problems: ∞, $-\infty$, singularities contained in the domain of integration.*
2. *Split the domain of integration into subintervals such that each subinterval has a singularity or $\infty/-\infty$ at* **exactly one end**.
3. *Solve the improper integral (type I or II) over each of these subintervals as in the previous sections.*
4. *If at any point you find a divergent integral, the original integral is divergent, so you don't need to solve further. Otherwise, the original integral is convergent, and its value is the sum of all the values of all the improper integrals you computed at the previous step.*

EXAMPLE **7.10.** *Consider the integral $I = \displaystyle\int_0^\infty \frac{1}{x-1} \, dx$. The function $f(x) = 1/(x-1)$ has a singularity at $x = 1$ (which is in the domain of integration) and ∞ is part of the domain of integration.*

Therefore we split the domain as follows:

$$I = \int_0^1 \frac{1}{x-1}\,dx + \int_1^2 \frac{1}{x-1}\,dx + \int_2^\infty \frac{1}{x-1}\,dx$$

and the integral on the left-hand-side is convergent if and only if all three integrals on the right-hand-side are convergent.

We cannot compute $\displaystyle\int_1^\infty \frac{1}{x-1}\,dx$ directly as this integral has two problems, one at each end. Therefore we need to split up $(1,\infty)$ in two subintervals, we arbitrarily chose to split at 2, but one could have split at any number larger than 1.

We saw in Example 7.8 that the first improper integral $\int_0^1 \frac{1}{x-1}\,dx$ *is divergent, hence the integral I is divergent.* □

8
Sequences and series

8.1 Sequences

By a *sequence* we mean an infinite sequence of real numbers:

$$a_1, a_2, a_3, \ldots, a_n, \ldots$$

We denote such a sequence by (a_n) or $(a_n)_{n=1}^{\infty}$. Sometimes our sequences will start with a_m for some $m \neq 1$.

Example **8.1.** *(Sequences)*

1. $$1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}, \ldots$$

   Here $a_n = \dfrac{1}{n}$ *for all integers* $n \geq 1$.

2. $b_n = (-1)^n n^3$ *for* $n \geq 1$, *defines the sequence*

   $$-1, 2^3, -3^3, 4^3, -5^3, 6^3, \ldots$$

3. *For any integer* $n \geq 1$, *define* $a_n = 1$ *when* n *is odd, and* $a_n = 0$ *when* n *is even. This gives the sequence*

   $$1, 0, 1, 0, 1, 0, 1, 0, \ldots$$

4. *The sequence of the so called* Fibonacci numbers *is defined recursively as follows:* $a_1 = a_2 = 1$, $a_{n+2} = a_n + a_{n+1}$ *for* $n \geq 1$. *This is then the sequence*

   $$1, 1, 2, 3, 5, 8, 13, \ldots$$

In the same way that we could define the limit as $x \to \infty$ of a function $f(x)$ we can also define the limit of a sequence. This is not surprising since a sequence can be regarded as a function with the domain being the set of positive integers.

Definition **8.2.** *(Intuitive definition of the limit of a sequence)*

Let (a_n) be a sequence and L be a real number. We say that (a_n) has a limit *L if we can make a_n arbitrarily close to L by taking n to be sufficiently large. We denote this situation by*

$$\lim_{n\to\infty} a_n = L.$$

We say that (a_n) is convergent *if $\lim_{n\to\infty} a_n$ exists; otherwise we say that (a_n) is* divergent.

This definition can be made more precise in the following manner: We say that (a_n) has a limit L if for every $\epsilon > 0$ there exists a positive integer N such that $|a_n - L| < \epsilon$ for all $n \geq N$. Most proofs in this chapter use this definition, and hence are a bit technical (some proofs will only be sketched or omitted).

EXAMPLE **8.3.** *(Limits of sequences)*

1. *Let b be a real number and consider the constant series (a_n) where $a_n = b$ for all $n \geq 1$. Then $\lim_{n\to\infty} a_n = b$.*

2. *Consider the sequence $a_n = \dfrac{1}{n}$ $(n \geq 1)$. Then $\lim_{n\to\infty} a_n = 0$.*

3. *If $\alpha > 0$ is a constant (that is, does not depend on n) and $a_n = \dfrac{1}{n^\alpha}$ for any $n \geq 1$, then $\lim_{n\to\infty} a_n = 0$. For instance, taking $\alpha = \dfrac{1}{2}$ gives*

$$1, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{3}}, \ldots, \frac{1}{\sqrt{n}}, \ldots \longrightarrow 0$$

4. *Consider the sequence (a_n) from Example 8.1(3) above, that is*

$$1, 0, 1, 0, 1, 0, 1, 0, \ldots$$

This sequence is divergent.

Just as for limits of functions of a real variable we have Limit Laws and a Squeeze Theorem:

THEOREM **8.4** (Limit laws)**.** *Let (a_n) and (b_n) be convergent sequences with $\lim_{n\to\infty} a_n = a$ and $\lim_{n\to\infty} b_n = b$. Then:*

1. $\lim_{n\to\infty} (a_n \pm b_n) = a \pm b$.

2. $\lim_{n\to\infty} (c\, a_n) = c\, a$ *for any constant* $c \in \mathbb{R}$.

3. $\lim_{n\to\infty} (a_n\, b_n) = a\, b$.

4. *If $b \neq 0$ and $b_n \neq 0$, for all n then $\lim_{n\to\infty} \dfrac{a_n}{b_n} = \dfrac{a}{b}$.*

THEOREM **8.5** (The squeeze theorem or the sandwich theorem)**.** *Let (a_n), (b_n) and (c_n) be sequences such that $\lim_{n\to\infty} a_n = \lim_{n\to\infty} c_n = a$ and*

$$a_n \leq b_n \leq c_n$$

for all sufficiently large n. Then the sequence (b_n) is also convergent and $\lim_{n\to\infty} b_n = a$.

We can use Theorems 8.4 and 8.5 to calculate limits of various sequences.

EXAMPLE **8.6.** *(Using Theorem 8.4 several times)*

$$\begin{aligned}\lim_{n\to\infty}\frac{n^2-n+1}{3n^2+2n-1} &= \lim_{n\to\infty}\frac{n^2\,(1-1/n+1/n^2)}{n^2\,(3+2/n-1/n^2)}\\ &= \frac{\lim\limits_{n\to\infty}(1-1/n+1/n^2)}{\lim\limits_{n\to\infty}(3+2/n-1/n^2)}\\ &= \frac{\lim\limits_{n\to\infty}1-\lim\limits_{n\to\infty}\frac{1}{n}+\lim\limits_{n\to\infty}\frac{1}{n^2}}{\lim\limits_{n\to\infty}3+2\lim\limits_{n\to\infty}\frac{1}{n}-\lim\limits_{n\to\infty}\frac{1}{n^2}}\\ &= \frac{1}{3}\end{aligned}$$

EXAMPLE **8.7.** *(Using the squeeze theorem) Find* $\lim\limits_{n\to\infty}\dfrac{\cos n}{n}$ *if it exists.*

Solution. *(Note: Theorem 8.4 is not applicable here, since* $\lim\limits_{n\to\infty}\cos n$ *does not exist.) Since* $-1\le\cos n\le 1$ *for all* n*, we have*

$$-\frac{1}{n}\le\frac{\cos n}{n}\le\frac{1}{n}$$

for all $n\ge 1$*. Using the Squeeze Theorem and the fact that*

$$\lim_{n\to\infty}-\frac{1}{n}=\lim_{n\to\infty}\frac{1}{n}=0$$

it follows that $\lim\limits_{n\to\infty}\dfrac{\cos n}{n}=0$.

A particular way for a sequence to be divergent is if it diverges to ∞ or $-\infty$.

DEFINITION **8.8.** *(Diverging to infinity)*
We say that the sequence (a_n) diverges to ∞ *if given any positive number* M *we can always find a point in the sequence after which all terms are greater than* M*. We denote this by* $\lim\limits_{n\to\infty}a_n=\infty$ *or* $a_n\to\infty$.

Similarly, we say that (a_n) diverges to $-\infty$ *if given any negative number* M *we can always find a point in the sequence after which all terms are less than* M*. We denote this by* $\lim\limits_{n\to\infty}a_n=-\infty$ *or* $a_n\to-\infty$.

Note that it follows from the definitions that if $a_n\to-\infty$, then $|a_n|\to\infty$.

EXAMPLE **8.9.** *(Sequences diverging to* ∞ *or* $-\infty$*)*

1. *Let* $a_n=n$ *and* $b_n=-n^2$ *for all* $n\ge 1$*. Then* $a_n\to\infty$*, while* $b_n\to-\infty$*, and* $|b_n|=n^2\to\infty$.
2. $a_n=(-1)^n\,n$ *does not diverge to* ∞ *and* a_n *does not diverge to* $-\infty$*, either. However,* $|a_n|=n\to\infty$.

3. *Let $r > 1$ be a constant. Then $r^n \to \infty$. If $r < -1$, then r^n does not diverge to ∞ or $-\infty$. However, $|r^n| = |r|^n \to \infty$.*

□

Note as in the case of the limit of a function, when we write $\lim_{n\to\infty} a_n = \infty$ we do not mean that the limit exists and is equal to some special number ∞. We are just using a combination of symbols which has been agreed will be taken to mean "the limit does not exist and the reason it does not exist is that the terms in the sequence increase without bound'. In particular, ∞ IS NOT a number and so do not try to do arithmetic with it, and do not use Theorem 8.4 nor Theorem 8.5 if any of the sequences diverges to ∞ or $-\infty$. If you do you will almost certainly end up with incorrect results, sooner or later, and you will always be writing nonsense.

EXAMPLE **8.10.** *(Problems with adding diverging sequences) The sequence $a_n = n$ diverges to ∞. The following sequences all diverge to $-\infty$ but have different behaviours when added to the sequence (a_n).*

1. *$b_n = -n$. Then $a_n + b_n = 0 \to 0$*
2. *$b_n = 1 - n$. Then $a_n + b_n = 1 \to 1$*
3. *$b_n = -n^2$. Then $a_n + b_n = n - n^2 = n(1-n) \to -\infty$*
4. *$b_n = -\sqrt{n}$. Then $a_n + b_n = n - \sqrt{n} = \sqrt{n}(\sqrt{n} - 1) \to \infty$*

□

The following properties can be easily derived from the definitions.

THEOREM **8.11.** *If $a_n \neq 0$ for all n, then $a_n \to 0$ if and only if $\dfrac{1}{|a_n|} \to \infty$. Similarly, if $a_n > 0$ for all $n \geq 1$, then $a_n \to \infty$ if and only if $\dfrac{1}{a_n} \to 0$.*

It follows from Example 8.9(3) and this theorem that for any constant r with $|r| < 1$ we have $r^n \to 0$.

8.1.1 Bounded sequences

DEFINITION **8.12.** *(Bounded set)*
A non-empty subset A of $\mathbb{R}$ is called bounded above *if there exists $N \in \mathbb{R}$ such that $x \leq N$ for every $x \in A$. Any such N is called an* upper bound *of A. Similarly, A is called* bounded below *if there exists $M \in \mathbb{R}$ such that $x \geq M$ for every $x \in A$. Any such M is called a* lower bound *of A.*

If A is bounded both below and above, then A is called bounded.

Clearly, the set A is bounded if and only if there exists a (finite) interval $[M, N]$ containing A.

Now we can apply this definition to a sequence by considering the set of elements in the sequence, which we denote by $\{a_n\}$.

Definition 8.13. *(Bounded sequence)*
A sequence $(a_n)_{n=1}^{\infty}$ is called bounded above *if the set $\{a_n\}$ is bounded above, that is if a_n is less than or equal to some number (an upper bound) for all n. Similarly, it is called* bounded below *if the set $\{a_n\}$ is bounded below (by a lower bound).*

We say that a sequence is bounded *if it has both an upper and lower bound.*

In Examples 8.1, the sequences in part (1) and (3) are bounded, the one in part (2) is bounded neither above nor below, while the sequence in part (4) is bounded below but not above.

Theorem 8.14. *Every convergent sequence is bounded.*

It is important to note that the converse statement in Theorem 8.14 is not true; there exist bounded sequences that are divergent, for instance Example 8.1(3) above. So this theorem is not used to prove that sequences are convergent but to prove that they are not, as in the example below.

Example 8.15. *(Unbounded sequence is divergent) The sequence $a_n = (-1)^n n^3$ is not bounded, so by Theorem 8.14. it is divergent.*

An upper bound of a sequence is not unique. For example, both 1 and 2 are upper bounds for the sequence $a_n = \dfrac{1}{n}$. This motivates the following definition.

Definition 8.16. *(Supremum and infimum)*
Let $A \subset \mathbb{R}$. The least upper bound of A (whenever A is bounded above) is called the supremum *of A and is denoted* sup A.

Similarly, the greatest lower bound of A (whenever A is bounded below) is called the infimum *of A and is denoted* inf A.

Note that finite sets always have a maximum and a minimum. The notions of infimum and supremum are most useful for infinite sets.

Notice that if the set A has a maximum (that is a largest element), then sup A is the largest element of A. Similarly, if A has a minimum (a smallest element), then inf A is the smallest element of A. However, sup A always exists when A is bounded above even when A has no maximal element. For example, if A is the open interval $(0,1)$ then A does not have a maximal element (for any $a \in A$ there is always $b \in A$ such that $a < b$). However, it has a supremum and sup $A = 1$. Similarly, inf A always exists when A is bounded below, regardless of whether A has a minimum or not. For example, if $A = (0,1)$ then inf $A = 0$.

Definition 8.17. *(Monotone)*
A sequence (a_n) is called monotone *if it is non-decreasing or non-increasing, that is, if either $a_n \leq a_{n+1}$ for all n or $a_n \geq a_{n+1}$ for all n.*

Note a constant sequence (see Example 8.3(1)) is both monotone non-decreasing and monotone non-increasing.

We can now state one important property of sequences.

Theorem 8.18 (The monotone sequences theorem)**.** *If the sequence $(a_n)_{n=1}^{\infty}$ is non-decreasing and bounded above for all sufficiently large n, then the sequence is convergent and*

$$\lim_{n\to\infty} a_n = \sup(\{a_n\}).$$

If $(a_n)_{n=1}^{\infty}$ is non-increasing and bounded below for all sufficiently large n, then (a_n) is convergent and

$$\lim_{n\to\infty} a_n = \inf(\{a_n\}).$$

That is, every monotone bounded sequence is convergent.

Theorem 8.18 and the definition of divergence to $\pm\infty$ implies the following:

Key Concept 8.19. *For any non-decreasing sequence*

$$a_1 \leq a_2 \leq a_3 \leq \cdots \leq a_n \leq a_{n+1} \leq \cdots$$

there are only two options:

Option 1. *The sequence is bounded above. Then by the Monotone Sequences Theorem the sequence is convergent, that is, $\lim_{n\to\infty} a_n$ exists.*

Option 2. *The sequence is not bounded above. It then follows from the definition of divergence to ∞ that $\lim_{n\to\infty} a_n = \infty$.*

Similarly, for any non-increasing sequence

$$b_1 \geq b_2 \geq b_3 \geq \cdots \geq b_n \geq b_{n+1} \geq \cdots$$

either the sequence is bounded below and then $\lim_{n\to\infty} b_n$ exists or the sequence is not bounded below and then $\lim_{n\to\infty} b_n = -\infty$.

Example 8.20. *Let $a_n = \dfrac{1}{n}$ for all $n \geq 1$. As we know,*

$$\lim_{n\to\infty} a_n = 0 = \inf\left\{\frac{1}{n} \mid n = 1, 2, \ldots\right\}$$

This just confirms Theorem 8.18.

The following limits of sequences can sometimes be useful (do not memorise them though). The first three sequences are non-increasing for n sufficiently large and the last one is non-decreasing for n sufficiently large (these facts are hard to prove).

THEOREM **8.21.** *1. For every real constant $\alpha > 0$*

$$\lim_{n\to\infty} \frac{\ln n}{n^\alpha} = 0 \, .$$

2.

$$\lim_{n\to\infty} \sqrt[n]{n} = \lim_{n\to\infty} n^{1/n} = 1 \, .$$

3. For every constant $a \in \mathbb{R}$

$$\lim_{n\to\infty} \frac{a^n}{n!} = 0 \, .$$

4. For every constant $a \in \mathbb{R}$

$$\lim_{n\to\infty} \left(1 + \frac{a}{n}\right)^n = e^a \, .$$

8.2 Infinite series

DEFINITION **8.22.** *(Infinite series)*
An infinite series *is by definition of the form*

$$\sum_{n=1}^{\infty} a_n = a_1 + a_2 + \cdots + a_n + \cdots \tag{8.1}$$

where $a_1, a_2, \ldots, a_n, \ldots$ is a sequence of real numbers.

It is important to be clear about the difference between a sequence and a series. These terms are often used interchangeably in ordinary English but in mathematics they have distinctly different and precise meanings.

As for sequences, the series will sometimes start with a_m with $m \neq 1$ (for instance $m = 0$).

EXAMPLE **8.23.** *(Infinite series)*

1. The geometric series *with common ratio r is*

$$\sum_{n=0}^{\infty} r^n = 1 + r + r^2 + \cdots + r^{n-1} + \ldots$$

2. The harmonic series *is*

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} + \ldots$$

*3. The p-*series *is*

$$\sum_{n=1}^{\infty} \frac{1}{n^p}$$

where $p \in \mathbb{R}$ is an arbitrary constant.

Note that the case $p = 1$ is the harmonic series.

□

Since an infinite series involves the sum of infinitely many terms this raises the issue of what the sum actually is. We can deal with this in a precise manner by the following definition.

DEFINITION **8.24.** *(Convergent and divergent series)*
For every $n \geq 1$,

$$s_n = a_1 + a_2 + \cdots + a_n$$

*is called the n*th partial sum *of the series in Equation (8.1). Then (s_n) is a sequence. If $\lim_{n\to\infty} s_n = s$ we say that the infinite series is* convergent *and write*

$$\sum_{n=1}^{\infty} a_n = s.$$

The number s is then called the sum of the series.

If $\lim_{n\to\infty} s_n$ does not exist, we say that the infinite series (8.1) is divergent.

In other words, when a series is convergent we have

$$\sum_{n=1}^{\infty} a_n = \lim_{n\to\infty} (a_1 + a_2 + \cdots + a_n).$$

EXAMPLE **8.25.** *(Convergent series) The interval $(0, 1]$ can be covered by subintervals as follows: first take the subinterval $[1/2, 1]$ (of length $1/2$), then the subinterval $[1/4, 1/2]$ (of length $1/4$); then the subinterval $[1/8, 1/4]$ (of length $1/8$), etc. The total length of these subintervals is 1, which implies the geometric series with common ratio $1/2$ converges to 2:*

$$1 = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots + \frac{1}{2^n} + \ldots$$

We now generalise this more formally.

EXAMPLE **8.26.** *(Convergence of geometric series) Consider the* geometric series *from Example 8.23(1):*

$$\sum_{n=0}^{\infty} r^n = 1 + r + r^2 + \cdots + r^{n-1} + \cdots$$

For the nth partial sum of the above series we have

$$s_n = 1 + r + r^2 + \cdots + r^{n-1} = \frac{1 - r^n}{1 - r}.$$

when $r \neq 1$.

1. *Let $|r| < 1$. Then $\lim_{n\to\infty} s_n = \lim_{n\to\infty} \dfrac{1 - r^n}{1 - r} = \dfrac{1}{1 - r}$, since $\lim r^n = 0$ whenever $|r| < 1$. Thus the geometric series is convergent in this case and we write*

$$1 + r + r^2 + \cdots + r^{n-1} + \cdots = \frac{1}{1 - r}$$

Going back to Example 8.25, now rigorously we have

$$\sum_{n=1}^{\infty} \frac{1}{2^n} = -1 + \sum_{n=0}^{\infty} \frac{1}{2^n} = -1 + \frac{1}{1 - 1/2} = 1.$$

2. *If $|r| > 1$, then $\lim_{n\to\infty} r^n$ does not exist (in fact $|r^n| \to \infty$ as $n \to \infty$), so $s_n = \dfrac{1-r^n}{1-r}$ does not have a limit; that is, the geometric series is divergent.*

3. *If $r = 1$, then $s_n = n \to \infty$ as $n \to \infty$, so the geometric series is again divergent.*

4. *Let $r = -1$. Then $s_n = 1$ for odd n and $s_n = 0$ for even n. Thus the sequence of partial sums is $1, 0, 1, 0, 1, 0, \dots$, which is a divergent sequence. Hence the geometric series is again divergent.*

□

In conclusion we get the following theorem.

THEOREM **8.27** (Convergence of the geometric series)**.** *The geometric series $\sum_{n=0}^{\infty} r^n$ is convergent if and only if $|r| < 1$, in which case its sum is $\dfrac{1}{1-r}$.*

If the series $\sum_{n=1}^{\infty} a_n$ is convergent, then by definition $\lim_{n\to\infty} s_n = s$. Of course we also have that $\lim_{n\to\infty} s_{n-1} = s$. Since $a_n = s_n - s_{n-1}$, we get that

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} s_n - s_{n-1} = \lim_{n\to\infty} s_n - \lim_{n\to\infty} s_{n-1} = s - s = 0$$

(we used the limit laws Theorem 8.4(1)). Therefore we get the following theorem.

THEOREM **8.28.** *If the series $\sum_{n=1}^{\infty} a_n$ is convergent, then $\lim_{n\to\infty} a_n = 0$.*

The above theorem says that $\lim_{n\to\infty} a_n = 0$ is necessary for convergence. The theorem is most useful in the following equivalent form.

THEOREM **8.29** (Test for Divergence)**.** *If the sequence (a_n) does not converge to 0, then the series $\sum_{n=1}^{\infty} a_n$ is divergent.*

EXAMPLE **8.30.** *(Using the Test for Divergence) The series $\sum_{n=1}^{\infty} \dfrac{n^2+2n+1}{-3n^2+4}$ is divergent by the Test for Divergence, since*

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} \frac{n^2+2n+1}{-3n^2+4} = \lim_{n\to\infty} \frac{1+2/n+1/n^2}{-3+4/n^2} = -\frac{1}{3} \neq 0\,.$$

Note that Theorem 8.28 does not say that $\lim_{n\to\infty} a_n = 0$ implies that the series is convergent. For instance the harmonic series is divergent even though $\lim_{n\to\infty} \dfrac{1}{n} = 0$ (see Theorem 8.35 below).

Convergent series behave well with regard to addition, subtraction, multiplying be a constant (but not multiplying/dividing series).

THEOREM **8.31** (Series laws). *If the infinite series* $\sum_{n=1}^{\infty} a_n$ *and* $\sum_{n=1}^{\infty} b_n$ *are convergent, then the series* $\sum_{n=1}^{\infty} (a_n \pm b_n)$ *and* $\sum_{n=1}^{\infty} c\,a_n$ *(for any constant* $c \in \mathbb{R}$*) are also convergent with*

$$\sum_{n=1}^{\infty} (a_n \pm b_n) = \sum_{n=1}^{\infty} a_n \pm \sum_{n=1}^{\infty} b_n \quad \text{and} \quad \sum_{n=1}^{\infty} c\,a_n = c \sum_{n=1}^{\infty} a_n.$$

EXAMPLE **8.32.** *Using Theorem 8.31 and the formula for the sum of a (convergent) geometric series (see Theorem 8.27), we get*

$$\begin{aligned}\sum_{n=0}^{\infty} \left[\left(-\frac{2}{3}\right)^n + \frac{3}{4^{n+1}} \right] &= \sum_{n=0}^{\infty} \left(-\frac{2}{3}\right)^n + \frac{3}{4} \sum_{n=0}^{\infty} \frac{1}{4^n} \\ &= \frac{1}{1+2/3} + \frac{3}{4} \cdot \frac{1}{1-1/4} = \frac{8}{5}.\end{aligned}$$

8.2.1 The integral test

Recall that we associate any series $\sum_{n=1}^{\infty} a_n$ with two sequences, namely the sequence (s_n) of its partial sums and the sequence (a_n) of its terms. The sequence (a_n) can be seen as a function whose domain is the set of positive integers

$$f : \mathbb{N} \longrightarrow \mathbb{R} : n \to a_n.$$

In other words $a_n = f(n)$ for each positive integer n.

THEOREM **8.33** (The integral test). *Suppose that* $\sum_{n=1}^{\infty} a_n$ *is an infinite series such that* $a_n > 0$ *for all* n *and* f *is a continuous, positive, decreasing*[1] *function for* $x \geq 1$*. If* $f(n) = a_n$ *for all integers* $n \geq 1$*, then the series and improper integral*

$$\sum_{n=1}^{\infty} a_n \quad \text{and} \quad \int_1^{\infty} f(x)\,dx$$

either both converge or both diverge.

[1] Recall that a function is decreasing if $f(x) > f(y)$ whenever $x < y$. If f is differentiable, then $f'(x) < 0$ is a necessary and sufficient condition.

Note that this theorem also holds for a series $\sum_{n=m}^{\infty} a_n$ if all the conditions on f hold for all $x \geq m$.

Proof. Because f is a decreasing function, the rectangular polygon with area

$$s_n = a_1 + a_2 + a_3 + \cdots + a_n$$

shown in Figure 8.1 contains the region under $y = f(x)$ from $x = 1$ to $x = n + 1$. Hence

$$\int_1^{n+1} f(x)\,dx \leq s_n. \tag{8.2}$$

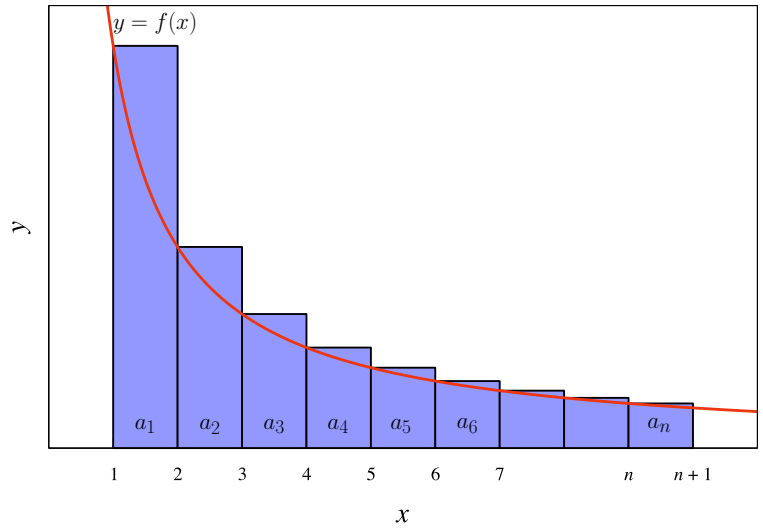


Figure 8.1: Underestimating the partial sums with an integral.

Similarly, the rectangular polygon with area

$$s_n - a_1 = a_2 + a_3 + \cdots + a_n$$

shown in Figure 8.2 is contained in the region under $y = f(x)$ from $x = 1$ to $x = n$. Hence

$$s_n - a_1 \leq \int_1^n f(x)\,dx. \tag{8.3}$$

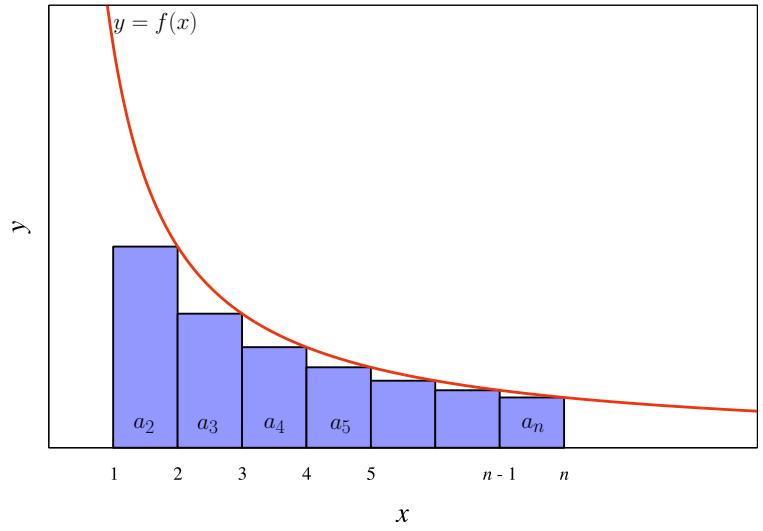


Figure 8.2: Overestimating the partial sums with an integral.

Notice that since $f(x) > 0$ for all $x \geq 1$, $\int_1^t f(x)\,dx$ is an increasing function of t. Either that function is bounded and the improper integral $\int_1^\infty f(x)\,dx$ is convergent, or it is unbounded and the integral diverges to ∞ (this is a similar idea to Key Concept 8.19).

Suppose first that the improper integral $\int_1^\infty f(x)\,dx$ diverges, then

$$\lim_{t\to\infty}\int_1^t f(x)\,dx = \infty,$$

so it follows from Equation (8.2) that $\lim_{n\to\infty} s_n = \infty$ as well, and hence the infinite series $\sum_{n=1}^{\infty} a_n$ likewise diverges.

Now suppose instead that the improper integral $\int_1^\infty f(x)\,dx$ converges to a finite value I. Then Equation (8.3) implies that

$$s_n \le a_1 + \int_1^n f(x)\,dx \le a_1 + I$$

so the increasing sequence (s_n) is bounded, and so converges by the monotone sequences theorem (Theorem 8.18). Thus the infinite series

$$\sum_{n=1}^{\infty} a_n = \lim_{n\to\infty} s_n$$

converges as well.

Hence we have shown that the infinite series and the improper integral either both converge or both diverge. □

REMARK **8.34.** *Unfortunately this theorem does not tell us what the sum of the series is, whenever it is convergent. In particular, it is not equal to* $\int_1^\infty f(x)\,dx$.

Consider again the p-series from Example 8.23(3), that is,

$$\sum_{n=1}^{\infty} \frac{1}{n^p}$$

where $p \in \mathbb{R}$ is an arbitrary constant.

THEOREM **8.35.** *The p-series is convergent if and only if* $p > 1$.

In particular, the harmonic series (a p-series with $p = 1$) diverges.

Proof. If $p \le 0$ then $\lim_{n\to\infty} \frac{1}{n^p} = \lim_{n\to\infty} n^{-p} \ne 0$, hence by the test for divergence the p-series diverges.

If $p > 0$ the function $f(x) = \frac{1}{x^p}$ is continuous, positive and decreasing for $x \ge 1$, since $f'(x) = -\frac{p}{x^{p+1}} < 0$ for $x \ge 1$. Thus we can apply the integral test.

Recall from Example 7.5 that the improper integral $\int_1^\infty \frac{1}{x^p}\,dx$ converges if $p > 1$ and diverges if $p \le 1$.

It therefore follows from the integral test that the p-series converges if $p > 1$ and diverges if $p \le 1$. □

This result does not tell us what the sum of the series is.

