NICHOLAS JACKSON

EC961 INTRODUCTORY MATHEMATICS AND STATISTICS

DEPARTMENT OF ECONOMICS, UNIVERSITY OF WARWICK

Contents

Introduction

THIS MODULE, *EC961 Introductory Mathematics and Statistics* is intended to provide a working background knowledge of the mathematical and statistical techniques necessary for MSc programmes in the Department of Economics at the University of Warwick.

What is this?

This is a fairly intensive pre-sessional module, running in the last two weeks of September, just before the start of the Autumn Term. The idea is to give everyone a basic working knowledge of a range of topics in mathematics and statistics that are necessary for fully engaging with subsequent modules in macroeconomics, microeconomics and econometrics.

The module is in four parts:

Calculus and Dynamics: This covers compound and exponential growth, differential calculus and its use in finding optimal solutions to problems in economics, elasticity, Taylor–Maclaurin series, some more detailed theoretical background in calculus (such as the Intermediate Value Theorem and the Fixed Point Theorem), solution of first order linear difference equations, and some basic concepts and facts about concave, convex, quasiconcave and quasiconvex functions.

Linear Algebra: This is concerned with the algebra of vectors and matrices, and the solution of problems involving linear functions. In particular, we will cover eigenvalues and eigenvectors, solution of simultaneous linear equations, matrix diagonalisation and its applications, and the classification of quadratic forms.

Multivariate Calculus: This section covers more advanced topics in calculus, including partial differentiation, finding optimal solutions to functions of more than one variable, using Lagrangian optimisation and the Karush–Kuhn–Tucker (KKT) conditions, solving systems of difference equations, and solving some classes of differential equations.

Statistics and Probability In this section, we will introduce the fundamental concepts of probability theory such as discrete and continuous random variables, conditional probability, Bayes' Theorem, standard probability distributions, and their use in statistical inference and hypothesis testing.

This document contains the lecture notes for the first three of these.

Who are we?

The two module lecturers are:

	Subject	Office
Dr Nicholas Jackson	Calc and Dynamics, Linear Algebra, Multivariate Calc	Zeeman B0.09, Economics S.084
Dr Nikhil Datta	Statistics	Economics S1.109

The class tutors are:

Group	Tutor	Location
1	**Hussain Abass**	FAB 3.33
2	**Kyle Boutilier**	FAB 4.80
3	**Andrew Brendon-Penn**	FAB 6.02
4	**Dr Juliana Cunha Carneiro Pinto**	FAB 3.31
5	**Dr Darina Dintcheva**	FAB 4.79
6	**George Ferridge**	FAB 3.30
7	**Dr Farzad Javidanrad**	FAB 6.01
8	**Minh Tung Le**	FAB 4.73
9	**Dr James Massey**	FAB 3.32
10	(TBC)	FAB 4.78
11	**Dr Deva Velivela**	FAB 3.25
12	**Dr Nicholas Jackson / Dr Neil Lloyd**	FAB 2.43

When?

Table 1 shows the provisional timetable.

Week 1	**Mon 18 Sep**	**Tue 19 Sep**	**Wed 20 Sep**	**Thu 21 Sep**	**Fri 22 Sep**
Morning (10am–1pm)		**Lecture 2** Calc and Dynamics	**Lecture 3** Linear Algebra	**Lecture 4** Linear Algebra	**Lecture 5** Statistics
Afternoon (2pm–5pm)	**Lecture 1** Calc and Dynamics	**Class 1** Calc and Dynamics	**Class 2** Calc and Dynamics	**Class 3** Linear Algebra	**Class 4** Linear Algebra

Week 2	**Mon 25 Sep**	**Tue 26 Sep**	**Wed 27 Sep**	**Thu 28 Sep**	**Fri 29 Sep**
Morning (10am–1pm)	**Test 1** (12pm–1pm)	**Lecture 6** Statistics	**Lecture 7** Statistics	**Lecture 8** Multivariate Calc	**Lecture 9** Multivariate Calc
Afternoon (2pm–5pm)		**Class 5** Statistics	**Class 6** Statistics	**Class 7** Multivariate Calc	**Class 8** Multivariate Calc

Week 3	**Mon 2 Oct**	**Tue 3 Oct**	**Wed 4 Oct**	**Thu 5 Oct**	**Fri 6 Oct**
Morning (10am–1pm)	**Test 1** (retake) (12pm–1pm)		**Test 2** (9am–12pm)		
Afternoon (2pm–5pm)	**Revision Lecture** (2pm–3pm)				

Table 1: Provisional timetable

There will be nine lectures (mostly in the mornings), in which we will cover the course material, and eight classes (in the afternoons) in which you will work through practice questions in smaller groups, under the supervision of a class tutor. The lectures will take place in lecture theatre MS.01 in the Zeeman Building, and the classes will take place in smaller rooms in the Faculty of Arts Building.

There will also be two online tests. The first of these will take place 12pm-1pm on Monday 25 September, and will cover the Calculus and Dynamics material, and the first half of the Linear Algebra topics. If you're not happy with your performance on this test, there will be a second opportunity to take (a similar but different version of) it on Monday 2 October.

The second test will cover all of the material, and will take place from 9am–12pm on Wednesday 4 October. There will also be a revision lecture and Q&A session at 2pm on Monday 2 October.

I

Calculus and Dynamics

1 Logic

In this section we introduce some basic concepts of logic: logical implication, necessary and sufficient conditions.

In the following, we will use letters such as P and Q as placeholders for statements that might be true or false. For example "4 is an even number" is a true statement, while "$\frac{1}{2}$ is an integer" is false.

Definition 1.1 Given two statements P and Q, each of which may be true or false, we say that P **implies** Q if, whenever P is true, Q is true as well. We write $P \Rightarrow Q$.

Note carefully what this definition says, and what it *doesn't* say. In particular, Q might also be true when P isn't, but it is always true when P is true.

Examples 1.2

(i) n is an integer $\Rightarrow n$ is a real number.

(ii) Suppose that $x \in \mathbb{R}$. Then $x \geqslant 10 \Rightarrow x > 0$.

(iii) Suppose that $y \in \mathbb{Z}$. Then $y \geqslant 10 \Rightarrow y > 9$.

In the first two examples, although $P \Rightarrow Q$, there are also cases (for example, $n = 2.5$ and $x = 3$) where Q is true but P isn't. In the third example, P is true exactly when Q is.

Formally, even if $P \Rightarrow Q$, it doesn't necessarily follow that $Q \Rightarrow P$.

If $P \Rightarrow Q$ and $Q \Rightarrow P$, that is, P is true *exactly* when Q is true, we write $P \Leftrightarrow Q$.

Definition 1.3 If $P \Rightarrow Q$, we say that P is a **sufficient condition** for Q.

If Q isn't true, then P couldn't have been true either, so we say that Q is a **necessary condition** for P.

If $P \Leftrightarrow Q$, they are necessary and sufficient conditions for each other. Equivalently, P is true if and only if Q is true.[1]

Definition 1.4 If $P \Rightarrow Q$, then the related compound statement $Q \Rightarrow P$ is called the **converse**.

Given some statement P that may be true or false, we sometimes want to refer to a statement that logically opposite to P.

Definition 1.5 Let P be a formal statement. The **negation** of P, denoted $\neg P$, is a statement that is false exactly when P is true, and true exactly when P is false.

Now suppose that we have two statements P and Q such that $P \Rightarrow Q$. Then if P is true, Q is true. But if Q is false then P must

[1] The phrase "if and only if" is sometimes abbreviated "iff".

have been false too. So the statement

$$(P \text{ is true}) \Rightarrow (Q \text{ is true})$$

can be rephrased as

$$(Q \text{ is false}) \Rightarrow (P \text{ is false}).$$

or, using the notation introduced above,

$$\neg Q \Rightarrow \neg P.$$

This is called the **contrapositive** statement, and is logically equivalent to the original implication $P \Rightarrow Q$.

If we want to formally prove some mathematical statement of the form $P \Rightarrow Q$, it is logically equivalent, and sometimes simpler, to prove the contrapositive $\neg Q \Rightarrow \neg P$.

Summary

Suppose that $P \Rightarrow Q$. Then

- P is a sufficient condition for Q: If P is true then Q is true.
- Q is a necessary condition for P: if Q isn't true, then P can't be true either.

But: Q might be true even if P is false, and P might be false even if Q is true.

2 Exponentiation and compounding

Now, we look in detail at exponential growth, compound interest and logarithms.

The exponential function

Consider the sequence

$$x_n = \left(1 + \tfrac{1}{n}\right)^n$$

for $n \in \mathbb{N}$. The first few terms of this sequence are as follows:

$$\begin{aligned} x_1 &= \left(1 + \tfrac{1}{1}\right)^1 = 2 \\ x_2 &= \left(1 + \tfrac{1}{2}\right)^2 = \tfrac{9}{4} = 2.25 \\ x_3 &= \left(1 + \tfrac{1}{3}\right)^3 = \tfrac{64}{27} = 2.37037\ldots \end{aligned}$$

It can be shown that this sequence is **strictly increasing**; that is, $x_{n+1} > x_n$ for all $n \in \mathbb{N}$. It can also be shown that $0 < x_n < 3$; that is, x_n is **bounded**.

A standard theorem in real analysis implies that this sequence converges to a finite limit. The limit of this particular sequence is the irrational constant

$$e = \lim_{n\to\infty} \left(1 + \tfrac{1}{n}\right)^n = 2.7182818284\ldots$$

Theorem 2.1 *Let (a_n) be a sequence. If a_n is bounded above (or below) and increasing (or decreasing) then it tends to a finite limit as $n \to \infty$.*

Now suppose we want to find the limit of the sequence

$$y_n = \left(1 + \tfrac{ax}{n}\right)^n$$

for some $a, x \in \mathbb{R}$ with $ax \neq 0$. To do this, we make use of another important theorem in real analysis, the **Sandwich Rule**.

We first note that

$$\left(1 + \tfrac{ax}{n}\right)^n = \left(1 + \tfrac{1}{n/ax}\right)^n = \left(\left(1 + \tfrac{1}{n/ax}\right)^{n/ax}\right)^{ax}. \tag{2.1}$$

Theorem 2.2 (The Sandwich Rule) *Let (a_n), (b_n) and (c_n) be sequences of real numbers, and suppose that there exists some $L \in \mathbb{R}$ such that $a_n \to L$ and $c_n \to L$ as $n \to \infty$. Suppose also that there exists some $N \in \mathbb{N}$ such that*

$$a_n \leqslant b_n \leqslant c_n$$

for all $n > N$. Then $b_n \to L$ as $n \to \infty$.

The function $f(y) = \left(1 + \tfrac{1}{y}\right)^y$ is strictly increasing, so we can form the following inequality:[1]

$$\left(1 + \tfrac{1}{\lfloor n/ax \rfloor}\right)^{\lfloor n/ax \rfloor} \leqslant \left(1 + \tfrac{1}{n/ax}\right)^{n/ax} \leqslant \left(1 + \tfrac{1}{\lceil n/ax \rceil}\right)^{\lceil n/ax \rceil} \tag{2.2}$$

Since $a_n = \lfloor n/ax \rfloor$ and $c_n = \lceil n/ax \rceil$ are both increasing sequences of integers, both the left and right sides of (2.2) converge to e as

[1] Here, $\lfloor x \rfloor$ is the **floor** of x; that is, the largest integer $\leqslant x$. And $\lceil x \rceil$ is the **ceiling** of x; that is, the smallest integer $\geqslant x$.

$n \to \infty$. And hence by the Sandwich Rule, this means that the term in the middle must also tend to e. That is,

$$\lim_{n\to\infty} \left(1 + \tfrac{1}{n/ax}\right) = e.$$

Substituting this into (2.1) we find that

$$\lim_{n\to\infty} \left(1 + \tfrac{ax}{n}\right)^n = e^{ax}.$$

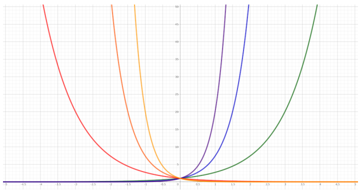


Figure 2.1: Graph of the function $f(x) = e^{ax}$ for $a = \pm 1, \pm 2, \pm 3$

We call this function $f(x) = e^{ax}$ the **exponential function**, for any $a, x \in \mathbb{R}$. See Figure 2.1.

Compound interest

We will now apply all this to study compound growth.

Example 2.3 Suppose we invest a sum $S_0 = £1000$ in an account earning $r = 12\%$ compound interest per year. The subsequent balance of the account depends on the frequency of compounding.

If the interest is paid annually, at the end of the year, then at the end of that first year we will have

$$S_1 = £1000 \times \left(1 + \tfrac{0.12}{1}\right)^{1\times 1} = £1120.$$

If, however, the interest is paid in quarterly instalments of $\frac{12}{4} = 3\%$, then at the end of the year we will have

$$S_1 = £1000 \times \left(1 + \tfrac{0.12}{4}\right)^{4\times 1} = £1125.51.$$

If the interest is paid monthly, in twelve equal instalments of $\frac{12}{12} = 1\%$, then the subsequent balance of the account will be

$$S_1 = £1000 \times \left(1 + \tfrac{0.12}{12}\right)^{12\times 1} = £1126.82.$$

And, more generally, if the interest is paid in m equal instalments of $\frac{12\%}{m} = \frac{0.12}{m}$ then the resulting balance will be

$$S_1 = £1000 \times \left(1 + \tfrac{0.12}{m}\right)^{m\times 1}$$

and at the end of year t, the total amount will be

$$S_t = £1000 \times \left(1 + \tfrac{0.12}{m}\right)^{mn}.$$

Generalising this example, suppose that we invest an initial capital amount S_0 in an account paying a nominal annual interest rate of r, in m equal instalments. Then at the end of year t, the balance of the account will be

$$S_t = S_0\left(1 + \tfrac{r}{m}\right)^{mt}.$$

The expression in parentheses should look familiar. If we let $m \to \infty$, then the limit of $\left(1 + \frac{r}{m}\right)^m$ is e^r. In the above example, we see that $S_1 \to £1000 \times e^{0.12} = £1127.50$.

More generally, for an annual interest rate of r, this **continuous compounding** process will result in a balance of

$$S_t = S_0 e^{rt}$$

at the end of year t.

Example 2.4 Now suppose we invest an initial sum of S_0 in an account paying a net interest rate of r, applied annually. What is the smallest value of r that enables the balance of the account to grow by a factor of 10 (that is, to be $10S_0$) in 23 years?

We have to solve

$$\begin{aligned} && S_0\left(1+\tfrac{r}{1}\right)^{23} &= 10S_0 \\ \Longrightarrow && (1+r)^{23} &= 10 \\ \Longrightarrow && r &= 10^{1/23} - 1 = 0.10529\ldots \end{aligned}$$

So the interest rate must be at least 10.529%.

Example 2.5 Now suppose that the interest is applied continuously. Suppose we have a nominal interest rate of $r = 4\%$. How many full years will it take for an account to grow by a factor of 100?

Here, we have to solve

$$\begin{aligned} && S_0 e^{0.04t} &= 100S_0 \\ \Longrightarrow && e^{0.04t} &= 100 \\ \Longrightarrow && 0.04t &= \ln(100) \\ \Longrightarrow && t &= \tfrac{\ln(100)}{0.04} = 115.129\ldots \end{aligned}$$

We want the number of full years, which is $\lceil t \rceil = 116$ years.

Here, ln denotes the **natural logarithm** function, which we will discuss in more depth later in this chapter.

Functions

First we introduce some details and terminology about functions.

Definition 2.6 Let $X, Y \subseteq \mathbb{R}$ be subsets of the set $\mathbb{R}$ of real numbers.

A **function** $f\colon X \to Y$ is a rule that assigns exactly one element of Y to each element of X. That is, for every element $x \in X$ there is some corresponding element $y \in Y$, determined according to some precise rule. We typically denote this element y as $f(x)$.

We call X the **domain** and Y the **codomain** of the function f.

Let's look at what this definition says, and what it doesn't say.

- Every element of X must map to something in Y.
- We don't allow any element of X to map to more than one element of Y. (That is, we don't allow "one to many" mappings.)
- Distinct elements of X can map to the same element of Y.
- Not every element of Y has to be mapped to by something in X.

The first two of these points rule out "one to many" and "one to nothing" correspondences. The third and fourth points lead us to consider two important classes of functions.

Definition 2.7 A function $f\colon X \to Y$ is **injective** or **one to one** if, for any $a, b \in X$ such that $f(a) = f(b)$, then $a = b$. That is, f maps distinct elements to distinct elements of Y. Equivalently, every element of Y is mapped to by *at most* one element of X.

Definition 2.8 A function $f\colon X \to Y$ is **surjective** or **onto** if, for every element $y \in Y$, there exists some element $x \in X$ such that $f(x) = y$. Equivalently, every element of Y is mapped to by *at least* one element of X.

It's also useful to consider which elements of the codomain are mapped to by some element of the domain:

Definition 2.9 The **image** or **range** of a function $f\colon X \to Y$ is the subset of the codomain consisting of elements that are mapped to by elements of the domain:

$$\operatorname{im}(f) = \{f(x) : x \in X\} \subseteq Y.$$

So a function $f\colon X \to Y$ is surjective exactly when $\operatorname{im}(f) = Y$.

Is it possible for a function to be both injective and surjective? Yes: these are the functions where every element of the codomain is mapped to by *exactly* one element of the domain.

Definition 2.10 A function $f\colon X \to Y$ is **bijective** if it is both injective and surjective.

Given two functions $f\colon A \to B$ and $g\colon C \to D$, where $B \subseteq C$, we can take the output of f and feed it into the input of g, thus chaining the functions together to make effectively a single function mapping from A to D.

Definition 2.11 Let $f\colon A \to B$ and $g\colon C \to D$ be functions, and suppose that $B \subseteq C$. Then the **composite** function $(g \circ f)\colon A \to D$ is the function defined by

$$(g \circ f)(a) = g(f(a))$$

for all $a \in A$.

We won't look in very much detail at function composition, but note that the order is important: $g \circ f$ means "apply f then apply g", not the other way round.[2]

Example 2.12 Let $f\colon \mathbb{R} \to \mathbb{R}$ and $g\colon \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = 3x + 1 \qquad \text{and} \qquad g(x) = x^2 - 4.$$

Then we can define the composite functions as follows:

$$(g \circ f)(x) = g(f(x)) = f(x)^2 - 4 = (3x+1)^2 - 4 = 9x^2 + 6x - 3$$
$$(f \circ g)(x) = f(g(x)) = 3g(x) + 1 = 3(x^2 - 4) + 1 = 3x^2 - 11$$

Note that, even if both composites $g \circ f$ and $f \circ g$ are defined, it needn't be the case that they're both equal to each other.

[2] The reason for this is that English is written from left to right, and that by convention we denote images of elements by $f(x)$.

But one important use of composition is in the definition of inverse functions. The idea here is that if we have a function $f\colon X \to Y$, we sometimes want to define a new function that takes every element in $\operatorname{im}(f)$ and maps it back to its original element in X. That is, we want to find (if possible) a function that reverses or undoes the action of f.

Definition 2.13 Let $f\colon X \to Y$ be a function. Then an **inverse** of f is a function $g\colon Y \to X$ such that

$$(g \circ f)(x) = g(f(x)) = x$$

for all $x \in X$, and

$$(f \circ g)(y) = f(g(y)) = y$$

for all $y \in Y$.

It turns out that a function $f\colon X \to Y$ has an inverse exactly when f is bijective. If f isn't injective, then for some elements in Y there will be more than one element in X that mapped to it. And if f isn't surjective, there will be elements in Y that didn't come from anything in X. It also turns out that if f has an inverse at all, it will be unique, and so we can safely talk about "the" inverse of f. Usually we will denote this function as f^{-1}.[3]

Example 2.14 Suppose $f\colon \mathbb{R} \to \mathbb{R}$ with $f(x) = \frac{1}{2}x + \frac{1}{3}$. This function is bijective, so it has an inverse, and we can calculate it as follows:

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

So we can define $f^{-1}\colon \mathbb{R} \to \mathbb{R}$ by $f^{-1}(y) = 2y - \frac{2}{3}$ for all $y \in \mathbb{R}$.

Logarithms

Now back to our discussion of compound growth and related topics. We will now introduce logarithms, whose original discovery is attributed to the Scottish mathematician John Napier.

John Napier (1550–1617)

Consider the exponential function $f\colon \mathbb{R} \to \mathbb{R}^+$ where $f(x) = e^x$. Its domain is the entire set $\mathbb{R}$ of real numbers, and its image is the set of strictly positive real numbers $\mathbb{R}^+ = \{x \in \mathbb{R} : x > 0\}$. If, as we've done here, we set the codomain of our function to be equal to the image, then it will be surjective. In fact, this function is injective as well, and hence bijective. It is therefore invertible.

We define this inverse function f^{-1} to be the **natural logarithm** function, and denote it ln. The natural logarithm $\ln(x)$ is the exponent we have to raise e to in order to get x.

More generally, given any positive real number $a > 0$ such that $a \neq 1$, the **logarithm to base** a, denoted $\log_a(x)$ is the power we

[3] If we plot the graph of an invertible function $f\colon X \to Y$, what will the graph of its inverse $f^{-1}\colon Y \to X$ look like?

have to raise a to, in order to get x.