8.2.2 More convergence tests for series

There are several other results and tests that allow you to determine if a series is convergent or not. We outline some of them in the theorems below.

An easy way to determine if a series converges is to compare it to a series whose behaviour we know. The first way to do this is in an analogous way to the Squeeze Theorem.

THEOREM **8.36** (The Comparison Test)**.** *Let* $\sum_{n=1}^{\infty} a_n$ *and* $\sum_{n=1}^{\infty} b_n$ *be infinite series such that*

$$0 \leq a_n \leq b_n$$

holds for all sufficiently large n.

1. *If* $\sum_{n=1}^{\infty} b_n$ *is convergent then* $\sum_{n=1}^{\infty} a_n$ *is convergent.*
2. *If* $\sum_{n=1}^{\infty} a_n$ *is divergent then* $\sum_{n=1}^{\infty} b_n$ *is divergent.*

Proof. (Sketch) Compare the two sequences of partial sums, notice they are non-decreasing and use Key Concept 8.19. □

Since we know the behaviour of a p-series (Theorem 8.35), we usually use one in our comparison.

EXAMPLE **8.37.** *(Using the Comparison Test)*

1. *Consider the series*
$$\sum_{n=1}^{\infty} \frac{1+\sin n}{n^2}.$$
Since $-1 \leq \sin n \leq 1$*, we have* $0 \leq 1+\sin n \leq 2$ *for all n. Thus*
$$0 \leq \frac{1+\sin n}{n^2} \leq \frac{2}{n^2} \tag{8.4}$$
for all integers $n \geq 1$*.*
The series $\sum_{n=1}^{\infty} \frac{1}{n^2}$ *is convergent, since it is a p-series with* $p = 2 > 1$*.*
Thus $\sum_{n=1}^{\infty} \frac{2}{n^2}$ *is convergent by the series laws, and now Equation (8.4) and the Comparison Test show that the series* $\sum_{n=1}^{\infty} \frac{1+\sin n}{n^2}$ *is also convergent.*
2. *Consider the series* $\sum_{n=1}^{\infty} \frac{\ln(n)}{n}$*. Since*
$$0 < \frac{1}{n} \leq \frac{\ln(n)}{n}$$

for all integers $n \geq 3$, and the series $\sum_{n=1}^{\infty} \frac{1}{n}$ is divergent (it is the harmonic series), the Comparison Test implies that the series $\sum_{n=1}^{\infty} \frac{\ln(n)}{n}$ is also divergent.

□

Another way to compare two series is to take the limit of the ratio of terms from the corresponding sequences. This allows us to compare the rates at which the two sequences go to 0.

THEOREM **8.38** (The Limit Comparison Test)**.** *Let $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$ be infinite series such that $a_n \geq 0$ and $b_n > 0$ for sufficiently large n, and let*

$$c = \lim_{n\to\infty} \frac{a_n}{b_n} \geq 0.$$

(a) If $0 < c < \infty$, then $\sum_{n=1}^{\infty} a_n$ is convergent if and only if $\sum_{n=1}^{\infty} b_n$ is convergent.

(b) If $c = 0$ and $\sum_{n=1}^{\infty} b_n$ is convergent then $\sum_{n=1}^{\infty} a_n$ is convergent.

(c) If $\lim_{n\to\infty} \frac{a_n}{b_n} = \infty$ and $\sum_{n=1}^{\infty} a_n$ is convergent then $\sum_{n=1}^{\infty} b_n$ is convergent.

The proof uses the formal definition of limits and the Comparison Test.

REMARK **8.39.** *1. Clearly in case (a) above we have that $\sum_{n=1}^{\infty} a_n$ is divergent whenever $\sum_{n=1}^{\infty} b_n$ is divergent. In case (b), if $\sum_{n=1}^{\infty} a_n$ is divergent, then $\sum_{n=1}^{\infty} b_n$ must be also divergent. And in case (c), if $\sum_{n=1}^{\infty} b_n$ is divergent, then $\sum_{n=1}^{\infty} a_n$ is divergent.*

2. Notice that in cases (b) and (c) we have implications (not equivalences). For example, in case (b) if we know that $\sum_{n=1}^{\infty} a_n$ is convergent, we cannot claim the same for $\sum_{n=1}^{\infty} b_n$. Similarly, in case (c) if $\sum_{n=1}^{\infty} b_n$ is convergent, we cannot conclude the same about $\sum_{n=1}^{\infty} a_n$.

Again we will often compare with a p-series.

EXAMPLE **8.40.** *(Using the Limit Comparison Test)*

1. Consider the series $\sum_{n=1}^{\infty} \frac{\sin^2 n + n}{2n^2 - 1}$. To check whether the series is convergent or divergent we will compare it with the series $\sum_{n=1}^{\infty} b_n$, where

$b_n = \dfrac{1}{n}$. For $a_n = \dfrac{\sin^2 n + n}{2n^2 - 1}$, we have

$$\frac{a_n}{b_n} = \frac{n(\sin^2 n + n)}{2n^2 - 1} = \frac{\frac{\sin^2 n}{n} + 1}{2 - \frac{1}{n^2}} \to \frac{1}{2}$$

as $n \to \infty$, since $0 \leq \dfrac{\sin^2 n}{n} \leq \dfrac{1}{n}$, so by the Squeeze Theorem $\lim\limits_{n\to\infty} \dfrac{\sin^2 n}{n} = 0$.

Now the series $\sum\limits_{n=1}^{\infty} b_n = \sum\limits_{n=1}^{\infty} \dfrac{1}{n}$ is divergent (it is the harmonic series), so part (a) of the Limit Comparison Test implies that $\sum\limits_{n=1}^{\infty} a_n$ is also divergent.

2. *Consider the series $\sum\limits_{n=1}^{\infty} \dfrac{2\sqrt{n} + 3}{3n^2 - 1}$. To check whether the series is convergent or divergent we will compare it with the series $\sum\limits_{n=1}^{\infty} b_n$, where $b_n = \dfrac{1}{n^{3/2}}$. For $a_n = \dfrac{2\sqrt{n} + 3}{3n^2 - 1}$, we have*

$$\frac{a_n}{b_n} = \frac{n^{3/2}(2\sqrt{n} + 3)}{3n^2 - 1} = \frac{2 + \frac{3}{\sqrt{n}}}{3 - \frac{1}{n^2}} \to \frac{2}{3}$$

as $n \to \infty$.

Since the series $\sum\limits_{n=1}^{\infty} b_n = \sum\limits_{n=1}^{\infty} \dfrac{1}{n^{3/2}}$ is convergent (it is a p-series with $p = 3/2 > 1$), part (a) of the Limit Comparison Test shows that $\sum\limits_{n=1}^{\infty} a_n$ is also convergent.

□

8.2.3 Alternating series

Alternating series are infinite series of the form

$$\sum_{n=1}^{\infty} (-1)^{n-1} a_n = a_1 - a_2 + a_3 - a_4 + a_5 - a_6 + \dots$$

where $a_1, a_2, \dots$ is a sequence with $a_n \geq 0$ for all n.

THEOREM **8.41** (The Alternating Series Test). *Let (a_n) be a non-increasing sequence such that $\lim\limits_{n\to\infty} a_n = 0$[2]. Then the alternating series $\sum\limits_{n=1}^{\infty} (-1)^{n-1} a_n$ is convergent. Moreover, if s is the sum of the alternating series and s_n its nth partial sum, then $|s - s_n| \leq a_{n+1}$ for all $n \geq 1$.*

[2] Such a sequence must satisfy $a_n \geq 0$ for all n.

The proof is quite technical and involves showing that the sequence of partial sums with even indices is non-decreasing and bounded above.

REMARK **8.42.** *The conclusion of Theorem 8.41 about the convergence of an alternating series remains true if we assume that (a_n) is non-increasing for sufficiently large n and of course that $\lim_{n\to\infty} a_n = 0$. In other words, the sequence can increase for a finite number of terms before becoming non-increasing.*

EXAMPLE **8.43.** *(Using the Alternating Series Test) The series*

$$\sum_{n=1}^{\infty}(-1)^{n-1}\frac{1}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6}\dots$$

is convergent according to the Alternating Series Test, since for $a_n = \dfrac{1}{n}$, the sequence (a_n) is decreasing and converges to 0. □

8.2.4 Absolute convergence and the ratio test

We begin with a theorem.

THEOREM **8.44.** *If the infinite series $\sum_{n=1}^{\infty} |a_n|$ is convergent, then the series $\sum_{n=1}^{\infty} a_n$ is also convergent.*

Proof. We apply the Comparison Test in an ingenious way.

Assume the infinite series $\sum_{n=1}^{\infty} |a_n|$ is convergent. Let $b_n = a_n + |a_n|$. Since $a_n = \pm|a_n|$, we have that $0 \le b_n \le 2|a_n|$. By the series laws, $\sum_{n=1}^{\infty} 2|a_n|$ is convergent, and so $\sum_{n=1}^{\infty} b_n$ is also convergent by the Comparison Test. Finally we use again the series laws:

$$\sum_{n=1}^{\infty} b_n - |a_n| = \sum_{n=1}^{\infty} b_n - \sum_{n=1}^{\infty} |a_n| = \sum_{n=1}^{\infty} a_n$$

is convergent. □

This motivates the following definition.

DEFINITION **8.45.** *(Absolute convergence)*
An infinite series $\sum_{n=1}^{\infty} a_n$ is called absolutely convergent *if the series $\sum_{n=1}^{\infty} |a_n|$ is convergent. If $\sum_{n=1}^{\infty} a_n$ is convergent but $\sum_{n=1}^{\infty} |a_n|$ is divergent then we say that $\sum_{n=1}^{\infty} a_n$ is* conditionally convergent.

As Theorem 8.44 shows, every absolutely convergent series is convergent. The next example shows that the converse is not true.

EXAMPLE **8.46.** *(Conditionally convergent) The series $\sum_{n=1}^{\infty}(-1)^{n-1}\frac{1}{n}$ is convergent (see Example 8.43). However,*

$$\sum_{n=1}^{\infty}\left|(-1)^{n-1}\frac{1}{n}\right| = \sum_{n=1}^{\infty}\frac{1}{n}$$

is divergent (it is the harmonic series). Thus $\sum_{n=1}^{\infty}(-1)^{n-1}\frac{1}{n}$ *is conditionally convergent.* □

EXAMPLE **8.47.** *(Absolutely convergent) Consider the series* $\sum_{n=1}^{\infty}\frac{\sin n}{n^2}$. *Since*

$$0 \leq \left|\frac{\sin n}{n^2}\right| \leq \frac{1}{n^2}$$

and the series $\sum_{n=1}^{\infty}\frac{1}{n^2}$ *is convergent (a p-series with* $p = 2 > 1$*), the Comparison Test implies that* $\sum_{n=1}^{\infty}\left|\frac{\sin n}{n^2}\right|$ *is convergent. Hence* $\sum_{n=1}^{\infty}\frac{\sin n}{n^2}$ *is absolutely convergent and therefore convergent.* □

The following test is very useful.

THEOREM **8.48** (The Ratio Test)**.** *Let* $\sum_{n=1}^{\infty} a_n$ *be such that*

$$\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| = L.$$

1. *If* $L < 1$*, then* $\sum_{n=1}^{\infty} a_n$ *is absolutely convergent.*
2. *If* $L > 1$*, then* $\sum_{n=1}^{\infty} a_n$ *is divergent.*

Note that when $L = 1$ the Ratio Test gives no information.

Proof. 1. Suppose $L < 1$. Choose a number r such that $L < r < 1$. For n large enough (say for $n \geq N$), the ratio $\left|\frac{a_{n+1}}{a_n}\right|$ will eventually be less than r (this follows from the formal definition of a limit). Therefore $|a_{n+1}| < r|a_n|$ for all $n \geq N$. In particular

$$\begin{aligned} |a_{N+1}| &< r|a_N| &&= r^1|a_N| \\ |a_{N+2}| &< r|a_{N+1}| &&< r^2|a_N| \\ |a_{N+3}| &< r|a_{N+3}| &&< r^3|a_N| \text{ etc} \end{aligned}$$

In general $|a_{N+t}| < r^t|a_N|$ for every positive integer t.

We are now going to use the Comparison Test with the two series

$$\sum_{n=1}^{\infty}|a_{N+n-1}| = |a_N| + |a_{N+1}| + |a_{N+2}| + \cdots \text{ and}$$

$$\sum_{n=1}^{\infty} r^{n-1}|a_N| = |a_N| + r|a_N| + r^2|a_N| + \cdots = |a_N|(1 + r + r^2 + \cdots) = |a_N|\frac{1}{1-r}.$$

We have $0 \leq |a_{N+n-1}| \leq r^{n-1}|a_N|$ for all n as seen above, and the second series converges by Theorem 8.27 and the series laws (it is a multiple of a geometric series with $|r| < 1$). Hence the first

series converges. Now $\sum_{n=1}^{\infty} |a_n|$ is just adding a finite number of terms to $\sum_{n=1}^{\infty} |a_{N+n-1}|$ (namely adding $|a_1| + |a_2| + \cdots + |a_{N-1}|$) so it is also convergent. Therefore, our series $\sum_{n=1}^{\infty} a_n$ is absolutely convergent (and therefore convergent).

2. If $L > 1$, then for n sufficiently large $\left|\frac{a_{n+1}}{a_n}\right| > 1$, that is, the sequence $(|a_n|)$ is increasing (this follows from the formal definition of a limit), so we cannot possibly have that $\lim_{n\to\infty} a_n = 0$. By the Test for Divergence (Theorem 8.28) the series $\sum_{n=1}^{\infty} a_n$ is divergent.

□

EXAMPLE **8.49.** *(Using the Ratio Test with $L < 1$)*

1. *Consider the series $\sum_{n=1}^{\infty} \frac{n^2}{2^n}$. We have*

$$\left|\frac{a_{n+1}}{a_n}\right| = \left|a_{n+1} \times \frac{1}{a_n}\right| = \frac{(n+1)^2}{2^{n+1}} \times \frac{2^n}{n^2} = \frac{1}{2}\left(\frac{n+1}{n}\right)^2 = \frac{1}{2}\left(1+\frac{1}{n}\right)^2 \to \frac{1}{2}$$

as $n \to \infty$. Since $L = \frac{1}{2} < 1$, the Ratio Test implies that $\sum_{n=1}^{\infty} \frac{n^2}{2^n}$ is absolutely convergent and hence convergent.

2. *For any constant $b \in \mathbb{R}$, consider the infinite series $\sum_{n=0}^{\infty} \frac{b^n}{n!}$. Then $a_n = \frac{b^n}{n!}$. We have*

$$\left|\frac{a_{n+1}}{a_n}\right| = \left|\frac{b^{n+1}}{(n+1)!} \times \frac{n!}{b^n}\right| = \frac{|b|}{n+1} \to 0$$

as $n \to \infty$. So, by the Ratio Test, the series $\sum_{n=0}^{\infty} \frac{b^n}{n!}$ is absolutely convergent. Note that by the Test for Divergence this implies Theorem 8.21(3).

□

EXAMPLE **8.50.** *(Using the Ratio Test with $L > 1$) Consider the series $\sum_{n=1}^{\infty} \frac{n4^n}{(-3)^n}$. We have*

$$\left|\frac{a_{n+1}}{a_n}\right| = \left|\frac{(n+1)4^{n+1}}{-3^{n+1}} \times \frac{(-3)^n}{n4^n}\right| = \frac{4}{3}(1+\frac{1}{n}) \to \frac{4}{3} > 1.$$

So, by the Ratio Test, the series $\sum_{n=1}^{\infty} \frac{n4^n}{(-3)^n}$ is divergent. □

8.3 Power series

Definition 8.51. *(Power series)*
Let $a \in \mathbb{R}$ be a given number, $(b_n)_{n=0}^{\infty}$ a given sequence of real numbers and $x \in \mathbb{R}$ a variable (parameter).

A series of the form

$$\sum_{n=0}^{\infty} b_n (x-a)^n = b_0 + b_1 (x-a) + b_2 (x-a)^2 + \cdots + b_n (x-a)^n + \cdots$$

is called a power series *centred at a. When $a = 0$, the series is simply called a power series.*

Clearly a power series can be regarded as a function of x defined for all $x \in \mathbb{R}$ for which the infinite series is convergent.

Let us apply the Ratio Test to this series. The terms of this series are of the form $a_n = b_n (x-a)^n$, so

$$L = \lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| = \lim_{n\to\infty} \frac{|b_{n+1}||x-a|^{n+1}}{|b_n||x-a|^n} = |x-a| \times \lim_{n\to\infty} \left| \frac{b_{n+1}}{b_n} \right|.$$

There are three cases, according to $\lim_{n\to\infty} \left| \frac{b_{n+1}}{b_n} \right|$.

(a) If $\lim_{n\to\infty} \left| \frac{b_{n+1}}{b_n} \right|$ is a positive real number, which we denote by $\frac{1}{R}$. Then $L = \frac{|x-a|}{R}$. By the Ratio Test, this diverges when $L > 1$, that is when $|x-a| > R$, and is absolutely convergent when $L < 1$, that is when $|x-a| < R$. Note that

$$R = \left(\lim_{n\to\infty} \left| \frac{b_{n+1}}{b_n} \right| \right)^{-1} = \lim_{n\to\infty} \left| \frac{b_n}{b_{n+1}} \right|.$$

(b) If $\lim_{n\to\infty} \left| \frac{b_{n+1}}{b_n} \right| = 0$, then $L = 0$ and so by the Ratio Test the power series is absolutely convergent for all x.

(c) If $\lim_{n\to\infty} \left| \frac{b_{n+1}}{b_n} \right| = \infty$, then $L = \infty$ and so by the Ratio Test the power series diverges EXCEPT if $x = a$, then $L = 0$ and so the series converges. We easily that the series reduces to just b_0 when $x = a$.

Thefore we proved the following theorem.

Theorem 8.52. *For a power series $\sum_{n=0}^{\infty} b_n (x-a)^n$, let $R = \lim_{n\to\infty} \left| \frac{b_n}{b_{n+1}} \right|$. Then one of the following three possibilities occurs:*

(a) R is a positive real number, and the series is absolutely convergent for $|x-a| < R$ and divergent for $|x-a| > R$.

(b) $R = \infty$ and the series is absolutely convergent for all $x \in \mathbb{R}$.

(c) *$R = 0$ and the series is absolutely convergent for $x = a$ and divergent for all $x \neq a$.*

In other words, a power series is convergent at only one point or everywhere or on an interval $(a - R, a + R)$ centred at a. It is not possible for it to only be convergent at several separated points or on several separate intervals. In case (a), when $|x - a| = R$ then the series may or not be convergent. It is even possible for a series to be convergent for $a + R$ but divergent for $a - R$ or vice versa.

DEFINITION **8.53.** *(Radius of convergence)*
The number $R = \lim_{n\to\infty} \left| \frac{b_n}{b_{n+1}} \right|$ is called the radius of convergence *of the power series $\sum_{n=0}^{\infty} b_n (x-a)^n$.*

EXAMPLE **8.54.** *(Convergence of power series) Find all $x \in \mathbb{R}$ for which the series*

$$\sum_{n=1}^{\infty} \frac{(-1)^n}{\sqrt{n}} (x+3)^n \tag{8.5}$$

is absolutely convergent, conditionally convergent or divergent.

Solution. *Here we have $b_n = \frac{(-1)^n}{\sqrt{n}}$ and $a = -3$. We compute*

$$R = \lim_{n\to\infty} \left| \frac{b_n}{b_{n+1}} \right| = \lim_{n\to\infty} \frac{1}{\sqrt{n}} \times \frac{\sqrt{n+1}}{1} = \lim_{n\to\infty} \sqrt{1 + \frac{1}{n}} = 1.$$

Therefore we are in case (a) and so the power series is absolutely convergent for $x \in (-3-1, -3+1) = (-4, -2)$ and is divergent for $x < -4$ and $x > -2$. It remains to check the points $x = -4$ and $x = -2$.

Substituting $x = -4$ in Equation (8.5) gives the series $\sum_{n=1}^{\infty} \frac{(-1)^n}{\sqrt{n}} (-1)^n = \sum_{n=1}^{\infty} \frac{1}{n^{1/2}}$ which is divergent (it is a p-series with $p = 1/2 < 1$).

When $x = -2$, the series (8.5) becomes $\sum_{n=1}^{\infty} \frac{(-1)^n}{\sqrt{n}}$, which is convergent by the Alternating Series Test since $(\frac{1}{\sqrt{n}})$ is non-increasing with limit 0. However, $\sum_{n=1}^{\infty} \left| \frac{(-1)^n}{\sqrt{n}} \right| = \sum_{n=1}^{\infty} \frac{1}{\sqrt{n}}$ is divergent (as we mentioned above), so $\sum_{n=1}^{\infty} \frac{(-1)^n}{\sqrt{n}}$ is conditionally convergent.

***Conclusion:** The series (8.5) is*

1. *absolutely convergent for $-4 < x < -2$;*
2. *conditionally convergent for $x = -2$*
3. *divergent for $x \leq -4$ and $x > -2$.*

□

Power series have the useful property that they can be differentiated term-by-term.

THEOREM **8.55** (Term-by-term differentiation of a power series)**.** *Assume that the power series* $\sum_{n=0}^{\infty} a_n (x-a)^n$ *has a radius of convergence* $R > 0$ *and let* $f(x)$ *be defined by*

$$f(x) = a_0 + a_1 (x-a) + a_2 (x-a)^2 + \cdots + a_n (x-a)^n + \cdots$$

for $|x-a| < R$*. Then* $f(x)$ *is differentiable (and so continuous) for* $|x-a| < R$ *and*

$$f'(x) = a_1 + 2a_2 (x-a) + 3a_3 (x-a)^2 + \cdots + n\, a_n (x-a)^{n-1} + \cdots$$

for $|x-a| < R$*. Moreover, the radius of convergence of the power series representation for* $f'(x)$ *is* R*.*

8.3.1 Taylor and MacLaurin series

DEFINITION **8.56.** *(Power series representation)*
If for a function $f(x)$ *we have*

$$f(x) = a_0 + a_1 (x-a) + a_2 (x-a)^2 + \cdots + a_n (x-a)^n + \cdots$$

for all x *in some interval* I *containing* a*, we say that the above is a* power series representation for f about a *on* I*. When* $a = 0$ *this is simply called a* power series representation *for* f *on* I*.*

For example, the formula for the sum of a geometric series gives

$$\frac{1}{1-x} = 1 + x + x^2 + \cdots + x^n + \cdots \qquad \text{for all } |x| < 1,$$

which provides a power series representation for $f(x) = \dfrac{1}{1-x}$ on $(-1, 1)$.

Suppose that a function $f(x)$ has a power series representation

$$f(x) = a_0 + a_1(x-a) + a_2(x-a)^2 + \cdots + a_n(x-a)^n + \ldots \tag{8.6}$$

for those x such that $|x-a| < R$ for some positive real number R. Substituting $x = a$ in Equation (8.6) implies that $f(a) = a_0$. Next, differentiating (8.6) using Theorem 8.55 implies

$$f'(x) = a_1 + 2a_2 (x-a) + 3a_3 (x-a)^2 + \cdots + n\, a_n (x-a)^{n-1} + \cdots \tag{8.7}$$

for all $|x-a| < R$. Then substituting $x = a$ in Equation (8.7) gives $f'(a) = a_1$.

Similarly, differentiating (8.7) yields

$$f''(x) = 2a_2 + 6a_3 (x-a) + \cdots + n\,(n-1)\, a_n (x-a)^{n-2} + \cdots$$

for all $|x-a| < R$ and substituting $x = a$ in this equality gives $f''(a) = 2a_2$, that is, $a_2 = \dfrac{f''(a)}{2!}$.

Continuing in this fashion we must have

$$a_n = \frac{f^{(n)}(a)}{n!} \qquad \text{for each } n.$$

Definition 8.57. *(Taylor series)*
Assume that $f(x)$ has derivatives of all orders on some interval I containing the point a in its interior. Then the power series

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots$$

is called the Taylor series of f about a. *When $a = 0$ this series is called the* MacLaurin Series *of f.*

We can find the radius of convergence of a Taylor series by using the formula

$$R = \lim_{n\to\infty}\left|\frac{b_n}{b_{n+1}}\right| = \lim_{n\to\infty}\left|\frac{f^{(n)}(a)}{n!} \times \frac{(n+1)!}{f^{(n+1)}(a)}\right| = \lim_{n\to\infty}(n+1)\left|\frac{f^{(n)}(a)}{f^{(n+1)}(a)}\right|$$

if this limit exists. Within the radius of convergence, we know by Theorem 8.52 that the power series is absolutely convergent. Note however that it is not guaranteed that it converges to $f(x)$, it could converge to something else (though in practice that is very rare).

To prove rigorously that the series converges to the function, we need to examine the error term for each partial sum. We denote the $(n+1)^{\text{st}}$ partial sum of the Taylor series of f at a by $T_{n,a}(x)$.

$$T_{n,a}(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n$$

Theorem 8.58. *For any given $x \in I$ we have*

$$f(x) = T_{n,a}(x) + R_{n,a}(x)$$

where $R_{n,a}(x)$ is the remainder (or error term) given by

$$R_{n,a}(x) = \frac{f^{(n+1)}(z)}{(n+1)!}(x-a)^{n+1}$$

for some z between a and x.

So, if for some $x \in I$ we have $R_{n,a}(x) \to 0$ as $n \to \infty$, then

$$f(x) = \lim_{n\to\infty} T_{n,a}(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n.$$

That is, when $R_{n,a}(x) \to 0$ as $n \to \infty$, the Taylor series is convergent at x and its sum is equal to $f(x)$.

To show that $R_{n,a}(x) \to 0$ we determine an upper bound $S_{n,a}(x)$ for $|R_{n,a}(x)|$. If the limit of the upper bound is 0, then $0 \le |R_{n,a}(x)| \le S_{n,a}(x) \to 0$ so we can use the squeeze theorem to get that $R_{n,a}(x) \to 0$.

EXAMPLE **8.59.** *(MacLaurin series)*

1. *Consider the function $f(x) = e^x$. Then $f^{(n)}(x) = e^x$ for all integers $n \ge 1$, so $f^{(n)}(0) = 1$ for all n. Thus, the Taylor series for e^x about 0 has the form*

$$\sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots . \tag{8.8}$$

We first determine the radius on convergence:

$$R = \lim_{n\to\infty} \left| \frac{b_n}{b_{n+1}} \right| = \lim_{n\to\infty} \frac{(n+1)!}{n!} = \lim_{n\to\infty} n+1 = \infty.$$

Therefore the series is absolutely convergent for all x, that is $I = \mathbb{R}$. We will now show that the sum of this series is equal to e^x for all $x \in \mathbb{R}$.

By Theorem 8.58, we have $f(x) = T_{n,0}(x) + R_{n,0}(x)$ for all x, where $R_{n,0}(x) = \dfrac{f^{(n+1)}(z)}{(n+1)!}\, x^{n+1} = \dfrac{e^z}{(n+1)!}\, x^{n+1}$ for some z between 0 and x. We now split the analysis into two cases.

If $x \ge 0$, then $0 \le z \le x$ so $e^z \le e^x$. Therefore

$$0 \le |R_{n,0}(x)| \le e^x \frac{x^{n+1}}{(n+1)!} \to 0 ,$$

as $n \to \infty$ by Theorem 8.21(3) (we consider x as a constant here).

Now if $x < 0$ then $x \le z < 0$ so $e^z \le 1$. Therefore

$$0 \le |R_{n,0}(x)| \le \frac{x^{n+1}}{(n+1)!} \to 0 ,$$

as $n \to \infty$ again by Theorem 8.21(3).

In both cases, by the Squeeze Theorem, $\lim\limits_{n\to\infty} |R_{n,0}(x)| = 0$ for any $x > 0$. Thus the MacLaurin series (8.8) is convergent and its sum is e^x for any x .

2. *We easily compute the MacLaurin series of $f(x) = \sin x$ to be*

$$\sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} = \frac{x}{1!} - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} + \cdots$$

In this case $\lim\limits_{n\to\infty} \left| \dfrac{b_n}{b_{n+1}} \right|$ does not exist (the sequence alternates between 0 and ∞), but we can use the Ratio Test, with $a_n = (-1)^n \dfrac{x^{2n+1}}{(2n+1)!}$.

Let

$$L = \lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| = \lim_{n\to\infty} \left| \frac{x^{2n+3}}{(2n+3)!} \times \frac{(2n+1)!}{x^{2n+1}} \right| = \lim_{n\to\infty} \frac{x^2}{(2n+2)(2n+3)} = 0$$

for all x. Thus the series is absolutely convergent for all x.

By Theorem 8.58, we have $f(x) = T_{n,0}(x) + R_{n,0}(x)$ for all x, where $R_{n,0}(x) = \dfrac{f^{(n+1)}(z)}{(n+1)!}\, x^{n+1}$ for some z between 0 and x, Now $f^{(n+1)}(z)$ is one of $\sin z$, $\cos z$, $-\sin z$, $-\cos z$. So $f^{(n+1)}(z) \leq 1$.

Therefore

$$0 \leq |R_{n,0}(x)| \leq \frac{x^{n+1}}{(n+1)!} \to 0\,,$$

as $n \to \infty$ by Theorem 8.21(3) and we conclude as in the previous exercise.

3. *Similarly,*

$$\cos x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} - \cdots + (-1)^n \frac{x^{2n}}{(2n)!} + \cdots$$

for all $x \in \mathbb{R}$. Here the right hand side is the Taylor series of $\cos x$ about 0.

4. *We easily compute the MacLaurin series of $f(x) = \ln(1+x)$ to be*

$$x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots + (-1)^{n-1} \frac{x^n}{n} + \cdots$$

We can compute the radius of convergence

$$R = \lim_{n\to\infty} \left| \frac{b_n}{b_{n+1}} \right| = \lim_{n\to\infty} \frac{n+1}{n} = \lim_{n\to\infty} 1 + \frac{1}{n} = 1.$$

So the series converges absolutely for $x \in (-1,1)$ and diverges for $x < -1$ and $x > 1$. We now examine the two points $x = \pm 1$. For $x = 1$ the series becomes

$$1 - \frac{1}{2} + \frac{1}{3} - \cdots + (-1)^{n-1} \frac{1}{n} + \cdots\,,$$

which is conditionally convergent by Example 8.46. For $x = 1$ the series becomes the harmonic series which is divergent. We conclude that the series converges for $x \in (-1,1]$.

In this case it is much harder to prove that the sum of the series is equal to $\ln(1+x)$ (that is, that the error term $R_{n,0}(x) \to 0$ as $n \to \infty$) and we omit that proof here.

□

9
Fourier series

Because waves and vibrations tend to have a periodic structure, that is, they repeat their basic shape in time or space, it is often convenient to approximate a periodic function by a linear combination of perfect waves (the sine and cosine functions, which arise in simple harmonic motion). Decomposing a general periodic function into a sum of trigonometrical functions is sometimes called harmonic analysis. This process is often necessary because most physical waves and vibrations do not have a perfect single sine or cosine form but rather they are made up of a number of different harmonics of the underlying system.