That is, $\log_a(x)$ is $b \in \mathbb{R}$ such that $a^b = x$. This is a well-defined function, as long as $a > 0$ and $a \neq 1$. In particular, we usually denote $\log_{10}(x)$ just as $\log(x)$, and $\log_e(x)$ as $\ln(x)$.

Properties of $\ln(x)$ ***and*** $\log_a(x)$

Suppose that $a, b, c, y \in \mathbb{R}^+$, that $a \neq 1$, and that $d \in \mathbb{R}$. Then:

(i) $y = a^{\log_a(y)}$.
(ii) $\log_a(b) + \log_a(c) = \log_a(bc)$.[4]
(iii) $\log_a(b^d) = d \log_a(b)$.
(iv) $\log_a(b) - \log_a(c) = \log_a(b/c)$.
(v) $(\log_a(b))(\log_b(c)) = \log_a(c)$.

(In particular, all of these properties hold for $\log_e = \ln$.)

[4] This is why slide rules work.

3 *Derivatives and Elasticity*

NOW WE WANT TO LOOK AT differential calculus, and its application to finding local and global maxima and minima of functions of a single real variable.

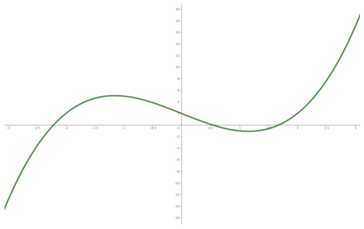


Figure 3.1: An example of a continuous function

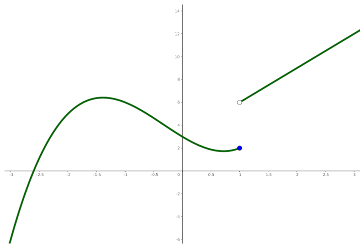


Figure 3.2: An example of a discontinuous function

Definitions and basic results

We will start by quickly reviewing the concepts of continuous and differentiable functions. We won't go into much detail, because this isn't a module on real analysis, but it's important to have some familiarity with the basic ideas.

Intuitively, we think of a function as being continuous if its graph has no breaks in it. For example, the function in Figure 3.1 is continuous in this sense, while the function in Figure 3.2 has a discontinuous point: the graph jumps abruptly at $x = 1$.

We want to formalise this idea and make it mathematically precise.

Definition 3.1 A function $f\colon \mathbb{R} \to \mathbb{R}$ is **continuous** at the point $x = a$ if the limit $\lim_{x\to a} f(x)$ exists, and is equal to the value $f(a)$.

We can extend this idea to continuity over an open interval in a relatively straightforward way:

Definition 3.2 A function $f\colon \mathbb{R} \to \mathbb{R}$ is **continuous** in the open interval (a, b) if it is continuous at every point $x \in (a, b)$.

Continuity over a closed interval is slightly more complicated, because we have to think about what happens at the endpoints. We want our definition to say exactly what we mean, no more, and no less. In particular, we don't need the limits at the endpoints to exist from both sides (although often they will).

Definition 3.3 A function $f\colon \mathbb{R} \to \mathbb{R}$ is **continuous** in the closed interval $[a, b]$ if it is continuous in the open interval (a, b), and if the one-sided limits

$$\lim_{x\to a^+} f(x) \qquad \text{and} \qquad \lim_{x\to b^-} f(x)$$

exist, and are equal to the values of $f(x)$ at the endpoints.

Now let's look at differentiation. The idea here is that the graph of the function must have a well-defined gradient at a given point. Geometrically, we define the gradient of the graph to be the gradient of the tangent to the graph at our chosen point.

In Figure 3.3 we want to find the gradient of the blue tangent line. In general, it's not obvious how we'd do this, but the gradient of the red chord is straightforward to calculate: it's the vertical height $f(x) - f(a)$ divided by the horizontal distance $x - a$.

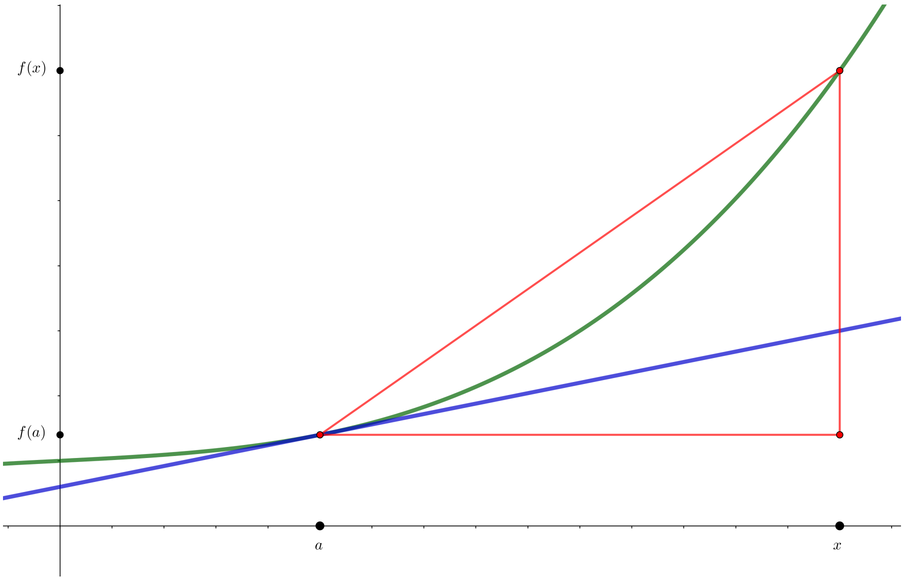


Figure 3.3: Illustration of the definition of the first derivative $f'(a)$

Joseph-Louis Lagrange (1736–1813)

Gottfried Wilhelm Leibniz (1646–1716)

If x is close to a, then the chord will be fairly close to the tangent. And the closer x gets to a, the better this value will approximate the gradient of the tangent. So we take the limit as $x \to a$ and define the **derivative** of f at a to be

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}.$$

This limit must exist: we need it to be the same if x approaches a from both a negative and a positive direction.

Definition 3.4 A function $f \colon \mathbb{R} \to \mathbb{R}$ is **differentiable** at $x = a$ if the limit

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

exists.

Furthermore, f is **differentiable** in the open interval (a, b) if it is differentiable at all $x \in (a, b)$.

And f is **differentiable** in the closed interval $[a, b]$ if it is differentiable over (a, b) and if the limits

$$f'(a) = \lim_{x \to a^+} \frac{f(x) - f(a)}{x - a} \quad \text{and} \quad f'(b) = \lim_{x \to b^-} \frac{f(x) - f(b)}{x - b}$$

exist.

This enables us to define the **first derivative** of a function $f \colon \mathbb{R} \to \mathbb{R}$, which we denote $f'(x)$ or $f^{(1)}(x)$ (this is sometimes called Lagrange's notation, after the Italian and French mathematician Joseph-Louis Lagrange), or $\frac{df}{dx}$ (sometimes called Leibniz' notation, after the German philosopher and mathematician Gottfried Wilhelm Leibniz).

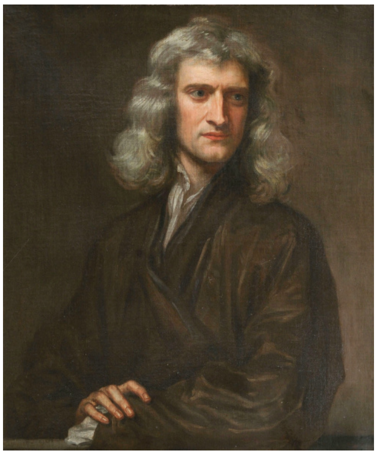
We can also define **higher derivatives**:

$$f''(x) = f^{(2)}(x) = (f'(x))' \qquad \text{or} \qquad \frac{d^2 f}{dx^2} = \frac{d}{dx}\left(\frac{df}{dx}\right)$$

and

$$\frac{d^n f}{dx^n} = \frac{d}{dx}\left(\frac{d^{n-1} f}{dx^{n-1}}\right).$$

For the nth derivative to exist, we need the first, second, third,. . . , $(n-1)$st derivatives to also exist.



Isaac Newton (1642–1727)

This raises a question: what does it mean for a function to *not* be differentiable? Well, formally speaking, the limit in Definition 3.4 would have to fail to exist at least somewhere. In particular, one or other of the left-hand and right-hand limits

$$f'(a) = \lim_{x \to a^-} \frac{f(x) - f(a)}{x - a} \qquad \text{and} \qquad f'(a) = \lim_{x \to a^+} \frac{f(x) - f(a)}{x - a}$$

might not exist. Or alternatively they might both exist but not be equal to each other.

Probably the simplest example of a non-differentiable function is the absolute value function $|x|$. The graph of this function is shown in Figure 3.4.

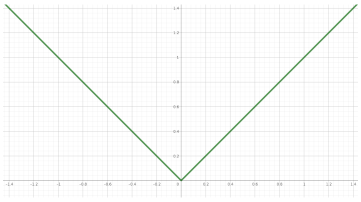


Figure 3.4: The graph of the absolute value function $|x|$

This function is not differentiable at $x = 0$, because

$$\lim_{x \to 0^-} \frac{|x| - |0|}{x - 0} = \lim_{x \to 0^-} \frac{-x}{x} = -1$$

but

$$\lim_{x \to 0^+} \frac{|x| - |0|}{x - 0} = \lim_{x \to 0^+} \frac{x}{x} = 1.$$

It is, however, differentiable at every other $x \in \mathbb{R}$.

Geometrically, non-differentiable functions tend to have some sort of kink in the graph: a point where the direction of the graph abruptly changes direction, rather than varying smoothly.

Another important detail is that differentiable functions must be continuous:

Proposition 3.5 *Let $f \colon \mathbb{R} \to \mathbb{R}$ be differentiable at $x = a$. Then f is continuous at $x = a$.*

Proof Since f is differentiable at $x = a$, the limit

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

exists. By standard algebra of limits, we have

$$f'(a) = \left(\lim_{x \to a}(f(x) - f(a))\right) \Big/ \left(\lim_{x \to a}(x - a)\right)$$

Rearranging this, we get

$$\lim_{x \to a}(f(x) - f(a)) = f'(a) \lim_{x \to a}(x - a) = 0.$$

So

$$\lim_{x \to a} f(x) - \lim_{x \to a} f(a) = 0$$

and hence

$$\lim_{x\to a} f(x) = f(a),$$

so f is continuous at $x = a$. □

However, not all continuous functions are differentiable: for example the absolute value function $|x|$ is continuous at $x = 0$, but as we've seen it isn't differentiable there.

Properties of derivatives

Suppose we have two functions $u\colon \mathbb{R} \to \mathbb{R}$ and $v\colon \mathbb{R} \to \mathbb{R}$. Then:

Linearity

(i) $\frac{d}{dx}(u+v) = \frac{du}{dx} + \frac{dv}{dx}$.

(ii) $\frac{d}{dx}(ku) = k\frac{du}{dx}$ for any constant $k \in \mathbb{R}$.

Product Rule $\frac{d}{dx}(uv) = u\frac{dv}{dx} + v\frac{du}{dx}$.

Quotient Rule $\frac{d}{dx}\left(\frac{u}{v}\right) = \frac{1}{v^2}\left(v\frac{du}{dx} - u\frac{dv}{dx}\right)$.

Chain Rule $\frac{d}{dx}u(v(x)) = \frac{du}{dv}\frac{dv}{dx}$.

The last of these, the Chain Rule, is for differentiating composite functions.

Standard derivatives

Table 3.1 lists derivatives of some standard functions.

Table 3.1: Table of standard derivatives

$f(x)$	$f'(x)$	$f(x)$	$f'(x)$	$f(x)$	$f'(x)$
k	0	$\sin(ax)$	$a\cos(ax)$	$\sin^{-1}\left(\frac{x}{a}\right)$	$\frac{1}{\sqrt{a^2-x^2}}$
x^n	nx^{n-1}	$\cos(ax)$	$-a\sin(ax)$	$\cos^{-1}\left(\frac{x}{a}\right)$	$-\frac{1}{\sqrt{a^2-x^2}}$
e^{ax}	ae^{ax}	$\tan(ax)$	$a\sec^2(ax)$	$\tan^{-1}\left(\frac{x}{a}\right)$	$\frac{a}{a^2+x^2}$
$\ln(ax)$	$\frac{1}{x}$				

Extreme points

Suppose we have a function $f\colon D \to \mathbb{R}$ (for some set $D \subseteq \mathbb{R}$), and we want to find the maximum or minimum value this function attains over its domain D.

Definition 3.6 A point $x^* \in D$ is a **global maximum** if $f(x) \leqslant f(x^*)$ for all $x \in D$. This is the point where f reaches its maximum value.

Similarly, a point $x^* \in D$ is a **global minimum** if $f(x) \geqslant f(x^*)$ for all $x \in D$. This is the point where f reaches its minimum value.

If x^* is either a maximum or a minimum, we say it is an **extreme point** or **optimal point**.

If $f(x) < f(x^*)$ for all $x \in D$, we call x^* a **strict maximum**, and if $f(x) > f(x^*)$ for all $x \in D$, we call x^* a **strict minimum**.

Sometimes there are values of x for which $f(x)$ isn't a global maximum or minimum over the entire domain D, but only in some smaller region:

Definition 3.7 A function $f\colon D \to \mathbb{R}$ has a **local maximum** (or a **local minimum**) at $x^* \in D$ if there exists some open interval $(a,b) \subseteq D$ containing x^*, such that $f(x) \leqslant f(x^*)$ (or $f(x) \geqslant f(x^*)$) for all $x \in (a,b)$.

Often in economics we will have a function, representing some relevant economic value, that we want to optimise in some way: we want to find a suitable value of the input variable that maximises or minimises the output value of the function. In effect, we want to be able to find local or global maxima or minima for the function in question.

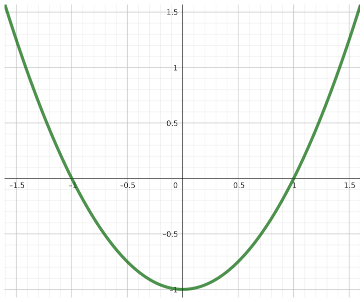


Figure 3.5: The graph of the function $f(x) = x^2 - 1$

Sometimes, finding extreme points is easy: we can just plot the graph of the function and inspect it. For example, consider the function $f(x) = x^2 - 1$. Plotting the graph (see Figure 3.5), we can see that f has a local (in fact, a global) minimum at $x = 0$, and the mininum value is $f(0) = -1$.

The key observation is that the tangent to the graph is horizontal at $x = 0$. Equivalently, $f'(x) = 2x = 0$ when $x = 0$.

Definition 3.8 A point $x^* \in (a,b) \subseteq D \subseteq \mathbb{R}$ is a **stationary point** or **turning point** of the function $f\colon D \to \mathbb{R}$ if $f'(x^*) = 0$.

This is called the **First Order Condition** (**FOC**).

Proposition 3.9 *Suppose a function $f\colon D \to \mathbb{R}$ is differentiable over an interval $(a,b) \subseteq D$. If $x^* \in (a,b)$ is an extreme point of f, then $f'(x^*) = 0$.*

The FOC is a *necessary* but not *sufficient* condition for x^* to be a local maximum or minimum. All local maxima or minima satisfy the FOC (that is, $f'(x^*) = 0$), but not all points satisfying the FOC are local maxima or minima.

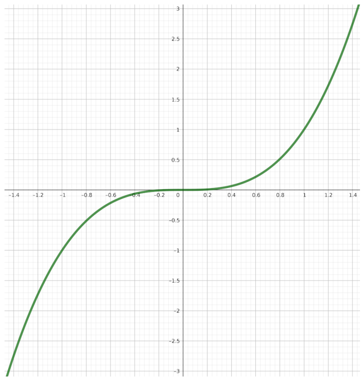


Figure 3.6: The graph of the function $f(x) = x^3$

For example, the function $f(x) = x^3$ has a stationary point at $x = 0$, since $f'(x) = 3x^2 = 0$ when $x = 0$. But looking at the graph (see Figure 3.6), we can see that this is neither a local minimum, nor a local maximum. In fact, it's what's called a **point of inflection**: the curve of the function crosses the tangent at that point.

In order to check whether a stationary point is a local minimum or maximum, we must check the **Second Order Condition** (**SOC**):

Proposition 3.10 *Suppose a function $f\colon D \to \mathbb{R}$ is twice continuously differentiable over an interval $(a,b) \subseteq D \subseteq \mathbb{R}$, and that $x^* \in (a,b)$ is a stationary point (that is, $f'(x^*) = 0$).*

(i) *If $f''(x^*) < 0$ then x^* is a local maximum.*
(ii) *If $f''(x^*) > 0$ then x^* is a local minimum.*

This is a *sufficient*, but not necessary condition to check whether x^* is a local maximum or minimum.

In general, to find local minima or maxima of functions over a given domain, we apply both the FOC and the SOC: the first of these (checking the first derivative) helps us find stationary points, while the second (checking the second derivative) helps us decide whether a given stationary point is actually a local minimum or maximum.

But sometimes this approach is inconclusive: The SOC is *sufficient* but not *necessary*. If the second derivative is zero, then our stationary point might be a point of inflection (as is the case with the function $f(x) = x^3$) but it might still be a local maximum or minimum.

This latter scenario occurs with the function $f(x) = x^4$: the FOC tells us that there is a stationary point at $x = 0$, since $f'(x) = 4x^3$ is zero there. But the SOC is inconclusive, since $f''(x) = 12x^2$ is also zero at that point.

This is another example of where we need to read a mathematical statement very carefully and think not just about what it *is* telling us, but also about what it *isn't* telling us.

If we find ourself in this situation, then one thing we can do is look at the sign of the first derivative a tiny distance either side of the stationary point. Choose some small positive value ε (for example, $\varepsilon = 0.1$), evaluate $f'(x^* - \varepsilon)$ and $f'(x^* + \varepsilon)$, and consult Table 3.2.

$f'(x^*-\varepsilon)$	$f'(x^*+\varepsilon)$	type
negative	negative	inflection
negative	positive	minimum
positive	negative	maximum
positive	positive	inflection

Table 3.2: Classifying stationary points

For example, the FOC tells us that $f(x) = x^4$ has a stationary point at $x = 0$. But checking the SOC we find that $f''(0) = 0$, so the test is inconclusive. So now we set $\varepsilon = 0.1$ and calculate $f'(-0.1) = -0.004 < 0$ and $f'(0.1) = 0.004 > 0$. Checking Table 3.2 we see that this must be a local minimum.

These are the first and second order conditions for functions of a single real variable. Things are more complicated with multivariate functions, as we will see later.

Elasticity

An issue that we run into when using derivatives is that they are *unit dependent*, which means they aren't so useful for measuring sensitivity of a function to changes in its input variable. The solution is to look at *relative* (for example, percentage) changes instead.

Consider a function $f\colon D \to \mathbb{R}$, where $D \subseteq \mathbb{R}$. Suppose that if x changes by an infinitesimal amount δx, then $f(x)$ changes by an amount δf. Then the relative change in f is $\frac{\delta f}{f}$ and the relative change in x is $\frac{\delta x}{x}$.

Hence

$$\frac{\delta f/f}{\delta x/x} = \frac{\delta f}{\delta x} \cdot \frac{x}{f}. \tag{3.1}$$

Taking the limit as $\delta x \to 0$, we have $\frac{\delta f}{\delta x} \to \frac{df}{dx}$. And the expression in (3.1) tends to $\frac{df}{dx} \cdot \frac{x}{f}$. We give this concept a special name:

Definition 3.11 The **elasticity** of f with respect to x is:

$$\mathrm{El}_f(x) = \frac{df}{dx}\frac{f}{x}.$$

Let's try an example.

Example 3.12 Let $f(x) = 3x^2 + 2x - 1$. Then $\frac{df}{dx} = 6x + 2$, and

$$\mathrm{El}_f(x) = \frac{df}{dx}\frac{x}{f} = (6x+2)\frac{x}{3x^2+2x-1} = \frac{6x^2+2x}{3x^2+2x-1}.$$

Economists sometimes use the following terminology:

Definition 3.13 Consider a function $f\colon D \to \mathbb{R}$ for some domain $D \subseteq \mathbb{R}$. Then f is said to be:

elastic at x	if	$\lvert \mathrm{El}_x(f) \rvert > 1$
unit elastic at x	if	$\lvert \mathrm{El}_x(f) \rvert = 1$
inelastic at x	if	$\lvert \mathrm{El}_x(f) \rvert < 1$
completely inelastic at x	if	$\lvert \mathrm{El}_x(f) \rvert = 0$

Logarithmic derivatives

Consider the derivative $\frac{d\ln(f)}{d\ln(x)}$. This is the derivative of $\ln(f(x))$ with respect to $\ln(x)$. Then, by applying the chain rule for derivatives,

$$\begin{aligned}
\frac{d\ln(f)}{d\ln(x)} &= \frac{d\ln(f)}{df}\cdot\frac{df}{d\ln(x)} \\
&= \frac{d\ln(f)}{df}\cdot\frac{df}{dx}\cdot\frac{dx}{d\ln(x)} \\
&= \frac{d\ln(f)}{df}\cdot\frac{df}{dx}\bigg/\frac{d\ln(x)}{dx} \\
&= \frac{1}{f}\cdot\frac{df}{dx}\bigg/\frac{1}{x} \\
&= \frac{df}{dx}\cdot\frac{x}{f} \\
&= \mathrm{El}_f(x)
\end{aligned}$$

So $\frac{d\ln(f)}{d\ln(x)} = \mathrm{El}_f(x) = \frac{df}{dx}\frac{x}{f}$ gives an alternative but equivalent formulation of elasticity that might simplify calculations in some cases.

Example 3.14 Consider an economic model

$$G_t = K_t^\alpha L_t^\alpha e^{\varepsilon_t}$$

where G_t, K_t and L_t denote (respectively) GDP, capital and labour at time t, and $\varepsilon_t \sim \mathrm{NIID}(0,1)$.

Take natural logarithms of both sides to get

$$\ln(G_t) = \alpha\ln(K_t) + \beta\ln(L_t) + \varepsilon_t$$

Differentiating with respect to $\ln(K_t)$ and $\ln(L_t)$ we see that

$$\mathrm{El}_{K_t}(G_t) = \frac{d\ln(G_t)}{d\ln(K_t)} = \alpha \qquad \mathrm{El}_{L_t}(G_t) = \frac{d\ln(G_t)}{d\ln(L_t)} = \beta$$

That is, the elasticity of GDP with respect to capital is α and the elasticity of GDP with respect to labour is β.

4 Taylor Series

Now we will study a technique for approximating functions by means of polynomials. This is useful because polynomials are relatively straightforward to deal with: they are easy to evaluate (since they only involve basic arithmetical operations) and they are easy to differentiate or integrate. And often we can derive useful insights or information from studying a quadratic or even a linear approximation to a given function.

Polynomial approximations

The approach we're going to use is to construct a polynomial whose first n derivatives agree with the first n derivatives of the given function, at some chosen point. The values of this polynomial should then be pretty close to the values of the actual function, at least for values of x close to that chosen point.

We'll illustrate this by working through a simple example.

Example 4.1 Let $f(x) = e^x$, and suppose we want a degree–n polynomial $p_n(x)$ that provides a good approximation to $f(x)$ for values of x close to 0.

Degree 0 Suppose $p_0(x) = a_0$ for some (constant) real number a_0. We want $p_0(x)$ to agree with $f(x)$ when $x = 0$. That is, $p_0(0) = a_0 = f(0) = e^0 = 1$. So we set $a_0 = 1$ and get

$$p_0(x) = 1.$$

Then, close to $x = 0$, $e^x \approx p_0(x) = 1$. This is the horizontal red line in Figure 4.1.

Degree 1 Here we want to find a linear polynomial $p_1(x) = a_0 + a_1 x$ such that $p_1(0) = f(0)$ and $p_1'(0) = f'(0)$. That is, p_1 and its first derivative agree with f and its first derivative f' at $x = 0$. Again, if $a_0 = 1$ we have $p_1(0) = f(0)$. But we want the first derivatives to agree as well. Here $p_1'(x) = a_1$ and $f'(x) = e^x$. So we require $a_1 = p_1'(0) = f'(0) = e^0 = 1$. So $a_1 = 1$, and

$$p_1(x) = 1 + x.$$

Then, close to $x = 0$ we have $e^x \approx p_1(x) = 1 + x$. We can see from the orange line in Figure 4.1 that this is a slightly better approximation, at least if $x \approx 0$.

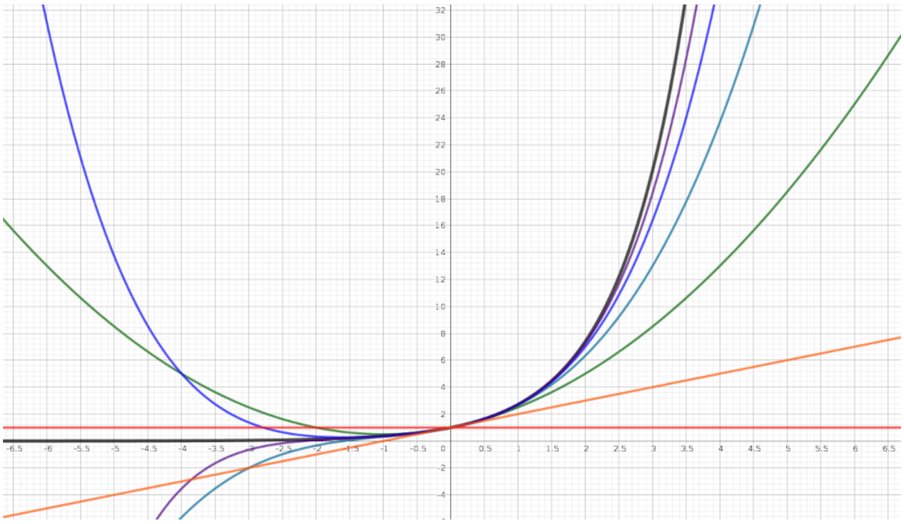


Figure 4.1: Taylor polynomial approximations $p_n(x)$ to the function $f(x) = e^x$ for $n = 0, \ldots, 5$

Degree 2 Now set $p_2(x) = a_0 + a_1 x + a_2 x^2$. Then

$$p_2''(x) = 2a_2 \qquad\qquad f''(x) = e^x$$

As before, we have $a_0 = e^0 = 1$ and $a_1 = e^0 = 1$. For the coefficient a_2 we have $2a_2 = e^0 = 1$, so $a_2 = \frac{1}{2}$, and hence

$$p_2(x) = 1 + x + \tfrac{1}{2}x^2.$$

This is the green curve in Figure 4.1 and we can see that it provides an even better approximation to the actual function.

Degree 3 Setting $p_3(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$, we find that

$$p_3'''(x) = 6a_3 \qquad\qquad f'''(x) = e^x$$

So $6a_3 = e^0 = 1$, and thus $a_3 = \frac{1}{6}$. Hence

$$p_3(x) = 1 + x + \tfrac{1}{2}x^2 + \tfrac{1}{6}x^3.$$

This is the lighter blue curve in Figure 4.1.

Degree *n* More generally, we can see that $a_n = \frac{1}{n!}$,[1] so

$$p_n(x) = 1 + x + \tfrac{1}{2}x^2 + \tfrac{1}{6}x^3 + \cdots + \tfrac{1}{n!}x^n.$$

The polynomials $p_n(x)$ in the above example provide increasingly good approximations to e^x: the higher n is, the closer $p_n(x)$ is to the actual value of e^x.

So if we just want to work with e^x numerically, up to some number of decimal places accuracy, we can set n to be sufficiently high and use $p_n(x)$ instead.

A couple of important points:

(i) We can do this with other differentiable functions, not just e^x.

(ii) Sometimes it will be more useful to approximate our chosen function close to some other value of x, instead of $x = 0$.

The full generalisation of this process is given by the following important result:

[1] Here, $n!$ denotes the **factorial** of n:

$$n! = \begin{cases} 1 & \text{if } n = 0, \\ n(n-1)(n-2)\ldots 2.1 & \text{if } n > 0. \end{cases}$$

This is defined for non-negative integers only.

Theorem 4.2 (Taylor's Theorem) *Suppose a function $f\colon D \to \mathbb{R}$ (where $D \subseteq \mathbb{R}$) is:*

(i) *differentiable (and thus continuous) in a closed interval $[a, b] \subseteq D$, up to order n, for some $n \in \mathbb{N}$, and*

(ii) *differentiable to order $(n+1)$ in the corresponding open interval (a, b).*

Then there exists some $c \in (a, b)$ such that:

$$f(b) = f(a) + (b-a)f'(a) + \tfrac{(b-a)^2}{2}f''(a) + \cdots + \tfrac{(b-a)^n}{n!}f^{(n)}(a) + \tfrac{(b-a)^{n+1}}{(n+1)!}f^{(n+1)}(c) \quad (4.1)$$

Brook Taylor (1685–1731)

Colin Maclaurin (1698–1746)

The first part of (4.1) gives us an approximation to $f(b)$ as a degree–n polynomial in $(b-a)$:

$$f(b) = f(a) + (b-a)f'(a) + \tfrac{(b-a)^2}{2}f''(a) + \cdots + \tfrac{(b-a)^n}{n!}f^{(n)}(a).$$

Setting $a = 0$ and $b = x$ this becomes

$$f(x) = f(0) + xf'(0) + \tfrac{x^2}{2}f''(0) + \cdots + \tfrac{x^n}{n!}f^{(n)}(0),$$

which is essentially the polynomial $p_n(x)$ that we constructed in Example 4.1.

The other part of (4.1),

$$\tfrac{(b-a)^{n+1}}{(n+1)!}f^{(n+1)}(c),$$

is called the **remainder term**, and is the discrepancy between the degree–n polynomial approximation to $f(b)$ and the actual value of $f(b)$. The idea is that as n increases, this should tend to zero.

The polynomial approximation $p_n(x)$ is called a **Taylor series** or **Taylor polynomial**; the special case where $a = 0$ is often called a **Maclaurin series** or **Maclaurin polynomial**.

Examples

Now we'll try a few more illustrative examples. First, we'll calculate the Maclaurin series for $\sin(x)$.

Example 4.3 Let $f(x) = \sin(x)$. This is continuous and differentiable to arbitrary order over the entirety of $\mathbb{R}$. We see that:

$$\begin{aligned}
f(x) &= \sin(x) & f(0) &= 0\\
f'(x) &= \cos(x) & f'(0) &= 1\\
f''(x) &= -\sin(x) & f''(0) &= 0\\
f'''(x) &= -\cos(x) & f'''(0) &= -1\\
&\vdots & &\vdots
\end{aligned}$$

(At this point, the pattern of derivatives repeats with period 4.)

So, the Maclaurin series (that is, the Taylor series around $x = 0$) is

$$f(x) = 0 + x + 0 - \tfrac{1}{3!}x^3 + 0 + \tfrac{1}{5!}x^5 + \cdots$$

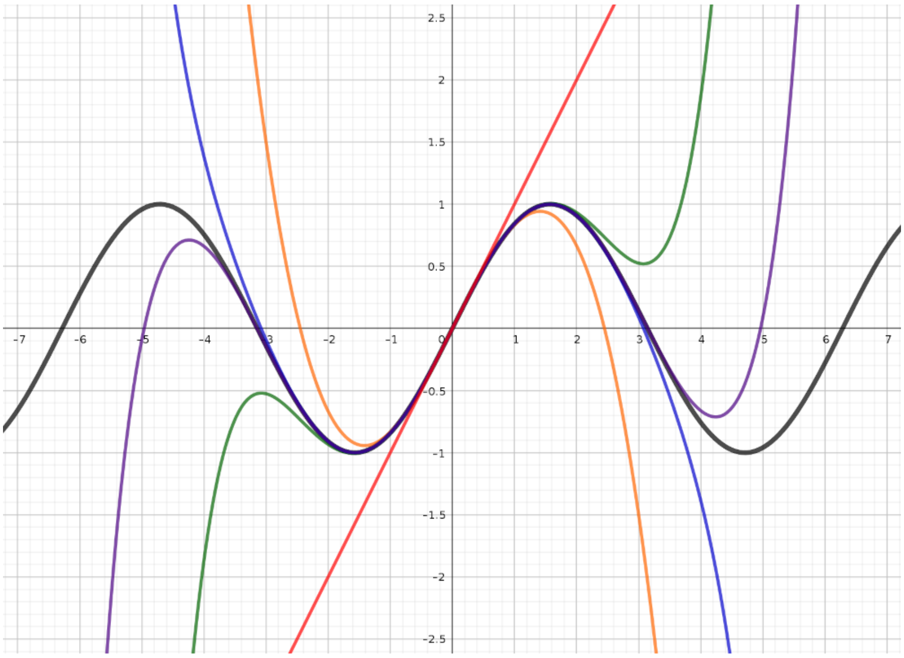


Figure 4.2: Taylor–Maclaurin polynomial approximations $p_n(x)$ to the function $f(x) = \sin(x)$, for $n = 1,3,5,7,9$

or, more generally,

$$p_n(x) = \sum_{k=0}^{n} \frac{(-1)^k x^{2k+1}}{(2k+1)!}.$$

We can see from the graph in Figure 4.2 that as n increases, $p_n(x)$ becomes an increasingly accurate approximation to $\sin(x)$.

As an exercise, try this for $f(x) = \cos(x)$.

Example 4.4 Now we will use the Maclaurin series for $f(x) = e^x$ to approximate $e = f(1)$. We already know from Example 4.1 that

$$e^x \approx \sum_{k=0}^{n} \frac{x^k}{k!}.$$

What is the smallest value of n that ensures the relative error is less than 1%?

The remainder term is

$$R_n = \frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1} = \frac{e^c x^{n+1}}{(n+1)!}.$$

The function $f(x) = e^x$ is strictly increasing on the interval $[0,1]$ (actually, it's strictly increasing everywhere in $\mathbb{R}$). We know that $a = 0$, $x = 1$ and $c \in (0,1)$, so

$$\left|\frac{e^0(1-0)^{n+1}}{(n+1)!}\right| < \left|\frac{e^c(1-0)^{n+1}}{(n+1)!}\right| < \left|\frac{e^1(1-0)^{n+1}}{(n+1)!}\right|$$

and hence

$$\left|\frac{1}{(n+1)!}\right| < \left|\frac{e^c}{(n+1)!}\right| < \left|\frac{e^x}{(n+1)!}\right|. \tag{4.2}$$

We want to find the value of n that ensures the remainder term $R_n = \frac{e^c}{(n+1)!}$ is at most 1% of the actual value of e. That is, $\frac{e^c}{(n+1)!} \leqslant 0.01e$.

The upper bound in (4.2) needs to be less than or equal to $0.01e$, because that will ensure that R_n definitely is.

So we want n such that $\frac{e}{(n+1)!} \leqslant 0.01e$; that is, $\frac{1}{(n+1)!} \leqslant 0.01$.

The smallest value of n satisfying this is $n = 4$, because then

$$\frac{1}{(n+1)!} = \frac{1}{5!} = \frac{1}{120} \leqslant 0.01.$$

Therefore, to approximate e to a relative accuracy of 1%, we need $n = 4$, and

$$p_4(x) = 1 + x + \tfrac{x^2}{2} + \tfrac{x^3}{6} + \tfrac{x^4}{24}.$$

Setting $x = 1$ this gives

$$e \approx p_4(1) = 1 + 1 + \tfrac{1}{2} + \tfrac{1}{6} + \tfrac{1}{24} = \tfrac{65}{24} = 2.7083\ldots.$$

Now let's try to calculate e with an *absolute* accuracy of 0.01; that is, accurate to two decimal places. Here we want to find n such that the upper bound on the remainder term R_n is at most 0.01 (not $0.01e$ as in the relative case).

So we want to solve $\frac{e}{(n+1)!} \leqslant 0.01$. Given that $e \approx 2.7$, this means that $(n+1)! \geqslant 100e \geqslant 270$. Here, $n = 5$ will suffice, because then $(n+1)! = 6! = 720 > 270$.

To calculate e to an accuracy of two decimal places, then, we need

$$p_5(x) = 1 + x + \tfrac{x^2}{2} + \tfrac{x^3}{6} + \tfrac{x^4}{24} + \tfrac{x^5}{120}.$$

Setting $x = 1$ this gives

$$e \approx p_5(1) = 1 + 1 + \tfrac{1}{2} + \tfrac{1}{6} + \tfrac{1}{24} + \tfrac{120}{=}\tfrac{326}{120} = 2.716\ldots.$$

Now let's try an example where we expand around some other value of x apart from zero.

Example 4.5 Find the fifth-order ($n = 5$) Taylor expansion $p_5(x)$ of $f(x) = x^2 \ln(x)$ around $a = 1$.

$$\begin{aligned}
f(x) &= x^2\ln(x) & f(1) &= 1\\
f'(x) &= 2x\ln(x) + x & f'(1) &= 1\\
f''(x) &= 2\ln(x) + 3 & f''(1) &= 3\\
f'''(x) &= \tfrac{2}{x} = 2x^{-1} & f'''(1) &= 2\\
f^{(4)}(x) &= -2x^{-2} & f^{(4)}(1) &= -2\\
f^{(5)}(x) &= 4x^{-3} & f^{(5)}(1) &= 4
\end{aligned}$$

So

$$p_5(x) = (x-1) + \tfrac{3}{2}(x-1)^2 + \tfrac{1}{3}(x-1)^3 - \tfrac{1}{12}(x-1)^4 - \tfrac{1}{30}(x-1)^5$$

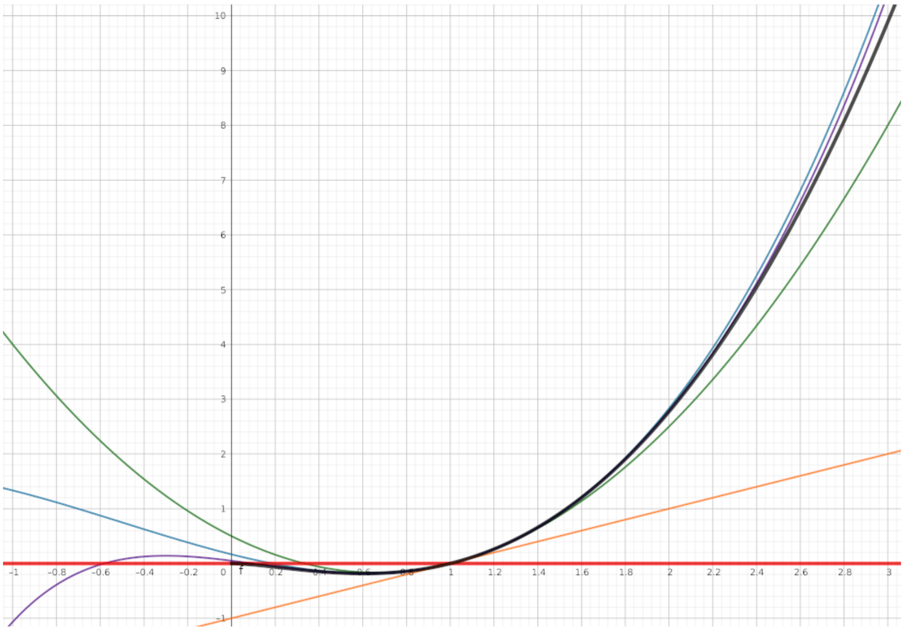


Figure 4.3: Taylor polynomials $p_n(x)$ for the function $f(x) = x^2 \ln(x)$ around x=1, for $n = 0, \ldots, 5$

This should approximate $f(x)$ close to $x = 1$, and we can see from Figure 4.3 that it does so pretty well.

If we want, we can multiply out the powers of $(x-1)$ in this expansion and regroup everything to get

$$p_5(x) = \tfrac{1}{20} - \tfrac{1}{2}x - \tfrac{1}{3}x^2 + x^3 - \tfrac{1}{4}x^4 + \tfrac{1}{30}x^5,$$

but that's just a matter of style, and the previous expression is fine for our purposes anyway.

Note that the function $f(x) = x^2 \ln(x)$ is only defined for $x > 0$, but the polynomial $p_5(x)$ is defined for all $x \in \mathbb{R}$. So we should only use $p_5(x)$ to approximate $f(x)$ for $x > 0$. We'll discuss this issue in more detail later.

New series from old

Suppose that we have a Taylor series for a function $f(x)$. We can often use this to find the Taylor series for a related function. Let

$$f(x) = \frac{1}{1+x} = \sum_{k=0}^{\infty} (-1)^k x^k = 1 - x + x^2 - x^3 + \cdots . \tag{4.3}$$

This is the infinite power series, rather than the degree–n approximation obtained by stopping after n terms. This series is only valid for $|x| < 1$, but we'll come back to that later.

Example 4.6 We can obtain the Taylor series for $f(x^2) = \frac{1}{1+x^2}$ by substituting x^2 for x in the series (4.3):

$$\frac{1}{1+x^2} = \sum_{k=0}^{\infty} (-1)^k x^{2k} = 1 - x^2 + x^4 - x^6 + \cdots .$$

(This is also only valid for $|x| < 1$, because the original series was.)

We can also differentiate Taylor series:

Example 4.7 Substituting $-x$ for x in (4.3) we get

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k = 1 + x + x^2 + x^3 + \cdots$$

for $|x| < 1$. Differentiating both sides of this, we get

$$\frac{1}{(1-x)^2} = \sum_{k=0}^{\infty} kx^{k-1} = 1 + 2x + 3x^2 + 4x^3 + \cdots .$$

Both of these series are also only valid for $|x| < 1$.