In order to develop the ideas required we need to formalise the process. For convenience we shall assume for now that the underlying periodicity (whether it be in space or time) is of length 2π. This eases the subsequent manipulation somewhat but turns out not to be unduly restrictive; if in practice our function has a periodicity of length different to 2π it is relatively easy to scale our results so as to account for this change. However, it is worth quickly reminding ourselves of the definition of the *period of a function*.

DEFINITION **9.1.** *(Period of a function)*
We say that a function f from $\mathbb{R}$ to $\mathbb{R}$ has period *P if $f(t+P) = f(t)$ for all t in $\mathbb{R}$. In this case, we also say that f is P-periodic.*

Graphically it is generally not difficult to identify a periodic function for its sketch takes the form of a curve that clearly repeats itself after an interval of length P. The rather peculiar function in Figure 9.1 possesses discontinuities at $t = (2n+1)\pi$ for $n \in \mathbb{Z}$ but nevertheless has a period 2π.

Recall that both $\sin(nt)$ and $\cos(nt)$ are 2π-periodic functions for positive integer values of n. Note that the smallest period of $\sin(nt)$ and $\cos(nt)$ is actually $2\pi/n,\ n \geq 1$.

We will define an infinite series using the functions $\sin(nt)$ and $\cos(nt)$ and a constant term, which will hopefully converge to our 2π-periodic function. In other words, our hope is to approximate

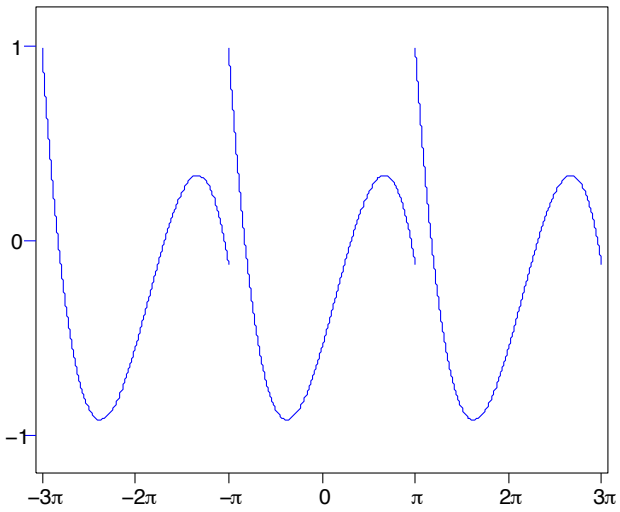


Figure 9.1: A periodic function with period 2π.

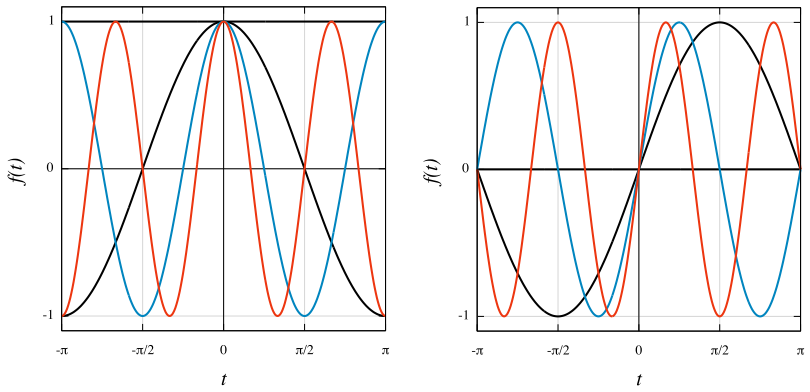


Figure 9.2: Graphs of $\cos(nt)$ and $\sin(nt)$ for $n = 0, 1, 2, 3$.

$f(t)$ in the form

$$S_N f(t) = \frac{a_0}{2} + a_1 \cos t + b_1 \sin t + a_2 \cos(2t) + b_2 \sin(2t) + \ldots + a_N \cos(Nt) + b_N \sin(Nt)$$
$$= \frac{a_0}{2} + \sum_{n=1}^{N} (a_n \cos(nt) + b_n \sin(nt)).$$

If our approximation is well-behaved we should expect that it improves as N goes to infinity.

Definition 9.2. *(Fourier series)*
The infinite series

$$\mathrm{FS}_f(t) = \lim_{N\to\infty} S_N f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos(nt) + b_n \sin(nt)) \qquad (9.1)$$

is called the Fourier series expansion *of f where the constants a_0, a_n, b_n ($n = 1, 2, 3, \ldots$) are known as the* Fourier coefficients. *The various $S_N f(t)$ functions are the* partial sums *of the Fourier series expansion.*

There is no need to consider n negative because $\sin(-nt) = -\sin(nt)$ and $\cos(-nt) = \cos(nt)$.

We will explain in the next section how to compute the Fourier coefficients. The following example illustrates how the approximations get better when N increases.

EXAMPLE **9.3.** *Consider the piecewise 2π-periodic function*

$$f(t) = \begin{cases} 1, & -\pi < t \leq 0, \\ -1, & 0 < t \leq \pi, \end{cases} \qquad \text{and} \qquad f(t+2\pi) = f(t).$$

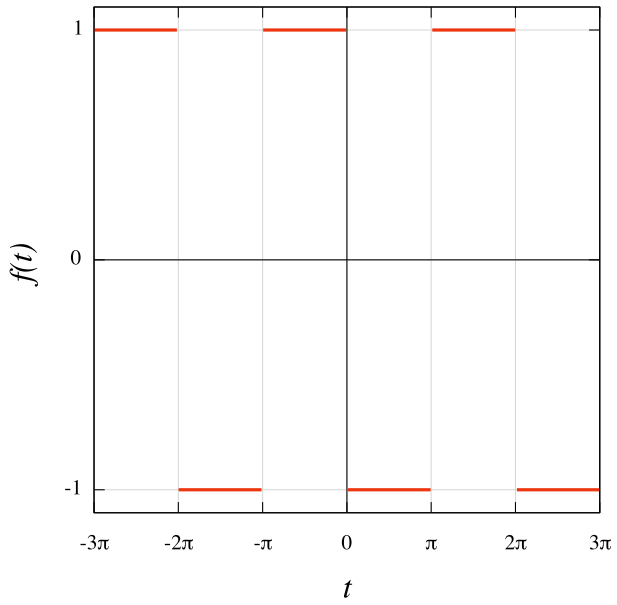


Figure 9.3: Graph of f for Example 9.3.

We will see in Example 9.13 that

$$S_N f(t) = -\frac{4}{\pi} \sum_{n=1, n \text{ odd}}^{N} \frac{\sin(nt)}{n}.$$

Figure 9.4 illustrates the Fourier sums for different values of N.

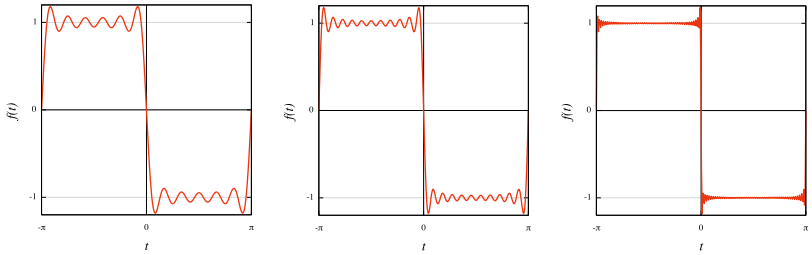


Figure 9.4: Graph of $S_N f$ for $N = 11$ (left), $N = 21$ (middle) and $N = 101$ (right), on $[-\pi, \pi]$.

Notice how the approximation improves as we increase N (so we are retaining more terms in the partial sum $S_N(f)$). We know that the true function is equal to -1 for $0 < t \leq \pi$ and the the approximation $S_N(f)$ oscillates about this value. When $N = 11$ this oscillation is quite noticeable and relatively large; by the time $N = 101$ it is far less pronounced. Notice also how the approximation $S_N f$ is relatively good away from discontinuities in the function $f(t)$ but poorer as these points are approached.

As an example of this look at the $N = 11$ result; it is clear that oscillations in the approximating function are small around $t = \pi/2$ but increase as either $t \to 0$ or $t \to \pi$. This is a well-known behaviour known as Gibbs' phenomena which tends to occur when the function $f(t)$ has points of discontinuity. □

9.1 Calculation of the Fourier coefficients

Recall that the Fourier series representation of $f(t)$ is

$$\mathrm{FS}_f(t) = \frac{a_0}{2} + a_1 \cos t + b_1 \sin t + a_2 \cos(2t) + b_2 \sin(2t) + \cdots \quad (9.1)$$

Note at this juncture it seems a little strange to define the constant term as $\frac{a_0}{2}$ rather than simply a_0. We will see why this is done presently.

We need a method of determining the values of the coefficients a_0, a_n, b_n, for $n = 1, 2, 3, \ldots$, so that $\mathrm{FS}_f(t)$ converges (if possible) to $f(t)$. To find a_0 we simply integrate both sides from $-\pi$ to π to get

$$\int_{-\pi}^{\pi} \mathrm{FS}_f(t)dt = \int_{-\pi}^{\pi} \frac{a_0}{2} dt = \pi a_0$$

because

$$\int_{-\pi}^{\pi} \cos(nt)\, dt = 0 \quad \text{and} \quad \int_{-\pi}^{\pi} \sin(nt)\, dt = 0.$$

So we set

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t)\, dt. \quad (9.1a)$$

This says that a_0 is twice the average value of $f(t)$, or equivalently, the zeroth order term $a_0/2$ is the average value of the function $f(t)$.

We can show by direct evaluation of the integrals that

$$\int_{-\pi}^{\pi} \sin(mt)\sin(nt)dt = 0, \quad \text{for any integers } m, n, \text{ with } m \neq n,$$

$$\int_{-\pi}^{\pi} \sin(mt)\cos(nt)dt = 0, \quad \text{for any integers } m, n, \text{ with } m \neq n,$$

$$\int_{-\pi}^{\pi} \cos(mt)\cos(nt)dt = 0, \quad \text{for any integers } m, n, \text{ with } m \neq n,$$

$$\int_{-\pi}^{\pi} \sin^2(nt)dt = \pi \quad , \quad \int_{-\pi}^{\pi} \cos^2(nt)dt = \pi \quad \text{for any integer } n.$$

It is an exercise to verify these statements, using the trigonometric formulae in the Appendix, or using integration by parts (twice).

This allows us to calculate the other Fourier coefficients. To obtain a_n we multiply Equation (9.1) by $\cos(nt)$ and integrate from $-\pi$ to π:

$$\int_{-\pi}^{\pi} \mathrm{FS}_f(t)\cos(nt)\, dt = \pi a_n$$

so we set

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos(nt)dt. \qquad (9.1b)$$

Note that only one term on the right-hand-side survives the integration. To obtain b_n we multiply Equation (9.1) by $\sin(nt)$ and integrate from $-\pi$ to π:

$$\int_{-\pi}^{\pi} \mathrm{FS}_f(t)\sin(nt)dt = \pi b_n,$$

so we set

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin(nt)\,dt. \qquad (9.1c)$$

Again, only one term on the right-hand-side survives. The above process is called expanding the function f as an infinite sum of orthogonal[1] functions.

Notice we need to assume here that f is sufficiently regular for all these integrals to be defined (for instance it is sufficient for f to be piecewise continuous[2] on $[-\pi,\pi]$).

[1] We can think of functions as vectors, and define a dot product $f_1 . f_2 = \int_{-\pi}^{\pi} f_1(t)f_2(t)\,dt$, so that the functions in the Fourier series are mutually orthogonal (dot product equal to 0).

[2] A function $f(x)$ is called *piecewise continuous* on a given interval $[a,b]$ if f has only finitely many points of discontinuity in $[a,b]$.

The expressions $(9.1a,b,c)$ are called ***Euler's formulae***.

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\,dt$$

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos(nt)\,dt$$

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin(nt)\,dt$$

It is at this point that we can appreciate the reason we defined the constant term in the Fourier series to be $\frac{a_0}{2}$ and not simply a_0. If we put $n = 0$ in Euler's formula for a_n this result collapses to the expression defining a_0. Had the factor $\frac{1}{2}$ not been inserted in the definition of the Fourier series then the universal formula that defines a_n for all values of n would not apply.

EXAMPLE **9.4.** *Define the 2π-periodic function*

$$f(t) = \begin{cases} 0, & -\pi < t \le 0, \\ \pi - t, & 0 < t \le \pi, \end{cases} \quad \textit{and} \quad f(t+2\pi) = f(t).$$

What is its Fourier series?

Solution: *A simple calculation yields*

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{0} 0\,dt + \frac{1}{\pi}\int_{0}^{\pi} \pi - t\,dt = \frac{1}{\pi}\int_{0}^{\pi}(\pi - t)\,dt = \frac{\pi}{2}.$$

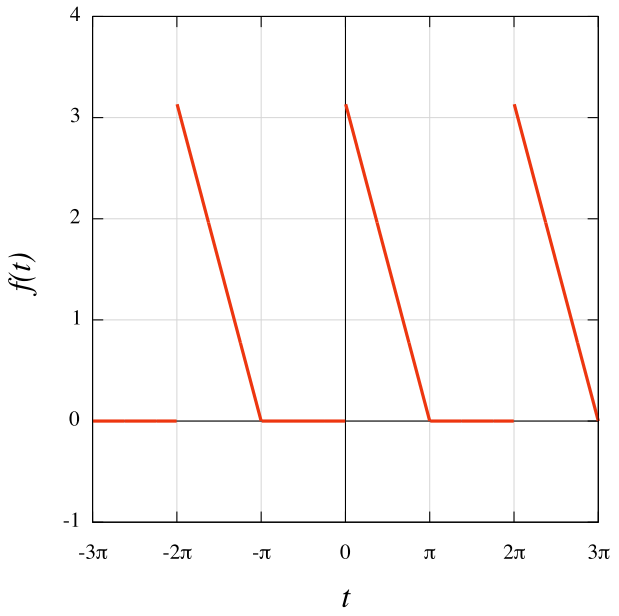


Figure 9.5: Graph of f for Example 9.4.

A more complicated calculation yields for $n > 0$,

$$a_n = \frac{1}{\pi}\int_0^{\pi} (\pi - t)\cos(nt)\,dt = \frac{1-\cos(n\pi)}{\pi n^2}$$

Useful anti-derivative formulas (derived via integration by parts):

$$\int t\cos(nt)\,dt = \frac{t}{n}\sin(nt) + \frac{1}{n^2}\cos(nt) + C$$

and

$$\int t\sin(nt)\,dt = -\frac{t}{n}\cos(nt) + \frac{1}{n^2}\sin(nt) + C.$$

Hence we can evaluate a_n as

$$a_n = \begin{cases} 0, & n > 0 \text{ even} \\ \dfrac{2}{\pi n^2}, & n \text{ odd} \end{cases}$$

or, for $k = 1, 2, \ldots$,

$$\begin{aligned} a_{2k-1} &= \frac{2}{\pi(2k-1)^2} \\ a_{2k} &= 0. \end{aligned}$$

A similarly complicated calculation yields

$$b_n = \frac{1}{\pi}\int_0^{\pi} (\pi - t)\sin(nt)\,dt = \frac{1}{n}.$$

Hence the Fourier series of the above function is

$$\begin{aligned} \mathrm{FS}_f(t) &= \frac{\pi}{4} + \frac{2}{\pi}\sum_{n=1,\ n \text{ odd}}^{\infty} \frac{\cos(nt)}{n^2} + \sum_{n=1}^{\infty}\frac{\sin(nt)}{n} \\ &= \frac{\pi}{4} + \frac{2}{\pi}\left(\cos t + \frac{\cos(3t)}{9} + \frac{\cos(5t)}{25} + \cdots\right) \\ &\quad + \left(\sin t + \frac{\sin(2t)}{2} + \frac{\sin(3t)}{3} + \cdots\right). \end{aligned}$$

□

9.2 *Functions of an arbitrary period*

The above analysis extends to functions of *arbitrary period* $2L$ rather than the special value 2π used above. In this case it turns out that

$$\mathrm{FS}_f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi t}{L}\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi t}{L}\right)$$

where

$$a_0 = \frac{1}{L}\int_{-L}^{L} f(t)\,dt$$

$$a_n = \frac{1}{L}\int_{-L}^{L} f(t)\cos\left(\frac{n\pi t}{L}\right)\,dt$$

$$b_n = \frac{1}{L}\int_{-L}^{L} f(t)\sin\left(\frac{n\pi t}{L}\right)\,dt$$

Notice that these results revert to the expressions $(9.1a, b, c)$ when $L = \pi$, as indeed they should.

9.3 *Convergence of Fourier series*

It is of importance to consider the convergence properties of Fourier series, in other words what happens to the partial sums $S_N(f)$ as $N \to \infty$. We would hope that in this limit the partial sums would approach $f(t)$ so that we could approximate the true value of $f(t)$ at any given point to a given accuracy just by taking enough terms in the requisite partial sum. Unfortunately, this cannot be guaranteed. Rather, instead of being able to prove that the partial sums converge to $f(t)$ at every point, it is only possible to be assured that the integral of the squared difference of the partial sums and the function goes to zero.

THEOREM **9.5.** *Assume* $\int_{-L}^{L} f(t)^2\,dt < \infty$. *Then*

$$\lim_{N\to\infty}\int_{-L}^{L} (S_N f(t) - f(t))^2\,dt = 0.$$

In other words, while we cannot be sure of the behaviour of the partial sums at any one single given point, we do know that the integral of the square of the difference does approach zero as $N \to \infty$. The consequence of this is that while $\mathrm{FS}_f(t) = f(t)$ at almost all points, there could be a finite number of locations in the interval $[-L, L]$, where $\mathrm{FS}_f(t) \neq f(t)$.

An additional issue arises for functions which possess a discontinuity. The function in Example 9.3 has a jump at $t = 0$; for $\pi < t \leq 0$ then $f(t) = 1$ but for $0 < t \leq \pi$ we have $f(t) = -1$. What does the Fourier series converge to at $t = 0$? This issue is settled by the following theorem.

THEOREM **9.6.** *Provided that $f(t)$ and $f'(t)$ are bounded and piecewise continuous on $[-L, L]$, the Fourier series will converge to (be equal to) $f(t)$* except at points of discontinuity, *where it will converge to the average of the right- and left-hand limits of $f(t)$ at that point, i.e.*

$$\frac{f(t^+) + f(t^-)}{2}$$

where $f(t^+)$ is the right-hand limit and $f(t^-)$ is the left-hand limit.

EXAMPLE **9.7.** *(Example 9.4 revisited) We see that $f(t)$ and $f'(t)$ are bounded and piecewise continuous on $[-\pi, \pi]$ (with the only discontinuity point being 0).*[3] *Since for this function we have $f(0^-) = 0$ and $f(0^+) = \pi$, then by Theorem 9.6 the Fourier series $\mathrm{FS}_f(t)$ converges to the average of these values, ie. $\pi/2$ at $t = \cdots, -2\pi, 0, 2\pi, \cdots$. The graph of the Fourier series is then as shown in Figure 9.6.*

[3] Note that $f'(t)$ considered on its full domain also has discontinuity points in the odd multiples of π.

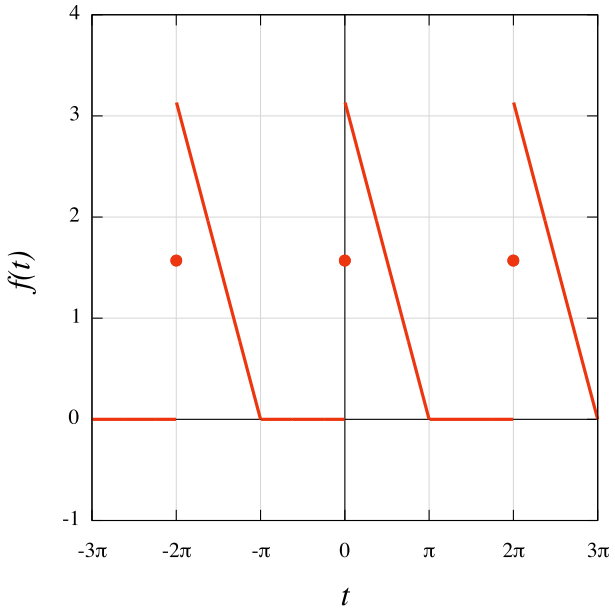


Figure 9.6: Fourier series function of Example 9.4.

Note that it is identical to the graph of $f(t)$ except that it takes the value $\frac{\pi}{2}$ at integer multiples of 2π whereas the function itself is 0 at these points. □

9.4 *Functions defined over a finite interval*

Writing a function in terms of a Fourier series is convenient for many calculations. Up to now we have only discussed strictly periodic functions but the ideas of Fourier series can be extended and applied to many functions that are defined on a finite interval but which appear to have no intrinsic periodic properties.

Suppose we have $f(t)$ defined on some finite interval of length $2L$ given by $-L < t \leq L$. (It might seem restrictive to assume that the interval is centred on $t = 0$. However if the interval is not centred on the origin it is straightforward to apply a translation and consider the function in terms of a new co-ordinate t' for which the centre is at $t' = 0$.)

We can now extend $f(t)$ to all real values of t by defining the *periodic extension* of $f(t)$.

Definition 9.8. *(Periodic extension)*
Let $f(t)$ be a function defined on the interval $(-L, L]$. The periodic extension *of $f(t)$ is the function $\phi(t)$ defined by:*

$$\phi(t) = f(t), \quad -L < t \leq L, \text{ and } \phi(t + 2L) = \phi(t) \quad \forall t.$$

Now $\phi(t)$ is defined for all values of t and is naturally a periodic function of period $2L$. Hence we are able to apply the theory of Fourier series to $\phi(t)$.

Example 9.9. *The graph of the Fourier series of the periodic extension of e^t, $-1 < t \leq 1$ is illustrated in Figure 9.7.*

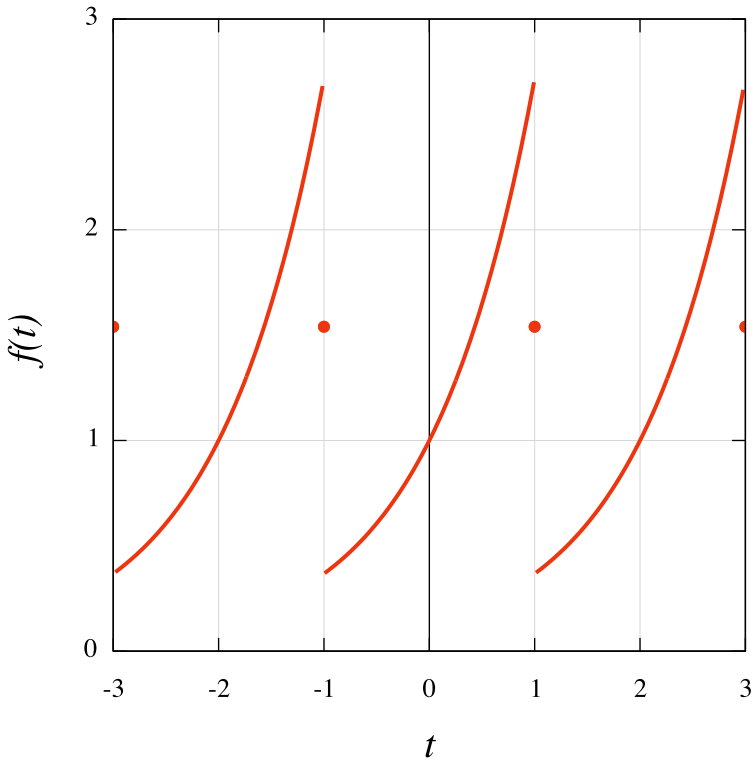


Figure 9.7: Fourier series of the periodic extension of e^t.

Note that the periodic extension is discontinuous at $t = \cdots, -5, -3, -1, 1, 3, 5, \cdots$. □

9.5 *Even and odd functions*

Definition 9.10. *(Even functions)*
A function $f(t)$ if even *if and only if $f(-t) = f(t)$ for all t. The graph of an* even function *is symmetrical in the vertical axis.*

Simple examples of even functions include

$$f(t) = 1 \quad , \quad f(t) = t^2 \quad , \quad f(t) = \cos t.$$

In particular all the **even** power functions $f(t) = t^{2n}$ are even.

Definition 9.11. *(Odd functions)*
A function $f(t)$ if odd *if and only if $f(-t) = -f(t)$ for all t. The graph of an* odd function is 180° *rotationally symmetric around the origin.*

Elementary examples of odd functions include

$$f(t) = t \quad , \quad f(t) = t^3 \quad , \quad f(t) = \sin t,$$

together with the function in Example 9.3. In particular all the **odd** power functions $f(t) = t^{2n+1}$ are odd.

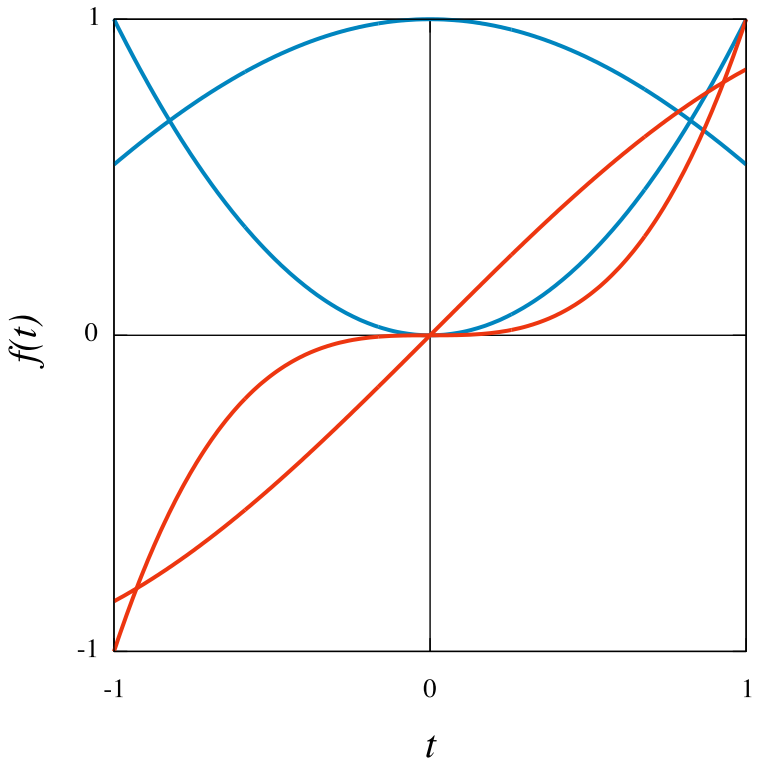


Figure 9.8: Graphs of the even functions $\cos t$ and t^2, and the odd functions $\sin t$ and t^3.

Properties of odd and even functions:

$$(\text{even}) + (\text{even}) = (\text{even}), \quad (\text{odd}) + (\text{odd}) = (\text{odd})$$
$$(\text{even}) \cdot (\text{even}) = (\text{even}), \quad (\text{odd}) \cdot (\text{odd}) = (\text{even})$$
$$(\text{odd}) \cdot (\text{even}) = (\text{odd}), \quad (\text{even}) \cdot (\text{odd}) = (\text{odd})$$

$$\int_{-L}^{L} (\text{odd}) dt = 0$$

If $f(t)$ is **even**: $\int_{-L}^{L} f(t)dt = 2\int_{0}^{L} f(t)dt$

Try proving them

In words, this tells us that the sum of two even (odd) functions is itself even (odd). The product of two even or two odd functions is even while the product of an odd and an even function is odd. The integral results are particularly important for they facilitate some great simplifications in the calculation of Fourier series.

9.6 *Fourier cosine series for even functions*

Even functions must have even Fourier series and hence $b_n = 0$ for all n, giving a *Fourier cosine series*. There is nothing particularly special about a Fourier cosine series; really it is little more than a standard Fourier series with the property that all its sine terms are absent because the coefficients b_n all happen to vanish. We can use the integration properties of odd and even functions to verify that $b_n = 0$ when $f(t)$ is an even function. From its definition

$$\begin{aligned} b_n &= \frac{1}{L}\int_{-L}^{L} f(t)\sin\left(\frac{n\pi t}{L}\right) dt \\ &= \frac{1}{L}\int_{-L}^{L} (\text{even})(\text{odd})\, dt \\ &= \frac{1}{L}\int_{-L}^{L} (\text{odd})\, dt \\ &= 0. \end{aligned}$$

The Fourier series of an even function $f(t)$ is the cosine series

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi t}{L}\right),$$

where

$$a_0 = \frac{2}{L}\int_0^L f(t)\,dt \quad \text{and} \quad a_n = \frac{2}{L}\int_0^L f(t)\cos\left(\frac{n\pi t}{L}\right)dt$$

EXAMPLE **9.12.** *Determine the Fourier series of the even ('Hats'[4]) function*

$$f(t) = \begin{cases} \pi + t, & -\pi < t \le 0 \\ \pi - t, & 0 < t \le \pi \end{cases} \quad \textit{and} \quad f(t+2\pi) = f(t).$$

[4] Note that the name 'hats' function derives from the from of its graph sketched in Figure 9.9.

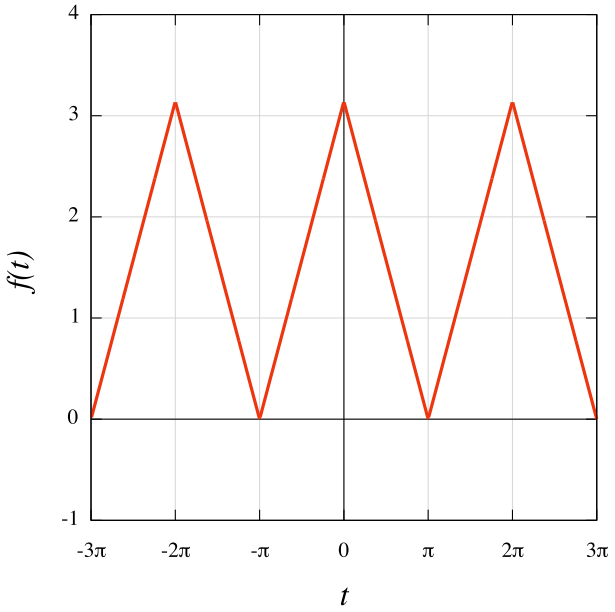


Figure 9.9: Fourier cosine series for Example 9.12.