We can also integrate Taylor series:

Example 4.8 Given

$$\frac{1}{1+x^2} = \sum_{k=0}^{\infty} (-1)^k x^{2k} = 1 - x^2 + x^4 - x^6 + \cdots$$

for $|x| < 1$, we can integrate both sides to get

$$\tan^{-1}(x) = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{2k+1} = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots .$$

We can multiply, divide, add and subtract Taylor series, and substitute one into another:

Example 4.9 Recall from Examples 4.1 and 4.3 that

$$\begin{aligned} e^x &= \sum_{k=0}^{\infty} \frac{x^k}{k!} &&= 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \cdots \\ \sin(x) &= \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!} &&= x - \frac{x^3}{6} + \frac{x^5}{120} - \cdots \end{aligned}$$

Then $e^x \sin(x)$ is the product of these:

$$e^x \sin(x) = \left(1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \cdots\right)\left(x - \frac{x^3}{6} + \frac{x^5}{120} - \cdots\right).$$

We can multiply these out to get the first few terms quite easily:

$$e^x \sin(x) \approx x + x^2 + \tfrac{1}{3}x^3 - \tfrac{1}{30}x^5$$

See Figure 4.4 for an illustration.

Convergence

So far we have quietly made the following assumptions:

(i) The functions we are interested in have enough derivatives.
(ii) The polynomials converge to the actual function:
  (a) as n increases, and
  (b) for all $x \in \mathbb{R}$.

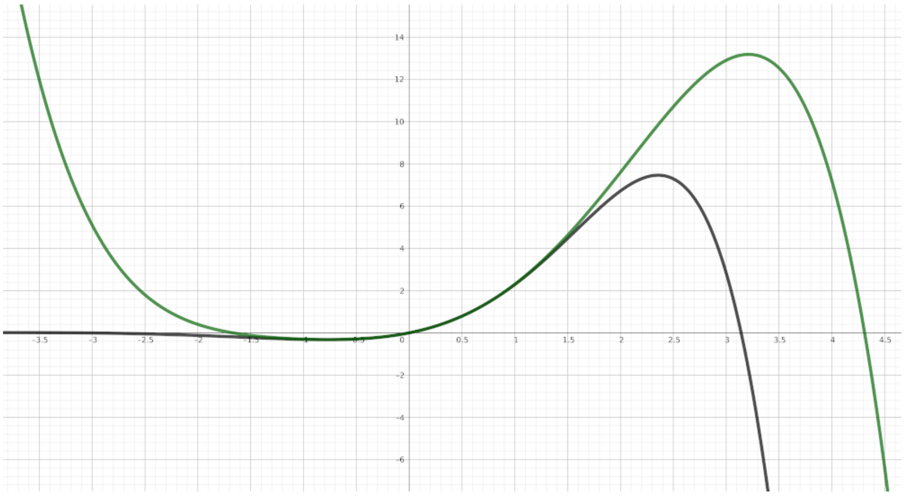


Figure 4.4: Degree–5 Taylor polynomial (green) for $f(x) = e^x \sin(x)$ (black) around $x = 0$

Neither of these need be the case. The first is something we'll need to check, but for most straightforward, familiar functions everything will be fine.

Now we'll address the second assumption. We want the polynomials $p_n(x) \to f(x)$ as $n \to \infty$, and also for $R_n \to 0$ as $n \to \infty$.

This is somewhat complicated, but we'll take a brief look at the questions involved.

For an infinite series $\sum_{k=0}^{\infty} a_k$ to **converge**, that is, sum to a finite, well-defined value, we need the sequence of **partial sums** $S_n = \sum_{k=0}^{n} a_k$ to converge to a finite limit.

We can often use the following test, first formulated by the 18th century French mathematician Jean-Baptiste d'Alembert:

Jean-Baptiste le Rond d'Alembert (1717–1783)

Theorem 4.10 (The Ratio Test) *Given a series $\sum_{k=0}^{\infty} a_k$, we consider the ratio of successive terms $\left|\frac{a_{k+1}}{a_k}\right|$. If this tends to a finite limit L as $k \to \infty$, then:*

- *if $L > 1$ the series doesn't converge,*
- *if $L < 1$ the series does converge, and*
- *if $L = 1$ the test is inconclusive.*

We'll illustrate this with an example:[2]

Example 4.11 Consider the geometric sequence $1, x, x^2, x^3, \ldots$.

The Taylor series for $\frac{1}{1+x} = 1 + x + x^2 + \cdots = \sum_{k=0}^{\infty} x^k$.

Applying the Ratio Test: $\left|\frac{a_{k+1}}{a_k}\right| = \left|\frac{x^{k+1}}{x^k}\right| = |x|$. This ratio tends to the limit $|x|$ as $k \to \infty$.

Now, if $L = |x| > 1$ then by the Ratio Test the series doesn't converge. But if $L = |x| < 1$ then the series does converge. (The Ratio Test is inconclusive for the case $|x| = 1$, and we would have to use other techniques here.)

So the series $\sum_{k=0}^{\infty} x^k$ converges to a finite value (in fact, to $\frac{1}{1+x}$) for $|x| < 1$; that is, when $-1 < x < 1$. This is called the **interval of convergence** for the series.

We would also need to check that $R_n \to 0$ as $n \to \infty$, at least for $|x| < 1$. If so, then this all works.

[2] As an exercise, try this process to confirm that the Maclaurin series for e^x derived in Example 4.1 actually is valid for all $x \in \mathbb{R}$.

5 *Analytical Theorems*

IN THIS CHAPTER we will state some important analytical concepts, terminology and theorems.

Open sets in $\mathbb{R}^n$

Suppose we are working in n–dimensional Euclidean space $\mathbb{R}^n$; for example, $\mathbb{R}$, $\mathbb{R}^2$, $\mathbb{R}^3$, and so forth.

We want to take the the concept of an open interval (a, b) in $\mathbb{R}$, and generalise it to $\mathbb{R}^n$. We do it as follows:

Definition 5.1 A **ball** in $\mathbb{R}^n$, centred on a point $\mathbf{x}$, with radius r, is denoted

$$B_r(\mathbf{x}) = \{\mathbf{v} \in \mathbb{R}^n : \|\mathbf{v} - \mathbf{x}\| \leqslant r\}.$$[1]

Geometrically, this is all the points in $\mathbb{R}^n$ that are a distance at most r from the point $\mathbf{x}$.

Definition 5.2 A set $S \subseteq \mathbb{R}^n$ is **open** if and only if every point $\mathbf{x} \in S$ is contained within a ball $B_r(\mathbf{x})$ for some $r > 0$ such that $B_r(\mathbf{x}) \subseteq S$.

Informally, an open set doesn't contain any of its boundary. So in $\mathbb{R}$, an open interval (a, b) doesn't include its endpoints a and b.

The idea is that for a set S to qualify as open, we should be able to put a ball of nonzero radius, however small, around any point $\mathbf{x} \in S$, that is also entirely within the set S. Think about this, and convince yourself that the only points you can't do this for are on the boundary of the set S.

Definition 5.3 A **neighbourhood** of a point $\mathbf{x} \in \mathbb{R}^n$ is an open set S containing $\mathbf{x}$.

Definition 5.4 The **complement** of a set $S \subseteq \mathbb{R}^n$ is the difference

$$S' = \mathbb{R}^n \setminus S = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \notin S\}.$$

That is, all the points in $\mathbb{R}^n$ that *aren't* in S.

For example, if $S = [a, b] \subset \mathbb{R}$, the complement

$$S' = \{x \in \mathbb{R} : x \notin [a, b]\} = (-\infty, a) \cup (b, \infty).$$

Definition 5.5 We say that a set B is **closed** if and only if its complement is open.

[1] Here $\|\cdot\|$ denotes the usual **norm** in $\mathbb{R}^n$; see the Linear Algebra section for more details.

However, some sets are neither open nor closed. Informally, a closed set contains all of its boundary. So in $\mathbb{R}$, a closed interval $[a, b]$ includes its endpoints a and b. As defined above, the ball $B_r(\mathbf{x})$ is closed.

Definition 5.6 A set $B \subseteq \mathbb{R}^n$ is **bounded** if and only if there exists some $r > 0$ and $\mathbf{x} \in \mathbb{R}^n$ such that $B \subseteq B_r(\mathbf{x})$. That is, B can be contained within some finite-radius ball.

Definition 5.7 If a set $B \subseteq \mathbb{R}^n$ is bounded and closed, we say it is **compact**.[2]

Important theorems

In this section, we state a few important theorems from real analysis. The first of these is the Intermediate Value Theorem, first proved in 1817 by the Bohemian mathematician and priest Bernard Bolzano, although an earlier form was postulated in the 5th century BCE by the Greek mathematician Bryson of Heraclea:

Bernard Bolzano (1781–1848)

Theorem 5.8 (Intermediate Value Theorem) *Let $f : D \to \mathbb{R}$, where $D \subseteq \mathbb{R}$. If f is continuous over a closed interval $[a, b]$, and if $f(a) \neq f(b)$, then for any $y \in (f(a), f(b))$ there exists $c \in (a, b)$ such that $f(c) = y$.*

This says that a value c exists for which $f(c)$ matches any level between $f(a)$ and $f(b)$. See Figure 5.1 for an illustration.

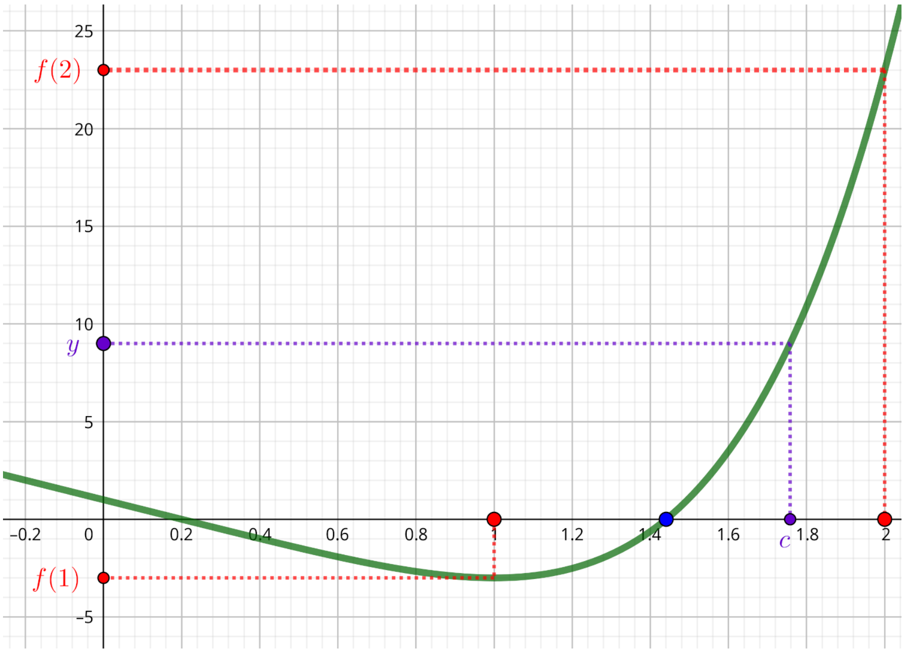


Figure 5.1: Illustration of the Intermediate Value Theorem over the interval $(1, 2)$

There is a simpler version in which we set the required level $y = 0$:

Corollary 5.9 *Let $f : D \to \mathbb{R}$, where $D \subseteq \mathbb{R}$. If f is continuous over a closed interval $[a, b]$ and if $f(a)$ and $f(b)$ have different signs (that is, one is positive and the other negative) then there exists $c \in (a, b)$ such that $f(c) = 0$.*

Geometrically, this result makes the (intuitively obvious) statement

[2] The more general framework for defining open, closed, bounded and compact sets is complicated, and these definitions will suffice for our purposes. For more details, see: W. A. Sutherland, *Introduction to Metric and Topological Spaces*, Clarendon Press, Oxford (1975).

that if we draw the graph of a continuous function (that is, one with no breaks in it) then if the graph is below the horizontal axis somewhere, and above it somewhere else, there must be a point somewhere in between where the graph crosses the horizontal axis. This theorem helps confirm that our formalised, precise definition of continuity agrees with our intuitive understanding.

We can use the Intermediate Value Theorem to prove the following important result, which says that every continuous function from a closed interval to itself must fix at least one point in place.

Theorem 5.10 (Brouwer's Fixed Point Theorem) *If $f\colon [a,b] \to [a,b]$ is continuous, then there exists $c \in [a,b]$ such that $f(c) = c$. We call c a* ***fixed point*** *of f.*

Luitzen Egbertus Jan Brouwer (1881–1966)

Proof Let $g(x) = f(x) - x$. This is continuous on $[a,b]$ because both f and x are. Now consider $g(a)$ and $g(b)$. Since $a \leqslant f(x) \leqslant b$ we have $f(a) \geqslant a$ and $f(b) \leqslant b$. Hence $g(a) = f(a) - a \geqslant 0$ and $g(b) = f(b) - b \leqslant 0$. There are three cases to consider:

Case 1 If $f(a) = a$ then a is a fixed point of f in $[a,b]$.

Case 2 If $f(b) = b$ then b is a fixed point of f in $[a,b]$.

Case 3 If $f(a) > a$ and $f(b) < b$ then we have $g(a) > 0$ and $g(b) < 0$. By Corollary 5.9, there must exist $c \in (a,b)$ such that $g(c) = 0$, and hence $f(c) - c = 0$, so $f(c) = c$ as required. □

The fixed point needn't be unique: there can be more than one. This theorem is required to prove the existence of Nash equilibria in game theory. This result also holds if we replace $[a,b]$ with any compact set in $\mathbb{R}^n$.[3]

The next result says that continuous functions are bounded over closed intervals; that is, they have a finite maximum and minimum value over that interval. And, importantly, those bounds are **attained**: there exist specific values of x in the interval that map to the maximum and minimum values.

Theorem 5.11 (Extreme Value Theorem) *Let $f\colon D \to \mathbb{R}$, where $D \subseteq \mathbb{R}$. If f is continuous over some closed interval $[a,b] \subseteq D$, then f is bounded over $[a,b]$, and attains its bounds. That is, there exist $m, M \in \mathbb{R}$ such that $m \leqslant f(x) \leqslant M$ for all $x \in [a,b]$, and furthermore there exist $c, d \in [a,b]$ such that $f(c) = m$ and $f(d) = M$.*

The following theorem is attributed to the French mathematician Michel Rolle, who proved it for polynomial functions in 1691. A more general and analytically rigorous version was proved by the French mathematician Augustin-Louis Cauchy in 1823. See Figure 5.2 for an illustration.

Michel Rolle (1652–1719)

Theorem 5.12 (Rolle's Theorem) *Let $f\colon D \to \mathbb{R}$, where $D \subseteq \mathbb{R}$. If f is continous over some closed interval $[a,b]$, differentiable over the corresponding open interval (a,b), and if $f(a) = f(b)$, then there exists $c \in (a,b)$ such that $f'(c) = 0$.*

This theorem says that if a continuous and differentiable function f has two values a and b for which $f(a) = f(b)$, then there must exist at least one stationary point between them. Intuitively, this makes sense: the only way we could avoid a stationary point (that

[3] The 2–dimensional analogue can be proved by using the **fundamental group** $\pi_1(X)$ in algebraic topology. One consequence of the 2–dimensional version is that if we take a map of the UK and put it on the ground anywhere in Britain, then there is a point on the map that is exactly over the point on the ground that it represents.

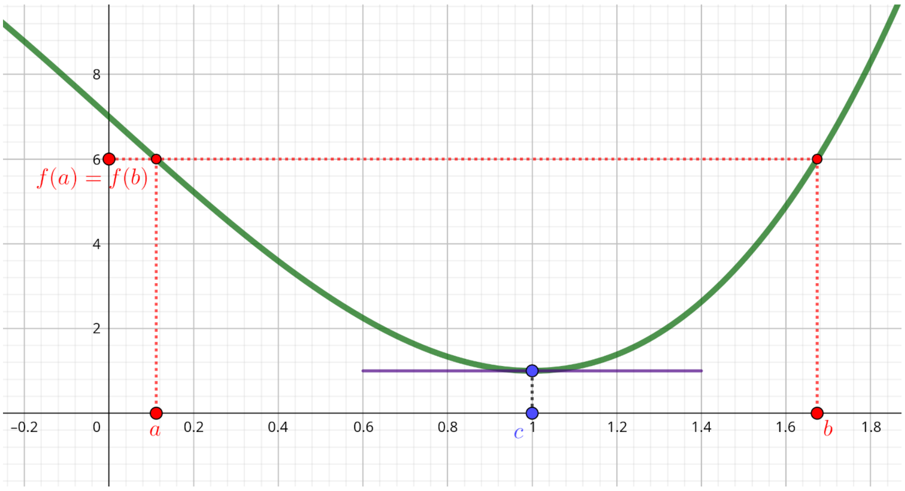


Figure 5.2: Illustration of Rolle's Theorem

is, a maximum, minimum or point of inflection) is if we allowed f to have a discontinuity, or fail to be differentiable somewhere.

Augustin-Louis Cauchy (1789–1857)

Theorem 5.13 (Mean Value Theorem) *Let $f: D \to \mathbb{R}$, where $D \subseteq \mathbb{R}$. If f is continuous over some closed interval $[a, b]$, and differentiable over the corresponding open interval (a, b), then there exists $c \in (a, b)$ such that*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Proof Consider the function $F: \mathbb{R} \to \mathbb{R}$ given by

$$F(x) = (b-a)(f(b)-f(x)) - (b-x)(f(b)-f(a)).$$

Note that $F(a) = F(b) = 0$. And since f is continuous on $[a, b]$ and differentiable on (a, b), so is F. Differentiating F, we get

$$F'(x) = -f'(x)(b-a) + f(b) - f(a).$$

Applying Rolle's Theorem to F over the interval $[a, b]$, there exists some $c \in (a, b)$ such that $F'(c) = 0$, and hence

$$f(b) - f(a) - f'(c)(b-a) = 0.$$

Rearranging this, we get

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

as claimed. □

The Mean Value Theorem says that if f is continuous and differentiable over some interval, then there exists at least one point in that interval where the first derivative is equal to the "average" of the first derivative. This is illustrated in Figure 5.3, where the blue gradient at c is parallel to the red chord between the points $(a, f(a))$ and $(b, f(b))$.

Corollary 5.14 *Let $f: D \to \mathbb{R}$, where $D \subseteq \mathbb{R}$. If f is continuous over some closed interval $[a, b]$ and differentiable over the corresponding open interval (a, b), then there exists some $c \in (a, b)$ such that*

$$f(b) = f(a) + (b-a)f'(c).$$

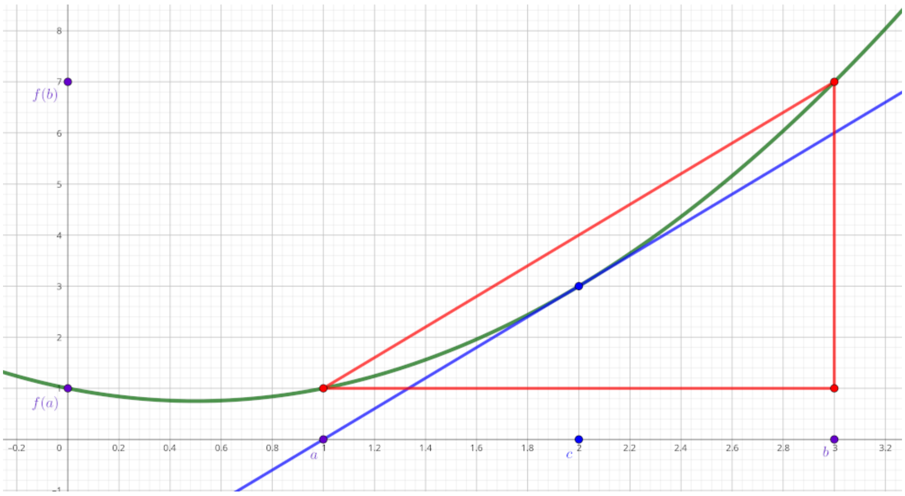


Figure 5.3: Illustration of the Mean Value Theorem

This corollary (which can be proved by a simple rearrangement of the statement of the Mean Value Theorem) is effectively the $n = 1$ case of Taylor's Theorem,[4] and so we can regard Taylor's Theorem as a generalisation of the Mean Value Theorem.

[4] Theorem 4.2, page 21.

6 *Single Difference Equations*

Many quantities of interest in economics change over time: for example, investment, GDP, unemployment, etc. We often know how the change of such a quantity relates to its current value. Sometimes we can write down this relationship as a mathematical expression, and we'd like to be able to use such a model to predict the value of a given quantity at some point in the future: for example, in five years' time.

In this chapter we will study **difference equations** (sometimes called **recurrence relations**). These are analogous to differential equations (which we will cover later on).

	Difference equations	Differential equations
Value depends on:	previous values	derivatives
Measure:	at fixed time intervals	over continuous time
Solution:	discrete function of t	continuous function of t
discrete or continuous:	discrete	continuous

Table 6.1: Analogies between difference equations and differential equations

Definitions and examples

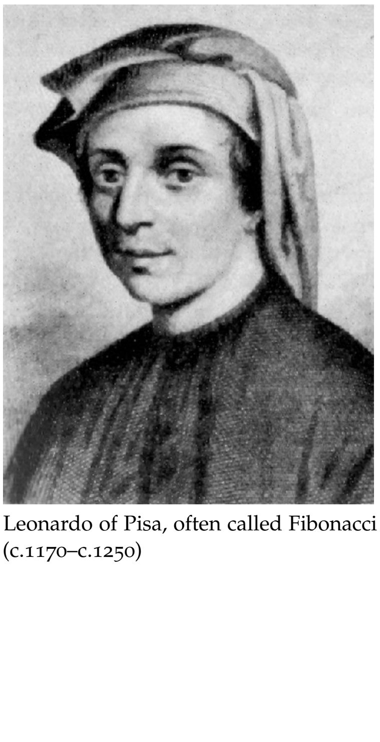
A difference equation is **linear** if the current value depends on a linear function of its previous value(s). Otherwise, it is **nonlinear**.