Solution: *Since f is even, its Fourier series is a cosine series. We can compute:*

$$a_0 = \frac{2}{\pi}\int_0^\pi (\pi - t)\,dt = \pi$$

and

$$a_n = \frac{2}{\pi}\int_0^\pi (\pi - t)\cos(nt)\,dt = \frac{2(1-(-1)^n)}{n^2\pi} = \begin{cases} 0 & \textit{if } n \textit{ even} \\ \dfrac{4}{n^2\pi} & \textit{if } n \textit{ odd} \end{cases}$$

and hence the Fourier cosine series is

$$\mathrm{FS}_f(t) = \frac{\pi}{2} + \frac{4}{\pi}\sum_{n=1,\ n\ \text{odd}}^{\infty}\frac{\cos(nt)}{n^2} = \frac{\pi}{2} + \frac{4}{\pi}\left(\cos t + \frac{\cos(3t)}{9} + \frac{\cos(5t)}{25} + \cdots\right).$$

□

9.7 *Fourier sine series for odd functions*

Odd functions must have odd Fourier series and hence $a_n = 0$ for

all n leading to a *Fourier sine series*. Again it is relatively straightforward to check that the $a_n = 0$ because from the definition

$$\begin{aligned} a_n &= \frac{1}{L}\int_{-L}^{L} f(t)\cos\left(\frac{n\pi t}{L}\right)\,dt \\ &= \frac{1}{L}\int_{-L}^{L} (\text{odd})(\text{evem})\,dt \\ &= \frac{1}{L}\int_{-L}^{L} (\text{odd})\,dt \\ &= 0. \end{aligned}$$

The Fourier series of an odd function $f(t)$ is the sine series

$$\sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi t}{L}\right),$$

where

$$b_n = \frac{2}{L}\int_{0}^{L} f(t)\sin\left(\frac{n\pi t}{L}\right)\,dt$$

Example **9.13.** *Determine the Fourier series of the function in Example 9.3.*

Solution: *Since $f(t)$ is an odd function, its Fourier series is a sine series. We compute*

$$b_n = \frac{2}{\pi}\int_{0}^{\pi}(-1)\sin(nt)\,dt = \frac{2(\cos(n\pi)-1)}{n\pi} = \begin{cases} 0 & \text{if } n \text{ even} \\ -\dfrac{4}{n\pi} & \text{if } n \text{ odd} \end{cases},$$

and hence the Fourier sine series is

$$\text{FS}_f(t) = -\frac{4}{\pi}\sum_{n=1,n\,\text{odd}}^{\infty}\frac{\sin(nt)}{n}.$$

By Theorem 9.6 we know this converges to function shown in Figure 9.10.

□

9.8 Half-range expansions

Suppose that a function $f(t)$ is only defined on $[0, L]$. We could use the ideas described above to create a periodic function and hence derive a Fourier series representation. However, with the function defined on $[0, L]$ it is possible to extend the function in such a way that the resulting series contains only cosine terms or, if

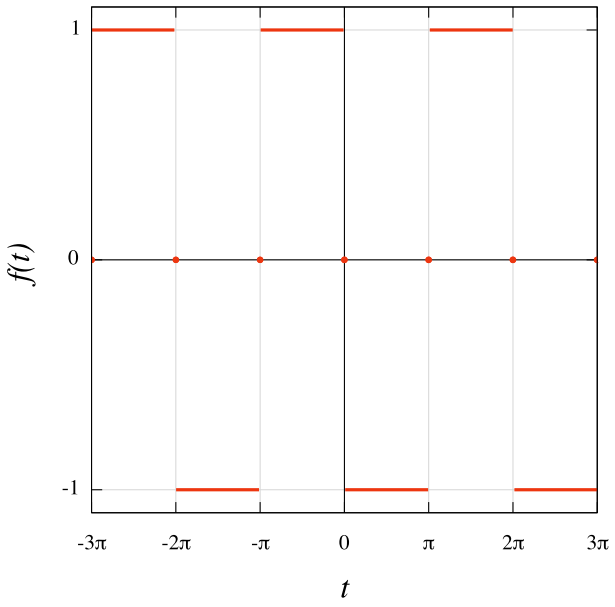


Figure 9.10: Fourier sine series for Example 9.13.

the extension is made in another way, such that the series contains only sine terms.

To see how to do this, we extend the domain of definition to $[-L, L]$, called a *half-range expansion*, in two ways. We accomplish this by defining two new functions $g(t)$ and $h(t)$ according to the following recipes.

Even expansion:

$$g(t) = \begin{cases} f(t) & \text{if } 0 \leq t \leq L \\ f(-t) & \text{if } -L \leq t \leq 0. \end{cases}$$

Now $g(t)$ is an even function by construction; therefore the series for $g(t)$ (or more precisely for its periodic extension) will be a Fourier cosine series.

Odd expansion:

$$h(t) = \begin{cases} f(t) & \text{if } 0 < t \leq L \\ 0 & \text{if } t = 0 \\ -f(-t) & \text{if } -L \leq t < 0. \end{cases}$$

This time our function is an odd one so will be given by a Fourier sine series.

As an example look at the sketches in Figure 9.11. Here a function $f(t)$ is defined for $0 < t < 2$ (left panel). In the centre is shown the even expansion of $f(t)$, that is the function $g(t)$ given above. This function is now defined on $-2 < t < 2$ and is clearly even (its graph is symmetric about the vertical axis). On the other hand, the right diagram illustrates the odd expansion $h(t)$. This time the graph possesses the characteristic 180^o rotational symmetry about the origin indicative of an odd function. The two extended functions $g(t)$ and $h(t)$ clearly must be given by Fourier cosine and Fourier sine series respectively. Notice that these two (different)

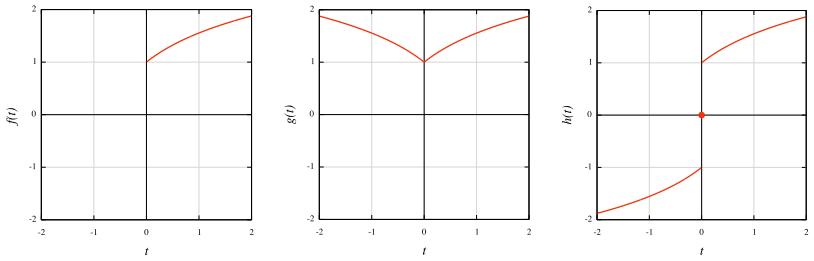


Figure 9.11: Original function and its even and odd expansions.

series converge to the same $f(t)$ for $0 < t < 2$ as both $g(t)$ and $h(t)$ equal $f(t)$ here but will naturally converge to different values for $t < 0$.

EXAMPLE **9.14.** *Find the Fourier series of the even and odd expansions of*

$$f(t) = t^2 \quad , \quad 0 \leq t \leq 1.$$

Solution:

Even expansion: The even expansion is just $g(t) = t^2$, $-1 \leq t \leq 1$. *We find the Fourier coefficients (as* $g(t)$ *is even,* $b_n = 0$*):*

$$a_0 = 2\int_0^1 t^2\,dt = \frac{2}{3},$$

$$a_n = 2\int_0^1 t^2\cos(n\pi t)\,dt = \frac{4(-1)^n}{\pi^2 n^2}.$$

Then the Fourier cosine series is

$$\frac{1}{3} + \frac{4}{\pi^2}\sum_{n=1}^{\infty}\frac{(-1)^n}{n^2}\cos(n\pi t).$$

This series converges to $g(t)$ *on* $[-1,1]$ *so in particular, it converges to* $f(t)$ *for* $0 \leq t \leq 1$.

Odd expansion: The odd expansion is $h(t) = \begin{cases} t^2 & \text{if } 0 \leq t \leq 1 \\ -t^2 & \text{if } -1 \leq t < 0. \end{cases}$

We find the Fourier coefficients (as $h(t)$ *is odd,* $a_0 = a_n = 0$*):*

$$b_n = 2\int_0^1 t^2\sin(n\pi t)\,dt = \frac{-4 - 2(n^2\pi^2 - 2)(-1)^n}{\pi^3 n^3}.$$

Hence the Fourier sine series is

$$-\frac{2}{\pi^3}\sum_{1}^{\infty}\frac{2 + (n^2\pi^2 - 2)(-1)^n}{n^3}\sin(n\pi t).$$

This series converges to $h(t)$ *on* $(-1,1)$ *so in particular, it converges to* $f(t)$ *for* $0 \leq t < 1$.

Notice these two series look very different, but they converge to the same value $f(t) = t^2$ *for* $0 \leq t < 1$*!* □

9.9 *Parseval's theorem (not for assessment)*

There is a relationship between the sum of the squares of all of the Fourier coefficients of a function and the integral of the square of the function itself over one period. This relationship turns out to be very useful in Engineering, Physics and other branches of Mathematics.

THEOREM **9.15** (Parseval's theorem). *If a 2π-periodic, piecewise continuous on $[-\pi, \pi]$, bounded function $f(t)$ has a Fourier series given by*

$$\mathrm{FS}_f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos(nt) + b_n \sin(nt)\right)$$

then

$$\frac{1}{\pi} \int_{-\pi}^{\pi} [f(t)]^2 \, dt = \frac{a_0^2}{2} + \sum_{n=1}^{\infty} \left(a_n^2 + b_n^2\right).$$

The proof of this result is omitted here. What is more important is to see how the theorem enables us to derive results concerning the sums of infinite series of terms. We do this via a few examples.

EXAMPLE **9.16.** *We shall apply Parseval's theorem to Example 9.13. Recall the function*

$$f(t) = \begin{cases} 1, & -\pi < t \leq 0, \\ -1, & 0 < t \leq \pi, \end{cases} \quad \textit{and} \quad f(t + 2\pi) = f(t).$$

We found that its Fourier series is

$$\mathrm{FS}_f(t) = \sum_{n=1, n \,\mathrm{odd}}^{\infty} \left(-\frac{4}{n\pi}\right) \sin(nt).$$

Parseval's theorem says that

$$\frac{1}{\pi} \int_{-\pi}^{0} 1^2 \, dt + \frac{1}{\pi} \int_{0}^{\pi} (-1)^2 \, dt = \sum_{n=1,\ n\ odd}^{\infty} \left(-\frac{4}{n\pi}\right)^2$$

$$\Rightarrow \quad 1 + 1 = \frac{16}{\pi^2} \sum_{n=1,\ n\ odd}^{\infty} \frac{1}{n^2} \quad \Rightarrow \quad \sum_{n=1,\ n\ odd}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{8}.$$

□

We can use the result of the previous example to find the value of $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^2}$ by noting that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \sum_{n=1,\ n\ odd}^{\infty} \frac{1}{n^2} + \sum_{n=1,\ n\ even}^{\infty} \frac{1}{n^2}$$

and realising that

$$\sum_{k=1}^{\infty} \frac{1}{(2k)^2} = \frac{1}{4} \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{1}{4} \sum_{n=1}^{\infty} \frac{1}{n^2}.$$

Making use of the result in the previous example gives

$$\sum_{n=1}^{\infty}\frac{1}{n^2}=\frac{\pi^2}{8}+\frac{1}{4}\sum_{n=1}^{\infty}\frac{1}{n^2}\qquad\Rightarrow\qquad\sum_{n=1}^{\infty}\frac{1}{n^2}=\frac{\pi^2}{6}.$$

Leonhard Euler first proved that equality in 1741 (by an entirely different method, though).

EXAMPLE **9.17.** *We shall apply Parseval's theorem to Example 9.4. Recall that the function is*

$$f(t)=\begin{cases}0, & -\pi<t\le 0,\\ \pi-t, & 0<t\le\pi,\end{cases}\quad\text{and}\quad f(t+2\pi)=f(t)$$

with Fourier series

$$\mathrm{FS}_f(t)=\frac{1}{2}\cdot\frac{\pi}{2}+\sum_{n=1,\ n\text{ odd}}^{\infty}\frac{2}{\pi n^2}\cos(nt)+\sum_{n=1}^{\infty}\frac{1}{n}\sin(nt).$$

Then Parseval's theorem tells us that

$$\frac{1}{\pi}\int_0^{\pi}(\pi-t)^2dt=\frac{\pi^2}{8}+\sum_{n=1,\ n\ odd}^{\infty}\left(\frac{2}{\pi n^2}\right)^2+\sum_{n=1}^{\infty}\left(\frac{1}{n}\right)^2$$

$$\Rightarrow\quad\frac{\pi^2}{3}=\frac{\pi^2}{8}+\frac{4}{\pi^2}\sum_{n=1,\ n\ odd}^{\infty}\frac{1}{n^4}+\sum_{n=1}^{\infty}\frac{1}{n^2}$$

$$\Rightarrow\quad\frac{5\pi^2}{24}=\frac{4}{\pi^2}\sum_{n=1,\ n\ odd}^{\infty}\frac{1}{n^4}+\frac{\pi^2}{6}.$$

$$\Rightarrow\quad\sum_{n=1,\ n\ odd}^{\infty}\frac{1}{n^4}=\frac{\pi^4}{96}.$$

□

An application of Parseval's theorem to a suitably chosen function can often yield equivalent results for other infinite sums that are often difficult to evaluate by other means.

EXERCISE **9.9.1.** *Use* $\displaystyle\sum_{n=1,\ n\ odd}^{\infty}\frac{1}{n^4}=\frac{\pi^4}{96}$ *to determine the sum of the p-series with* $p=4$.

9.10 Differentiation of Fourier series

If we wish to differentiate a function expressed as a Fourier series it is tempting to simply differentiate each term in the infinite series one by one. Very often in mathematics we need to be ultra-careful when dealing with infinite series because results that look as if they ought to be reasonable and sensible are not always true! Therefore it is not an obvious result that the differentiation of a Fourier series of a function is possible *term by term* but, fortunately, it can be proved to yield something that is useful. In particular, for a 2π-periodic function, if

$$\mathrm{FS}_f(t)=\frac{a_0}{2}+\sum_{n=1}^{\infty}a_n\cos(nt)+\sum_{n=1}^{\infty}b_n\sin(nt)$$

then

$$(\mathrm{FS}_f)'(t) = -\sum_{n=1}^{\infty} na_n \sin(nt) + \sum_{n=1}^{\infty} nb_n \cos(nt).$$

Moreover, the following theorem tells us what $f'(t)$ is, if f is continuous.

THEOREM **9.18.** *If f is* **continuous**, *then $(\mathrm{FS}_f)'(t) = f'(t)$ at points t where $f(t)$ is differentiable.*

In particular, note the requirement that the function $f(t)$ be continuous; if it is not continuous then the result does not necessarily follow.

EXAMPLE **9.19.** *Recall the 'Hats' function in Example 9.12; this function is continuous. It is differentiable except at points an integer multiple of π. Previously we showed that*

$$\mathrm{FS}_f(t) = \frac{\pi}{2} + \frac{4}{\pi} \sum_{n=1,\ n \text{ odd}}^{\infty} \frac{\cos(nt)}{n^2}.$$

By Theorem 9.18, we have

$$f'(t) = (\mathrm{FS}_f)'(t) = -\frac{4}{\pi} \sum_{n=1,\ n \text{ odd}}^{\infty} \frac{\sin(nt)}{n},$$

for all t not a multiple of π.

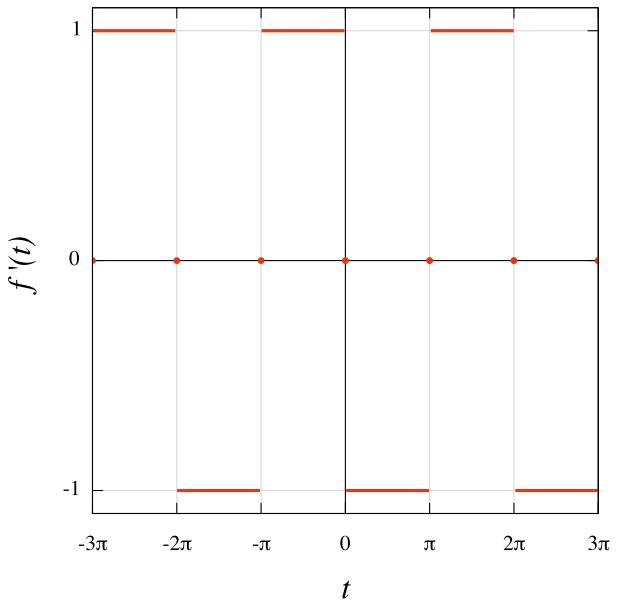


Figure 9.12: Derivative of the Fourier series of the Hats function.

The graph of the original function (and of its Fourier series) appeared in Example 9.12. We recognise from Example 9.13 that $(\mathrm{FS}_f)'(t)$ is the Fourier series of the function from Example 9.3, so the graph of $(\mathrm{FS}_f)'(t)$ is shown in Figure 9.12. Note that it has the value 0 at multiples of π, the average of the left and right limits, but that $f'(t)$ is not defined at multiples of π.

□

Recall that a function cannot be differentiated at points of discontinuity. There are also problems with the convergence of the differentiated series if $f(t)$ is not continuous, as illustrated by the following example.

EXAMPLE **9.20.** *Let $f(t)$ be the 'Slopes' function:*

$$f(t) = \frac{t}{2} \quad , \quad -\pi < t \leq \pi \quad , \quad f(t+2\pi) = f(t).$$

The graph of its Fourier series is shown in Figure 9.13. Notice it is discontinuous at odd multiples of π.

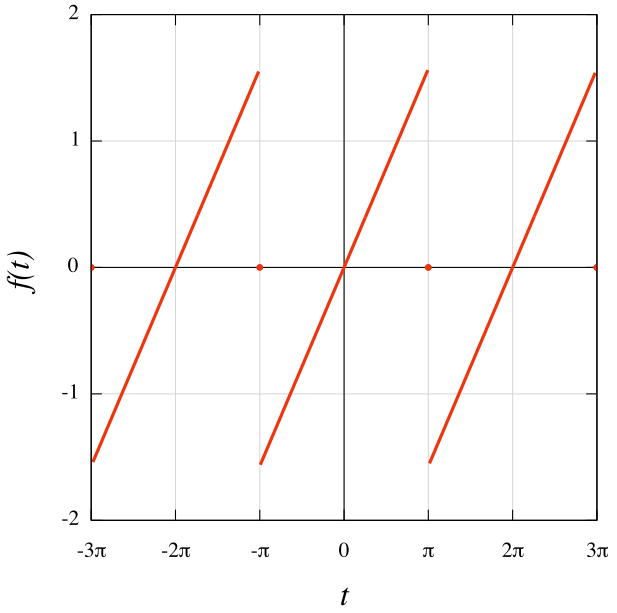


Figure 9.13: Fourier series of the Slopes function (Example 9.20)

Since f is an odd function, it has a sine Fourier series and we can show (exercise) that

$$\text{FS}_f = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \sin(nt).$$

If we naively find the derivative of the Fourier series term-by-term we deduce that

$$(\text{FS}_f)'(t) = \sum_{n=1}^{\infty} (-1)^{n+1} \cos(nt) = \cos t - \cos(2t) + \cos(3t) + \cdots . \tag{9.2}$$

We have an obvious problem here. We know[5] *For example, if we try to evaluate Equation* (9.2) *at $t = 0$ we have*

$$1 - 1 + 1 - 1 + \ldots$$

and this series clearly does not converge. Moreover, note that $t = 0$ is not a problem point for f and $f'(0) = 1/2$ so it is not the case that the differentiated series only fails at points where $f(t)$ is not differentiable or has some other problem. If we evaluate Equation (9.2) *at $t = \pi$ we have $-1 - 1 - 1 - \cdots$ which is also clearly nonsense. The actual derivative function is shown in Figure 9.14 and it is not defined at odd multiples of π.*

□

[5] Recall the 'Test for Divergence' Theorem 8.28 that if an infinite series is to converge then necessarily the n^{th} term in the series must go to zero as $n \to \infty$.

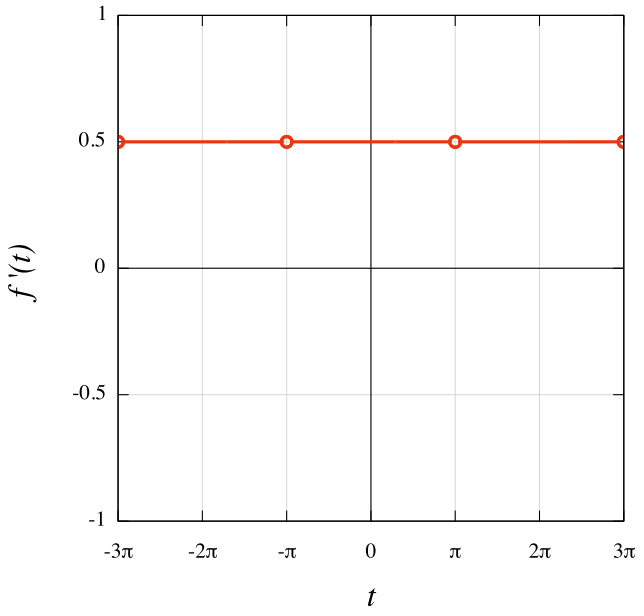


Figure 9.14: Actual derivative of the slopes function.

The conclusion is that we must not differentiate the Fourier series of a non-continuous function and expect to obtain results with any meaning.

9.11 *Integration of Fourier series*

It turns out that the Integration of Fourier series is more stable than differentiation in the sense that fewer potential problems tend to arise.

THEOREM **9.21.** *Let $f(t)$ be a 2π-periodic, piecewise continuous on $[-\pi, \pi]$, bounded function. If $a_0 = 0$ then*

$$\int_{-\pi}^{t} f(\alpha)\, d\alpha = \sum_{n=1}^{\infty} \frac{a_n}{n} \sin(nt) - \sum_{n=1}^{\infty} \frac{b_n}{n} \left(\cos(nt) - \cos(n\pi)\right).$$

Recall that $\sin(n\pi) = 0$ for integer values of n so the first right-hand side term is simpler than the second. We must have $a_0 = 0$ because

$$\int_{-\pi}^{t} \frac{a_0}{2}\, d\alpha = \frac{a_0}{2}(t + \pi)$$

which is not a Fourier series component.

EXAMPLE **9.22.** *Recall the Slopes function from Example 9.20 and its Fourier series*

$$\text{FS}_f = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \sin(nt).$$

Notice $a_0 = 0$ and the function is bounded and piecewise continuous on $[-\pi, \pi]$. Thus by Theorem 9.21:

$$\int_{-\pi}^{t} \frac{\alpha}{2}\, d\alpha = \sum_{n=1}^{\infty} (-1)^n \left(\frac{\cos(nt) - \cos(n\pi)}{n^2}\right) = \sum_{n=1}^{\infty} (-1)^n \left(\frac{\cos(nt) - (-1)^n}{n^2}\right).$$

From this we get

$$\frac{t^2-\pi^2}{4} = \sum_{n=1}^{\infty} \frac{(-1)^n \cos(nt)}{n^2} - \sum_{n=1}^{\infty} \frac{1}{n^2} = -\frac{\pi^2}{6} + \sum_{n=1}^{\infty} \frac{(-1)^n \cos(nt)}{n^2}.$$

Note that the average value of function $\frac{t^2-\pi^2}{4}$ *is* $-\frac{\pi^2}{6}$*, and that the integral function is continuous.*

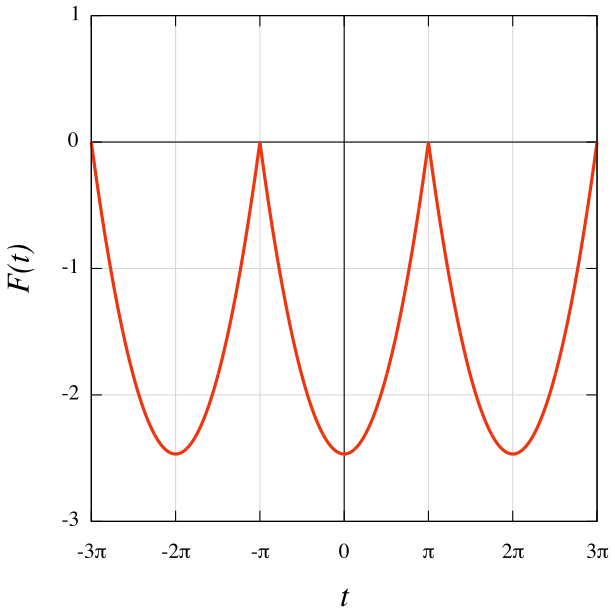


Figure 9.15: The integral of the Fourier series of the slopes function.

□

10
Differential equations

10.1 Introduction

In a vast number of situations a mathematical model of a system or process will result in an equation (or set of equations) involving not only functions of the dependent variables but also derivatives of some or all of those functions with respect to one or more of the variables. Such equations are called *differential equations*.

The simplest situation is that of a single function of a single independent variable, in which case the equation is referred to as an *ordinary differential equation*. A situation in which there is more than one independent variable will involve a function of those variables and an equation involving partial derivatives of that function is called a *partial differential equation*.

Notationally, it is easy to tell the difference. For example, the equation

$$\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} = f^2 \tag{10.1}$$

is a partial differential equation to be solved for $f(x, y)$, whereas

$$\frac{d^2 f}{dx^2} + 3\frac{df}{dx} + 2f = x^4 \tag{10.2}$$

is an ordinary differential equation to be solved for $f(x)$.

The *order* of a differential equation is the degree of the highest derivative that occurs in it. The partial differential equation 10.1 is first-order and the ordinary differential equation 10.2 is second-order. For partial differential equations the degree of a mixed derivative is the total number of derivatives taken. For example, the following partial differential equation for $f(x, t)$ has order five:

$$\frac{\partial^5 f}{\partial x^3 \partial t^2} + \frac{\partial^2 f}{\partial x^2} + \frac{\partial f}{\partial t} = 0. \tag{10.3}$$

An important class of differential equations are those referred to as *linear*. Roughly speaking, linear differential equations are those in which neither the function nor its derivatives occur in products, powers or nonlinear functions. Differential equations that are not linear are referred to as *nonlinear*. Equation 10.1 is nonlinear, whereas Equations 10.2 and 10.3 are both linear.

EXAMPLE **10.1.** *Classify the following differential equations with respect to (i) their nature (ordinary or partial), (ii) their order, and (iii) linear or nonlinear:*

$$(a)\quad \frac{\partial f}{\partial x} - \frac{\partial f}{\partial y} = 1. \qquad (d)\quad \frac{\partial u}{\partial x}\frac{\partial u}{\partial t} = x + t.$$

$$(b)\quad \frac{\partial^2 g}{\partial t^2} + g = \sin t. \qquad (e)\quad P^2\frac{d^2P}{dx^2} = x^5 + 1.$$

$$(c)\quad \frac{d^3y}{dx^3} + 8y = x\sin x. \qquad (f)\quad \frac{\partial^4 F}{\partial x \partial y^3} = t^2 F.$$

Solution:

1. *Equations (a), (b), (d) and (f) involve partial derivatives and are hence partial differential equations, whereas equations (c) and (e) involve ordinary derivatives and are hence ordinary differential equations.*

2. *Recall that the order of a differential equation is the degree of the highest derivative that occurs in it. The orders of the differential equations are as follows:*

   (a) *First-order.* (d) *First-order.*
   (b) *Second-order.* (e) *Second-order.*
   (c) *Third-order.* (f) *Fourth-order.*

3. *Recall that linear differential equations are those in which neither the function nor its derivatives occur in products, powers or nonlinear functions. It doesn't matter how the independent variables appear. We observe that equations (a), (b), (c) and (f) are linear whereas equations (d) and (e) are nonlinear.*

10.1.1 Solutions of differential equations

When asked to solve an algebraic equation, for example $x^2 - 3x + 2 = 0$, we expect the answers to be numbers. The situation with differential equations is much more difficult because we are being asked to find functions that will satisfy the given equation, for example in Example 10.1(a) we are asked for a function $f(x, y)$ that will satisfy the partial differential equation $\dfrac{\partial f}{\partial x} - \dfrac{\partial f}{\partial y} = 1$, and in Example 10.1(c) we are asked to find a function $y(x)$ that will satisfy $\dfrac{d^3y}{dx^3} + 8y = x\sin x$.

Unlike algebraic equations, which only have a discrete set of solutions (for example $x^2 - 3x + 2 = 0$ only has the solutions $x = 1$ or 2) differential equations can have whole families of solutions. For example, $y = Ce^{3x}$ satisfies the ordinary differential equation $\dfrac{dy}{dx} = 3y$ for *any* value of C.

If a differential equation is linear then there is a well-established procedure for finding solutions and we shall cover this in detail for ordinary differential equations. If an ordinary differential equation is nonlinear but is of first-order then we may also be able to find solutions.

The theory of partial differential equations is outside the scope of this unit.

10.1.2 Verification of solutions of differential equations

To get a feel for things (and to practice our algebra) we'll have a quick look at the relatively simple procedure of verifying solutions of differential equations by way of a few examples.

EXAMPLE **10.2.** *Verify that*

$$y(x) = C_1 e^{2x} + C_2 e^{-2x} - 2\cos x - 5x\sin x \tag{10.4}$$

is a solution of the ordinary differential equation

$$\frac{d^2y}{dx^2} - 4y = 25x\sin x \tag{10.5}$$

for any value of the constants C_1 and C_2.

Solution: *We need to calculate $\dfrac{d^2y}{dx^2}$. In order to do this we need the product rule to differentiate $x\sin x$. It gives*

$$\frac{d}{dx}(x\sin x) = \sin x + x\cos x$$

and

$$\frac{d^2}{dx^2}(x\sin x) = \frac{d}{dx}(\sin x + x\cos x) = 2\cos x - x\sin x.$$

Hence

$$\frac{d^2y}{dx^2} = 4C_1e^{2x} + 4C_2e^{-2x} - 8\cos x + 5x\sin x$$

and substitution of this and Equation 10.4 into Equation 10.5 quickly yields the required verification.

EXAMPLE **10.3.** *Verify that both*

$$f(x,y) = xy - \frac{1}{2}y^2 \quad \text{and} \quad f(x,y) = \sin(y-x) + \frac{1}{2}x^2$$

are solutions of the partial differential equation

$$\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} = x.$$

Solution: *In each case we need to calculate $\dfrac{\partial f}{\partial x}$ and $\dfrac{\partial f}{\partial y}$. For $f(x,y) = xy - \dfrac{1}{2}y^2$ we have*

$$\frac{\partial f}{\partial x} = y \quad \text{and} \quad \frac{\partial f}{\partial y} = x - y \quad \Rightarrow \quad \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} = y + x - y = x.$$

For $f(x,y) = \sin(y-x) + \frac{1}{2}x^2$ *we have*

$$\frac{\partial f}{\partial x} = -\cos(y-x) + x \quad \text{and} \quad \frac{\partial f}{\partial y} = \cos(y-x)$$

$$\Rightarrow \quad \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} = -\cos(y-x) + x + \cos(y-x) = x.$$

In both cases we have verified the solution of the partial differential equation.

10.2 *Mathematical modelling with ordinary differential equations*

For real-world systems changing continuously in time we can use derivatives to model the rates of change of quantities. Our mathematical models are thus differential equations.

EXAMPLE **10.4.** ***Modelling population growth.***

The simplest model of population growth is to assume that the rate of change of population is proportional to the population at that time. Let $P(t)$ *represent the population at time* t*. Then the mathematical model is*

$$\frac{dP}{dt} = rP \quad \textit{for some constant} \quad r > 0.$$

It can be shown (using a method called separation of variables*, which we shall learn shortly) that the function* $P(t)$ *that satisfies this differential equation is*

$$P(t) = P_0 e^{rt} \quad \textit{where} \quad P_0 \quad \textit{is the population at} \quad t = 0.$$

This model is clearly inadequate in that it predicts that the population will increase without bound if $r > 0$*. A more realistic model is the* logistic growth model

$$\frac{dP}{dt} = rP(C-P) \quad \textit{where} \quad r > 0 \quad \textit{and} \quad C > 0 \quad \textit{are constants.}$$

The method of separation of variables can be used to show that the solution of this differential equation is

$$P(t) = \frac{CP_0}{P_0 + (C-P_0)e^{-rt}} \quad \textit{where} \quad P_0 = P(0).$$

This model predicts that as time goes on, the population will tend towards the constant value C*, called the* carrying capacity*.*

EXAMPLE **10.5.** ***Newton's law of cooling***

The rate at which heat is lost from an object is proportional to the difference between the temperature of the object and the ambient temperature.

Let $H(t)$ be the temperature of the object (in $^\circ C$) at time t and suppose the fixed ambient temperature is $A^\circ C$. Newton's law of cooling says that

$$\frac{dH}{dt} = \alpha(A - H) \quad \text{for some constant} \quad \alpha > 0.$$

The method of separation of variables can be used to show that the solution of this differential equation is

$$H(t) = A + (H_0 - A)e^{-\alpha t} \quad \text{where} \quad H_0 = H(0).$$

This model predicts that as time goes on, the temperature of the object will approach that of its surroundings, which agrees with our intuition.

EXAMPLE **10.6.** ***One tank mixing process***

Suppose we have a tank of salt water and we allow fresh water into the tank at a rate of F m^3/sec, and allow salt water out of the tank at the same rate, as illustrated in Figure 10.1. Note this is a volume rate, and that the volume V of the tank is maintained constant. We assume instantaneous mixing so that the tank has a uniform concentration.

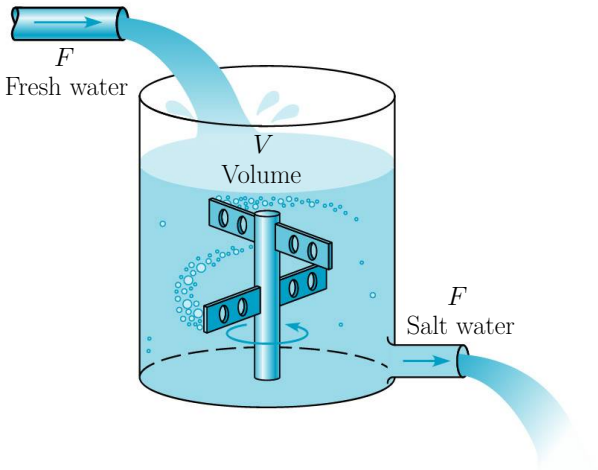


Figure 10.1: A mixing tank.

Let $y(t)$ represent the salt concentration of the water (kg/m^3) in the tank at time t and $a(t)$ represent the amount of salt (kg). We have $y(t) = \frac{a(t)}{V}$. The tank starts with an amount of salt a_0 kg.

The rate at which salt is being removed from the tank at time t is given by

$$\frac{da}{dt} = -y(t) \times (\text{flow rate}) = -Fy(t) = -\frac{F}{V}a(t) = -\alpha a(t)$$

where $\alpha = \frac{F}{V}$ is a positive constant. This equation has the solution $a(t) = a_0 e^{-\alpha t}$, which approaches zero as $t \to \infty$ (as expected).

Consider the same tank which is now filled with fresh water. Water polluted with q kg/m^3 of some chemical enters the tank at a rate of F m^3/sec, and polluted water exits the tank at the same rate. We again assume instantaneous mixing so that the tank has a uniform concentration.

Let $y(t)$ represent the concentration of pollutant (kg/m^3) in the water in the tank at time t and $a(t)$ represent the amount of pollutant (kg). We again have $y(t) = \frac{a(t)}{V}$. The rate at which pollutant is being added to the tank at time t is given by

$$\frac{da}{dt} = (\text{amount of pollutant added per second}) - (\text{amount of pollutant removed per second}).$$

That is,

$$\frac{da}{dt} = qF - Fy(t) = qF - \frac{F}{V}a(t).$$

Alternatively, we can obtain a differential equation for the concentration $x(t)$ by dividing through the above equation by V to give

$$\frac{dy}{dt} = \frac{F}{V}(q - y) \quad \Rightarrow \quad \frac{dy}{dt} = \alpha(q - x)$$

where $\alpha = \frac{F}{V}$ is a positive constant. Notice that this is essentially the same as the differential equation that we obtained for Newton's law of cooling.

10.3 First-order ordinary differential equations

Most first-order ordinary differential equations can be expressed (by algebraic re-arrangement if necessary) in the form

$$\frac{dy}{dx} = f(x, y) \tag{10.6}$$

where the function $f(x, y)$ is known, and we are asked to find the solution $y(x)$.

10.3.1 Direction fields

Equation 10.6 means that for any point in the $xy-$plane (for which f is defined) we can evaluate the gradient $\frac{dy}{dx}$ and represent this graphically by means of a small arrow representing the vector $\left(1, \frac{dy}{dx}\right)$. If we do this for a whole grid of points in the xy-plane and place all of the arrows on the same plot we produce what is called a *direction field* or *slope field*. Figure 10.2 displays the direction field in the case where $f(x, y) = y^2 - x^2$.

A solution of Equation 10.6 is a function relating y and x which geometrically is a curve in the $xy-$plane. Since this solution satisfies the differential equation, the curve is such that its gradient is the same as the direction field vector at any point on the curve.

That is, the direction field is a collection of arrows that are tangential to the solution curves. This observation enables us to roughly sketch solution curves without actually solving the differential equation, as long as we have a device to plot the direction

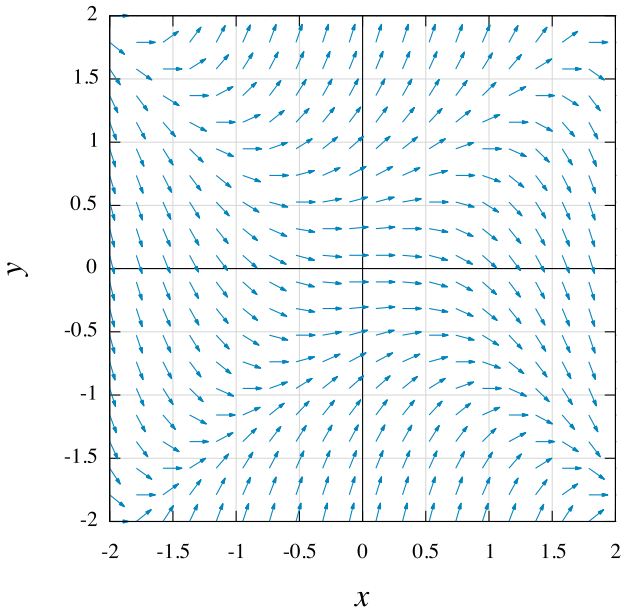


Figure 10.2: The direction field of $\frac{dy}{dx} = y^2 - x^2$.

field. We can indeed sketch many such curves (called a *family of solution curves*) superimposed on the same direction field.

EXAMPLE **10.7.** *The direction field of* $\frac{dy}{dx} = y^2 - x^2$ *along with three (disjoint) solution curves through the points* $(x, y) = (0, 1)$, $(0, 0)$ *and* $(0, -2)$ *is shown in Figure 10.3.*

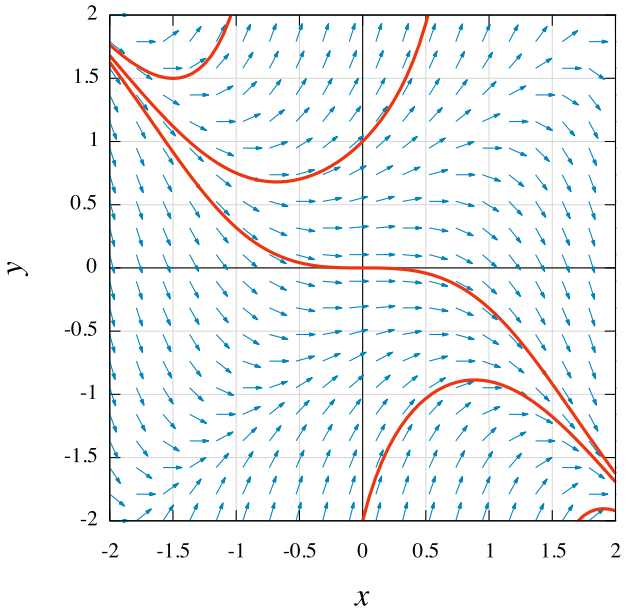


Figure 10.3: Three solution curves for $\frac{dy}{dx} = y^2 - x^2$.

REMARK **10.8.** *Note that we will* not *be able to solve the differential equation in Example 10.7 using the techniques we will cover in this unit – this differential equation is known as a* Ricatti differential equation, *which are notoriously difficult to solve.*

EXAMPLE **10.9.** *The direction field of* $\dfrac{dy}{dx} = 3y + e^x$ *along with three solution curves are shown in Figure 10.4. The top curve is the solution that goes through* $(x, y) = (0, 1)$*, the middle curve is the solution that goes through* $(x, y) = (0, 0)$ *and the bottom curve is the solution that goes through* $(x, y) = (0, -1)$*.*

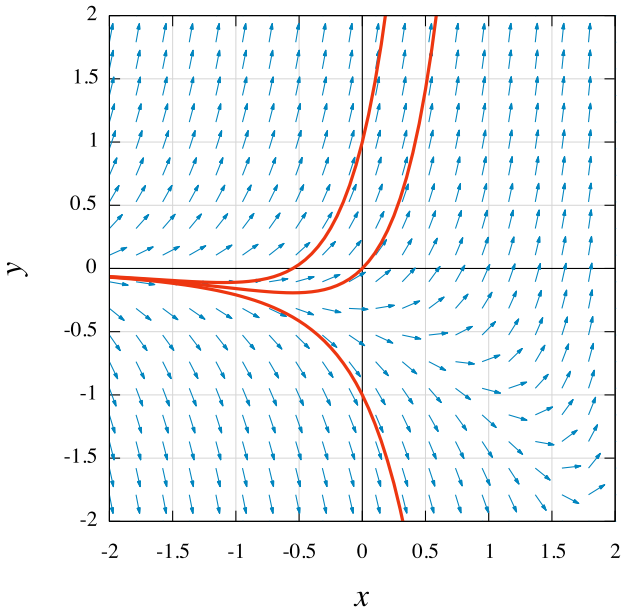


Figure 10.4: Three solution curves for $\dfrac{dy}{dx} = 3y + e^x$.

10.3.2 *Separation of variables*

A first-order differential equation is called *separable* provided that the function $f(x, y)$ may be written as the *product* of a function of x and a function of y, that is $f(x, y) = F(x)G(y)$.

Thus the variables x and y can be "separated" and placed on opposite sides of the equation; that is, given

$$\frac{dy}{dx} = F(x)G(y),$$

then by thinking of the derivative $\dfrac{dy}{dx}$ as a fraction[1] we have

[1] Just like we do with the chain rule.

$$\frac{1}{G(y)}\,dy = F(x)\,dx,$$

and then each side can be integrated, so that

$$\int \frac{1}{G(y)}\,dy = \int F(x)\,dx + C,$$

where the arbitrary integration constant C includes the constants from both integrals.

We then solve this equation (if possible) for y, which yields the *general solution* of the differential equation.

If we can *uniquely* solve for y, then the solution is called the *explicit solution* of the differential equation, but if we cannot uniquely

solve for y, then the solution is called the *implicit solution* of the differential equation.

EXAMPLE **10.10.** *Solve the first-order differential equation* $\dfrac{dy}{dx} = y^2 \sin x$.

Solution: *The differential equation is separable. The solution is given by*

$$\int y^{-2}\, dy = \int \sin x\, dx \Rightarrow -\frac{1}{y} = -\cos x + C \tag{10.7}$$

which is the implicit solution, where C is a constant of integration. We can re-arrange Equation 10.7 to get the explicit solution

$$y(x) = \frac{1}{\cos x + C}, \tag{10.8}$$

where we have arbitrarily re-named the integration constant from $-C$ to $+C$. Note that Equation 10.7 does not hold if $x = 0$. In such situations we have to investigate the original differential equation $\dfrac{dy}{dx} = y^2 \sin x$. In this case it turns out that $y(x) = 0$ is in fact a solution, but not of the form of Equation 10.8. Special situations like this are something that we should be aware of.

10.3.3 The integrating factor method

A first-order *linear* differential equation is one that may be written in the following *standard form*

$$\frac{dy}{dx} + f(x)y = g(x),$$

where $f(x)$ and $g(x)$ are arbitrary functions of x only. Note that if $g(x) \neq 0$, the differential equation is <u>not</u> separable.

To solve such a differential equation, we multiply both sides by a function $I(x)$ such that the left-hand-side may be written

$$I\left(\frac{dy}{dx} + fy\right) = \frac{d}{dx}(Iy),$$

thus allowing the left-hand-side to be integrated – hence the function $I(x)$ is called an *integrating factor*.

<u>If</u> an integrating factor $I(x)$ can be found, then the general solution is

$$\frac{d}{dx}(Iy) = Ig \Rightarrow Iy = \int Ig\, dx + C$$

which implies

$$y(x) = \frac{1}{I(x)} \int I(x)g(x)\, dx + \frac{C}{I(x)}. \tag{10.9}$$

How to we find the function $I(x)$? Since

$$I\left(\frac{dy}{dx} + fy\right) = \frac{d}{dx}(Iy),$$

we have by expanding the left-hand-side and using the product rule on the right-hand-side that

$$I\frac{dy}{dx} + Ify = y\frac{dI}{dx} + I\frac{dy}{dx} \Rightarrow Ify = y\frac{dI}{dx} \Rightarrow \frac{dI}{dx} = If.$$

This is a separable differential equation for $I(x)$, with solution

$$\frac{1}{I}\,dI = f\,dx \Rightarrow \ln(I) = \int f\,dx + C \Rightarrow I = \exp\left(\int f\,dx + C\right).$$

We want the simplest possible solution for $I(x)$, so we set $C = 0$.

Hence the integrating factor is[2]

$$I(x) = \exp\left(\int f(x)\,dx\right).$$

[2] Note that $\exp x$ is just another way of writing e^x but has the advantage that the "power" x is not a small superscript.

EXAMPLE **10.11.** . *Solve the first-order linear differential equation*

$$\frac{dy}{dx} - 3y = e^x. \tag{10.10}$$

As a guide to the shape of the solutions, the direction field for this differential equation along with a number of solution curves appears in Figure 10.4.

Solution: *The integrating factor is*

$$I(x) = \exp\left(\int -3\,dx\right) = e^{-3x}.$$

Multiplying the differential equation through by $I(x)$ gives

$$e^{-3x}\frac{dy}{dx} - 3e^{-3x}y = e^{-2x}.$$

We know that now the left-hand side of this can be rewritten in product form and so we obtain:

$$\frac{d}{dx}\left(e^{-3x}y\right) = e^{-2x},$$

which can be integrated to give

$$e^{-3x}y = \int e^{-2x}\,dx = -\frac{1}{2}e^{-2x} + C$$

hence

$$y(x) = -\frac{1}{2}e^x + Ce^{3x}.$$

REMARK **10.12.** *Note that we could have written down the solution immediately by appealing to Equation 10.9 but when learning the method it is instructive to follow through each step in the process in order to gain a better understanding of how it works.*

However, the general solution strategy is as follows:

1. *Write the linear first-order differential equation in standard form* $\frac{dy}{dx} + f(x)y = g(x)$ *and identify the functions* $f(x)$ *and* $g(x)$.
2. *Find the integrating factor* $I(x) = \exp\left(\int f(x)\,dx\right)$, *omitting the integration constant.*
3. *Find* $\int I(x)g(x)\,dx$, *omitting the integration constant.*
4. *The general solution is then* $y(x) = \frac{1}{I(x)}\int I(x)g(x)\,dx + \frac{C}{I(x)}$.

EXAMPLE **10.13.** *Solve the first-order linear differential equation*

$$\frac{dy}{dx} - \frac{1}{x}y = xe^x.$$

Solution: *Here we have* $f(x) = -\frac{1}{x}$ *and* $g(x) = xe^x$. *Then*

$$\int f(x)\,dx = \int -\frac{1}{x}\,dx = -\ln x = \ln\left(x^{-1}\right),$$

and hence

$$I(x) = \exp\left(\int f(x)\,dx\right) = e^{\ln\left(x^{-1}\right)} = x^{-1}.$$

Then

$$\int I(x)g(x)\,dx = \int \left(x^{-1}\right)\left(xe^x\right)\,dx = \int e^x\,dx = e^x,$$

and the general solution is therefore

$$\begin{aligned} y(x) &= \frac{1}{I(x)}\int I(x)g(x)\,dx + \frac{C}{I(x)} \\ &= \left(\frac{1}{x^{-1}}\right)\left(e^x\right) + \frac{C}{x^{-1}} \\ &= xe^x + Cx. \end{aligned}$$

10.3.4 Initial conditions

The values of constants of integration that arise when we solve differential equations can be determined by making use of other conditions (or restrictions) placed on the problem. For first-order differential equations, these conditions are called *initial conditions* and the combined differential equation plus initial condition is called an *initial value problem*.

EXAMPLE **10.14.** *Solve* $\dfrac{dy}{dx} - 3y = e^x$ *subject to* $y(0) = 1$, *that is,* $y = 1$ *when* $x = 0$.

Solution: *We have already seen this differential equation. It is Equation 10.10 and we have determined that its (most general) solution is given by*

$$y(x) = -\frac{1}{2}e^x + Ce^{3x}.$$

All we have to do is substitute $y = 1$ *and* $x = 0$ *and solve the resulting algebraic equation for* C. *We have*

$$1 = -\frac{1}{2}e^0 + Ce^0 \quad \Rightarrow \quad C = \frac{3}{2},$$

so the required solution is

$$y(x) = \frac{3e^{3x} - e^x}{2}. \tag{10.11}$$

The solution curve of Equation 10.11 appears in Figure 10.4.

EXAMPLE **10.15.** *Solve the initial value problem*

$$\frac{dy}{dx} = \frac{x^2}{y} \quad , \quad y(1) = 4.$$

Solution: *We observe that the differential equation is separable. The solution is:*

$$\int y\,dy = \int x^2\,dx \quad \Rightarrow \quad \frac{1}{2}y^2 = \frac{1}{3}x^3 + C$$

which implies

$$y(x) = \pm\sqrt{\frac{2}{3}x^3 + C},$$

where we have arbitrarily re-named the integration constant.

Notice that we have two different solutions to the differential equation, one positive and one negative. The initial condition $y(1) = 4$ *allows us to eliminate the negative solution, so we are left with*

$$y(x) = \sqrt{\frac{2}{3}x^3 + C},$$

and substituting into this $y = 4$ *and* $x = 1$ *gives*

$$4 = \sqrt{\frac{2}{3} + C} \quad \Rightarrow \quad C = \frac{46}{3} \quad \Rightarrow \quad y(x) = \sqrt{\frac{2x^3 + 46}{3}}.$$

10.4 Second-order ordinary differential equations

A general second-order differential equation may be written in the form

$$\frac{d^2y}{dx^2} = f\left(x, y, \frac{dy}{dx}\right),$$

where $f(x, y, y')$ is an arbitrary (but known!) function of x, y and y', and we wish to find a solution $y(x)$ that satisfies the given differential equation.

A second-order *linear* differential equation is one that may be written

$$\frac{d^2y}{dx^2} + p(x)\frac{dy}{dx} + q(x)y = g(x),$$

where $p(x)$, $q(x)$ and $g(x)$ are arbitrary functions of x. A second-order linear differential equation is said to be *homogeneous* if $g(x) = 0$, otherwise the differential equation is *nonhomogeneous* and $g(x)$ is called the *nonhomogeneous term*.

If $p(x) = p$ and $q(x) = q$ are constant functions, then we have a second-order linear differential equation with *constant coefficients*

$$\frac{d^2y}{dx^2} + p\frac{dy}{dx} + qy = g(x),$$

otherwise the differential equation is said to have *variable coefficients*.

Since two integration's are required to find a solution of a second-order differential equation and each integration produces an arbitrary integration constant, the general solution $y(x)$ will contain two integration constants, C_1 and C_2.

THEOREM **10.16.** *(Principal of Superposition)*

If y_1 and y_2 are two solutions of the second-order linear homogeneous differential equation

$$\frac{d^2y}{dx^2} + p(x)\frac{dy}{dx} + q(x)y = 0,$$

then the linear combination $C_1y_1 + C_2y_2$ is also a solution for any values of the constants C_1 and C_2.

Proof. If y_1 and y_2 are both solutions of $\frac{d^2y}{dx^2} + p(x)\frac{dy}{dx} + q(x)y = 0$, then

$$\frac{d^2y_1}{dx^2} + p(x)\frac{dy_1}{dx} + q(x)y_1 = 0 \quad \text{and} \quad \frac{d^2y_2}{dx^2} + p(x)\frac{dy_2}{dx} + q(x)y_2 = 0.$$

Now,

$$\frac{d^2}{dx^2}(C_1y_1 + C_2y_2) + p(x)\frac{d}{dx}(C_1y_1 + C_2y_2) + q(x)(C_1y_1 + C_2y_2)$$

$$= C_1\frac{d^2y_1}{dx^2} + C_2\frac{d^2y_2}{dx^2} + C_1p(x)\frac{dy_1}{dx} + C_2p(x)\frac{dy_2}{dx} + C_1q(x)y_1 + C_2q(x)y_2$$

$$= C_1\underbrace{\left(\frac{d^2y_1}{dx^2} + p(x)\frac{dy_1}{dx} + q(x)y_1\right)}_{0} + C_2\underbrace{\left(\frac{d^2y_2}{dx^2} + p(x)\frac{dy_2}{dx} + q(x)y_2\right)}_{0}$$

$$= 0.$$

□

Definition 10.17. *(Linear dependence)*
Let $y_1(x), y_2(x)$ be a set of functions. The set $y_1(x), y_2(x)$ is linearly dependent *on an interval I if there are constants C_1 and C_2, not both zero, so that*

$$C_1 y_1(x) + C_2 y_2(x) = 0$$

for every value of x in I.

The set $y_1(x), y_2(x)$ is linearly independent *if it is not linearly dependent (that is, the only possible way the above equation is satisfied is if $C_1 = 0 = C_2$).*

A simple way to check if two solutions y_1 and y_2 are linearly independent is to calculate a function called the *Wronskian* of y_1 and y_2, denoted by $W[y_1, y_2](x)$, which is defined below.

Definition 10.18. *(Wronskian of a set of two functions)*
Let $y_1(x), y_2(x)$ be a set of differentiable functions. The Wronskian *of the set $y_1(x), y_2(x)$, denoted by $W[y_1, y_2](x)$, is*

$$W[y_1, y_2](x) = \det \begin{bmatrix} y_1 & y_2 \\ y_1' & y_2' \end{bmatrix} = y_1 \frac{dy_2}{dx} - y_2 \frac{dy_1}{dx}.$$

Theorem 10.19. *(Wronskian and Linear Dependence)*
Let $y_1(x), y_2(x)$ be a set of differentiable functions. If $W[y_1, y_2](x) \neq 0$ for all x in some interval I, then y_1 and y_2 are linearly independent on I. If $W[y_1, y_2](x) = 0$ for every x in some interval I, then y_1 and y_2 are linearly dependent on I.

Remark 10.20. *To prove Theorem 10.19 and the next Theorem, we need some concepts of* linear algebra *which will not be covered until the next unit.*

Theorem 10.21. *(General Solution)*
Consider the second-order linear homogeneous differential equation

$$\frac{d^2y}{dx^2} + p(x)\frac{dy}{dx} + q(x)y = 0,$$

and suppose that $y_1(x)$ and $y_2(x)$ are linear independent solutions of this differential equation.

Then the general solution

$$y(x) = C_1 y_1(x) + C_2 y_2(x)$$

with arbitrary constants C_1 and C_2 includes every possible solution of the differential equation.

KEY CONCEPT **10.22.** *The conclusion from all this is: given a second-order linear homogeneous differential equation*

$$\frac{d^2y}{dx^2} + p(x)\frac{dy}{dx} + q(x)y = 0,$$

the general solution is

$$y(x) = C_1y_1(x) + C_2y_2(x)$$

*for arbitrary constants C_1 and C_2, where both y_1 and y_2 are solutions of the differential equation **and** the functions y_1 and y_2 are linearly independent, that is the Wronskian*

$$W[y_1, y_2](x) = \det\begin{bmatrix} y_1 & y_2 \\ y_1' & y_2' \end{bmatrix} = y_1\frac{dy_2}{dx} - y_2\frac{dy_1}{dx} \neq 0.$$

10.5 *Linear homogeneous second-order ordinary differential equations with constant coefficients*

For the remainder of this Chapter, we will only consider linear second-order ordinary differential equations with constant coefficients, which when homogeneous have the general form

$$\frac{d^2y}{dx^2} + p\frac{dy}{dx} + qy = 0,$$

where p and q are constants.

In seeking a solution technique, consider the first-order equation

$$p\frac{dy}{dx} + qy = 0,$$

which is a separable first-order differential equation with solution

$$\begin{aligned} \frac{dy}{dx} = -\frac{qy}{p} \quad &\Rightarrow \quad \int \frac{1}{y}\,dy = \int -\frac{q}{p}\,dx \\ &\Rightarrow \quad \ln y = -\frac{q}{p}x + C \\ &\Rightarrow \quad y(x) = Ce^{mt} \end{aligned}$$

where $m = -\dfrac{q}{p}$ and C is the constant of integration that we have arbitrarily re-named from e^C.

By analogy we attempt to find a solution to the second-order differential equation by assuming a solution of the form $y = e^{mx}$, and the differential equation becomes

$$e^{mx}(m^2 + pm + q) = 0 \Rightarrow m^2 + pm + q = 0,$$

which is the *characteristic equation* or *auxiliary equation* of the differential equation. Since it is a quadratic in m, it has two roots

$$m_1 = \frac{-p + \sqrt{p^2 - 4q}}{2} \quad , \quad m_2 = \frac{-p - \sqrt{p^2 - 4q}}{2}.$$

Hence there are three cases to consider, depending on whether the *discriminant* $p^2 - 4q$ is positive, negative or zero.

Case 1. Two real roots

In this case the discriminant is positive and we have two real distinct roots m_1 and m_2. Then $y_1 = e^{m_1 x}$ and $y_2 = e^{m_2 x}$ are two solutions of the differential equation, and the Wronskian is

$$W[y_1, y_2](x) = \det \begin{bmatrix} e^{m_1 x} & e^{m_2 x} \\ m_1 e^{m_1 x} & m_2 e^{m_2 x} \end{bmatrix} = (m_2 - m_1) e^{(m_1 + m_2)x},$$

which is never zero since $m_1 \neq m_2$. Hence the general solution of the differential equation is

$$y(x) = C_1 e^{m_1 x} + C_2 e^{m_2 x}.$$

EXAMPLE **10.23.** *Solve the differential equation*

$$\frac{d^2y}{dx^2} - 5\frac{dy}{dx} + 4y = 0.$$

Solution: *The characteristic equation is* $m^2 - 5m + 4 = 0$ *which factorizes into* $(m-1)(m-4) = 0$ *and hence the required solutions are* $m_1 = 1$ *and* $m_2 = 4$. *Then the general solution is*

$$y(x) = C_1 e^x + C_2 e^{4x}.$$

Case 2. Complex conjugate roots

In this case the discriminant is negative and we have two complex roots $m_1 = a + ib$ and $m_2 = a - ib$ that are complex conjugates of each other, where $a = -\frac{1}{2}p$ and $b = \frac{1}{2}\sqrt{4q - p^2}$. Then $y_1 = e^{m_1 x}$ and $y_2 = e^{m_2 x}$ are two solutions of the differential equation, and the Wronskian is again never zero since $m_1 \neq m_2$. The general solution of the differential equation is then

$$y(x) = C_1 e^{(a+ib)x} + C_2 e^{(a-ib)x}.$$

Recalling *Euler's formula* $e^{ix} = \cos x + i \sin x$, we have

$$\begin{aligned} y(x) &= C_1 e^{ax} e^{ibx} + C_2 e^{ax} e^{-ibx} \\ &= C_1 e^{ax} \left[\cos(bx) + i \sin(bx)\right] + C_2 e^{ax} \left[\cos(bx) - i \sin(bx)\right] \\ &= (C_1 + C_2) e^{ax} \cos(bx) + i(C_1 - C_2) e^{ax} \sin(bx) \\ &= C_1 e^{ax} \cos(bx) + C_2 e^{ax} \sin(bx), \end{aligned}$$

where we have arbitrarily re-named the two integration constants. Hence the general solution of the differential equation is

$$y(x) = C_1 e^{ax} \cos(bx) + C_2 e^{ax} \sin(bx).$$

EXAMPLE **10.24.** *Solve the differential equation*

$$\frac{d^2y}{dx^2} - 4\frac{dy}{dx} + 13y = 0.$$

Solution: *The characteristic equation is* $m^2 - 4m + 13 = 0$. *The quadratic formula gives the roots as* $m_{1,2} = 2 \pm 3i$ *and hence the general solution is*

$$y(x) = C_1 e^{2x}\cos(3x) + C_2 e^{2x}\sin(3x).$$

Case 3. Equal roots

In this case the discriminant is zero and we have one repeated root $m = -\frac{1}{2}p$, so we only know "half" of the general solution.

How to we find the other "half" of the solution, namely y_2? If we let $y(x) = v(x)y_1(x) = v(x)e^{-\frac{1}{2}px}$ for some function $v(x)$ to be found, then using the product rule we have

$$\frac{dy}{dx} = \left(e^{-\frac{1}{2}px}\right)\frac{dv}{dx} + v\left(-\frac{1}{2}pe^{-\frac{1}{2}px}\right) = e^{-\frac{1}{2}px}\left(\frac{dv}{dx} - \frac{1}{2}pv\right),$$

and using the product rule again we find that

$$\frac{d^2y}{dx^2} = e^{-\frac{1}{2}px}\left(\frac{d^2v}{dx^2} - p\frac{dv}{dx} + \frac{1}{4}p^2v\right).$$

Then the differential equation becomes

$$e^{-\frac{1}{2}px}\left(\frac{d^2v}{dx^2} - p\frac{dv}{dx} + \frac{1}{4}p^2v\right) + pe^{-\frac{1}{2}px}\left(\frac{dv}{dx} - \frac{1}{2}pv\right) + qve^{-\frac{1}{2}px} = 0,$$

which simplifies to

$$\frac{d^2v}{dx^2} + \left(-\frac{1}{4}p^2 + q\right)v = 0.$$

Since $p^2 - 4q = 0$, the coefficient of v in the above equation is zero, so we have

$$\frac{d^2v}{dx^2} = 0,$$

which can be integrated twice to give

$$v(x) = C_1 + C_2x.$$

Therefore

$$y(x) = v(x)y_1(x) = (C_1 + C_2x)e^{-\frac{1}{2}px} = C_1e^{-\frac{1}{2}px} + C_2xe^{-\frac{1}{2}px},$$

and the second linearly independent solution of the differential equation is therefore $y_2 = xe^{-\frac{1}{2}px}$.