Example 6.1 The **Fibonacci sequence** (named after the Italian mathematician Leonardo of Pisa, nicknamed Fibonacci) is a linear difference equation:

$$x_{t+1} = x_t + x_{t-1}$$

The **logistic equation**, which turns up in population dynamics, is a nonlinear difference equation:

$$x_{t+1} = rx_t(1 - x_t) = rx_t - rx_t^2$$


Leonardo of Pisa, often called Fibonacci (c.1170–c.1250)

Often, we want to solve a difference equation given some **initial value** x_0. For example, the usual starting point for the Fibonacci sequence is to set x_0 and x_1 both equal to 1, and then use the formula to find x_2 and so on.

The **order** of a difference equation is the number of previous values the next value depends on. So the logistic equation is a **first order** equation because x_{t+1} only depends on the previous value x_t. The Fibonacci equation is **second order**, because x_{t+1} depends on the previous *two* values x_t and x_{t-1}.

We say that a difference equation is **homogeneous** if it has no terms that don't depend on previous values. The Fibonacci and logistic equations are both homogeneous in this sense. However, the equation

$$x_{t+1} = 2x_t + 7x_{t-1} + 4$$

is **inhomogeneous** because it has a term (the constant 4) that doesn't depend on x_t, x_{t-1} and so on.

Solving first-order linear difference equations

Now we will work through a full solution for the general first-order linear difference equation

$$x_{t+1} = ax_t + b \tag{6.1}$$

where a and b are real constant parameters, and the initial value x_0 is a real number.

We need to consider two cases:

Case 1 If $a \neq 1$, we have

$$\begin{aligned} x_1 &= ax_0 + b \\ x_2 &= ax_1 + b = a(ax_0 + b) + b & &= a^2x_0 + ab + b \\ x_3 &= ax_2 + b = a(a^2x_0 + ab + b) + b & &= a^3x_0 + a^2b + ab + b \\ &\vdots \\ x_t &= a^t x_0 + (a^{t-1} + a^{t-2} + \cdots + a + 1)b \\ &= a^t x_0 + \frac{1-a^t}{1-a} b \end{aligned}$$

The last step uses the fact that

$$a^{t-1} + a^{t-2} + \cdots + a + 1 = \frac{1-a^t}{1-a} = \frac{a^t - 1}{a - 1}$$

for all $a \neq 1$.[1]

Case 2 If $a = 1$, then we have

$$\begin{aligned} x_1 &= x_0 + b \\ x_2 &= x_1 + b = (x_0 + b) + b & &= x_0 + 2b \\ x_3 &= x_2 + b = (x_0 + 2b) + b & &= x_0 + 3b \\ &\vdots \\ x_t &= x_0 + tb \end{aligned}$$

So we can write the solution to (6.1) as:

$$x_t = \begin{cases} x_0 + tb & \text{if } a = 1 \\ a^t x_0 + \frac{1-a^t}{1-a} b & \text{if } a \neq 1 \end{cases} \tag{6.2}$$

This is the **general solution** to the original problem (6.1). Think of this as a family of all the sequences that satisfy the given difference equation.

[1] Check this by expanding

$$(1-a)(a^{t-1} + \cdots + a + 1)$$

and verifying that almost all the terms cancel, leaving $1 - a^t$.

The particular solution

The following is also a solution to (6.1):

$$\overline{x}_t = \begin{cases} bt & \text{if } a = 1 \\ \frac{b}{1-a} & \text{if } a \neq 1 \end{cases} \tag{6.3}$$

We can verify this by substituting both expressions into (6.1):

Case 1 If $a \neq 1$ then the two sides of (6.1) give:

$$\begin{aligned} &\text{LHS:} && \overline{x}_{t+1} = \frac{b}{1-a} \\ &\text{RHS:} && a\overline{x}_t + b = a\left(\frac{b}{1-a}\right) + b = \frac{b}{1-a} \end{aligned}$$

These are equal, so this is a valid solution..

Case 2 If $a = 1$ then the two sides of (6.1) are:

$$\begin{aligned} &\text{LHS:} && \overline{x}_{t+1} = b(t+1) \\ &\text{RHS:} && \overline{x}_t + b = bt + b = b(t+1) \end{aligned}$$

These are also equal, so this is also a valid solution when $a = 1$.

The solution in the case $a \neq 1$ is constant: it is the same for any value of t. We call this a **steady state** solution, and will often denote it by x^*.

Homogeneous equations

Suppose we have two different solutions to (6.1): a particular solution $\overline{x}_t$ and another solution x_t. Then the sequence of their differences, $y_t = x_t - \overline{x}_t$ satisfies the related homogeneous equation

$$y_{t+1} = ay_t \tag{6.4}$$

since if

$$x_{t+1} = ax_t + b \qquad \overline{x}_{t+1} = a\overline{x}_t + b$$

we have

$$y_{t+1} = x_{t+1} - \overline{x}_{t+1} = (ax_t + b) - (a\overline{x}_t + b) = a(x_t - \overline{x}_t) = ay_t.$$

This works because (6.1) is a linear equation.

Rearranging $y_t = x_t - \overline{x}_t$ we get

$$x_t = y_t + \overline{x}_t.$$

Here, $\overline{x}_t$ is a particular solution to the original inhomogeneous equation (6.1), while y_t is the general solution to the related homogeneous equation (6.4).

So, to solve the original (inhomogeneous) equation (6.1) one method is to find a **particular solution** (such as a steady state solution, if one exists), then adding it to the general solution for the related homogeneous equation (6.4).

To solve the homogeneous equation (6.4), we note that this is the same as (6.1), but with $b = 0$. The general solution is thus

$$y_t = \begin{cases} y_0 & \text{if } a = 1 \\ a^t y_0 & \text{if } a \neq 1 \end{cases}. \tag{6.5}$$

Now let $c = y_0$. Then the full solution of the original (inhomogeneous) equation (6.1) is the general solution of the homogeneous equation (6.4) plus a particular solution of (6.1):

$$x_t = \begin{cases} c + bt & \text{if } a = 1 \\ ca^t + \frac{b}{1-a} & \text{if } a \neq 1 \end{cases} \tag{6.6}$$

where

$$c = \begin{cases} x_0 & \text{if } a = 1 \\ x_0 - \frac{b}{1-a} & \text{if } a \neq 1 \end{cases} \tag{6.7}$$

That is,

$$x_t = \begin{cases} x_0 + bt & \text{if } a = 1 \\ \left(x_0 - \frac{b}{1-a}\right)a^t + \frac{b}{1-a} & \text{if } a \neq 1 \end{cases} \tag{6.8}$$

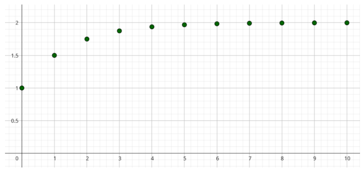


Figure 6.1: The first several values of the difference equation in Example 6.2

Examples

To make sense of all this, we'll look at some examples.

Example 6.2 Consider the equation

$$x_{t+1} = \tfrac{1}{2}x_t + 1$$

with $x_0 = 1$. This gives the sequence $1, \frac{3}{2}, \frac{7}{4}, \frac{15}{8}, \ldots$

First we find the particular solution, using (6.3) with $a = \frac{1}{2} \neq 1$. This is

$$\overline{x}_t = \frac{b}{1-a} = \frac{1}{1-1/2} = 2.$$

Now we solve the homogenous equation

$$y_{t+1} = \tfrac{1}{2}y_t$$

Using (6.5) we have

$$y_t = a^t y_0 = \left(\tfrac{1}{2}\right)^t y_0$$

where $y_0 = x_0 - \frac{b}{1-a} = 1 - 2 = -1$. So $y_t = -\left(\frac{1}{2}\right)^t$ and the general solution for the inhomogeneous equation is therefore

$$x_t = \overline{x}_t + y_t = -\left(\tfrac{1}{2}\right)^t + 2.$$

As $t \to \infty$, we see that $\left(\frac{1}{2}\right) \to 0$ and hence $x_t \to 2$. The solution of this equation is therefore converging to a steady state solution $x^* = 2$, as seen in Figure 6.1.

Now we consider the same equation with a different initial value:

Example 6.3 Again, consider

$$x_{t+1} = \tfrac{1}{2}x_t + 1$$

but this time with $x_0 = 2$. Again, we get the particular solution $\overline{x}_t = \frac{b}{1-a} = \frac{1}{1-1/2} = 2$.

The general solution to the homogeneous equation

$$y_{t+1} = \tfrac{1}{2}y_t$$

is

$$y_t = ca^t = c\left(\tfrac{1}{2}\right)^t$$

where $c = x_0 - \frac{b}{1-a} = 2 - 2 = 0$. So $y_t = 0$ and hence the solution to the original inhomogeneous equation is

$$x_t = y_t + \overline{x}_t = \overline{x}_t = 2 = x^*,$$

the steady state solution.

Example 6.4 Now consider the equation

$$x_{t+1} = -\tfrac{1}{2}x_t - 3$$

with $x_0 = 0$. Since $a = -\frac{1}{2} \neq 1$ we have

$$\overline{x}_t = \frac{b}{1-a} = \frac{-3}{1+1/2} = -2.$$

This is the steady state solution. The homogeneous equation

$$y_{t+1} = -\tfrac{1}{2}y_t$$

has solution

$$y_t = ca^t = c\left(-\tfrac{1}{2}\right)^t$$

where $c = x_0 - \frac{b}{1-a} = 2$. Hence

$$x_t = y_t + \overline{x}_t = 2\left(\tfrac{1}{2}\right)^t - 2.$$

As $t \to \infty$, this tends to the steady state solution $x^* = -2$, as shown in Figure 6.2

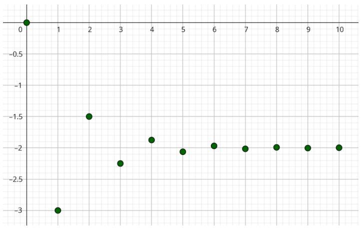


Figure 6.2: The first several values of the difference equation in Example 6.4

Example 6.5 Consider

$$x_{t+1} = -2x_t - 3$$

with $x_0 = -\frac{1}{2}$. Again, $a \neq 1$ so we have $\overline{x}_t = \frac{b}{1-a} = \frac{-3}{1+2} = -1$. So $\overline{x}_t = x^* = -1$ is the steady state solution.

Solving $y_{t+1} = -2y_t$ gives $y_t = ca^t = c(-2)^t$, where $c = x_0 - \frac{b}{1-a} = -\frac{1}{2} + 1 = \frac{1}{2}$. So $y_t = \frac{1}{2}(-2)^t$, and the full solution is

$$x_t = y_t + \overline{x}_t = \tfrac{1}{2}(-2)^t = 1$$

which diverges as $t \to \infty$. See Figure 6.3

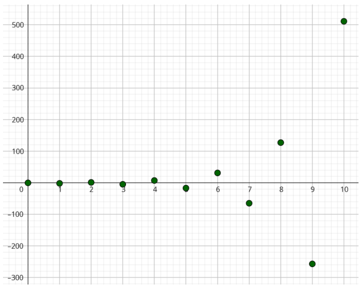


Figure 6.3: The first several values of the difference equation in Example 6.5

Stability of solutions

In the examples just discussed, we saw a few different types of long-term behaviour as $t \to \infty$: in Examples 6.2 and 6.4, the values of x_t converged to a finite limit as $t \to \infty$, while in Example 6.3, x_t was a constant, steady state sequence, and in Example 6.5 the values of x_t diverged.

If x_t converges to a finite limit as $t \to \infty$, we say the equation is **stable**, otherwise it is **unstable**.

As we saw earlier, the equation

$$x_{t+1} = ax_t + b$$

has general solution

$$x_t = \begin{cases} c + bt & \text{if } a = 1 \\ ca^t + \frac{b}{1-a} & \text{if } a \neq 1 \end{cases}$$

where

$$c = \begin{cases} x_0 & \text{if } a = 1 \\ x_0 - \frac{b}{1-a} & \text{if } a \neq 1 \end{cases}$$

The stability of this equation depends on the various parameters:

- If $|a| < 1$ we have $\lim_{t\to\infty}\left(ca^t + \frac{b}{1-a}\right) = \frac{b}{1-a}$, so the equation is stable.
- If $|a| > 1$ we have $\left(ca^t + \frac{b}{1-a}\right) \to \pm\infty$ as $t \to \infty$. This equation is unstable.
- If $a = 1$ and $b = 0$, the limit is $\lim_{t\to\infty}(c + bt) = c$, so this equation is stable.
- If $a = 1$ and $b \neq 0$, then $(c + bt) \to \pm\infty$, so the equation is unstable.
- If $a = -1$ then $x_t = c(-1)^t + \frac{b}{1-a}$, which alternates between $\frac{b}{1-a} \pm c$, depending on whether t is even or odd. This equation is **oscillatory** and unstable.

Summary

To summarise, the general method for solving first-order linear difference equations is as follows:

- Find a particular solution to the original inhomogeneous problem (for example, a steady state solution, if one exists).
- Find a general solution to the related homogeneous problem.
- Add them together to get the general solution to the original inhomogeneous problem.

7 Concavity and Convexity

So far we have concentrated on functions of a single input variable. Now we want to start thinking about functions with more than one input variable: **multivariate** or **multivariable functions**. These are functions of the form $f\colon \mathbb{R}^n \to \mathbb{R}$, where $\mathbb{R}^n$ denotes n–dimensional, real, Euclidean space. We'll study this more in linear algebra, but for now the following definition will do:

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

This is the set of **ordered n–tuples** of real numbers. That is, ordered lists of n real numbers, which we will often regard as representing points in n–dimensional space.[1] For the moment, think of $\mathbb{R}^1$ as the real number line, $\mathbb{R}^2$ as a flat, infinite plane, and $\mathbb{R}^3$ as three-dimensional space.

So a multivariate function $f\colon \mathbb{R}^n \to \mathbb{R}$ maps n real input variables to one real value. For our purposes, we will often consider functions $f\colon D \to \mathbb{R}$ defined on some domain $D \subseteq \mathbb{R}^n$. We will usually require D to be a **convex set**.

Convex sets

The idea is that we want to work with regions of $\mathbb{R}^n$ such that the straight line between any two points also lies inside that region.

Definition 7.1 Given two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we define the **section** between them to be the set

$$I(\mathbf{x}, \mathbf{y}) = \{\mathbf{z} = \alpha\mathbf{x} + (1-\alpha)\mathbf{y} : \alpha \in [0, 1]\}$$

Geometrically, this is the straight line joining the points $\mathbf{x}$ and $\mathbf{y}$.[2]

When $\alpha = 0$ we have $\mathbf{z} = 0\mathbf{x} + (1-0)\mathbf{y} = \mathbf{y}$, and when $\alpha = 1$ we have $\mathbf{z} = 1\mathbf{x} + (1-1)\mathbf{y} = \mathbf{x}$. As α ranges between 0 and 1, the point $\mathbf{z}$ ranges between $\mathbf{y}$ and $\mathbf{x}$. The set $I(\mathbf{x}, \mathbf{y})$ consists of all points of this form, that is, the straight line segment joining $\mathbf{x}$ and $\mathbf{y}$.

Definition 7.2 A set $D \subseteq \mathbb{R}^n$ is **convex** if, for all $\mathbf{x}, \mathbf{y} \in D$ the section $I(\mathbf{x}, \mathbf{y}) \subseteq D$.

Informally, a set is convex if the straight line between any two points doesn't go outside the set.[3]

[1] This is difficult to geometrically visualise for $n > 3$, but if you're interested in trying, a good place to start is: Rudy Rucker, *The Fourth Dimension: Toward a Geometry of Higher Reality*, Dover (2014).

[2] In printed notes and books, we usually denote vectors or points in $\mathbb{R}^n$ by **bold** letters. In handwritten notes, we will usually underline them, for example $\underline{x}$, $\underline{y}$, etc.

[3] Note that the empty set $\emptyset$ is convex, because it satisfies the definition trivially.

Concave and convex functions

We've just defined what we mean for a *set* to be **convex**. Slightly confusingly, we're about to use the same word in a different sense, to describe a particular sort of function.[4]

We'll state the definition and then discuss what it means.

Definition 7.3 Let $D \subseteq \mathbb{R}^n$ be a convex set. A function $f\colon D \to \mathbb{R}$ is **concave** if, for all $\mathbf{x}, \mathbf{y} \in D$, and $\alpha \in (0,1)$, we have

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \geqslant \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y})$$

and **strictly concave** if for all $\mathbf{x}, \mathbf{y} \in D$, and $\alpha \in (0,1)$, we have

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) > \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y})$$

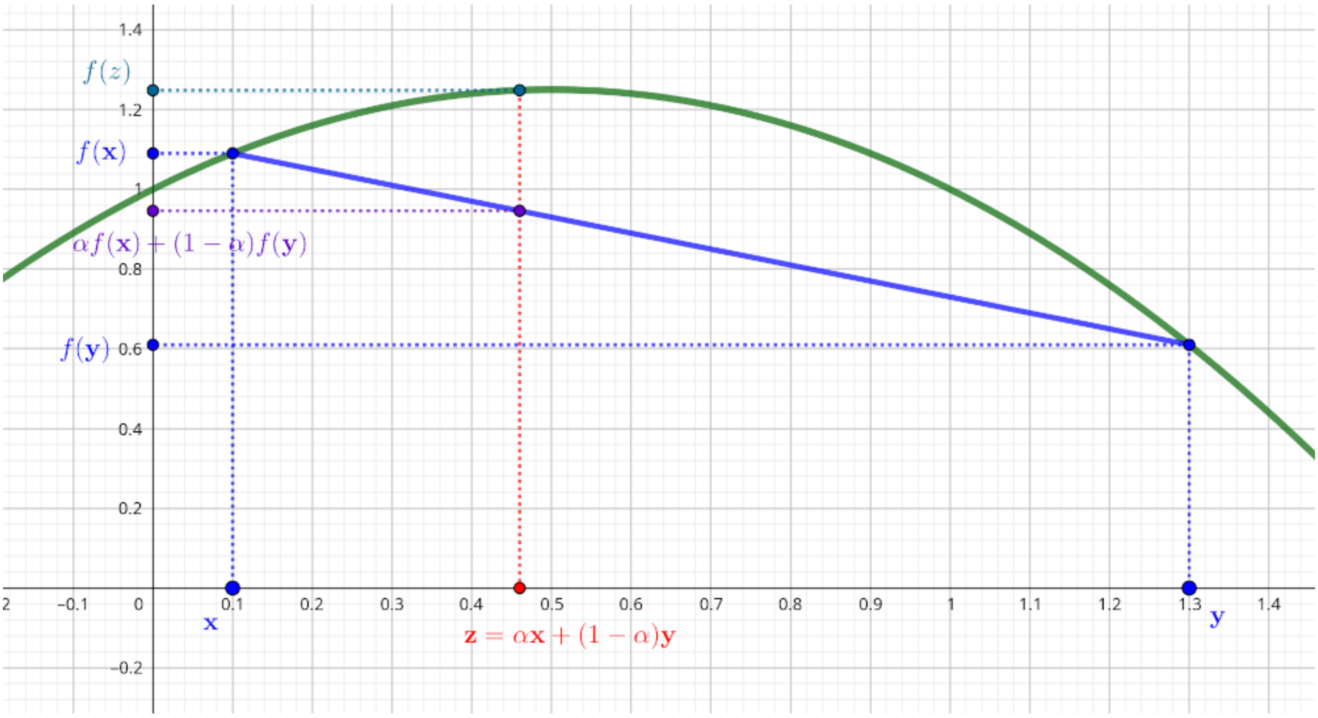


Figure 7.1: Illustration of the condition for a function to be concave

Figure 7.1 shows an illustration of this definition. Informally, for a function to be concave, we want the graph of the function to be above, or tangent to, the chord between any two points on the graph. (For strict concavity, we don't allow tangency.)

We need the domain D to be a convex set, to ensure that the midpoints $\mathbf{z} = \alpha\mathbf{x} + (1-\alpha)\mathbf{y}$ belong to D, for any $\mathbf{x}, \mathbf{y} \in D$. Otherwise, $f(\mathbf{z})$ might not be defined.

We can flip this upside down, replacing $\geqslant$ and $>$ with $\leqslant$ and $<$ to define **convex** functions. These are similar, but upside down.[5]

Definition 7.4 Let $D \subseteq \mathbb{R}^n$ be a convex set. A function $f\colon D \to \mathbb{R}$ is **convex** if, for all $\mathbf{x}, \mathbf{y} \in D$, and $\alpha \in (0,1)$, we have

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leqslant \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y})$$

and **strictly convex** if for all $\mathbf{x}, \mathbf{y} \in D$, and $\alpha \in (0,1)$, we have

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) < \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y})$$

[4] Sometimes this happens in mathematics: the same word will be used to mean two different (but possibly related) things. Soon, when we study linear algebra, we'll use the word **orthogonal** to refer to particular sorts of vectors, and then in a slightly different sense to refer to matrices with a specific property. This is a bit awkward, but hopefully we'll all cope.

[5] A function $f\colon D \to \mathbb{R}^n$ is **convex** if $-f$ is concave, and **strictly convex** if $-f$ is strictly concave.

Level curves and level sets

Often, given some function $f\colon D \to \mathbb{R}$, we want to know what points in the domain D map to some fixed value in the codomain $\mathbb{R}$. For example, if f represents a budget constraint and we want to find the possible bundles of goods we can buy with exactly some fixed amount of money. We formalise this as follows:

Definition 7.5 Let $f\colon D \to \mathbb{R}$ where $D \subseteq \mathbb{R}^n$, and let $c \in \mathbb{R}$. Then the **level curve** of f at c is the set

$$\{\mathbf{x} \in D : f(\mathbf{x}) = c\}.$$

We may also want to know which points in D map to less than, or greater than the chosen value. In the budget constraint example just mentioned, the points in the domain mapping to less than c correspond to all the bundles of goods we can afford.

Definition 7.6 Let $f\colon D \to \mathbb{R}$ where $D \subseteq \mathbb{R}^n$, and let $c \in \mathbb{R}$. Then the set

$$L_f(c) = \{\mathbf{x} \in D : f(\mathbf{x}) \leqslant c\}$$

is the **lower level set** for the value c, and

$$U_f(c) = \{\mathbf{x} \in D : f(\mathbf{x}) \geqslant c\}$$

is the **upper level set** for the value c.

Let's look at a concrete example.

Example 7.7 Let $D = \{(x,y) \in \mathbb{R}^2 : x,y > 0\}$ be the open set consisting of the upper right-hand quadrant of the plane, and define $f\colon D \to \mathbb{R}$ by $f(x,y) = x^{0.2}y^{0.2}$.

The level curve at a given value c consists of all points $(x,y) \in \mathbb{R}^2$ such that $x^{0.2}y^{0.2} = c$. We can rearrange this to give $y = \frac{c^5}{x}$.

Figure 7.2 shows the level curve (green), upper level set (blue) and lower level set (red) in the (x,y) plane.

In Figure 7.3, the blue surface is the graph of the function f, the red plane is the level $c = 2$, and the green curve is the level curve of f for $c = 2$ (it is the shadow cast by the intersection of the blue surface and the red plane).

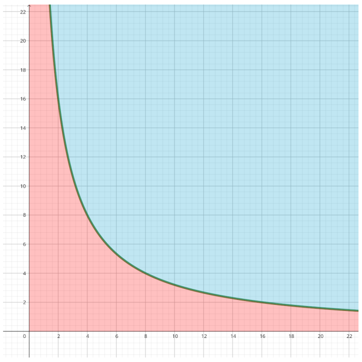


Figure 7.2: Illustration of the level curve, and the upper and lower level sets for $c = 2$ in Example 7.7

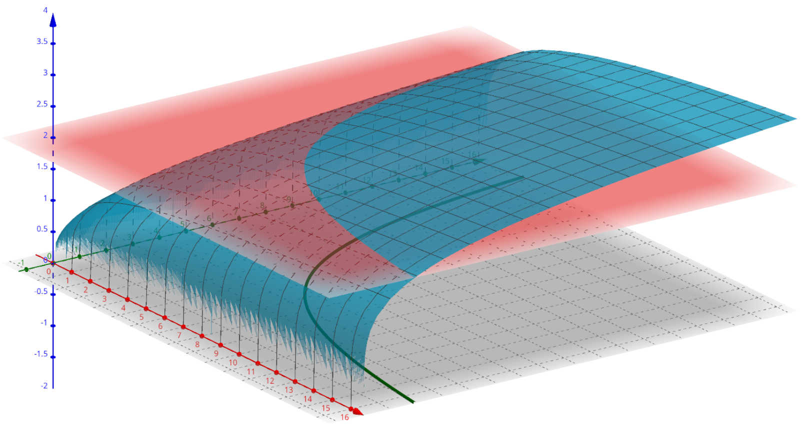


Figure 7.3: Illustration of the level curve for $c = 2$ in Example 7.7

The following theorem gives an important connection between convexity and concavity of a function, and convexity of its level sets.

Theorem 7.8 *Let $f\colon D \to \mathbb{R}$, with $D \subseteq \mathbb{R}^n$, and let $c \in \mathbb{R}$. Then:*

(i) *If f is concave, then the upper level set $U_f(c)$ is convex.*

(ii) *If f is convex, then the lower level set $L_f(c)$ is convex.*

Proof

(i) Suppose that $\mathbf{x}, \mathbf{y} \in U_f(c)$. Then $f(\mathbf{x}) \geqslant c$ and $f(\mathbf{y}) \geqslant c$, by the definition of the upper level set. Now choose $\alpha \in (0,1)$ and set $\mathbf{z} = \alpha\mathbf{x} + (1-\alpha)\mathbf{y}$.
Then

$$\begin{aligned} f(\mathbf{z}) &= f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \\ &\geqslant \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}) && \text{because } f \text{ is concave} \\ &\geqslant \alpha c + (1-\alpha)c && \text{because } \mathbf{x}, \mathbf{y} \in U_f(c) \\ &= c \end{aligned}$$

So $f(\mathbf{z}) \geqslant c$, and hence $\mathbf{z}$ is also in the upper level set $U_f(c)$. Thus $U_f(c)$ is convex.

(ii) This can be proved by a very similar method, and is left as an exercise.

Hence if f is concave, then the upper level sets are convex, and if f is convex, the lower level sets are convex. □

The converse isn't true, however: There are non-concave functions with convex upper level sets, and non-convex functions with convex lower level sets. This leads to the concepts in the next section.

Quasiconcavity and quasiconvexity

Given that the converse of Theorem 7.8 isn't true in general, nevertheless it's sometimes useful to study functions whose upper or lower level sets are convex, and to that end we introduce the following definition:

Definition 7.9 Let $f\colon D \to \mathbb{R}$, where $D \subseteq \mathbb{R}^n$. We say that f is **quasiconcave** if the upper level set $U_f(c)$ is convex for all $c \in \mathbb{R}$. And we say that f is **quasiconvex** if the lower level set $L_f(c)$ is convex for all $c \in \mathbb{R}$.

So, by Theorem 7.8, we have the following implications:

$$\begin{array}{lllll} f \text{ is concave} & \Longrightarrow & U_f(c) \text{ is convex} & \Longrightarrow & f \text{ is quasiconcave} \\ f \text{ is convex} & \Longrightarrow & L_f(c) \text{ is convex} & \Longrightarrow & f \text{ is quasiconvex} \end{array}$$

But the implications don't go the other way: in general quasiconcavity $\not\Rightarrow$ concavity, and quasiconvexity $\not\Rightarrow$ convexity. Concavity is a *sufficient* but not *necessary* condition for quasiconcavity, and convexity is a *sufficient* but not *necessary* condition for quasiconvexity.

There is a different but equivalent way of defining quasiconcavity and quasiconvexity:

Definition 7.10 Let $f\colon D \to \mathbb{R}^n$, where $D \subseteq \mathbb{R}^n$. Then:

(i) f is **quasiconcave** if, for all $\mathbf{x}, \mathbf{y} \in D$ and $\alpha \in (0,1)$, we have

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \geqslant \min\{f(\mathbf{x}), f(\mathbf{y})\}$$

(ii) f is **strictly quasiconcave** if, for all $\mathbf{x}, \mathbf{y} \in D$ and $\alpha \in (0,1)$, we have

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) > \min\{f(\mathbf{x}), f(\mathbf{y})\}$$

(iii) f is **quasiconvex** if $-f$ is quasiconcave. Or, equivalently, for all $\mathbf{x}, \mathbf{y} \in D$ and $\alpha \in (0,1)$, we have

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leqslant \min\{f(\mathbf{x}), f(\mathbf{y})\}$$

(iv) f is **strictly quasiconvex** if $-f$ is strictly quasiconcave. Or, equivalently, for all $\mathbf{x}, \mathbf{y} \in D$ and $\alpha \in (0,1)$, we have

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) < \min\{f(\mathbf{x}), f(\mathbf{y})\}$$

To summarise all this, we present the following theorem:

Theorem 7.11 *Let $f\colon D \to \mathbb{R}$, where $D \subseteq \mathbb{R}^n$. Then:*

(i) *f is strictly concave $\Rightarrow$ f is concave.*

(ii) *f is concave $\Rightarrow$ f is quasiconcave.*

(iii) *f is strictly concave $\Rightarrow$ f is strictly quasiconcave.*

(iv) *f is strictly quasiconcave $\Rightarrow$ f is quasiconcave.*

The proof is relatively straightforward, and is mostly a case of checking the various definitions.

Proof Let $\mathbf{x}, \mathbf{y} \in D$ and $\alpha \in (0,1)$.

(i) If f is strictly concave then

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) > \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}),$$

from which it follows that

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \geqslant \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}),$$

so f is concave.

(ii) If f is concave, then

$$\begin{aligned} f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) &\geqslant \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}) \\ &\geqslant \alpha \min\{f(\mathbf{x}), f(\mathbf{y})\} + (1-\alpha)\min\{f(\mathbf{x}), f(\mathbf{y})\} \\ &= \min\{f(\mathbf{x}), f(\mathbf{y})\}, \end{aligned}$$

so f is quasiconcave.

(iii) If f is strictly concave, then

$$\begin{aligned} f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) &> \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}) \\ &\geqslant \alpha \min\{f(\mathbf{x}), f(\mathbf{y})\} + (1-\alpha)\min\{f(\mathbf{x}), f(\mathbf{y})\} \\ &= \min\{f(\mathbf{x}), f(\mathbf{y})\}, \end{aligned}$$

hence f is strictly quasiconcave.

(iv) If f is strictly quasiconcave, then

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) > \min\{f(\mathbf{x}), f(\mathbf{y})\},$$

from which it follows that

$$f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \geqslant \min\{f(\mathbf{x}), f(\mathbf{y})\},$$

hence f is quasiconcave.

This completes the proof. □

An analogous result can be proved for (strictly) convex and (strictly) quasiconvex functions, and is left as an exercise.

We finish this section with some notes. Think about each of them and convince yourself that they are correct.

- Linear functions are the only ones that are both concave and convex.
- A sum of concave functions is concave, and a sum of convex functions if convex.
- But the corresponding statements for (strictly) quasiconcave or quasiconvex functions don't hold in general.
- If $f\colon D \to \mathbb{R}$ is strictly concave, it can't be convex or strictly convex.
- However, a strictly concave function might be strictly quasiconcave or quasiconcave.
- A function that has a level curve which includes a section (that is, if the function is flat on some section) cannot be strictly quasiconcave or strictly quasiconvex.

II

Linear Algebra

8 Vectors

To begin with, we introduce the basic definitions and algebra of vectors.

Definition 8.1 A **vector** is a quantity that is determined by its magnitude and direction.

A **scalar** is a quantity that is determined by its magnitude only.

A vector can be represented geometrically by a directed line segment, whose length represents the magnitude of the vector, and the direction is the same as direction of the vector. The direction is indicated by an arrow. A line segment from A to B is often written $\overrightarrow{AB}$. This definition specifies only the direction and magnitude of the vector, but not its position in space.

Vectors are often represented by a single bold lower case letter such as $\mathbf{u}$ or $\mathbf{v}$, but when handwritten we often underline, for example $\underline{u}$ or $\underline{v}$.

We will usually consider vectors in two- or three-dimensional space, but it is possible to generalise all of the following to four- or higher-dimensional space too. And although we will be working exclusively with vectors whose components are real numbers, all of the following works just as well (and in most cases almost identically) if we replace the real numbers $\mathbb{R}$ with the rational numbers $\mathbb{Q}$ or the complex numbers $\mathbb{C}$.

We can represent vectors in various ways: either geometrically as directed line segments, or as ordered sequences of real numbers. With the latter approach, we may write a given vector as a **column vector**:

$$\begin{bmatrix}1\\2\end{bmatrix}, \qquad \begin{bmatrix}-1\\7\end{bmatrix}, \qquad \begin{bmatrix}1\\0\\4\end{bmatrix},$$

as a **row vector**:

$$\begin{bmatrix}1 & 2\end{bmatrix}, \qquad \begin{bmatrix}-1 & 7\end{bmatrix}, \qquad \begin{bmatrix}1 & 0 & 4\end{bmatrix},$$

or in **coordinate** form, as ordered pairs, triples or n–tuples:

$$(1,2), \qquad (-1,7), \qquad (1,0,4).$$

Definition 8.2 We denote by $\mathbb{R}^2$ the real two-dimensional Euclidean vector space consisting of two-component column vectors (or, equivalently, two-component row vectors, or ordered pairs of real numbers).

$$\begin{aligned}\mathbb{R}^2 &= \left\{ \left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] : x, y \in \mathbb{R} \right\} \\ &= \left\{ \begin{bmatrix} x & y \end{bmatrix} : x, y \in \mathbb{R} \right\} \\ &= \{ (x, y) : x, y \in \mathbb{R} \}\end{aligned}$$

Geometrically, this is the infinite Euclidean plane.

Definition 8.3 We denote by $\mathbb{R}^3$ the real three-dimensional Euclidean vector space consisting of three-component column vectors (or, equivalently, three-component row vectors, or ordered triples of real numbers).

$$\begin{aligned}\mathbb{R}^3 &= \left\{ \left[\begin{smallmatrix} x \\ y \\ z \end{smallmatrix}\right] : x, y, z \in \mathbb{R} \right\} \\ &= \left\{ \begin{bmatrix} x & y & z \end{bmatrix} : x, y, z \in \mathbb{R} \right\} \\ &= \{ (x, y, z) : x, y, z \in \mathbb{R} \}\end{aligned}$$

Geometrically, this is ordinary three-dimensional Euclidean space.

Definition 8.4 The **position vector** $\mathbf{u} = (x, y)$ is the vector in $\mathbb{R}^2$ which starts at the origin and ends at the point with coordinates (x, y).

Similarly, in $\mathbb{R}^3$, the position vector $\mathbf{u} = (x, y, z)$ is the vector which starts at the origin and ends at the point with coordinates (x, y, z).

In any vector space, there is a special vector corresponding to the origin.

Definition 8.5 The **zero vector** is any vector whose magnitude is zero, and whose direction is arbitrary. It is denoted $\mathbf{0}$ (or, when handwritten, $\underline{0}$). In column and coordinate form this is written

$$\mathbf{0} = \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right] = (0, 0)$$

in $\mathbb{R}^2$ and

$$\mathbf{0} = \left[\begin{smallmatrix} 0 \\ 0 \\ 0 \end{smallmatrix}\right] = (0, 0, 0)$$

in $\mathbb{R}^3$. The zero vector is the position vector of the origin.

Vector addition

Given two vectors in $\mathbb{R}^2$ or $\mathbb{R}^3$, we can add them together to obtain another vector of the same type.

With a bit of thought, it makes sense to add two vectors that are in the same direction, for example, $\mathbf{u} + 2\mathbf{u} = 3\mathbf{u}$. But what if the two vectors are in different directions?

In that case, the sum $\mathbf{u} + \mathbf{v}$ is obtained as follows.

(i) Starting at the origin, draw $\mathbf{u}$.
(ii) Starting from the tip of the arrow representing $\mathbf{u}$, draw $\mathbf{v}$.
(iii) Then the sum is the vector from the origin to the finishing point of $\mathbf{v}$.

This is depicted in Figure 8.1

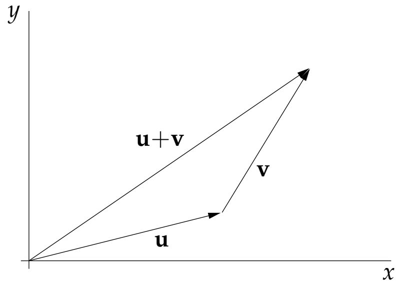


Figure 8.1: Vector addition

For vectors represented in column, row or coordinate form, we use componentwise addition. In $\mathbb{R}^2$ this is as follows:

$$\begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} a+c \\ b+d \end{bmatrix}$$
$$\begin{bmatrix} a & b \end{bmatrix} + \begin{bmatrix} c & d \end{bmatrix} = \begin{bmatrix} (a+c) & (b+d) \end{bmatrix}$$
$$(a,b) + (c,d) = (a+b, c+d)$$

And in $\mathbb{R}^3$ it works like this:

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} + \begin{bmatrix} d \\ e \\ f \end{bmatrix} = \begin{bmatrix} a+d \\ b+e \\ c+f \end{bmatrix}$$
$$\begin{bmatrix} a & b & c \end{bmatrix} + \begin{bmatrix} d & e & f \end{bmatrix} = \begin{bmatrix} (a+c) & (b+d) & (c+f) \end{bmatrix}$$
$$(a,b,c) + (d,e,f) = (a+d, b+e, c+f)$$

Vector addition is only defined for two vectors of the same dimension: it doesn't make sense, for example, to add a two-dimensional vector to a three-dimensional vector.

Scalar multiplication

Given a vector $\mathbf{v}$ in $\mathbb{R}^2$ or $\mathbb{R}^3$, and a scalar k in $\mathbb{R}$, we can form the vector $k\mathbf{v}$. This is the vector whose magnitude is $|k|$ times the magnitude of $\mathbf{v}$, and which points either in the same direction as $\mathbf{v}$ if $k > 0$, in the opposite direction if $k < 0$, or is the zero vector $\mathbf{0}$ if $k = 0$.

For vectors represented in column, row or coordinate form, we multiply each component by the given scalar:

$$k \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} ka \\ kb \end{bmatrix} \qquad k \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} ka \\ kb \\ kc \end{bmatrix}$$
$$k \begin{bmatrix} a & b \end{bmatrix} = \begin{bmatrix} ka & kb \end{bmatrix} \qquad k \begin{bmatrix} a & b & c \end{bmatrix} = \begin{bmatrix} ka & kb & kc \end{bmatrix}$$
$$k(a,b) = (ka, kb) \qquad k(a,b,c) = (ka, kb, kc)$$

The negative of a vector $\mathbf{v}$ is the vector of the same magnitude but the opposite sense to $\mathbf{v}$. It is denoted $-\mathbf{v}$.

Norm and inner product

The length or **norm** of a vector $\mathbf{u}$ is denoted by $\|\mathbf{u}\|$. In particular,

(i) $\|\mathbf{u}\| \geqslant 0$ for any vector $\mathbf{v}$,
(ii) $\|\mathbf{u}\| = 0$ if and only if $\mathbf{u} = \mathbf{0}$, and
(iii) $\|k\mathbf{u}\| = |k|\|\mathbf{u}\|$.

Two vectors $\mathbf{a}$ and $\mathbf{b}$ are equal if they have the same magnitudes ($\|\mathbf{a}\| = \|\mathbf{b}\|$) and they are in the same direction. We write $\mathbf{a} = \mathbf{b}$.

Definition 8.6 A **unit vector** is a vector of norm 1.

Given any non-zero vector $\mathbf{u}$ we can create a unit vector, written $\hat{\mathbf{u}}$, which is in the same direction as $\mathbf{u}$ but is of unit length, by dividing $\mathbf{u}$ by its norm (or, more precisely, multiplying $\mathbf{u}$ by $1/\|\mathbf{u}\|$):

$$\hat{\mathbf{u}} := \mathbf{u}/\|\mathbf{u}\|$$

This process is called **normalising** or **normalisation**. We will meet it later when we study diagonalisation of quadratic forms.

In $\mathbb{R}^2$ and $\mathbb{R}^3$ we use Pythagoras' Theorem to calculate the norm:

$$\|(x,y)\| = \sqrt{x^2+y^2} \qquad \|(x,y,z)\| = \sqrt{x^2+y^2+z^2}$$

An additional property of the norm is the **triangle inequality**:

Proposition 8.7 (The triangle inequality) *Let* $\mathbf{u}$ *and* $\mathbf{v}$ *be two vectors in* $V = \mathbb{R}^2$ *or* $\mathbb{R}^3$*. Then*

$$\|\mathbf{u}+\mathbf{v}\| \leqslant \|\mathbf{u}\| + \|\mathbf{v}\|.$$

Given two vectors $\mathbf{u}$ and $\mathbf{v}$ in either $\mathbb{R}^2$ or $\mathbb{R}^3$, we can define their **dot product**, **scalar product** or **inner product**. This is a well-defined way of combining two vectors of the same dimension to get a single scalar, and is denoted either by $\mathbf{u}\cdot\mathbf{v}$ or $\langle\mathbf{u},\mathbf{v}\rangle$ (or, in the Dirac notation used in quantum physics, $\langle\mathbf{u}|\mathbf{v}\rangle$). It is calculated as follows:

$$\langle(a,b),(c,d)\rangle = (a,b)\cdot(c,d) = ac+bd,$$
$$\langle(a,b,c),(d,e,f)\rangle = (a,b,c)\cdot(d,e,f) = ad+be+cf$$

If $\mathbf{u},\mathbf{v},\mathbf{w} \in V$, where V is either $\mathbb{R}^2$ or $\mathbb{R}^3$, and $k \in \mathbb{R}$ is any real scalar, then:

(i) $\langle\mathbf{u},\mathbf{v}\rangle = \langle\mathbf{v},\mathbf{u}\rangle$,
(ii) $\langle\mathbf{u}+\mathbf{v},\mathbf{w}\rangle = \langle\mathbf{u},\mathbf{w}\rangle + \langle\mathbf{v},\mathbf{w}\rangle$,
(iii) $\langle k\mathbf{u},\mathbf{v}\rangle = k\langle\mathbf{u},\mathbf{v}\rangle = \langle\mathbf{u},k\mathbf{v}\rangle$, and
(iv) $\langle\mathbf{u},\mathbf{u}\rangle \geqslant 0$ with $\langle\mathbf{u},\mathbf{u}\rangle = 0$ only when $\mathbf{u} = \mathbf{0}$.