REMARK **10.25.** *This is an example of a process called* reduction of order, *which is way to "build" the general solution of a second-order linear differential equation provided we can find a single solution* y_1.

So, if the characteristic equation has only one root m, and the general solution of the differential equation is

$$y(x) = C_1 e^{mx} + C_2 x e^{mx}.$$

Notice that the Wronskian is

$$W[y_1, y_2](x) = \det \begin{bmatrix} e^{mx} & xe^{mx} \\ me^{mx} & (1+mx)e^{mx} \end{bmatrix} = e^{2mx},$$

which is never zero since $m \neq 0$.

EXAMPLE **10.26.** *Solve the differential equation*

$$\frac{d^2y}{dx^2} + 6\frac{dy}{dx} + 9y = 0.$$

Solution: *The characteristic equation is $m^2 + 6m + 9 = 0$, which factorizes into $(m+3)^2 = 0$ and hence the required component solutions are e^{-3x} and xe^{-3x}. Hence the general solution of the differential equation is*

$$y(x) = C_1 e^{-3x} + C_2 x e^{-3x}.$$

KEY CONCEPT **10.27.** *In summary, to find the general solution of a linear homogeneous second-order ordinary homogeneous differential equation with constant coefficients of general form*

$$\frac{d^2y}{dx^2} + p\frac{dy}{dx} + qy = 0$$

where p and q are constants, find the roots of the characteristic equation

$$m^2 + pm + q = 0.$$

1. *If the roots m_1 and m_2 are real and unequal, then the general solution is*
$$y(x) = C_1 e^{m_1 x} + C_2 e^{m_2 x}.$$
2. *If the roots are complex conjugates $a \pm ib$, then the general solution is*
$$y(x) = C_1 e^{ax}\cos(bx) + C_2 e^{ax}\sin(bx).$$
3. *If there is a single (or repeated) root m, then the general solution is*
$$y(x) = C_1 e^{mx} + C_2 x e^{mx}.$$

10.6 Linear nonhomogeneous second-order ordinary differential equations with constant coefficients

Consider a linear nonhomogeneous second-order ordinary differential equations with constant coefficients

$$\frac{d^2y}{dx^2} + p\frac{dy}{dx} + qy = g(x),$$

where p and q are constants and the nonhomogeneous term $g(x)$ is an arbitrary function of x. For this differential equation we also consider the corresponding homogeneous differential equation

$$\frac{d^2y}{dx^2} + p\frac{dy}{dx} + qy = 0,$$

with general solution y_c, which we call the *complimentary solution*.

Definition 10.28. *(Particular solution)*
A particular solution y_p *of the nonhomogeneous differential equation*

$$\frac{d^2y}{dx^2} + p\frac{dy}{dx} + qy = g(x)$$

is a specific function that contains no arbitrary constants and satisfies the differential equation.

Example 10.29. *Recall Example 10.2, where we showed that*

$$y(x) = C_1e^{2x} + C_2e^{-2x} - 2\cos x - 5x\sin x$$

was a solution of the ordinary differential equation

$$\frac{d^2y}{dx^2} - 4y = 25x\sin x$$

for any value of the constants C_1 and C_2. With $C_1 = 0 = C_2$, a particular solution of this differential equation is simply

$$y_p(x) = -2\cos x - 5x\sin x.$$

Theorem 10.30. *(General Solution of a Nonhomogeneous Differential Equation)*

The general solution of a linear nonhomogeneous second-order ordinary differential equations with constant coefficients

$$\frac{d^2y}{dx^2} + p\frac{dy}{dx} + qy = g(x)$$

is

$$y(x) = y_c(x) + y_p(x),$$

where y_p is a particular solution of the nonhomogeneous differential equation and y_c is the general solution of the corresponding homogeneous differential equation

$$\frac{d^2y}{dx^2} + p\frac{dy}{dx} + qy = 0.$$

Proof.

$$\begin{aligned}\frac{d^2y}{dx^2} + p\frac{dy}{dx} + qy &= \left(\frac{d^2y_c}{dx^2} + \frac{d^2y_p}{dx^2}\right) + p\left(\frac{dy_c}{dx} + \frac{dy_p}{dx}\right) + q\left(y_c + y_p\right)\\ &= \left(\frac{d^2y_c}{dx^2} + p\frac{dy_c}{dx} + qy_c\right) + \left(\frac{d^2y_p}{dx^2} + p\frac{dy_p}{dx} + qy_p\right)\\ &= 0 + g(x) = g(x).\end{aligned}$$

□

There are two methods to find $y_p(x)$, either the *method of undetermined coefficients* (which is a very specific method) or *variation of parameters* (which is a much more general method).

10.6.1 Method of undetermined coefficients

The method of undetermined coefficients can be applied when the nonhomogeneous term $g(x)$ is:

- A polynomial;
- A linear combination of sines and cosines;
- An exponential function; or
- A combination of sums, differences and products of the above functions.

The idea behind this method is that the derivative of a polynomial is a polynomial, that of a trigonometric function is a trigonometric function, and that of an exponential function is an exponential function, meaning that we can make an intelligent guess for the form of $y_p(x)$.

Nonhomogeneous term $g(x)$	Form of trial particular solution $y_p(x)$
$a_n(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$	$A_n(x) = A_nx^n + A_{n-1}x^{n-1} + \cdots + A_1x + A_0$
$a_n(x)e^{\alpha x}$	$A_n(x)e^{\alpha x}$
$a_n(x)\sin(\beta x)$ or $a_n(x)\cos(\beta x)$	$A_n(x)\sin(\beta x) + B_n(x)\cos(\beta x)$
$a_n(x)e^{\alpha x}\sin(\beta x)$ or $a_n(x)e^{\alpha x}\cos(\beta x)$	$e^{\alpha x}\left[A_n(x)\sin(\beta x) + B_n(x)\cos(\beta x)\right]$

We formulate a guess for y_p using the above table and the following rules:

- **Basic rule:** If $g(x)$ is one of the functions listed in the first column, substitute the corresponding function from the second column and determine the unknown constants by equating coefficients.
- **Modification rule:** If a term in the choice for y_p is a solution of the homogeneous equation, then multiply this term by x.
- **Sum rule:** If $g(x)$ is a sum of functions listed in the first column, then substitute the corresponding sum of functions from the second column and solve for the unknown coefficients by equating coefficients.

This should come as no surprise – remember from Section 10.5 that the second solution for the case of an equal root of the characteristic equation was just x times the first solution.

EXAMPLE **10.31.** *Solve the differential equation* $\dfrac{d^2y}{dx^2} + 6\dfrac{dy}{dx} + 9y = 4x^2 + 5$.

Solution: *In Example 10.26, we found the complementary solution was* $y_c(x) = C_1e^{-3x} + C_2xe^{-3x}$. *We now need to determine a particular solution. Based on the Table of guesses we try*

$$y_p(x) = A_2x^2 + A_1x + A_0, \tag{10.12}$$

and we note that y_p *is not included already in* y_c. *Substitution of Equation 10.12 into the differential equation gives*

$$\begin{aligned} &2A_2 + 6(2A_2x + A_1) + 9(A_2x^2 + A_1x + A_0) = 4x^2 + 5 \\ \Rightarrow\quad &9A_2x^2 + (9A_1 + 12A_2)x + (9A_0 + 6A_1 + 2A_2) = 4a^2 + 5 \end{aligned}$$

and equating the coefficients of the powers of x on each side of this equation leads to a set of algebraic equations to solve for the unknowns A_0, A_1 *and* A_2*:*

$$9A_2 = 4 \quad , \quad 9A_1 + 12A_2 = 0 \quad , \quad 9A_0 + 6A_1 + 2A_2 = 5.$$

The solution of this set of equations is

$$A_0 = \frac{23}{27} \quad , \quad A_1 = -\frac{16}{27} \quad , \quad A_2 = \frac{4}{9}.$$

Hence the particular solution is

$$y_p(x) = \frac{4}{9}x^2 - \frac{16}{27}x + \frac{23}{27}$$

and finally, the general solution of the nonhomogeneous differential equation is

$$y(x) = C_1e^{-3x} + C_2xe^{-3x} + \frac{4}{9}x^2 - \frac{16}{27}x + \frac{23}{27}.$$

EXAMPLE **10.32.** *Solve the differential equation*

$$\frac{d^2y}{dx^2} - 5\frac{dy}{dx} + 4y = 7\cos(3x).$$

Solution: *In Example 10.23 we found the complementary solution was* $y_c(x) = C_1e^x + C_2e^{4x}$. *Based on the Table of guesses we try*

$$y_p(x) = A\cos(3x) + B\sin(3x) \tag{10.13}$$

as our particular solution y_p, *and we note that* y_p *is not included already in* y_c. *Substitution of Equation 10.13 into the differential equation gives*

$$-9A\cos(3x) - 9B\sin(3x) - 5(-3A\sin(3x) + 3B\cos(3x)) \\ + 4(A\cos(3x) + B\sin(3x)) = 7\cos(3x),$$

and equating coefficients of $\cos(3x)$ *and* $\sin(3x)$ *on both sides of this equation leads to a set of algebraic equations to solve for the unknowns* A *and* B:

$$-9A - 15B + 4A = 7 \quad , \quad -9B + 15A + 4B = 0.$$

The solution of this set of equations is

$$A = -\frac{7}{50} \quad , \quad B = -\frac{21}{50}$$

and so

$$y_p(x) = -\frac{7}{50}\cos(3x) - \frac{21}{50}\sin(3x),$$

and finally, the general solution of the nonhomogeneous differential equation is

$$y(x) = C_1e^x + C_2e^{4x} - \frac{7}{50}\cos(3x) - \frac{21}{50}\sin(3x).$$

EXAMPLE **10.33.** *Solve the differential equation*

$$\frac{d^2y}{dx^2} - 4\frac{dy}{dx} + 13y = 8e^{-3x}.$$

Solution: *In Example 10.24 we found the complementary solution was* $y_c(x) = C_1e^{2x}\cos(3x) + C_2e^{2x}\sin(3x)$. *Based on the Table of guesses we try*

$$y_p(x) = Ae^{-3x} \tag{10.14}$$

as our particular solution y_p, *and we note that* y_p *is not included already in* y_c. *Substitution of Equation 10.14 into the differential equation gives*

$$9Ae^{-3x} + 12Ae^{-3x} + 13Ae^{-3x} = 8e^{-3x}$$

and dividing through by e^{-3x}

$$34A = 8 \quad \Rightarrow \quad A = \frac{4}{17} \quad \Rightarrow \quad y_p(x) = \frac{4}{17}e^{-3x},$$

and hence the general solution of the nonhomogeneous differential equation is

$$y(x) = C_1e^{2x}\cos(3x) + C_2e^{2x}\sin(3x) + \frac{4}{17}e^{-3x}.$$

EXAMPLE **10.34.** *Solve the differential equation*

$$\frac{d^2y}{dx^2} + 5\frac{dy}{dx} + 6y = 3e^{-2x}.$$

Solution: *The corresponding homogeneous differential equation*

$$\frac{d^2y}{dx^2} + 5\frac{dy}{dx} + 6y = 0$$

has characteristic equation

$$m^2 + 5m + 6 = (m+2)(m+3) = 0 \quad \Rightarrow \quad m = -2,\ -3$$

so the complementary solution is

$$y_c(x) = C_1e^{-2x} + C_2e^{-3x}.$$

Based on the Table of guesses we try

$$y_p(x) = Ae^{-2x}$$

as our particular solution y_p, but in this case we note that y_p is included already in y_c, so instead we try

$$y_p(x) = Axe^{-2x} \tag{10.15}$$

as our particular solution y_p. Substitution of Equation 10.15 into the differential equation gives

$$4A(x-1)e^{-2x} + 5A(1-2x)e^{-2x} + 6Axe^{-2x} = 3e^{-2x}$$

and dividing through by e^{-2x} and expanding yields

$$4Ax - 4A + 5A - 10Ax + 6Ax = 3 \quad \Rightarrow \quad A = 3 \quad \Rightarrow \quad y_p(x) = 3xe^{-3x},$$

and hence the general solution of the nonhomogeneous differential equation is

$$y(x) = C_1e^{-2x} + C_2e^{-3x} + 3xe^{-2x}.$$

REMARK **10.35.** *You should use $y_p(x) = Ae^{-2x}$ as as the particular solution y_p in Example 10.34, and investigate what happens when you attempt to find the value of the constant A.*

10.6.2 Variation of parameters

In cases where the nonhomogeneous term is not of the right type and the method of undetermined coefficients cannot be applied, a more general method called variation of parameters may be used to find y_p.

Consider the complimentary solution $y_c = C_1y_1 + C_2y_2$ of the homogeneous differential equation

$$\frac{d^2y}{dx^2} + p(x)\frac{dy}{dx} + q(x)y = 0.$$

To find a particular solution y_p of the corresponding nonhomogeneous differential equation

$$\frac{d^2y}{dx^2} + p(x)\frac{dy}{dx} + q(x)y = g(x),$$

we replace the integration constants C_1 and C_2 in the complementary solution with unknown functions $u_1(x)$ and $u_2(x)$ and suppose that this is y_p; that is, we set

$$y_p(x) = u_1(x)y_1(x) + u_2(x)y_2(x).$$

From the product rule we have

$$\begin{aligned}\frac{dy_p}{dx} &= y_1\frac{du_1}{dx} + u_1\frac{dy_1}{dx} + y_2\frac{du_2}{dx} + u_2\frac{dy_2}{dx} \\ &= u_1\frac{dy_1}{dx} + u_2\frac{dy_2}{dx} + \left(y_1\frac{du_1}{dx} + y_2\frac{du_2}{dx}\right).\end{aligned}$$

This is a nasty-looking expression, so lets set the term in brackets equal to zero, that is

$$y_1\frac{du_1}{dx} + y_2\frac{du_2}{dx} = 0,$$

then the first derivative of y_p is the not-so-nasty looking

$$\frac{dy_p}{dx} = u_1\frac{dy_1}{dx} + u_2\frac{dy_2}{dx}.$$

Differentiating again using the product rule we have

$$\frac{d^2y_p}{dx^2} = \frac{dy_1}{dx}\frac{du_1}{dx} + u_1\frac{d^2y_1}{dx^2} + \frac{dy_2}{dx}\frac{du_2}{dx} + u_2\frac{d^2y_2}{dx^2},$$

and hence the nonhomogeneous differential equation becomes

$$\begin{aligned}&\frac{dy_1}{dx}\frac{du_1}{dx} + u_1\frac{d^2y_1}{dx^2} + \frac{dy_2}{dx}\frac{du_2}{dx} + u_2\frac{d^2y_2}{dx^2} \\ &\qquad + p\left(u_1\frac{dy_1}{dx} + u_2\frac{dy_2}{dx}\right) + q\left(u_1y_1 + u_2y_2\right) = g,\end{aligned}$$

which may be written

$$\begin{aligned}u_1\underbrace{\left(\frac{d^2y_1}{dx^2} + p\frac{dy_1}{dx} + qy_1\right)}_{0} + u_2\underbrace{\left(\frac{d^2y_2}{dx^2} + p\frac{dy_2}{dx} + qy_2\right)}_{0} \\ + \frac{dy_1}{dx}\frac{du_1}{dx} + \frac{dy_2}{dx}\frac{du_2}{dx} = g.\end{aligned}$$

Hence we have two equations for the derivatives of the unknown functions u_1 and u_2, namely

$$y_1\frac{du_1}{dx} + y_2\frac{du_2}{dx} = 0 \quad \text{and} \quad \frac{dy_1}{dx}\frac{du_1}{dx} + \frac{dy_2}{dx}\frac{du_2}{dx} = g.$$

Solving these equations for u_1' and u_2' we find

$$\frac{du_1}{dx} = -\frac{y_2g}{W[y_1, y_2]} \quad \text{and} \quad \frac{du_2}{dx} = \frac{y_1g}{W[y_1, y_2]},$$

where $W[y_1, y_2]$ is the Wronskian of the homogeneous solutions y_1 and y_2.

By integrating[3] these equations (omitting the integration constants) we can obtain the particular solution $y_p = u_1 y_1 + u_2 y_2$, and therefore the general solution $y = y_c + y_p$.

[3] Although we have said this is a general method, there is no guarantee that these two equations *can* be integrated to find $u_1(x)$ and/or $u_2(x)$.

REMARK **10.36.** *1. If one of the terms in y_p is already in y_c, we can "absorb" it into the integration constants contained in y_c.*

2. Of course, the Examples that were solved in the previous section by the method of undetermined coefficients can also be solved by variation of parameters. However, in the vast majority of cases if a nonhomogeneous differential equation can be solved by the method of undetermined coefficients, it will be much easier to use that method than to solve than the same problem using variation of parameters.

KEY CONCEPT **10.37.** *In summary, to find the general solution of a linear nonhomogeneous second-order ordinary differential equation with constant coefficients of general form*

$$\frac{d^2y}{dx^2} + p\frac{dy}{dx} + qy = g(x)$$

where p and q are constants and $g(x)$ is an arbitrary function of x by the method of variation of parameters:

1. *Find the general solution $y_c(x) = C_1 y_1(x) + C_2 y_2(x)$ of the corresponding homogeneous differential equation*

$$\frac{d^2y}{dx^2} + p\frac{dy}{dx} + qy = 0.$$

2. *Calculate the Wronskian*

$$W[y_1, y_2](x) = \det \begin{bmatrix} y_1 & y_2 \\ y_1' & y_2' \end{bmatrix} = y_1 \frac{dy_2}{dx} - y_2 \frac{dy_1}{dx}.$$

3. *Let $\frac{du_1}{dx} = -\frac{y_2(x)g(x)}{W[y_1, y_2]}$ and $\frac{du_2}{dx} = \frac{y_1(x)g(x)}{W[y_1, y_2]}$.*

4. *Integrate these two equations to find $u_1(x)$ and $u_2(x)$, omitting the integration constants.*

5. *A particular solution of the nonhomogeneous differential is then $y_p(x) = u_1(x)y_1(x) + u_2(x)y_2(x)$.*

6. *The general solution of the nonhomogeneous differential equation is then $y(x) = y_c(x) + y_p(x)$.*

EXAMPLE **10.38.** *Solve the differential equation*

$$\frac{d^2y}{dx^2} - 2\frac{dy}{dx} + y = e^x \ln x.$$

Solution: *Note that the nonhomogeneous term $g(x) = e^x \ln x$ is not of a form that we can apply the method of undetermined coefficients, because we cannot make an intelligent guess for $y_p(x)$. The corresponding homogeneous differential equation*

$$\frac{d^2y}{dx^2} - 2\frac{dy}{dx} + y = 0$$

has characteristic equation

$$m^2 - 2m + 1 = (m-1)^2 = 0 \quad \Rightarrow \quad m = 1,$$

so the complementary solution is

$$y_c(x) = C_1 e^x + C_2 x e^x.$$

With $y_1(x) = e^x$ and $y_2(x) = xe^x$ we find

$$W[y_1, y_2](x) = \det \begin{bmatrix} e^x & xe^x \\ e^x & e^x + xe^x \end{bmatrix} = e^{2x}.$$

Then

$$\frac{du_1}{dx} = -\frac{y_2(x)g(x)}{W[y_1, y_2]} = -\frac{(xe^x)(e^x \ln x)}{e^{2x}} = -x \ln x,$$

and using integration by parts we find that

$$u_1(x) = -\int x \ln x \, dx = \frac{1}{4}x^2 - \frac{1}{2}x^2 \ln x.$$

Similarly

$$\frac{du_2}{dx} = \frac{y_1(x)g(x)}{W[y_1, y_2]} = \frac{(e^x)(e^x \ln x)}{e^{2x}} = \ln x,$$

hence

$$u_2(x) = \int \ln x \, dx = x \ln x - x.$$

Then the particular solution is

$$\begin{aligned} y_p(x) &= u_1(x)y_1(x) + u_2(x)y_2(x) \\ &= \left(\frac{1}{4}x^2 - \frac{1}{2}x^2 \ln x\right)(e^x) + (x \ln x - x)(xe^x) \\ &= \frac{1}{2}x^2 e^x \ln x - \frac{3}{4}x^2 e^x, \end{aligned}$$

and hence the general solution of the nonhomogeneous differential equation is

$$y(x) = C_1 e^x + C_2 x e^x + \frac{1}{2}x^2 e^x \ln x - \frac{3}{4}x^2 e^x.$$

REMARK **10.39.** *It should be obvious from this example that it would have been extremely difficult (if not impossible) to "guess" the form of the particular solution y_p.*

10.7 Initial and boundary conditions

As mentioned earlier, the values of the integration constants that arise when we solve differential equations can be determined by making use of other conditions (or restrictions) placed on the problem. If there are n unknown constants then we will need n extra conditions.

If all of the extra conditions are given at one value of the independent variable then the extra conditions are called *initial conditions* and the combined differential equation plus initial conditions is called an *initial value problem*.

If the extra conditions are given at different values of the independent variable then they are called *boundary conditions* and the combined differential equation plus boundary conditions is called a *boundary value problem*.

EXAMPLE **10.40.** *Solve the initial value problem*

$$\frac{d^2y}{dx^2} - 5\frac{dy}{dx} + 4y = 7\cos(3x) \quad , \quad y(0) = 1 \quad , \quad y'(0) = 2.$$

Solution: *We have already seen this differential equation in Example 10.32 and determined that its general solution is given by*

$$y(x) = C_1e^x + C_2e^{4x} - \frac{7}{50}\cos(3x) - \frac{21}{50}\sin(3x).$$

The initial conditions will give two equations to solve for the unknowns, C_1 and C_2. Firstly,

$$y(0) = 1 \quad \Rightarrow \quad 1 = C_1 + C_2 - \frac{7}{50}. \tag{10.16}$$

The second initial condition involves the derivative, so:

$$\frac{dy}{dx} = C_1e^t + 4C_2e^{4x} + \frac{21}{50}\sin(3x) - \frac{63}{50}\cos(3x),$$

and the second initial condition then gives

$$y'(0) = 2 \quad \Rightarrow \quad 2 = C_1 + 4C_2 - \frac{63}{50}. \tag{10.17}$$

Solving the pair of algebraic equations 10.16 and 10.17 gives

$$C_1 = \frac{13}{30} \quad \text{and} \quad C_2 = \frac{53}{75},$$

so the required solution is

$$y(x) = \frac{13}{30}e^x + \frac{53}{75}e^{4x} - \frac{7}{50}\cos(3x) - \frac{21}{50}\sin(3x).$$

EXAMPLE **10.41.** *Solve the boundary value problem*

$$\frac{d^2y}{dx^2} + 6\frac{dy}{dx} + 9y = 4x^2 + 5 \quad , \quad y(0) = 7 \quad , \quad y(1) = -3.$$

Solution: *We have already seen this differential equation in Example 10.31 and determined that its general solution is given by*

$$y(x) = C_1 e^{-3x} + C_2 x e^{-3x} + \frac{4}{9}x^2 - \frac{16}{27}x + \frac{23}{27}.$$

The boundary conditions give two equations to solve for the unknowns, C_1 and C_2:

$$\begin{aligned} y(0) = 7 \quad &\Rightarrow \quad 7 = C_1 + \frac{23}{27}, \\ y(1) = -3 \quad &\Rightarrow \quad -3 = C_1 e^{-3} + C_2 e^{-3} + \frac{4}{9} - \frac{16}{27} + \frac{23}{27}. \end{aligned}$$

Solving this pair of algebraic equations gives

$$C_1 = \frac{166}{27} \quad \text{and} \quad C_2 = -\frac{166 + 100e^3}{75},$$

so the required solution is

$$y(x) = \frac{166}{27}e^{-3x} - \left(\frac{166 + 100e^3}{75}\right) x e^{-3x} + \frac{4}{9}x^2 - \frac{16}{27}x + \frac{23}{27}.$$

11
Laplace transforms

Laplace transforms represent a powerful method for tackling various problems that arise in engineering and physical sciences. Most often they are used for solving differential equations that cannot be solved via standard methods. An introduction to the concepts of and the language relating to Laplace transforms is our plan for this chapter. More advanced theory and uses of the transform are postponed until later units.

11.1 The Laplace transform and its inverse

We begin with a definition of the *Laplace transform* of a scalar function $f(t)$ defined for $t \geq 0$.

Definition 11.1. *(Laplace transform)*
Given a function $f(t)$ defined for all $t \geq 0$, the Laplace transform *(LT) of $f(t)$ is the function*

$$F(s) = \int_0^\infty e^{-st} f(t)\, dt$$

defined for all $s \in \mathbb{R}$ for which the above improper integral is convergent. We often write $F(s)$ as $\mathcal{L}(f)$, or, more precisely $\mathcal{L}(f)(s)$.

It is worth remarking that here we are following traditional notation and denoting the variable of the initial function t (this is motivated by regarding the function $f(t)$ as defined for all 'time' $t \geq 0$). Performing the transformation will of course yield a function $F(s)$ and the usual designation of this Laplace transform variable is s (although some texts might use p instead). Lastly, we point out that the Laplace transforms of functions $f(t)$, $g(t)$, $h(t)$, etc. are normally denoted by their corresponding capital letters $F(s)$, $G(s)$, $H(s)$, etc.

If $F = \mathcal{L}(f)$ is the Laplace transform of $f(t)$, we say that $f(t)$ is the *inverse Laplace transform* (ILT) of $F(s)$, written as $f = \mathcal{L}^{-1}(F)$. In slightly cumbersome terms, this is saying that the inverse transform of $F(s)$ is that function $f(t)$ whose Laplace transform is $F(s)$.

We now determine the Laplace transforms of some simple functions.

EXAMPLE **11.2.** *If $f(t) = 1$ for $t \geq 0$, then if $s > 0$*

$$F(s) = \int_0^\infty e^{-st}\,dt = \left[-\frac{e^{-st}}{s}\right]_0^\infty = -\frac{1}{s}\lim_{t\to\infty} e^{-st} + \frac{1}{s} = \frac{1}{s}.$$

This is so since $\lim_{t\to\infty} e^{-st} = 0$ for $s > 0$ (notice this limit does not exist if $s \leq 0$). Thus, the integral exists for $s > 0$, giving that

$$\mathcal{L}(1) = \frac{1}{s}.$$

Notice that here $F(s)$ is not defined for all real values of s, just for $s > 0$. The definition of the ILT now implies that

$$\mathcal{L}^{-1}\left(\frac{1}{s}\right) = 1.$$

□

Throughout this chapter we use the following notational convention: If for a function $f(t)$ the improper integral $\int_c^\infty f(t)\,dt$ exists and if $g(t)$ is an anti-derivative for $f(t)$, then we write $[g(t)]_c^\infty$ for $\lim_{t\to\infty}(g(t) - g(c))$.

EXAMPLE **11.3.** *For $f(t) = t^n$ for some integer $n \geq 0$ then*

$$\mathcal{L}(t^n) = \int_0^\infty e^{-st}t^n\,dt.$$

Substituting $u = ts$ gives

$$\mathcal{L}(t^n) = \int_0^\infty e^{-u}\left(\frac{u}{s}\right)^n \frac{du}{s} = \frac{1}{s^{n+1}}\int_0^\infty u^n e^{-u}du = \frac{n!}{s^{n+1}}\ (\textit{for } s > 0)$$

where the integral can be evaluated using the principle of mathematical induction and integration by parts. □

EXAMPLE **11.4.** *Consider $f(t) = e^{at}$ for $t \geq 0$, where a is a constant. Then for $s > a$ we have*

$$F(s) = \int_0^\infty e^{-st}e^{at}\,dt = \left[\frac{e^{(a-s)t}}{a-s}\right]_0^\infty$$

$$= -\frac{1}{a-s} + \frac{1}{a-s}\lim_{t\to\infty} e^{(a-s)t} = \frac{1}{s-a}.$$

Thus, the integral exists for $s > a$ (note $F(s)$ does not exist for $s \leq a$) and

$$\mathcal{L}(e^{at}) = \frac{1}{s-a} \quad (s > a).$$

Hence we can deduce that

$$\mathcal{L}^{-1}\left(\frac{1}{s-a}\right) = e^{at} \quad (t \geq 0).$$

For $a = 0$ this result is consistent with Example 11.2. □

EXAMPLE **11.5.** *Let $f(t) = \sin(at)$ for some $a \neq 0$, and let $F = \mathcal{L}(f)$. Notice that for $s > 0$ we have $\lim_{t\to\infty} e^{-st}\sin(at) = 0$ (by the Squeeze Theorem); similarly, $\lim_{t\to\infty} e^{-st}\cos(at) = 0$. Using this and two integrations by parts, we get*

$$\begin{aligned} F(s) &= \int_0^\infty e^{-st}\sin(at)\,dt = -\frac{1}{s}\int_0^\infty (e^{-st})'\sin(at)\,dt \\ &= -\frac{1}{s}[e^{-st}\sin(at)]_0^\infty + \frac{a}{s}\int_0^\infty e^{-st}\cos(at)\,dt \\ &= 0 - \frac{a}{s^2}\int_0^\infty (e^{-st})'\cos(at)\,dt \\ &= -\frac{a}{s^2}[e^{-st}\cos(at)]_0^\infty - \frac{a^2}{s^2}\int_0^\infty e^{-st}\sin(at)\,dt \\ &= \frac{a}{s^2} - \frac{a^2}{s^2}F(s). \end{aligned}$$

This gives an equation for $F(s)$:

$$F(s) = \frac{a}{s^2} - \frac{a^2}{s^2}F(s).$$

It is a matter of simple algebra to rearrange to deduce that

$$\mathcal{L}(\sin(at)) = F(s) = \frac{a}{s^2+a^2} \qquad (s > 0).$$

It is then immediately obvious that

$$\mathcal{L}^{-1}\left(\frac{a}{s^2+a^2}\right) = \sin(at).$$

□

EXERCISE **11.1.1.** *Use similar methods to show that*

$$\mathcal{L}(\cos(at)) = \frac{s}{s^2+a^2} \text{ and } \mathcal{L}^{-1}\left(\frac{s}{s^2+a^2}\right) = \cos(at) \quad \text{for} \quad s > 0.$$

These are just a few of the more straightforward examples of the Laplace transform. To obtain others we can use some of the properties of the Laplace transform operation.

EXERCISE **11.1.2.** *Use integration by parts to show that for any constants $a > 0$ and $\omega \in \mathbb{R}$ we have*

(a) $\mathcal{L}(e^{at}\sin(\omega t)) = \dfrac{\omega}{(s-a)^2+\omega^2}$, for $s > 0$

(b) $\mathcal{L}(e^{at}\cos(\omega t)) = \dfrac{s+a}{(s-a)^2+\omega^2}$, for $s > 0$

(Hint: Write down the definition of the Laplace transform in each case. A suitable substitution will reduce the integrals to those in Example 11.5 thereby circumventing the need to do pages of laborious calculation.)

11.1.1 Linearity of the Laplace transform

The first part of the theorem below is an immediate consequence of the definition of the Laplace transform. The second part follows immediately from the first one.

THEOREM **11.6.** *1. If the Laplace transforms $\mathcal{L}(f)(s)$ and $\mathcal{L}(g)(s)$ of two functions $f(t)$ and $g(t)$ exist for $s \geq a$ for some $a \in \mathbb{R}$, then for any constants $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}$ we have*

$$\mathcal{L}\left(\alpha f + \beta g\right)(s) = \alpha \mathcal{L}(f)(s) + \beta \mathcal{L}(g)(s)$$

for $s \geq a$.

This theorem says that both the Laplace transform and the inverse Laplace transform act as linear transformations on the space of functions.

2. *Let $F(s)$ and $G(s)$ be functions. If the inverse Laplace transforms $f(t) = \mathcal{L}^{-1}(F)(t)$ and $g(t) = \mathcal{L}^{-1}(G)(t)$ exist, then for any constants $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}$ we have*

$$\mathcal{L}^{-1}\left(\alpha F + \beta G\right)(t) = \alpha \mathcal{L}^{-1}(F)(t) + \beta \mathcal{L}^{-1}(G)(t).$$

EXAMPLE **11.7.** *Find* $\mathcal{L}^{-1}\left(\dfrac{1}{s(s-1)}\right)$.

Solution: *We decompose in partial fractions:* $\dfrac{1}{s(s-1)} = \dfrac{1}{s-1} - \dfrac{1}{s}$.
Hence

$$\mathcal{L}^{-1}\left(\frac{1}{s(s-1)}\right) = \mathcal{L}^{-1}\left(\frac{1}{s-1} - \frac{1}{s}\right) = \mathcal{L}^{-1}\left(\frac{1}{s-1}\right) - \mathcal{L}^{-1}\left(\frac{1}{s}\right) = e^t - 1.$$

□

EXERCISE **11.1.3.** *Use the linearity of the Laplace transform and some of the above examples to find the Laplace transforms of:*

(a) $f(t) = \cos t - \sin t$
(b) $f(t) = t^2 - 3t + 5$
(c) $f(t) = 3e^{-t} + \sin(6t)$

EXERCISE **11.1.4.** *Use the linearity of the inverse Laplace transform and some of the above examples to find the inverse Laplace transforms of:*

(a) $F(s) = -\dfrac{2}{s+16}, \ s > -16$
(b) $F(s) = \dfrac{4s}{s^2-9}, \ s > 3$
(c) $F(s) = \dfrac{3}{s-7} + \dfrac{1}{s^2}, \ s > 7$

11.1.2 Existence of Laplace transforms

Recall that a function $f(x)$ is called piecewise continuous on a given interval $[a, b]$ if f has only finitely many points of discontinuity in $[a, b]$. Piecewise continuous functions possess a Laplace transform if they are of exponential order:

DEFINITION **11.8.** *(Exponential order)*
A function $f(t)$, $t \geq 0$, is of exponential order *if $f(t)$ is piecewise continuous and bounded on every interval $[0, T]$ with $T > 0$ and there exist constants $M > 0$ and $\gamma \in \mathbb{R}$ such that*

$$|f(t)| \leq Me^{\gamma t} \quad \text{for all } t \geq 0.$$

When this holds we will say that the exponential order of f is $\leq \gamma$.

Given a function is of exponential order $\leq \gamma$ we can then deduce for what values of s its Laplace transform is defined:

THEOREM **11.9.** *If $f(t)$ is of exponential order $\leq \gamma$, then the Laplace transform $F(s) = \mathcal{L}(f)(s)$ exists for all $s > \gamma$.*

The Laplace transform of a given function is *unique*. Conversely, if two functions have the same Laplace transform then they can differ only at isolated points.

EXAMPLE **11.10.** *For $f(t) = e^{at}$, we saw in Example 11.4 that the transform exists for $s > a$. This is consistent with Theorem 11.9 since $f(t)$ is of exponential order $\leq a$: taking $M = 1$ and $\gamma = a$ we see that $|f(t)| \leq Me^{at}$ for all $t \geq 0$.* □

EXAMPLE **11.11.** *For $f(t) = e^{t^2}$ there are no M and γ for which $e^{t^2} \leq Me^{\gamma t}$ for all $t \geq 0$. In very informal terms, e^{t^2} grows more quickly than $e^{\gamma t}$ for any γ. The Laplace transform $\mathcal{L}(e^{t^2})$ does not exist in this case. This example proves that not every well-defined function necessarily has a Laplace transform.* □

11.2 *Inverse Laplace transforms of rational functions*

If the Laplace transform $F(s) = \mathcal{L}(f)$ of some function $f(t)$ has the special form

$$F(s) = \frac{P(s)}{Q(s)},$$

where $P(s)$ and $Q(s)$ are polynomials with $\deg(P(x)) < \deg(Q(x))$, then we can find the *inverse Laplace transform* $f(t) = \mathcal{L}^{-1}(F)$ using partial fractions, which were covered in MATH1011 and is summarised in the Appendix.

Notice that at this stage we know the inverse Laplace transforms of the following basic rational functions (see Examples 11.4, 11.5 and Exercise 11.1.1)

$$\mathcal{L}^{-1}\left(\frac{1}{s-a}\right) = e^{at} \text{ for } s > a$$

$$\mathcal{L}^{-1}\left(\frac{1}{s^2+a^2}\right) = \frac{1}{a}\sin(at), \text{ for } s > 0$$

$$\mathcal{L}^{-1}\left(\frac{s}{s^2+a^2}\right) = \cos(at), \text{ for } s > 0.$$

To recall the method of partial fractions and demonstrate how it applies to problems involving inverse Laplace transforms, we look at two examples.

EXAMPLE **11.12.** *Suppose $F(s) = \dfrac{2s-1}{(s^2-1)(s+3)}$, $s > 1$, and we want to find $f(t) = \mathcal{L}^{-1}(F)$. First, using partial fractions, we write*

$$F(s) = \frac{2s-1}{(s-1)(s+1)(s+3)} = \frac{A}{s-1} + \frac{B}{s+1} + \frac{C}{s+3}.$$

This is equivalent to

$$2s-1 = A(s+1)(s+3) + B(s-1)(s+3) + C(s-1)(s+1).$$

From this with $s = 1$ we get $A = 1/8$. Similarly, $s = -1$ gives $B = 3/4$, while $s = -3$ implies $C = -7/8$. Thus,

$$F(s) = \frac{1}{8(s-1)} + \frac{3}{4(s+1)} - \frac{7}{8(s+3)}$$

and therefore

$$\begin{aligned} f(t) &= \mathcal{L}^{-1}(F(s)) \\ &= \frac{1}{8}\mathcal{L}^{-1}\left(\frac{1}{s-1}\right) + \frac{3}{4}\mathcal{L}^{-1}\left(\frac{1}{s+1}\right) - \frac{7}{8}\mathcal{L}^{-1}\left(\frac{1}{s+3}\right) \\ &= \frac{1}{8}e^{t} + \frac{3}{4}e^{-t} - \frac{7}{8}e^{-3t}. \end{aligned}$$

□

EXAMPLE **11.13.** *Suppose $F(s) = \dfrac{2s^2 - s + 4}{s^3 + 4s}$, for $s \geq 0$. To find $f(t) = \mathcal{L}^{-1}(F)$, we first use partial fractions:*

$$F(s) = \frac{2s^2 - s + 4}{s(s^2+4)} = \frac{A}{s} + \frac{Bs + C}{s^2 + 4}.$$

This is equivalent to

$$2s^2 - s + 4 = A(s^2+4) + (Bs + C)s = (A + B)s^2 + Cs + 4A,$$

so we must have $A + B = 2$, $C = -1$ and $4A = 4$. This gives $A = 1$, $B = 1$, $C = -1$.

Thus,

$$F(s) = \frac{1}{s} + \frac{s-1}{s^2+4} = \frac{1}{s} + \frac{s}{s^2+4} - \frac{1}{s^2+4}$$

and therefore

$$\begin{aligned} f(t) = \mathcal{L}^{-1}(F) &= \mathcal{L}^{-1}\left(\frac{1}{s}\right) + \mathcal{L}^{-1}\left(\frac{s}{s^2+4}\right) - \mathcal{L}^{-1}\left(\frac{1}{s^2+4}\right) \\ &= 1 + \cos(2t) - \frac{1}{2}\sin(2t). \end{aligned}$$

□

EXERCISE **11.2.1.** *Use partial fractions to find the inverse Laplace transforms of:*

(a) $F(s) = -\dfrac{2s}{(s+1)(s^2+1)}$

(b) $F(s) = \dfrac{1}{s^4 - 16}$

11.3 The Laplace transform of derivatives and integrals of $f(t)$

Later we shall show that some of the most important applications of Laplace transforms are to the solutions of differential equations. To that end it is important to know the forms of the Laplace transforms of the derivatives and integral of $f(t)$. The form of the Laplace transform of $f'(t)$ is given in the following theorem:

THEOREM **11.14.** *If $f(t)$ is continuous and of exponential order $\leq \gamma$ and if $f'(t)$ exists and is piecewise continuous and bounded over $[0, T]$ for all $T \geq 0$, then the Laplace transform of $f'(t)$ exists for $s > \gamma$ and*

$$\mathcal{L}(f')(s) = s\mathcal{L}(f)(s) - f(0)\,.$$

Proof. This is easy to verify using integration by parts. Indeed, if we denote by $F(s)$ and $G(s)$ the Laplace transforms of $f(t)$ and $f'(t)$, respectively, then

$$\begin{aligned} G(s) &= \int_0^\infty e^{-st} f'(t)\,dt = \left[e^{-st} f(t)\right]_0^\infty + s\int_0^\infty e^{-st} f(t)\,dt \\ &= -f(0) + \lim_{t\to\infty} e^{-st} f(t) + sF(s) = sF(s) - f(0) \end{aligned}$$

since for $s > \gamma$ we have $|e^{-st} f(t)| \leq Me^{(\gamma - s)t} \to 0$ as $t \to \infty$. □

It is worth pausing at this juncture and spending a few moments reflecting on this result. There are a few very important properties that need to be appreciated before we proceed. First, note how $G(s) = \mathcal{L}(f')(s)$ is a multiple of $F(s)$ minus a constant. Before working through Theorem 11.14 we might well have expected that the Laplace transform of $f'(t)$ would have involved the derivative of the Laplace transform of $f(t)$, in other words $\dfrac{dF}{ds}$. But that clearly is not the case; $G(s)$ is related to $F(s)$ by simple algebraic multiplication and this is the first clue as to the usefulness of the Laplace transform in solving differential equations. Laplace transforms of derivatives of $y(t)$ are changed to algebraic multiples of its Laplace transform $Y(s)$ and, as we shall see, this means that ultimately the solution of the problem reduces to a task of solving an algebraic equation which is normally a much easier prospect than analysing the original differential equation.

The second aspect of note in the result of Theorem 11.14 is the presence of the value $f(0)$. This too is not expected – normally when dealing with the derivative of a function we are not concerned with the value of the function itself at any given point. But that is not the case for the Laplace transform; in order to find the Laplace transform of $f'(t)$ completely some knowledge of $f(0)$ is required.

EXERCISE **11.3.1.** *Use the formula for the Laplace transform of a derivative to find:*

(a) $\mathcal{L}(te^{at})$

(b) $\mathcal{L}(t^n e^{at})$

Given the technique used to deduce the form of $G(s)$ we can repeat the process to obtain the Laplace transforms of higher derivatives of $f(t)$ (provided some regularity conditions are satisfied). For example

$$\mathcal{L}(f'')(s) = s\mathcal{L}(f')(s) - f'(0) = s[s\mathcal{L}(f)(s) - f(0)] - f'(0)$$

so that

$$\mathcal{L}(f'')(s) = s^2\mathcal{L}(f)(s) - sf(0) - f'(0).$$

Similarly,

$$\mathcal{L}(f''')(s) = s^3\mathcal{L}(f)(s) - s^2f(0) - sf'(0) - f''(0)$$

and, more generally (this can be proved using the principle of mathematical induction),

$$\mathcal{L}(f^{(n)})(s) = s^n\mathcal{L}(f)(s) - s^{n-1}f(0) - \cdots - sf^{(n-2)}(0) - f^{(n-1)}(0).$$

Once again we remark that $\mathcal{L}(f^{(n)})(s)$ involves no derivatives of $F(s)$ at all; it is given by a multiple s^n of $F(s)$ plus a polynomial of degree $n-1$ in s. The coefficients of this polynomial involve the values of the first $n-1$ derivatives of $f(t)$ at $t = 0$.

EXAMPLE **11.15.** *The above can be used to find $\mathcal{L}(\sin(at))$ by an alternative route than that taken in Example 11.5. Let $f(t) = \sin(at)$. Then $f(t)$ is continuous and of exponential order ≤ 0 (since $|f(t)| \leq 1e^{0t}$) and $f'(t)$ is continuous, so we can apply Theorem 11.14. We have $f''(t) = -a^2f(t)$ and so*

$$-a^2\mathcal{L}(f) = \mathcal{L}(f'') = s^2\mathcal{L}(f) - sf(0) - f'(0), \text{ for } s > 0.$$

Collecting the two terms involving $\mathcal{L}(f)$,

$$(a^2 + s^2)\mathcal{L}(f) = sf(0) + f'(0)$$

and, using that $f(0) = 0$ and $f'(0) = a$, we obtain

$$\mathcal{L}(f) = \frac{a}{s^2 + a^2}, \text{ for } s > 0.$$

□

Using all of the techniques above (and some more to come later) we can construct a table of Laplace transforms of frequently-encountered functions. Such a table is provided on page 207.

EXERCISE **11.3.2.** *Use the formula for the Laplace transform of a double derivative to show that*

(a) $\mathcal{L}(t\sin(\omega t)) = \dfrac{2\omega s}{(s^2 + \omega^2)^2}$

(b) $\mathcal{L}(t\cos(\omega t)) = \dfrac{s^2 - \omega^2}{(s^2 + \omega^2)^2}$

Next we consider how one can derive a formula for the Laplace transform of an integral.

THEOREM **11.16.** *If $f(t)$ is of exponential order (so that $\mathcal{L}(f)$ exists) then $g(t) = \int_0^t f(u)du$ is of exponential order $\leq \gamma$ for some γ. Moreover, for $s > \gamma$ and $s \neq 0$ we have*

$$\mathcal{L}(g)(s) = \frac{1}{s}\mathcal{L}(f)(s).$$

In other words, if the Laplace transform of $f(t)$ is $F(s)$, then

$$\mathcal{L}(g)(s) = \frac{F(s)}{s} \quad \text{and} \quad \mathcal{L}^{-1}\left(\frac{F(s)}{s}\right) = \int_0^t \mathcal{L}^{-1}(F)(u)du.$$

Sketch of Proof. Denote the Laplace transform of $g(t)$ by $G(s)$. By definition of $g(t)$, it is continuous and it can be proved that it is of exponential order, say $\leq \gamma$. Since $g'(t) = f(t)$ by the Fundamental theorem of Calculus and since $g(0) = 0$, Theorem 11.14 implies (for $s > \gamma$)

$$F(s) = \mathcal{L}(f)(s) = \mathcal{L}(g')(s) = s\mathcal{L}(g)(s) - g(0) = sG(s).$$

Thus, for $s > \gamma$, $s \neq 0$, we have $G(s) = F(s)/s$. □

This can be particularly useful in helping to determine the inverse transform of functions which have a factor s appearing in the denominator.

EXAMPLE **11.17.** *Find the inverse Laplace transform $g(t)$ of $G(s) = \dfrac{1}{s(s^2+\omega^2)}$, using Theorem 11.16.*

We could also use the partial fraction method.

Solution: *Notice that $G(s) = \dfrac{F(s)}{s}$, where for $F(s) = \dfrac{1}{s^2+\omega^2}$ we know*

$$f(t) = \mathcal{L}^{-1}(F) = \frac{1}{\omega}\sin(\omega t).$$

Then Theorem 11.16 yields

$$g(t) = \mathcal{L}^{-1}(G(s)) = \int_0^t f(u)du = \frac{1}{\omega}\int_0^t \sin(\omega u)du = \frac{1}{\omega^2}[1-\cos(\omega t)].$$

□

EXERCISE **11.3.3.** *Use the formula for the Laplace transform of an integral to find $\mathcal{L}^{-1}\left(\dfrac{1}{s(s+3)}\right)$ and $\mathcal{L}^{-1}\left(\dfrac{1}{s^2(s+3)}\right)$ (no partial fractions required).*

11.4 Solving differential equations

Laplace transforms can be applied to initial-value problems for linear ordinary differential equations by reducing them to the task of solving an algebraic equation.

However it should be realised that Laplace transform methods will only be able to detect solutions of ordinary differential equations that have Laplace transforms. While most solutions will satisfy this requirement, not all will. We saw in Example 11.11 that $f(t) = e^{t^2}$ does not have a Laplace transform but this is a solution of the differential equation

$$\frac{dy}{dt} - 2ty = 0;$$

we could not therefore expect to derive a meaningful solution of this equation using the Laplace transform. We need to bear in mind

that although the Laplace transform will find most solutions of differential equations there are isolated cases when it will fail.

Despite this caution, it can be shown that all solutions of constant coefficient differential equations are of exponential order so we can use Laplace transform methods to seek a solution $y(t)$ of

$$y''(t) + ay'(t) + by(t) = r(t) \quad t \geq 0$$

where a, b are constants and $r(t)$ is a given function, such that $y(t)$ satisfies the initial conditions

$$y(0) = K_0 \ , \ y'(0) = K_1.$$

To solve this, first transform the differential equation, writing

$$\mathcal{L}(y'') + a\mathcal{L}(y') + b\mathcal{L}(y) = R(s)$$

where $R(s) = \mathcal{L}(r)(s)$. In terms of $Y(s) = \mathcal{L}(y)$, this gives the equation

$$[s^2Y(s) - sy(0) - y'(0)] + a[sY(s) - y(0)] + bY(s) = R(s).$$

This can be written in the form

$$(s^2 + as + b)Y(s) = R(s) + (s + a)y(0) + y'(0) = R(s) + (s + a)K_0 + K_1$$

so we have

$$Y(s) = \frac{R(s) + (s + a)K_0 + K_1}{s^2 + as + b}.$$

Therefore

$$y(t) = \mathcal{L}^{-1}(Y) = \mathcal{L}^{-1}\left(\frac{R(s) + (s + a)K_0 + K_1}{s^2 + as + b}\right).$$

This method will become clearer with some examples.

EXAMPLE **11.18.** *Solve the initial value problem*

$$y''(t) - y(t) = t \ , \ y(0) = 1 \ , \ y'(0) = 1.$$

Solution: *Applying the Laplace transform and denoting $Y(s) = \mathcal{L}(y)$, we get*

$$[s^2Y(s) - sy(0) - y'(0)] - Y(s) = \mathcal{L}(t) = \frac{1}{s^2}.$$

Using the initial conditions $y(0) = y'(0) = 1$, we write it in the form

$$(s^2 - 1)Y = s + 1 + \frac{1}{s^2}.$$

Solving for $Y(s)$ gives

$$Y(s) = \frac{s+1}{s^2-1} + \frac{1}{s^2(s^2-1)} = \frac{s^3+s^2+1}{s^2(s-1)(s+1)} = -\frac{1}{s^2} + \frac{3}{2}\cdot\frac{1}{s-1} - \frac{1}{2}\cdot\frac{1}{s+1}$$

using partial fractions. So, from the table of Laplace transforms,

$$y(t) = \mathcal{L}^{-1}(Y)(t) = -\mathcal{L}^{-1}\left(\frac{1}{s^2}\right) + \frac{3}{2}\mathcal{L}^{-1}\left(\frac{1}{s-1}\right) - \frac{1}{2}\mathcal{L}^{-1}\left(\frac{1}{s+1}\right)$$

$$= -t + \frac{3}{2}e^t - \frac{1}{2}e^{-t}.$$

□

EXAMPLE **11.19.** *Solve the initial value problem*

$$y^{(4)}(t) - y(t) = 0 \ , \ y(0) = 0 \ , \ y'(0) = 1, y''(0) = y'''(0) = 0.$$

Solution: *Let* $Y(s) = \mathcal{L}(y)(s)$. *Applying the Laplace transform to the DE, we get*

$$[s^4Y(s) - s^3y(0) - s^2y'(0) - sy''(0) - y'''(0)] - Y(s) = 0.$$

Using the initial conditions, this gives $s^4Y(s) - s^2 - Y(s) = 0$, *and therefore*

$$Y(s) = \frac{s^2}{s^4 - 1}.$$

To find $y(t) = \mathcal{L}^{-1}(Y)$ *we need to find a convenient partial fraction expansion for* $Y(s)$. *The following will be adequate:*

$$Y(s) = \frac{s^2}{(s-1)(s+1)(s^2+1)} = \frac{1}{4} \cdot \frac{1}{s-1} - \frac{1}{4} \cdot \frac{1}{s+1} + \frac{1}{2} \cdot \frac{1}{s^2+1}.$$

Hence

$$\begin{aligned} y(t) = \mathcal{L}^{-1}(Y) &= \frac{1}{4}\mathcal{L}^{-1}\left(\frac{1}{s-1}\right) - \frac{1}{4}\mathcal{L}^{-1}\left(\frac{1}{s+1}\right) + \frac{1}{2}\mathcal{L}^{-1}\left(\frac{1}{s^2+1}\right) \\ &= \frac{1}{4}e^t - \frac{1}{4}e^{-t} + \frac{1}{2}\sin t. \end{aligned}$$

□

An advantage of this technique is that it is not necessary to solve for the general solution of the homogeneous differential equation and then determine the arbitrary constants in that solution. Apart from this it can also be used for higher order differential equations, as we saw in Example 11.19.

EXERCISE **11.4.1.** *Solve the initial value problems using the Laplace transform :*

(a) $y'(t) - 9y(t) = t$, $y(0) = 5$
(b) $y''(t) - 4y'(t) + 4y(t) = \cos t$, $y(0) = 1$, $y'(0) = -1$
(c) $y''(t) - 5y'(t) + 6y(t) = e^{-t}$, $y(0) = 0$, $y'(0) = 2$
(d) $y^{(4)}(t) - 4y(t) = 0$, $y(0) = 1$, $y'(0) = 0$, $y''(0) = -2$, $y'''(0) = 0$

11.5 Shift theorems

We have now seen the general strategy for solving differential equations using Laplace transforms; we transform the differential problem to an algebraic one for $Y(s)$ and then, given our knowledge of inverse Laplace transforms, we attempt to reconstruct the form of $y(t)$. It is this last step that is potentially the tricky one for there is always the possibility that $Y(s)$ is of a form we do not recognise.

The situation gets worse. It is relatively straightforward to find the Laplace transform of a function in as much that given an $f(t)$ we can, at least theoretically, compute $F(s)$ using the definition of a Laplace transform but, unfortunately, there is no easy equivalent definition for going in the reverse direction (ie. given $F(s)$, deduce

$f(t)$). Thus it is of importance to expand our repertoire of easily identifiable inverse functions and this is facilitated using two so-called shift theorems.

THEOREM **11.20.** *If $F(s)$ is the Laplace transform of $f(t)$ for $s > b$, then the Laplace transform of $e^{at}f(t)$ is*

$$\mathcal{L}(e^{at}f(t)) = F(s-a)$$

Note that there is no restriction on a in this theorem: a can be positive or negative.

for $s - a > b$. Equivalently,

$$\mathcal{L}^{-1}(F(s-a)) = e^{at}f(t).$$

Proof. We have

$$\mathcal{L}\left(e^{at}f(t)\right) = \int_0^\infty e^{-st}\left(e^{at}f(t)\right)\,dt = \int_0^\infty e^{-(s-a)t}f(t)\,dt = F(s-a)$$

which proves the statement. □

This is called s-shifting, as the graph of the function $F(s-a)$ is obtained from that of $F(s)$ by shifting a units (to the right if $a > 0$ and to the left if $a < 0$) on the s-axis. Putting this result in words it tells us that if the Laplace transform of $f(t)$ is $F(s)$, then the shifted function $F(s-a)$ is the transform of $e^{at}f(t)$.

EXAMPLE **11.21.** *Find the Laplace transform of $e^{at}t^n$.*

Solution: *Recall Example 11.3: $\mathcal{L}(t^n)(s) = \dfrac{n!}{s^{n+1}}$ for $s > 0$. Using this and Theorem 11.20 we get*

$$\mathcal{L}(e^{at}t^n)(s) = \mathcal{L}(t^n)(s-a) = \frac{n!}{(s-a)^{n+1}} \qquad (s > a).$$

For example,

$$\mathcal{L}(e^{2t}t^4)(s) = \mathcal{L}(t^4)(s-2) = \frac{4!}{(s-2)^5} \qquad (s > 2).$$

□

EXAMPLE **11.22.** *Find the Laplace transform of $e^{at}\cos(\omega t)$.*

Solution: *The Laplace transform of $f(t) = \cos(\omega t)$ is $F(s) = \dfrac{s}{s^2+\omega^2}$ $(s > 0)$. Hence*

$$\mathcal{L}\left(e^{at}\cos(\omega t)\right) = \mathcal{L}(\cos(\omega t))(s-a) = \frac{s-a}{(s-a)^2+\omega^2} \qquad (s > a).$$

□

EXERCISE **11.5.1.** *Find the Laplace transform of the functions*
(a) $(t^3 - 3t + 2)e^{-2t}$
(b) $e^{4t}(t - \cos t)$

EXAMPLE **11.23.** *Find the inverse Laplace transform of $\dfrac{1}{(s-a)^n}$.*

Solution: *Notice* $\dfrac{1}{(s-a)^n} = F(s-a)$ *for* $F(s) = \dfrac{1}{s^n}$. *We know that* $f(t) = \mathcal{L}^{-1}(F)(s) = \dfrac{t^{n-1}}{(n-1)!}$.

It follows that, for any integer $n \geq 1$, *we have*

$$\mathcal{L}^{-1}\left(\frac{1}{(s-a)^n}\right)(t) = e^{at}f(t) = \frac{e^{at}t^{n-1}}{(n-1)!}.$$

For example,

$$\mathcal{L}^{-1}\left(\frac{1}{(s+2)^3}\right)(t) = \frac{e^{-2t}t^2}{2!} = \frac{t^2e^{-2t}}{2}.$$

□

Using s-shifting, we can find the inverse Laplace transform of any function of the form $G(s) = \dfrac{as+b}{ps^2+qs+r}$, where $ps^2+qs+r = 0$ has no real roots.

EXAMPLE **11.24.** *Find the inverse Laplace transform of*

$$G(s) = \frac{1}{s^2-4s+7} = \frac{1}{(s-2)^2+3}.$$

Solution: *We have* $G(s) = F(s-2)$, *where* $F(s) = \dfrac{1}{s^2+3}$, *so*

$$\mathcal{L}^{-1}(F)(t) = \frac{1}{\sqrt{3}}\sin(\sqrt{3}t).$$

By Theorem 11.20, $\mathcal{L}^{-1}(G)(t) = \dfrac{1}{\sqrt{3}}e^{2t}\sin(\sqrt{3}t)$. □

Here is a more complicated example.

EXAMPLE **11.25.** *Find the inverse Laplace transform of*

$$G(s) = \frac{2s}{s^2+2s+5} = \frac{2s}{(s+1)^2+4}.$$

Solution: *We use a similar method to the previous example.*

$$\begin{aligned}
\mathcal{L}^{-1}(G)(t) &= \mathcal{L}^{-1}\left(\frac{2(s+1)-2}{(s+1)^2+4}\right) \\
&= e^{-t}\mathcal{L}^{-1}\left(\frac{2s-2}{s^2+4}\right) \\
&= e^{-t}\mathcal{L}^{-1}\left(2\frac{s}{s^2+2^2} - 2\frac{1}{s^2+2^2}\right) \\
&= e^{-t}[2\cos(2t) - \sin(2t)].
\end{aligned}$$

□

The second shifting theorem is related to the so-called ***Heaviside function*** $H(t)$ defined by

$$H(t) = \begin{cases} 0 & , \ t < 0 \\ 1 & , \ t \geq 0 \end{cases}$$

This is also called the unit step function.

Notice that for any $a \in \mathbb{R}$ the graph of $H(t-a)$ is obtained from the graph of $H(t)$ by shifting a units (to the right if $a > 0$ and to the left if $a < 0$), that is:

$$H(t-a) = \begin{cases} 0 & , \ t < a \\ 1 & , \ t \geq a \end{cases}$$

We point out that multiplying a given function $g(t)$ by $H(t-a)$ has the effect of turning the function off until time $t = a$ and then activating it. More precisely, we have

$$g(t)H(t-a) = \begin{cases} 0 & , \ t < a \\ g(t) & , \ t \geq a \end{cases}$$

Multiplication by the ***pulse function*** $H(t-a) - H(t-b)$, where $a < b$ has the effect of a switch. This function has value one for $a \leq t < b$ and is zero for times $t < a$ and $t \geq b$. Thus the application of this function is equivalent to turning on a switch at $t = a$ then turning it off again at a later $t = b$.

$$g(t)[H(t-a) - H(t-b)] = \begin{cases} 0 & , \ t < a \\ g(t) & , \ a \leq t < b \\ 0 & , \ t \geq b \end{cases}$$

Because the Heaviside function is quite so important in real problems it is helpful to note the result of the following theorem, called t-shifting theorem.

THEOREM **11.26.** *If the Laplace transform of $f(t)$ is $F(s)$ for $s > b$, then for any $a \geq 0$ we have*

$$\mathcal{L}[f(t-a)H(t-a)] = e^{-as}F(s)$$

There is a restriction on a in this theorem: these results are *not* valid if a is negative.

for $s > b$. Consequently,

$$\mathcal{L}^{-1}\left(e^{-as}F(s)\right) = f(t-a)H(t-a).$$

You can try to prove this result (it involves the definition of the Laplace transform of H and one change of coordinate).