There is a strong connection between the norm and the scalar product: for any vector $\mathbf{v}$ in $\mathbb{R}^2$ or $\mathbb{R}^3$, we have

$$\|\mathbf{v}\| = \sqrt{\langle\mathbf{v},\mathbf{v}\rangle}.$$

(Check this using coordinate form.)

There is another, more geometric way of calculating the scalar product of two vectors in $\mathbb{R}^2$ or $\mathbb{R}^3$. Given $\mathbf{u}$ and $\mathbf{v}$, let $u = \|\mathbf{u}\|$ and $v = \|\mathbf{v}\|$, and let θ be the angle between $\mathbf{u}$ and $\mathbf{v}$. Then

$$\langle\mathbf{u},\mathbf{v}\rangle = uv\cos\theta.$$

Combining these two, for any two nonzero vectors $\mathbf{u}$ and $\mathbf{v}$ we have $\langle\mathbf{u},\mathbf{v}\rangle = 0$ exactly when $uv\cos\theta = 0$, which can only happen when $\cos\theta = 0$, which is true only when $\theta = \frac{\pi}{2}$ or $\frac{3\pi}{2}$. Geometrically, this means that the scalar product of two nonzero vectors is only zero when those vectors are **perpendicular** (at right angles) to each other.

Definition 8.8 Two vectors $\mathbf{u}$ and $\mathbf{v}$ are **orthogonal** if $\langle\mathbf{u},\mathbf{v}\rangle = 0$.

A set of vectors $\{\mathbf{v}_1,\ldots,\mathbf{v}_n\}$ is an **orthogonal set** if the vectors are pairwise orthogonal; that is, if $\langle\mathbf{v}_i,\mathbf{v}_j\rangle = 0$ for $1 \leqslant i \neq j \leqslant n$.

An orthogonal set is **orthonormal** if, additionally, each vector has unit norm: $\|\mathbf{v}_i\| = 1$ for $1 \leqslant i \leqslant n$.

Standard unit vectors

Geometrically, the vector $\left[\begin{smallmatrix}2\\3\end{smallmatrix}\right]$ in $\mathbb{R}^2$ is the position vector of the point we get to by starting at the origin, moving 2 units along the x–axis, and then moving 3 units parallel to the y–axis. (Or, equivalently, moving 3 units along the y–axis and then moving 2 units parallel to the x–axis.)

We can make this more precise by introducing the **standard unit vectors** or **standard coordinate vectors**. These are the unit vectors which point in the positive direction along each of the coordinate axes in $\mathbb{R}^2$ and $\mathbb{R}^3$.

In $\mathbb{R}^2$ we let

$$\mathbf{i} = (1,0) = \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right], \qquad \mathbf{j} = (0,1) = \left[\begin{smallmatrix}0\\1\end{smallmatrix}\right].$$

In $\mathbb{R}^3$ we let

$$\mathbf{i} = (1,0,0) = \left[\begin{smallmatrix}1\\0\\0\end{smallmatrix}\right], \quad \mathbf{j} = (0,1,0) = \left[\begin{smallmatrix}0\\1\\0\end{smallmatrix}\right], \quad \mathbf{k} = (0,0,1) = \left[\begin{smallmatrix}0\\0\\1\end{smallmatrix}\right].$$

So, any vector (x,y) in $\mathbb{R}^2$ can be written uniquely in the form $x\mathbf{i} + y\mathbf{j}$, and any vector (x,y,z) in $\mathbb{R}^3$ can be written uniquely as $x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$.

Each of these standard coordinate vectors is a unit vector, and they are all orthogonal to each other. The sets $\{\mathbf{i},\mathbf{j}\}$ and $\{\mathbf{i},\mathbf{j},\mathbf{k}\}$ are thus orthonormal sets.

We will return to these later when we study coordinate systems.

Vector algebra in $\mathbb{R}^2$ *and* $\mathbb{R}^3$

To summarise, in $\mathbb{R}^2$ we have the facts listed in Table 8.1.

Zero vector	$\mathbf{0}$	$=$	$(0,0)$
Addition	$(u_1,u_2)+(v_1,v_2)$	$=$	$(u_1{+}v_1, u_2{+}v_2)$
Negative	$-(u_1,u_2)$	$=$	$(-u_1,-u_2)$
Scalar multiplication	$k(u_1,u_2)$	$=$	(ku_1,ku_2)
Norm	$\\|(u_1,u_2)\\|$	$=$	$\sqrt{u_1^2+u_2^2}$

Table 8.1: Vector algebra in $\mathbb{R}^2$

And in $\mathbb{R}^3$ we have the facts listed in Table 8.2.

Zero vector	$\mathbf{0}$	$=$	$(0,0,0)$
Addition	$(u_1,u_2,u_3)+(v_1,v_2,v_3)$	$=$	$(u_1{+}v_1, u_2{+}v_2, u_3{+}v_3)$
Negative	$-(u_1,u_2,u_3)$	$=$	$(-u_1,-u_2,-u_3)$
Scalar multiplication	$k(u_1,u_2,u_3)$	$=$	(ku_1,ku_2,ku_3)
Norm	$\\|(u_1,u_2,u_3)\\|$	$=$	$\sqrt{u_1^2+u_2^2+u_3^2}$

Table 8.2: Vector algebra in $\mathbb{R}^3$

If we denote either $\mathbb{R}^2$ or $\mathbb{R}^3$ by V, and if $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ are vectors in V, and k and l are real scalars, then all of the properties listed in Table 8.3 are satisfied. It is very easy to check all these results using the component form of the vectors and the properties of the real numbers.

(i)	$\mathbf{u}+\mathbf{v}$	$\in$	V	(Closure under addition)
(ii)	$\mathbf{u}+\mathbf{v}$	$=$	$\mathbf{v}+\mathbf{u}$	(Commutative law of addition)
(iii)	$\mathbf{u}+(\mathbf{v}+\mathbf{w})$	$=$	$(\mathbf{u}+\mathbf{v})+\mathbf{w}$	(Associative law of addition)
(iv)	$\mathbf{0}+\mathbf{u}$	$=$	$\mathbf{u}+\mathbf{0}=\mathbf{u}$	(Existence of a zero vector: 'additive identity')
(v)	$\mathbf{u}+(-\mathbf{u})$	$=$	$\mathbf{0}$	(Existence of a negative vector: 'additive inverse')
(vi)	$k\mathbf{u}$	$\in$	V	(Closure under scalar multiplication)
(vii)	$k(\mathbf{u}+\mathbf{v})$	$=$	$k\mathbf{u}+k\mathbf{v}$	(First distributive law)
(viii)	$(k+l)\mathbf{u}$	$=$	$k\mathbf{u}+l\mathbf{u}$	(Second distributive law)
(ix)	$k(l\mathbf{u})$	$=$	$(kl)\mathbf{u}$	(Associative law for scalar multiplication)
(x)	$1\mathbf{u}$	$=$	$\mathbf{u}$	(Existence of identity for scalar multiplication)

Table 8.3: General properties of vectors in $\mathbb{R}^2$, $\mathbb{R}^3$ and $\mathbb{R}^n$

- Properties (i) and (vi) say that $\mathbb{R}^2$ and $\mathbb{R}^3$ are **closed** under vector addition and scalar multiplication: the result of adding two vectors together is another vector, and the result of multiplying a vector by a scalar is another vector.
- The **commutative law** for vector addition (ii) says that it doesn't matter what order we add two vectors together, the result is the same either way.
- The **associative law** (iii) says, effectively, that we can ignore brackets when adding together three or more vectors.
- Property (iv) asserts the existence of the zero vector: the unique vector in $\mathbb{R}^2$ or $\mathbb{R}^3$ which doesn't change any other vector we add it to (algebraists call this an **additive identity**).
- Property (v) says that for any vector $\mathbf{v}$ there is a corresponding negative vector $-\mathbf{v}$ of the same length but pointing in the opposite direction: if we add $\mathbf{v}$ and $-\mathbf{v}$ we get the zero vector.
- Properties (vii) and (viii), the **distributive laws** describe how the vector addition and scalar multiplication operations interact with each other.
- Property (ix) says that multiplying a vector $\mathbf{v}$ by a scalar l and then by another scalar k is the same as multiplying k and l together and multiplying $\mathbf{v}$ by the result.
- And property (x) says that multiplying a vector $\mathbf{v}$ by the scalar $k=1$ leaves $\mathbf{v}$ unchanged (algebraists call this a **multiplicative identity**).

Higher dimensional vector spaces $\mathbb{R}^n$

We can generalise all we've seen so far with $\mathbb{R}^2$ and $\mathbb{R}^3$ to higher dimensional spaces. It's not obvious how we'd geometrically visualise such spaces, but the maths works out fine anyway, and higher-dimensional spaces are often very useful for representing and understanding solutions to real problems.

Definition 8.9 Let $\mathbb{R}^n$ be the set of ordered n–tuples of real numbers (or, equivalently, n–component column or row vectors):

$$\mathbb{R}^n = \{(x_1,\dots,x_n) : x_1,\dots,x_n \in \mathbb{R}\} = \left\{ \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} : x_1,\dots,x_n \in \mathbb{R} \right\}$$

Define vector addition and scalar multiplication as follows:

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1+y_1 \\ \vdots \\ x_n+y_n \end{bmatrix} \qquad \text{and} \qquad k\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} kx_1 \\ \vdots \\ kx_n \end{bmatrix}.$$

This is the n–dimensional real Euclidean vector space.

The obvious analogues of the ten properties in Table 8.3 also hold in $\mathbb{R}^n$.

9 Matrices

In this chapter, we introduce the basic definitions and algebra of matrices.

Definition 9.1 An $m \times n$ (real) **matrix** is a rectangular array of the form

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

with m rows and n columns.

Matrix operations

Two $m \times n$ matrices may be added together by adding their corresponding entries:

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{bmatrix} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \dots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \dots & a_{2n}+b_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \dots & a_{mn}+b_{mn} \end{bmatrix}$$

We may also define the difference $A - B$ in the obvious way. Matrix addition is commutative: $A + B = B + A$ for arbitrary matrices A and B (as long as the sum is defined).

An $m \times n$ matrix A may be multiplied by an element k of $\mathbb{R}$ (a **scalar**) by multiplying each entry of A by k:

$$kA = k \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} = \begin{bmatrix} ka_{11} & ka_{12} & \dots & ka_{1n} \\ ka_{21} & ka_{22} & \dots & ka_{2n} \\ \vdots & \vdots & & \vdots \\ ka_{m1} & ka_{m2} & \dots & ka_{mn} \end{bmatrix}$$

An $m \times k$ matrix A may be multiplied by a $k \times n$ matrix B to give an $m \times n$ matrix $C = AB$ by setting

$$c_{ij} = \sum_{t=1}^{k} a_{it} b_{tj}$$

That is, the entry c_{ij} is given by multiplying the corresponding elements of the ith row of A and the jth row of B, and then adding the results together. Matrix multiplication is not, in general, commutative: it is not always the case that $AB = BA$ for arbitrary matrices A and B; indeed, one or both of these products may not even be defined.

Transposition

Given an $m \times n$ matrix A, we form the **transpose** A^T by interchanging rows and columns:

$$A^T = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}^T = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{bmatrix}$$

In general, where the relevant addition and multiplication operations are defined, $(A + B)^T = A^T + B^T$ and $(AB)^T = B^T A^T$.

A square $n \times n$ matrix is **symmetric** if it is equal to its transpose; that is, $A = A^T$.

Triangular and diagonal matrices

The $n \times n$ **identity matrix**

$$I_n = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

has the property that for any $m \times n$ matrix A, and any $n \times m$ matrix B, $AI_n = A$ and $I_n B = B$.

An $n \times n$ matrix is **upper triangular** if it is of the form

$$A = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ 0 & a_{22} & \dots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix}$$

That is, all the entries below the leading (top-left to bottom-right) diagonal are zero. Similarly, an $n \times n$ matrix is **lower triangular** if all of the entries above the leading diagonal are zero:

$$B = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ a_{12} & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{nn} \end{bmatrix}$$

An $n \times n$ matrix is **diagonal** if every off-diagonal entry is zero:

$$D = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix}$$

Matrices acting on vectors

We may also regard vectors in $\mathbb{R}^2$ or $\mathbb{R}^3$ as 1×2 or 1×3 matrices over $\mathbb{R}$. In this case, matrix addition reduces to ordinary coordinate-wise vector addition. Matrix multiplication, however, has the effect of transforming a vector into another of the same type. For example:

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

Some standard matrix transformations in $\mathbb{R}^2$ can be seen in Table 9.1.

Transformation	$f(x,y)$	Matrix
identity map	(x,y)	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
rotation through angle θ	$(x\cos\theta - y\sin\theta, x\sin\theta + y\cos\theta)$	$\begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$
reflection about y axis	$(-x,y)$	$\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$
reflection about x axis	$(x,-y)$	$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$
reflection about $y = x$	(y,x)	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$
reflection in $y = (\tan\alpha)x$	$(x\cos(2\alpha)+y\sin(2\alpha), x\sin(2\alpha)-y\cos(2\alpha))$	$\begin{bmatrix} \cos(2\alpha) & \sin(2\alpha) \\ \sin(2\alpha) & -\cos(2\alpha) \end{bmatrix}$
scaling, factor k	(kx,ky)	$\begin{bmatrix} k & 0 \\ 0 & k \end{bmatrix}$
scalings, factors k_1, k_2	(k_1x, k_2y)	$\begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix}$

Table 9.1: Some standard matrix transformations in $\mathbb{R}^2$

Specifically, a 2×2 matrix maps a vector in $\mathbb{R}^2$ to another vector in $\mathbb{R}^2$, and a 3×3 matrix maps a vector in $\mathbb{R}^3$ to another vector in $\mathbb{R}^3$. In addition, a 2×3 matrix maps a vector in $\mathbb{R}^3$ to a vector in $\mathbb{R}^2$, while a 3×2 matrix maps a vector in $\mathbb{R}^3$ to a vector in $\mathbb{R}^2$, and so forth. Now

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} a \\ c \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} b \\ d \end{bmatrix}$$

so $\mathbf{i} = (1,0)$ is mapped to the first column of A, and $\mathbf{j} = (0,1)$ is mapped to the second column of the matrix.

Conversely, if $f(\mathbf{x}) = A\mathbf{x}$ and we know that

$$f\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} a \\ c \end{bmatrix} \quad \text{and} \quad f\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} b \\ d \end{bmatrix},$$

then we can construct the appropriate matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Notice that for any matrix $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ and vectors $\mathbf{v}_1 = \left[\begin{smallmatrix} x_1 \\ y_1 \end{smallmatrix}\right]$ and $\mathbf{v}_2 = \left[\begin{smallmatrix} x_2 \\ y_2 \end{smallmatrix}\right]$ we have

$$\begin{aligned} A(\mathbf{v}_1 + \mathbf{v}_2) &= \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_1 + x_2 \\ y_1 + y_2 \end{bmatrix} \\ &= \begin{bmatrix} ax_1 + ax_2 + by_1 + by_2 \\ cx_1 + cx_2 + dy_1 + dy_2 \end{bmatrix} \\ &= \begin{bmatrix} ax_1 + by_1 \\ cx_1 + dy_1 \end{bmatrix} + \begin{bmatrix} ax_2 + by_2 \\ cx_2 + dy_2 \end{bmatrix} \\ &= \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_1 + x_2 \\ y_1 + y_2 \end{bmatrix} + \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_2 + x_2 \\ y_2 + y_2 \end{bmatrix} \\ &= A\mathbf{v}_1 + A\mathbf{v}_2 \end{aligned}$$

and

$$A(k\mathbf{v}_1) = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} kx_1 \\ ky_1 \end{bmatrix} = \begin{bmatrix} kax_1 + kbx_1 \\ kcy_1 + kdy_1 \end{bmatrix} = k \begin{bmatrix} ax_1 + bx_1 \\ cy_1 + dy_1 \end{bmatrix} = kA\mathbf{v}_1.$$

In other words, matrix multiplication behaves nicely with respect to sums and scalar multiples of vectors. In the rest of these notes, we will be particularly concerned with transformations which satisfy these nice properties.

Definition 9.2 A function $f\colon \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is said to be a **linear transformation** if it satisfies the criteria

$$\begin{aligned} f(\mathbf{u}+\mathbf{v}) &= f(x_1+y_1, x_2+y_2) \\ &= f(x_1,x_2) + f(y_1,y_2) \\ &= f(\mathbf{u}) + f(\mathbf{v}) \end{aligned}$$

and

$$f(k\mathbf{u}) = f(kx_1, kx_2) = kf(x_1,x_2) = kf(\mathbf{u})$$

for any $k \in \mathbb{R}$ and $\mathbf{u} = (x_1, x_2)$ and $\mathbf{v} = (y_1, y_2) \in \mathbb{R}^2$.

We can extend this definition to $\mathbb{R}^n$ too:

Definition 9.3 A function $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is said to be a **linear transformation** if it satisfies the criteria

$$\begin{aligned} f(\mathbf{u}+\mathbf{v}) &= f(x_1+y_1, \ldots, x_n+y_n) \\ &= f(x_1,\ldots,x_n) + f(y_1,\ldots,y_n) \\ &= f(\mathbf{u}) + f(\mathbf{v}) \end{aligned}$$

and

$$f(k\mathbf{u} = f(kx_1, \ldots, kx_n) = kf(x_1,\ldots,x_n) = kf(\mathbf{u})$$

for any $k \in \mathbb{R}$ and $\mathbf{u} = (x_1, \ldots, x_n)$ and $\mathbf{v} = (y_1, \ldots, y_n) \in \mathbb{R}^n$.

What the observation at the beginning of this section means is that any function which is representable by a 2×2 (or $n\times n$) matrix acting on vectors in $\mathbb{R}^2$ (or $\mathbb{R}^n$) is one of these special "linear transformations". This includes reflections, rotations and scaling transformations, but not translations. A translation by a fixed vector $\mathbf{v} = \left[\begin{smallmatrix} a \\ b \end{smallmatrix}\right]$ yields a function $f(x,y) = (x+a, y+b)$, but this doesn't satisfy either of the linearity criteria:

$$\begin{aligned} f(x_1+x_2, y_1+y_2) &= (x_1+x_2+a, y_1+y_2+b) \\ &\neq (x_1+a, y_1+b) + (x_2+a, y_2+b) \\ &= f(x_1,y_1) + f(x_2,y_2) \end{aligned}$$

$$f(kx_1, ky_1) = (kx_1+a, ky_1+b) \neq k(x_1+a, y_1+b) = kf(x_1,y_1)$$

Actually, it follows almost immediately from the definition that a linear transformation must leave the origin $(0,0)$ or $(0,0,0)$ fixed. Rotations, reflections (in lines or planes passing through the origin) or scaling transformations all do this, but translations don't.

In general, a given linear transformation $f\colon \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ will be of the form

$$f(x,y) = (ax+by, cx+dy)$$

for any $(x,y) \in \mathbb{R}^2$ and some fixed $a,b,c,d \in \mathbb{R}$.

So, any matrix yields a linear transformation, and any linear transformation can be represented by a matrix. The connection between matrices and linear transformations is not, however, a straightforward bijection: for any linear transformation acting on, say, the plane $\mathbb{R}^2$, there are many matrices (actually, uncountably infinitely many of them) which represent that transformation. We will study this conundrum further in the next section, on coordinate systems.