EXAMPLE **11.27.** *Find $\mathcal{L}(H(t-a))$ where $a \geq 0$.*

Solution: *We take $f(t) = 1$ and apply the theorem. We saw in Example 11.2 that $F(s) = \mathcal{L}(f) = \dfrac{1}{s}$ for $s > 0$. Thus*

$$\mathcal{L}\left(H(t-a)\right) = \mathcal{L}\left(f(t-a)H(t-a)\right) = e^{-as}F(s) = \frac{e^{-as}}{s}, \text{ for } s > 0.$$

□

EXAMPLE **11.28.** *Find $\mathcal{L}(g(t))$ where*

$$g(t) = \begin{cases} t & \text{if } 0 \le t < 3 \\ 1 - 3t & \text{if } t \ge 3 \end{cases}$$

Solution: *We can express $g(t)$ with Heaviside functions:*

$$\begin{aligned} g(t) &= t[H(t) - H(t-3)] + (1-3t)H(t-3) \\ &= tH(t) + (1-4t)H(t-3) \\ &= t + (1-4t)H(t-3). \end{aligned}$$

Recall we are only concerned with functions defined on $[0, \infty)$. On that interval, $H(t)$ is nothing else than the constant function 1.

This is still not in the form required in order to use Theorem 11.26: in the second term we have to write $1 - 4t$ as a function of $t - 3$. Since $t = (t-3) + 3$, we have $1 - 4t = 1 - 4(t-3) - 12 = -4(t-3) - 11$. Thus

$$g(t) = t - [4(t-3) + 11]H(t-3).$$

We now apply Theorem 11.26 with $f(t) = 4t + 11$.

$$\mathcal{L}([4(t-3)+11]H(t-3)) = \mathcal{L}(f(t-3)H(t-3))(s) = e^{-3s}F(s) = e^{-3s}\left(\frac{4}{s^2} + \frac{11}{s}\right).$$

Thus

$$\mathcal{L}(g)(s) = \frac{1}{s^2} - e^{-3s}\left(\frac{4}{s^2} + \frac{11}{s}\right).$$

□

We now solve a differential equation where the right-hand side uses Heavyside functions.

EXERCISE **11.5.2.** *Find the Laplace transform of the function*

$$f(t) = \begin{cases} 2t + 1 & \text{if } 0 \le t < 2 \\ 2 - 3t & \text{if } t \ge 2. \end{cases}$$

EXAMPLE **11.29.** *Find $\mathcal{L}^{-1}\left(\dfrac{e^{-4s}}{s^3}\right)$.*

Solution: *We apply the theorem with $a = 4$ and $F(s) = \dfrac{1}{s^3}$, so that $\mathcal{L}^{-1}\left(\dfrac{e^{-4s}}{s^3}\right) = f(t-4)H(t-4)$. All we have to do is determine $f(t)$. From the table we get that $f(t) = \dfrac{1}{2}t^2$, so that,*

$$\mathcal{L}^{-1}\left(\frac{e^{-4s}}{s^3}\right) = H(t-4)\frac{(t-4)^2}{2} = \begin{cases} 0 & \text{if } t < 4 \\ \dfrac{1}{2}(t-4)^2 & \text{if } t \ge 4 \end{cases}$$

□

EXERCISE **11.5.3.** *Find the inverse Laplace transforms of*

(a) $\dfrac{e^{-s}}{(s-5)^3}$

(b) $\dfrac{se^{-2s}}{s^2+9}$

EXAMPLE **11.30.** *Solve the initial value problem*

$$y'' + y = H(t-1) - H(t-2)$$

with initial conditions $y(0) = 0$ *and* $y'(0) = 1$.

Solution: *Taking transforms,* $Y(s) = \mathcal{L}(y)$ *satisfies*

$$s^2Y(s) - sy(0) - y'(0) + Y(s) = \frac{e^{-s}}{s} - \frac{e^{-2s}}{s}$$

and solving for Y *yields that*

$$\begin{aligned} Y(s) &= \frac{1}{s^2+1} + (e^{-s} - e^{-2s})\frac{1}{s(s^2+1)} \\ &= \frac{1}{s^2+1} + (e^{-s} - e^{-2s})\left(\frac{1}{s} - \frac{s}{s^2+1}\right) \\ &= \frac{1}{s^2+1} + e^{-s}\left(\frac{1}{s} - \frac{s}{s^2+1}\right) - e^{-2s}\left(\frac{1}{s} - \frac{s}{s^2+1}\right). \end{aligned}$$

We used partial fractions here.

We now need to find the inverse Laplace transform of $Y(s)$. *For the last two terms, we will apply Theorem 11.26 with* $F(s) = \left(\frac{1}{s} - \frac{s}{s^2+1}\right)$, *so that* $f(t) = 1 - \cos t$, *and with* $a = 1$ *or* 2. *We get*

$$y(t) = \sin t + H(t-1)[1 - \cos(t-1)] - H(t-2)[1 - \cos(t-2)].$$

Hence

$$y(t) = \begin{cases} \sin t, & 0 \le t < 1 \\ \sin t + 1 - \cos(t-1), & 1 \le t < 2 \\ \sin t - \cos(t-1) + \cos(t-2), & 2 \le t \end{cases}$$

Note that in this solution both y *and* y' *are continuous at* $t = 1$ *and* $t = 2$. □

EXERCISE **11.5.4.** *Solve the initial value problem*

$$y''(t) - 2y'(t) - 3y(t) = f(t),$$

where

$$f(t) = \begin{cases} 0 & \text{if } 0 \le t < 4 \\ 12 & \text{if } t \ge 4, \end{cases}$$

such that $y(0) = 1$ *and* $y'(0) = 0$.

11.6 Derivatives of transforms

If $f(t)$ is of exponential order $\le \gamma$ and piecewise continuous and bounded on $[0, T]$ for any $T > 0$, then by Theorem 11.9,

$$F(s) = \mathcal{L}(f)(s) = \int_0^\infty e^{-st} f(t)\, dt$$

exists for $s > \gamma$. Moreover we have the following:

THEOREM **11.31.** ***Derivative of transform.*** *Under the above assumptions, $F'(s)$ exists for all $s > \gamma$, and*

$$-F'(s) = \mathcal{L}\left(tf(t)\right). \tag{11.1}$$

Consequently,

$$\mathcal{L}^{-1}\left(F'(s)\right) = -tf(t) \quad (t \geq 0).$$

The proof of this result is omitted here as our focus is on how we might use Theorem 11.31 to find more function-transform pairs.

EXAMPLE **11.32.** *In Exercise 11.3.2, we found the transform of $g(t) = t\sin(\omega t)$ by differentiating twice. We now have an easier method since by taking $f(t) = \sin(\omega t)$ in Equation (11.1), so that $F(s) = \mathcal{L}(\sin(\omega t)) = \dfrac{\omega}{s^2+\omega^2}$, we get:*

$$\mathcal{L}(t\sin(\omega t)) = -F'(s) = -\frac{d}{ds}\left(\frac{\omega}{s^2+\omega^2}\right) = \frac{2\omega s}{(s^2+\omega^2)^2}.$$

□

EXERCISE **11.6.1.** *Use Theorem 11.31 twice to find the Laplace transform of $f(t) = t^2\cos(\omega t)$.*

11.7 Convolution

One last idea relevant to the theory of Laplace transforms is that of the convolution of two functions. Very often the solution of the transformed problem can be written in the form $Y(s) = F(s)G(s)$; it is extremely tempting to suppose that $y(t) = f(t)g(t)$ which would say that the Laplace transform of a product of two functions is the product of the Laplace transforms of the two functions. Unfortunately things are not that simple and instead requires the introduction of a concept known as the convolution.

DEFINITION **11.33.** *(Convolution)*
Given two functions $f(t)$ and $g(t)$, both of them being piecewise continuous and bounded on every finite interval $[0, T]$, the convolution *$f * g$ of f and g is defined by*

$$(f * g)(t) = \int_0^t f(u)g(t-u)\,du.$$

Main properties of the convolution:

$$\begin{aligned} f * g &= g * f && \text{commutative} \\ f * (g_1 + g_2) &= f * g_1 + f * g_2 && \text{distributive} \\ (f * g) * h &= f * (g * h) && \text{associative} \\ f * 0 &= 0 * f = 0 \end{aligned}$$

However, note that $f * 1$ is not equal to f in general and that $f * f$ can be negative.

THEOREM **11.34.** *(The Convolution Theorem) Let $f(t)$ and $g(t)$ be as above and let $F(s) = \mathcal{L}(f)$ and $G(s) = \mathcal{L}(g)$ be defined for $s > \gamma$. Then*

$$\mathcal{L}(f * g)(s) = F(s)G(s) \quad (s > \gamma).$$

Equivalently, $\mathcal{L}^{-1}(F(s)G(s)) = (f * g)(t)$.

This result tells us that if $f(t)$ and $g(t)$ have Laplace transforms $F(s)$ and $G(s)$ respectively then the function $(f * g)(t)$ has Laplace transform $F(s)G(s)$.

EXAMPLE **11.35.** *Find the inverse Laplace transform of*

$$K(s) = \frac{1}{(s-1)^2(s-3)^2}$$

using the Convolution Theorem.

Another method would be to use partial fractions as in Section 11.2.

Solution: *Notice that, by Example 11.23,*

$$\mathcal{L}^{-1}\left(\frac{1}{(s-1)^2}\right) = te^t = f(t) \quad , \quad \mathcal{L}^{-1}\left(\frac{1}{(s-3)^2}\right) = te^{3t} = g(t).$$

Therefore by the Convolution Theorem,

$$\begin{aligned}
\mathcal{L}^{-1}(K) &= \mathcal{L}^{-1}\left(\frac{1}{(s-1)^2} \cdot \frac{1}{(s-3)^2}\right) = f(t) * g(t) \\
&= \int_0^t f(u)g(t-u)\,du = \int_0^t ue^u(t-u)e^{3(t-u)}\,du \\
&= e^{3t}\int_0^t (tu - u^2)e^{-2u}\,du.
\end{aligned}$$

To evaluate the latter integral we have to use several integrations by parts which shows that

$$\int_0^t (tu - u^2)e^{-2u}\,du = \frac{t - 1 + (t+1)e^{-2t}}{4}.$$

Thus, $\mathcal{L}^{-1}(K) = e^{3t}\dfrac{t - 1 + (t+1)e^{-2t}}{4} = \dfrac{(t-1)e^{3t} + (t+1)e^t}{4}$. □

EXAMPLE **11.36.** *Find the inverse Laplace transform of $F(s) = \dfrac{1}{(s^2+1)^2}$ using the Convolution Theorem.*

Solution: *Since $F(s) = \dfrac{1}{(s^2+1)} \cdot \dfrac{1}{(s^2+1)}$ and $\mathcal{L}^{-1}(\dfrac{1}{s^2+1}) = \sin(t)$,*

by the Convolution Theorem,

$$f(t) = \mathcal{L}^{-1}(F) = \sin(t) * \sin(t) = \int_0^t \sin(u)\sin(t-u)\,du$$

$$= \frac{1}{2}\int_0^t [\cos(u-(t-u)) - \cos(u+(t-u))]\,du \quad \text{(using the cosine sum formula)}$$

$$= \frac{1}{2}\int_0^t [\cos(2u-t) - \cos t]\,du$$

$$= \frac{1}{2}\left[\frac{1}{2}\sin(2u-t) - u\cos t\right]_0^t$$

$$= \frac{1}{2}\sin t - \frac{1}{2}t\cos t.$$

□

EXERCISE **11.7.1.** *In the same manner as the previous example, show that*

$$\mathcal{L}^{-1}\left(\frac{s^2}{(s^2+1)^2}\right) = \frac{1}{2}\sin t + \frac{1}{2}t\cos t\ .$$

EXERCISE **11.7.2.** *Find the inverse Laplace transform of the functions using the Convolution Theorem:*

(a) $\dfrac{1}{(s^2+4)(s^2-4)}$

(b) $\dfrac{s}{(s^2+a^2)(s^2+b^2)}$

(c) $\dfrac{e^{-2s}}{s^2+16}$

We will use an example to demonstrate how the convolution Theorem can be applied to solve differential equations.

EXAMPLE **11.37.** *Solve the initial value problem*

$$y''(t) + y(t) = f(t) = \begin{cases} 1 & \text{if } 0 \le t < 1 \\ 0 & \text{if } t \ge 1 \end{cases}$$

with initial conditions $y(0) = 0$, $y'(0) = 1$.

Solution: *First we write the right-hand side using Heaviside functions:* $f(t) = H(t) - H(t-1)$. *Applying the Laplace transform to the DE, we get*

$$s^2Y(s) - 1 + Y(s) = \frac{1-e^{-s}}{s}$$

where $Y(s) = \mathcal{L}(y)$. *This gives (using partial fractions)*

$$Y(s) = \frac{1}{s^2+1} + (1-e^{-s})\frac{1}{s(s^2+1)}$$
$$= \frac{1}{s^2+1} + \frac{1}{s} - \frac{s}{s^2+1} - \frac{e^{-s}}{s}\cdot\frac{1}{s^2+1}.$$

Taking the inverse Laplace transform of this and using the Convolution Theorem for the last term, we get

$$y(t) = \mathcal{L}^{-1}(Y) = \sin t + 1 - \cos t - H(t-1) * (\sin t).$$

To evaluate the convolution integral, notice that when $0 \leq u \leq t < 1$ *we have* $H(u-1) = 0$*. Thus,*

$$H(t-1) * (\sin t) = \int_0^t H(u-1)\sin(t-u)\,du = 0, \quad \text{for } t < 1$$

while for $t \geq 1$ *we have:*

$$\begin{aligned} H(t-1) * (\sin t) &= \int_0^t H(u-1)\sin(t-u)\,du \\ &= \int_1^t \sin(t-u)\,du \\ &= [\cos(t-u)]_1^t \\ &= [1 - \cos(t-1)]. \end{aligned}$$

Hence $H(t-1) * (\sin t) = [1 - \cos(t-1)]H(t-1)$.

Finally we get the solution to the initial value problem:

$$y(t) = \sin t + 1 - \cos t - H(t-1)[1 - \cos(t-1)].$$

□

EXERCISE **11.7.3.** *Solve the following initial value problems using Laplace transforms:*

(a) $y''(t) + 4y'(t) + 13y(t) = f(t)$, $y(0) = y'(0) = 0$, *where* $f(t) = 1$ *for* $0 \leq t < \pi$ *and* $f(t) = 0$ *for* $t \geq \pi$.

(b) $y''(t) + 2y'(t) + 2y(t) = \sin t$, $y(0) = y'(0) = 0$

11.8 Laplace transforms table

$$\mathcal{L}\left(f(t)\right) = F(s) = \int_0^\infty f(t)e^{-st}\,dt$$

SPECIFIC FUNCTIONS		GENERAL RULES	
$F(s)$	$f(t)$	$F(s)$	$f(t)$
$\frac{1}{s}$	1	$\frac{e^{-as}}{s}$	$H(t-a)$
$\frac{1}{s^n}, \quad n \in \mathbb{Z}^+$	$\frac{t^{n-1}}{(n-1)!}$	$e^{-as}F(s)$	$f(t-a)H(t-a)$
$\frac{1}{s-a}$	e^{at}	$F(s-a)$	$e^{at}f(t)$
$\frac{1}{(s-a)^n}, \quad n \in Z^+$	$e^{at}\frac{t^{n-1}}{(n-1)!}$	$sF(s)-f(0)$	$f'(t)$
$\frac{1}{s^2+\omega^2}$	$\frac{\sin(\omega t)}{\omega}$	$s^2F(s)-sf(0)-f'(0)$	$f''(t)$
$\frac{s}{s^2+\omega^2}$	$\cos(\omega t)$	$F'(s)$	$-tf(t)$
$\frac{1}{(s-a)^2+\omega^2}$	$\frac{e^{at}\sin(\omega t)}{\omega}$	$F^{(n)}(s)$	$(-t)^n f(t)$
$\frac{s-a}{(s-a)^2+\omega^2}$	$e^{at}\cos(\omega t)$	$\frac{F(s)}{s}$	$\int_0^t f(u)\,du$
$\frac{1}{(s^2+\omega^2)^2}$	$\frac{\sin(\omega t)-\omega t\cos(\omega t)}{2\omega^3}$	$F(s)G(s)$	$\big(f*g\big)(t)$
$\frac{s}{(s^2+\omega^2)^2}$	$\frac{t\sin(\omega t)}{2\omega}$		

Higher derivatives:

$$\mathcal{L}\left(f^{(n)}(t)\right) = s^nF(s) - s^{n-1}f(0) - s^{n-2}f'(0) - \cdots - sf^{(n-2)}(0) - f^{(n-1)}(0)$$

The Convolution Theorem:

$$\mathcal{L}\left(f*g\right) = \mathcal{L}\left(f\right)\mathcal{L}\left(g\right) \quad \text{where} \quad \big(f*g\big)(t) = \int_0^t f(u)g(t-u)\,du$$

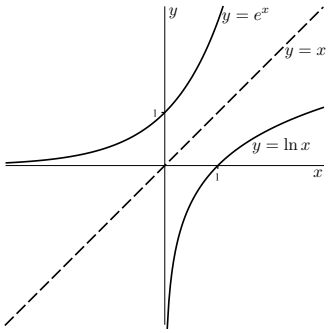
12
Appendix - Useful formulas

Exponential and logarithmic functions

(Natural) exponential function:
$y = e^x$, Domain $\mathbb{R}$, Range $(0, \infty)$.

(Natural) logarithmic function:
$y = \ln x$, Domain $(0, \infty)$, Range $\mathbb{R}$.

Cancellation equations:
$\ln(e^x) = x$ and $e^{\ln x} = x$.



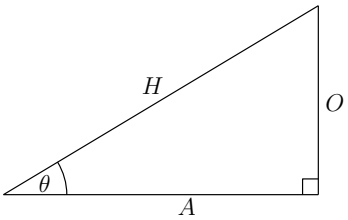
Note that $\ln x$ is shorthand for $\log_e x$.

Index and log laws:

$$e^x e^y = e^{x+y} \qquad \frac{e^x}{e^y} = e^{x-y} \qquad (e^x)^y = e^{xy}$$

$$\ln(xy) = \ln x + \ln y \qquad \ln\left(\frac{x}{y}\right) = \ln x - \ln y \qquad \ln(x^y) = y \ln x$$

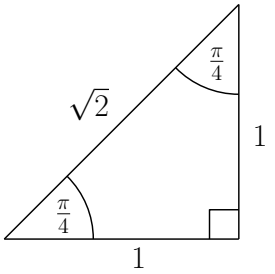
Trigonometry

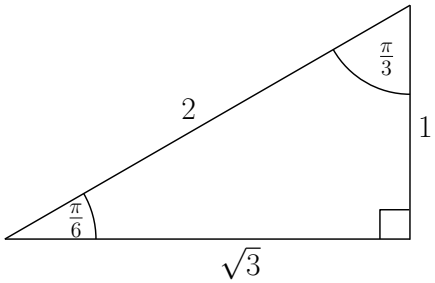


$$\sin\theta = \frac{O}{H}; \qquad \operatorname{cosec}\theta = \frac{1}{\sin\theta} = \frac{H}{O};$$
$$\cos\theta = \frac{A}{H}; \qquad \sec\theta = \frac{1}{\cos\theta} = \frac{H}{A};$$
$$\tan\theta = \frac{O}{A} \qquad \cot\theta = \frac{1}{\tan\theta} = \frac{A}{O}$$
$$= \frac{O/H}{A/H}$$
$$= \frac{\sin\theta}{\cos\theta}.$$

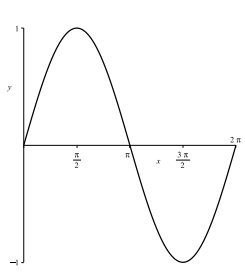
Note that $\operatorname{cosec}\theta$ is also known as just $\csc\theta$.

Reference triangles for common angles:





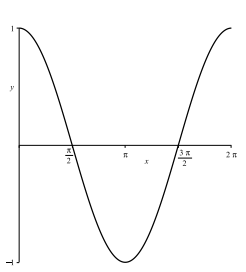
Trigonometric functions:



(a) $y = \sin x$

Domain $\mathbb{R}$

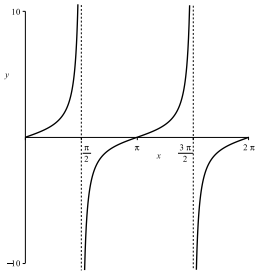
Range $[-1, 1]$



(b) $y = \cos x$

Domain $\mathbb{R}$

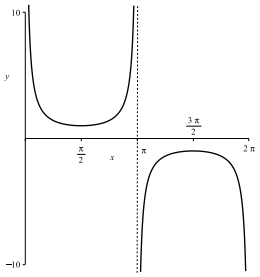
Range $[-1, 1]$



(c) $y = \tan x$

Domain $x \neq \frac{\pi}{2} + n\pi$

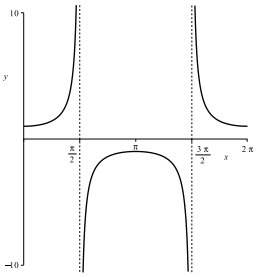
Range $\mathbb{R}$



(d) $y = \operatorname{cosec} x$

Domain $x \neq n\pi$

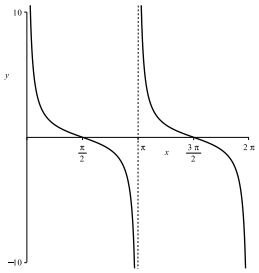
Range $\mathbb{R}$



(e) $y = \sec x$

Domain $x \neq \frac{\pi}{2} + n\pi$

Range $\mathbb{R}$



(f) $y = \cot x$

Domain $x \neq n\pi$

Range $\mathbb{R}$

Trigonometric properties:

Fundamental properties: $\sin^2 x + \cos^2 x = 1$,

$$\tan^2 x + 1 = \sec^2 x,$$

$$1 + \cot^2 x = \operatorname{cosec}^2 x.$$

Odd/even properties: $\sin(-x) = -\sin x$, $\cos(-x) = \cos x$.

Addition formula: $\sin(x + y) = \sin x \cos y + \cos x \sin y$,

$$\cos(x + y) = \cos x \cos y - \sin x \sin y,$$

$$\tan(x + y) = \frac{\tan x + \tan y}{1 - \tan x \tan y}.$$

Half-angle formula: $\sin(2x) = 2 \sin x \cos x$,

$$\cos(2x) = \cos^2 x - \sin^2 x = 2\cos^2 x - 1 = 1 - 2\sin^2 x,$$

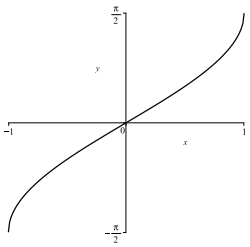
$$\tan(2x) = \frac{2 \tan x}{1 - \tan^2 x}.$$

Product formula: $\sin x \cos y = \frac{1}{2}[\sin(x + y) + \sin(x - y)]$,

$$\sin x \sin y = \tfrac{1}{2}[\cos(x - y) - \cos(x + y)],$$

$$\cos x \cos y = \tfrac{1}{2}[\cos(x + y) + \cos(x - y)].$$

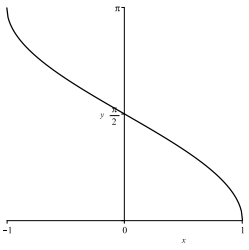
Inverse trigonometric functions:



(g) $y = \sin^{-1} x$

Domain $[-1, 1]$

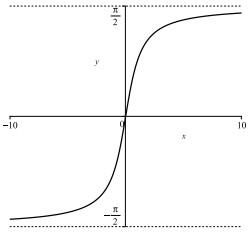
Range $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$



(h) $y = \cos^{-1} x$

Domain $[-1, 1]$

Range $[0, \pi]$



(i) $y = \tan^{-1} x$

Domain $\mathbb{R}$

Range $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$

Differentiation

The product rule: If $y = uv$ then $\dfrac{dy}{dx} = v\dfrac{du}{dx} + u\dfrac{dv}{dx}$.

The quotient rule: If $y = \dfrac{u}{v}$ then $\dfrac{dy}{dx} = \dfrac{v\dfrac{du}{dx} - u\dfrac{dv}{dx}}{v^2}$.

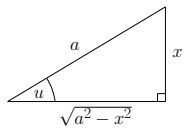
The chain rule: If $y = f(u)$ and $u = g(x)$ then $\dfrac{dy}{dx} = \dfrac{dy}{du} \times \dfrac{du}{dx}$.

Integration

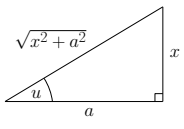
Integration by inverse trigonometric substitution:

Integral involves	Then substitute	Restriction on u	Use the identity
$\sqrt{a^2 - x^2}$	$x = a\sin u$	$-\dfrac{\pi}{2} \le u \le \dfrac{\pi}{2}$	$1 - \sin^2 u = \cos^2 u$
$\sqrt{a^2 + x^2}$	$x = a\tan u$	$-\dfrac{\pi}{2} < u < \dfrac{\pi}{2}$	$1 + \tan^2 u = \sec^2 u$
$\sqrt{x^2 - a^2}$	$x = a\sec u$	$0 \le u < \dfrac{\pi}{2}$	$\sec^2 u - 1 = \tan^2 u$

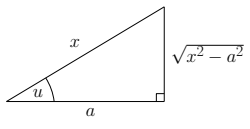
To return to the original variable x use the *reference triangles* illustrated below.



(j) Reference triangle for $x = a\sin u$



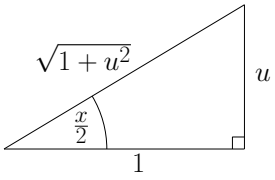
(k) Reference triangle for $x = a\tan u$



(l) Reference triangle for $x = a\sec u$

Integration by half-angle substitution:

The substitution $u = \tan\left(\frac{x}{2}\right) \Rightarrow x = 2\tan^{-1} u$ with reference triangle shown to the right turns an integral with a quotient involving $\sin x$ and/or $\cos x$ into an integral of a rational function of u, where $\sin x = \dfrac{2u}{1+u^2}$ and $\cos x = \dfrac{1-u^2}{1+u^2}$.



Integration by partial fractions:

A rational function $f(x) = \dfrac{P(x)}{Q(x)}$ with $\deg(P(x)) < \deg(Q(x))$ can be decomposed into partial fractions as follows:

Case 1: Denominator has distinct linear factors

$$f(x) = \frac{P(x)}{(x-a_1)\cdots(x-a_k)} = \frac{A_1}{x-a_1} + \cdots + \frac{A_k}{x-a_k},$$

where $a_1, \ldots, a_k$ are pairwise distinct.

Case 2: Denominator has repeated linear factors

$$f(x) = \frac{P(x)}{(x-a)^c} = \frac{B_1}{x-a} + \frac{B_2}{(x-a)^2} + \cdots + \frac{B_{c-1}}{(x-a)^{c-1}} + \frac{B_c}{(x-a)^c}.$$

Case 3: Denominator has an irreducible factor of degree 2

$$f(x) = \frac{P(x)}{(x-a)(x^2+bx+c)} = \frac{A_1}{x-a} + \frac{C_1x+C_2}{x^2+bx+c}.$$

If $\deg(P(x)) \geq \deg(Q(x))$ use *polynomial division* on the rational function before decomposing into partial fractions.

Integration by parts:

$$\int u\,dv = uv - \int v\,du.$$

Use the following table as a *guide*:

u	dv
Polynomial	Exponential Trigonometric
Logarithmic Inverse trigonometric	Polynomial

Differentiation and integration formulas

$\frac{dy}{dx}$	y	$\int y\,dx$
0	a (constant)	$ax + C$
nx^{n-1}	x^n $(n \neq -1)$	$\frac{x^{n+1}}{n+1} + C$
$-\frac{1}{x^2}$ or $-x^{-2}$	$\frac{1}{x}$ or x^{-1}	$\ln x + C$
e^x	e^x	$e^x + C$
$\frac{1}{x}$	$\ln x$	$x \ln x - x + C$
$\cos x$	$\sin x$	$-\cos x + C$
$-\sin x$	$\cos x$	$\sin x + C$
$\sec^2 x$	$\tan x$	$\ln(\sec x) + C$
$-\cot x \operatorname{cosec} x$	$\operatorname{cosec} x$	$\ln(\operatorname{cosec} x - \cot x) + C$
$\tan x \sec x$	$\sec x$	$\ln(\sec x + \tan x) + C$
$-\operatorname{cosec}^2 x$	$\cot x$	$\ln(\sin x) + C$
$\frac{1}{\sqrt{1-x^2}}$	$\sin^{-1} x$	$x \sin^{-1} x + \sqrt{1-x^2} + C$
$-\frac{1}{\sqrt{1-x^2}}$	$\cos^{-1} x$	$x \cos^{-1} x - \sqrt{1-x^2} + C$
$\frac{1}{1+x^2}$	$\tan^{-1} x$	$x \tan^{-1} x - \frac{1}{2}\ln(1+x^2) + C$

Laplace transforms table

$$\mathcal{L}\left(f(t)\right) = F(s) = \int_0^\infty f(t)e^{-st}\,dt$$

SPECIFIC FUNCTIONS		GENERAL RULES	
$F(s)$	$f(t)$	$F(s)$	$f(t)$
$\frac{1}{s}$	1	$\frac{e^{-as}}{s}$	$H(t-a)$
$\frac{1}{s^n}, \quad n \in \mathbb{Z}^+$	$\frac{t^{n-1}}{(n-1)!}$	$e^{-as}F(s)$	$f(t-a)H(t-a)$
$\frac{1}{s-a}$	e^{at}	$F(s-a)$	$e^{at}f(t)$
$\frac{1}{(s-a)^n}, \quad n \in Z^+$	$e^{at}\frac{t^{n-1}}{(n-1)!}$	$sF(s) - f(0)$	$f'(t)$
$\frac{1}{s^2+\omega^2}$	$\frac{\sin(\omega t)}{\omega}$	$s^2F(s) - sf(0) - f'(0)$	$f''(t)$
$\frac{s}{s^2+\omega^2}$	$\cos(\omega t)$	$F'(s)$	$-tf(t)$
$\frac{1}{(s-a)^2+\omega^2}$	$\frac{e^{at}\sin(\omega t)}{\omega}$	$F^{(n)}(s)$	$(-t)^n f(t)$
$\frac{s-a}{(s-a)^2+\omega^2}$	$e^{at}\cos(\omega t)$	$\frac{F(s)}{s}$	$\int_0^t f(u)\,du$
$\frac{1}{(s^2+\omega^2)^2}$	$\frac{\sin(\omega t) - \omega t\cos(\omega t)}{2\omega^3}$	$F(s)G(s)$	$\big(f*g\big)(t)$
$\frac{s}{(s^2+\omega^2)^2}$	$\frac{t\sin(\omega t)}{2\omega}$		

Higher derivatives:

$$\mathcal{L}\left(f^{(n)}(t)\right) = s^nF(s) - s^{n-1}f(0) - s^{n-2}f'(0) - \cdots - sf^{(n-2)}(0) - f^{(n-1)}(0)$$

The Convolution Theorem:

$$\mathcal{L}\left(f*g\right) = \mathcal{L}\left(f\right)\mathcal{L}\left(g\right) \quad \text{where} \quad \big(f*g\big)(t) = \int_0^t f(u)g(t-u)\,du$$

13
Index