Example 9.4 None of the following transformations are linear.

$$\begin{aligned} &f(x) = x^2 && f\colon \mathbb{R} \to \mathbb{R} \\ &f(x,y) = x^2 + y^2 && f\colon \mathbb{R}^2 \to \mathbb{R} \\ &f(x,y,z) = x^2 + y^2 + z^2 && f\colon \mathbb{R}^3 \to \mathbb{R} \\ &f(x_1, x_2, \dots x_n) = x_1^2 + x_2^2 + \cdots + x_n^2 && f\colon \mathbb{R}^n \to \mathbb{R} \end{aligned}$$

Determinants and traces

Definition 9.5 Let $A = [a_{ij}]$ be an $n \times n$ matrix. The **determinant** $\det A$ or $|A|$ of A is defined recursively as

$$\begin{aligned} |A| &= \sum_{k=1}^{n} a_{1k}C_{1k} = \sum_{k=1}^{n} a_{2k}C_{2k} = \cdots = \sum_{k=1}^{n} a_{nk}C_{nk} \\ &= \sum_{k=1}^{n} a_{k1}C_{k1} = \sum_{k=1}^{n} a_{k2}C_{k2} = \cdots = \sum_{k=1}^{n} a_{kn}C_{kn} \end{aligned}$$

where C_{ij} denotes the (i,j) **cofactor** of A; this is defined to be $(-1)^{i+j}|M_{ij}|$, where M_{ij} is the matrix obtained by deleting the ith row and jth column from A. The determinant of a 1×1 matrix $[x]$ is simply x.

Applying this definition to a 2×2 matrix $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ yields the usual expression $ad - bc$; higher-order matrices have more complicated forms. In general,

$$f\begin{bmatrix} x \\ y \end{bmatrix} = A\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax + by \\ cx + dy \end{bmatrix},$$

so the point (x,y) is mapped to the point $(ax + by, cx + dy)$.

In particular, $(1,1)$ is mapped to $(a + b, c + d)$, so that the unit square is mapped to a parallelogram with vertices at $(0,0)$, (a,c), $(a + b, c + d)$, (b,d). This is shown in Figure 9.1

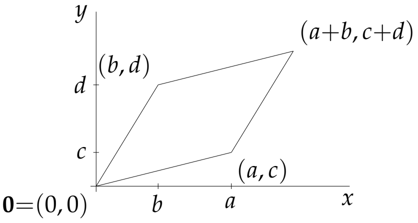


Figure 9.1: Image of the unit square under the matrix transformation $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$

More generally, this means that the whole Cartesian coordinate system with perpendicular axes $0xy$ is transformed to a new grid system with axes which are generally not at right angles, and each unit square of the grid (of area 1 unit) is transformed into a parallelogram. We will study this concept further in the next chapter.

It may be shown, however, that the area of the parallelogram in the above diagram is equal to the absolute value of the determinant of the matrix $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$. That is,

$$\text{Area} = |ad - bc| = |\det A|.$$

The sign of $|A|$ tells us the **orientation** of the parallelogram relative to the original unit square: a negative determinant indicates that the orientation has been reversed (broadly speaking, the parallelogram has been "flipped over" in some sense) while a positive determinant indicates that the orientation is unchanged.

Consider the matrix $R_\theta = \left[\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right]$, which represents a rotation of angle θ around the origin. Then

$$|R_\theta| = \cos^2\theta + \sin^2\theta = 1$$

which indicates that rotations leave area and orientation unchanged, as expected.

Now consider the matrix $Q_\alpha = \left[\begin{smallmatrix} \cos 2\alpha & \sin 2\alpha \\ \sin 2\alpha & \cos 2\alpha \end{smallmatrix}\right]$, which represents a reflection in the line $y = x\tan\alpha$. Then

$$|Q_\alpha| = -\cos^2 2\alpha - \sin^2 2\alpha = -1$$

which confirms that reflections leave area unchanged but reverse orientation.

The matrix $D = \left[\begin{smallmatrix} k_1 & 0 \\ 0 & k_2 \end{smallmatrix}\right]$ represents a scaling by factor of k_1 in the horizontal direction, and a factor k_2 in the vertical direction. Then

$$|D| = k_1 k_2,$$

which tells us that the area is scaled by a factor of k_1k_2. The change of orientation depends on the signs of the factors k_1 and k_2: if both are positive or both negative, then the orientation is preserved, but if one is negative then the orientation is reversed.

Definition 9.6 Let $A{=}[a_{ij}]$ be an $n{\times}n$ matrix. The **trace** $\operatorname{tr} A$ of A is the sum of the diagonal elements of A:

$$\operatorname{tr} A = \sum_{i=1}^{n} a_{ii} = a_{11} + a_{22} + \cdots + a_{nn}.$$

The trace has the following properties:

$$\operatorname{tr}(A+B) = \operatorname{tr}(A) + \operatorname{tr}(B), \qquad \operatorname{tr}(kA) = k\operatorname{tr}(A),$$
$$\operatorname{tr}(AB) = \operatorname{tr}(BA), \qquad \operatorname{tr}(A) = \operatorname{tr}(A^T).$$

Inverse matrices

Definition 9.7 Let $A = [a_{ij}]$ be a square $n{\times}n$ matrix, and let C_{ij} denote the (i,j) cofactor (as introduced in Definition 9.5). Then the matrix

$$\operatorname{adj} A = \begin{bmatrix} C_{11} & C_{21} & \dots & C_{n1} \\ C_{12} & C_{22} & \dots & C_{n2} \\ \vdots & \vdots & & \vdots \\ C_{1n} & C_{2n} & \dots & C_{nn} \end{bmatrix},$$

that is, the transpose of the matrix of cofactors of A, is called the **adjoint** of A.

If the matrix A has nonzero determinant, we may construct its inverse to be

$$A^{-1} = \tfrac{1}{|A|} \operatorname{adj} A.$$

This, as one might expect, has the property that

$$A^{-1}A = AA^{-1} = I_n.$$

If A has zero determinant, then it doesn't have an inverse. Such matrices are said to be **singular** or **noninvertible**, while matrices with nonzero determinants are **nonsingular** or **invertible**.

The inverse matrix A^{-1} effectively reverses the action of A, in the sense that if $\mathbf{v} = A\mathbf{u}$ then $u = A^{-1}\mathbf{v}$.

Orthogonal matrices

Definition 9.8 An $n \times n$ square matrix A is **orthogonal** if $A^T A = AA^T = I_n$ or, equivalently, if $A^T = A^{-1}$.

10 Coordinate systems

In this section, we will investigate linear coordinate systems in $\mathbb{R}^2$ and $\mathbb{R}^3$, and learn how to transform between them.

Linear combinations

We can add two vectors in $\mathbb{R}^2$, $\mathbb{R}^3$ or $\mathbb{R}^n$, and multiply a vector by a (real) scalar. Generalising this, we obtain the following:

Definition 10.1 Given a set of m vectors

$$\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$$

where each vector $\mathbf{v}_i \in \mathbb{R}^n$, then a **linear combination** of them is any vector of the form

$$\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_m\mathbf{v}_m$$

where each scalar constant $\alpha_i \in \mathbb{R}$.

Example 10.2 If $\mathbf{v}_1$ and $\mathbf{v}_2$ are vectors in $\mathbb{R}^n$, then

$$2\mathbf{v}_1 + 3\mathbf{v}_2, \quad \mathbf{v}_1 - \mathbf{v}_2, \quad \tfrac{1}{2}\mathbf{v}_1, \quad \mathbf{v}_2, \quad \pi\mathbf{v}_1 + \sqrt{2}\mathbf{v}_2$$

are all linear combinations of the set $\{\mathbf{v}_1, \mathbf{v}_2\}$.
Note that $\frac{1}{2}\mathbf{v}_1 = \frac{1}{2}\mathbf{v}_1 + 0\mathbf{v}_2$ and $\mathbf{v}_2 = 0\mathbf{v}_1 + 1\mathbf{v}_2$.

Example 10.3 $\begin{bmatrix} 0 \\ 27 \end{bmatrix}$ is a linear combination of $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} -1 \\ 4 \end{bmatrix}$ since

$$\begin{bmatrix} 0 \\ 27 \end{bmatrix} = 3\begin{bmatrix} 2 \\ 1 \end{bmatrix} + 6\begin{bmatrix} -1 \\ 4 \end{bmatrix}.$$

Definition 10.4 Given a set $S = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$ of vectors, we define the **span** of S to be the set of all possible linear combinations of vectors from S:

$$\operatorname{span} S = \{\alpha_1\mathbf{v}_1 + \cdots + \alpha_m\mathbf{v}_m : \alpha_1, \ldots, \alpha_m \in \mathbb{R}\}$$

Example 10.5 Let $S = \left\{\begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 4 \end{bmatrix}\right\}$. Then

$$\operatorname{span} S = \left\{\alpha_1\begin{bmatrix} 2 \\ 1 \end{bmatrix} + \alpha_2\begin{bmatrix} -1 \\ 4 \end{bmatrix} : \alpha_1, \alpha_2 \in \mathbb{R}\right\} = \mathbb{R}^2.$$

However,

$$\operatorname{span}\left\{\begin{bmatrix} 2 \\ 1 \end{bmatrix}\right\} = \left\{\alpha_1\begin{bmatrix} 2 \\ 1 \end{bmatrix} : \alpha_1 \in \mathbb{R}\right\},$$

which is the line $y = \frac{1}{2}x$ in $\mathbb{R}^2$.

Linear independence

Example 10.3 says that we can form the vector $\begin{bmatrix} 0 \\ 27 \end{bmatrix}$ as a specific (unique, as it happens) linear combination of $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} -1 \\ 4 \end{bmatrix}$. In other words, we can get to the point $(0, 27)$ in the plane with just those two vectors. So, the vector $\begin{bmatrix} 0 \\ 27 \end{bmatrix}$ is in some sense redundant.

Equivalently, there exist nonzero constants α_1, α_2 and α_3 such that

$$\alpha_1 \begin{bmatrix} 0 \\ 27 \end{bmatrix} + \alpha_2 \begin{bmatrix} 2 \\ 1 \end{bmatrix} + \alpha_3 \begin{bmatrix} -1 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix};$$

the triple $(\alpha_1, \alpha_2, \alpha_3) = (-1, 3, 6)$ satisfies this condition as, for that matter, does $(k\alpha_1, k\alpha_2, k\alpha_3)$ for any nonzero scalar $k \in \mathbb{R}$. Also, trivially, does the triple $(0, 0, 0)$.

Now consider the vectors

$$\mathbf{v}_1 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} -1 \\ 0 \\ 3 \end{bmatrix}$$

in $\mathbb{R}^3$. Apart from the trivial case $(\alpha_1, \alpha_2, \alpha_3) = (0, 0, 0)$, there exists no triple of real scalar constants satisfying the equation

$$\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \alpha_3\mathbf{v}_3 = \alpha_1 \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} + \alpha_3 \begin{bmatrix} -1 \\ 0 \\ 3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \mathbf{0}.$$

Equivalently, we can't get to the point $(-1, 0, 3)$ just by using linear combinations of the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, so the vector $\mathbf{v}_3$ isn't redundant in the same sense. More generally:

Definition 10.6 A set $S = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$ of vectors in $\mathbb{R}^n$ is **linearly dependent** if there exist scalar constants $\alpha_1, \ldots, \alpha_m$, *not all of which are zero* such that

$$\alpha_1\mathbf{v}_1 + \cdots + \alpha_m\mathbf{v}_m = \mathbf{0}. \tag{10.1}$$

If no such nontrivial m–tuple of scalar constants exists then the vectors in S are said to be **linearly independent**. That is, if the only values of $\alpha_1, \ldots, \alpha_m$ which satisfy (10.1) are $\alpha_1 = \cdots = \alpha_m = 0$.

Example 10.7 The vectors $\begin{bmatrix} 0 \\ 27 \end{bmatrix}$, $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} -1 \\ 4 \end{bmatrix}$ in $\mathbb{R}^3$ are linearly dependent, because (as noted a few paragraphs ago) the triple $(\alpha_1, \alpha_2, \alpha_3) = (-1, 3, 6)$ satisfies equation (10.1).

Example 10.8 Suppose $\mathbf{v}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$. Set up equation (10.1):

$$a\mathbf{v}_1 + b\mathbf{v}_2 = a\begin{bmatrix} 1 \\ 2 \end{bmatrix} + b\begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \mathbf{0}.$$

this means that

$$\begin{bmatrix} a+2b \\ 2a+3b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

and hence we obtain two linear simultaneous equations in the variables a and b:

$$\left.\begin{aligned} a + 2b &= 0 \\ 2a + 3b &= 0 \end{aligned}\right\} \Longrightarrow a = b = 0$$

and hence (since this is the only solution of those simultaneous equations) the vectors are linearly independent.

Example 10.9 Consider the set

$$\left\{ \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \\ 0 \end{bmatrix} \right\}.$$

Set up equation (10.1):

$$\alpha \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix} + \beta \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} + \gamma \begin{bmatrix} 1 \\ 3 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Then

$$\begin{aligned} 3\alpha + \beta + \gamma &= 0 \\ \alpha - \beta + 3\gamma &= 0 \\ 2\alpha + \beta &= 0 \end{aligned}$$

We can solve this system of equations using any of the usual methods, and find that this is the 'infinite number of solutions' case. Parametrically, the solution is $\alpha = t$, $\beta = -2t$, $\gamma = -t$.

The system has one parameter (or *one degree of freedom*). We just need one nonzero choice of t, so we'll choose $t = 1$ (that is, we only need the ratios $\alpha : \beta : \gamma$), so $\alpha = 1$, $\beta = -2$, and $\gamma = -1$. The point is that we've found values for the constants which are not all zero, so the vectors are linearly dependent.

Note We can often 'spot' values for the constants which make (10.1) true, which saves us from having to solve the linear equations formally.

We now formally state a remark from earlier:

Proposition 10.10 *A set $\{\mathbf{v}_1, \mathbf{v}_2, \dots \mathbf{v}_m\}$ of vectors in $\mathbb{R}^n$ is linearly dependent if and only if any one of the vectors may be written as a linear combination of the others.*

Basis vectors

In Definition 10.4 we introduced the concept of the **span** of a set of vectors in $\mathbb{R}^n$. Note (see Example 10.5) that in some cases the span is the entirety of $\mathbb{R}^n$, and in other cases we only get a subset of $\mathbb{R}^n$.

Definition 10.11 A set $S = \{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ of vectors in $\mathbb{R}^n$ is said to **span** or **generate** $\mathbb{R}^n$ if *every* vector in $\mathbb{R}^n$ can be expressed as a linear combination of the vectors in S.

In other words, for any $\mathbf{v} \in \mathbb{R}^n$ we can find scalars $\alpha_1, \dots, \alpha_m \in \mathbb{R}$ such that

$$\mathbf{v} = \alpha_1 \mathbf{v}_1 + \cdots + \alpha_m \mathbf{v}_2.$$

Example 10.12 The set $\{\mathbf{i}, \mathbf{j}\}$ spans $\mathbb{R}^2$, since if $\mathbf{v} = \begin{bmatrix} x \\ y \end{bmatrix}$ is any vector in $\mathbb{R}^2$, then we can write $\mathbf{v} = x\mathbf{i} + y\mathbf{j}$.

Similarly, the set $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$ spans $\mathbb{R}^3$, since we can write any vector

$$\mathbf{v} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}.$$

Example 10.13 The set $\left\{ \left[\begin{smallmatrix}3\\1\end{smallmatrix}\right], \left[\begin{smallmatrix}1\\2\end{smallmatrix}\right] \right\}$ also spans $\mathbb{R}^2$, but this time it isn't quite so evident. Suppose

$$\begin{bmatrix}x\\y\end{bmatrix} = \alpha \begin{bmatrix}3\\1\end{bmatrix} + \beta \begin{bmatrix}1\\2\end{bmatrix}.$$

We have to find α and β:

$$\begin{aligned} x &= 3\alpha + \beta, \\ y &= \alpha + 2\beta. \end{aligned}$$

Solving these for α and β gives

$$\alpha = \tfrac{1}{5}(2x - y), \quad \beta = \tfrac{1}{5}(3y - x),$$

so that any vector $\left[\begin{smallmatrix}x\\y\end{smallmatrix}\right]$ can be expressed in terms of the two given vectors:

$$\begin{bmatrix}x\\y\end{bmatrix} = \tfrac{1}{5}(2x - y) \begin{bmatrix}3\\1\end{bmatrix} + \tfrac{1}{5}(3y - x) \begin{bmatrix}1\\2\end{bmatrix},$$

and, for example, $\left[\begin{smallmatrix}2\\3\end{smallmatrix}\right] = \frac{1}{5}\left[\begin{smallmatrix}3\\1\end{smallmatrix}\right] + \frac{7}{5}\left[\begin{smallmatrix}1\\2\end{smallmatrix}\right]$.

Example 10.14 The vectors $\left\{ \left[\begin{smallmatrix}0\\27\end{smallmatrix}\right], \left[\begin{smallmatrix}2\\1\end{smallmatrix}\right], \left[\begin{smallmatrix}-1\\4\end{smallmatrix}\right] \right\}$ span $\mathbb{R}^2$, but not uniquely. As before, suppose

$$\begin{bmatrix}x\\y\end{bmatrix} = \alpha \begin{bmatrix}0\\27\end{bmatrix} + \beta \begin{bmatrix}2\\1\end{bmatrix} + \gamma \begin{bmatrix}-1\\4\end{bmatrix}$$

This yields the simultaneous equations

$$\begin{aligned} x &= 2\beta - \gamma \\ y &= 27\alpha + \beta + 4\gamma \end{aligned}$$

which solve to give

$$\alpha = y - \tfrac{1}{3}\beta + \tfrac{4}{27}x, \qquad \gamma = 2\beta - x$$

So this gives us infinitely many solutions

$$\begin{bmatrix}x\\y\end{bmatrix} = (\tfrac{1}{27}y + \tfrac{4}{27}x - \tfrac{1}{3}\beta) \begin{bmatrix}0\\27\end{bmatrix} + \beta \begin{bmatrix}2\\1\end{bmatrix} + (2\beta - x) \begin{bmatrix}-1\\4\end{bmatrix}$$

parametrised by the single variable β. (There are other valid sets of solutions parametrised by either α or γ.) So, for example,

$$\begin{bmatrix}2\\3\end{bmatrix} = (\tfrac{11}{27} - \tfrac{1}{3}\beta) \begin{bmatrix}0\\27\end{bmatrix} + \beta \begin{bmatrix}2\\1\end{bmatrix} + (2\beta - 2) \begin{bmatrix}-1\\4\end{bmatrix}.$$

Setting $\beta = 0$ yields $\alpha = \frac{11}{27}$ and $\gamma = -2$, and hence

$$\begin{bmatrix}2\\3\end{bmatrix} = \tfrac{11}{27} \begin{bmatrix}0\\27\end{bmatrix} - 2 \begin{bmatrix}-1\\4\end{bmatrix},$$

which is certainly valid, but setting $\beta = 1$ yields

$$\begin{bmatrix}2\\3\end{bmatrix} = \tfrac{2}{27} \begin{bmatrix}0\\27\end{bmatrix} + \begin{bmatrix}2\\1\end{bmatrix}.$$

In this last example, the three vectors are linearly dependent, whereas in the previous two examples the spaces are spanned by linearly independent vectors. Crucially, in the linearly independent case, the solutions we obtained for the simultaneous equations were unique, so any given vector $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right]$ admits a unique description as a linear combination of our chosen vectors. In the linearly dependent case, though, that description is not unique, but instead depends on one or more additional parameter.

It would be more useful for our purposes, certainly for defining coordinate systems, if we could rely on uniqueness in these circumstances. So, we give a linearly independent spanning set a special name:

Definition 10.15 A set $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ which spans $\mathbb{R}^n$ and is linearly independent is said to be a **basis** for $\mathbb{R}^n$.

The basis $\left\{\mathbf{i} = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right], \mathbf{j} = \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]\right\}$ is the **standard basis** for $\mathbb{R}^2$. Similarly, the basis

$$\left\{\mathbf{i} = \left[\begin{smallmatrix} 1 \\ 0 \\ 0 \end{smallmatrix}\right], \mathbf{j} = \left[\begin{smallmatrix} 0 \\ 1 \\ 0 \end{smallmatrix}\right], \mathbf{k} = \left[\begin{smallmatrix} 0 \\ 0 \\ 1 \end{smallmatrix}\right]\right\}$$

is the **standard basis** for $\mathbb{R}^3$.

Example 10.16 The set $\left\{\left[\begin{smallmatrix} 3 \\ 1 \end{smallmatrix}\right], \left[\begin{smallmatrix} 1 \\ 2 \end{smallmatrix}\right]\right\}$ is a basis for $\mathbb{R}^2$.

Example 10.17 Does the set $\left\{\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right], \left[\begin{smallmatrix} -2 \\ 3 \end{smallmatrix}\right]\right\}$ form a basis for $\mathbb{R}^2$?

We need to check that it spans $\mathbb{R}^2$ and is linearly independent.

Spanning Solve for a, b and c:

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

This is equivalent to solving

$$\begin{aligned} a - 2c &= x \\ b + 3c &= y \end{aligned}$$

for a and b.

This system has non-unique solutions (this can be seen by the fact it has more variables than equations). If we use the parameter t, then if $c = t$, $b = y - 3t$, $a = x + 2t$. We can set t to be anything we like, for example $t = 1$, in which case $a = x + 2$, $b = y - 3$, $c = 1$, and

$$\begin{bmatrix} x \\ y \end{bmatrix} = (x+2)\begin{bmatrix} 1 \\ 0 \end{bmatrix} + (y-3)\begin{bmatrix} 0 \\ + \end{bmatrix}\begin{bmatrix} -2 \\ 3 \end{bmatrix}, \tag{10.2}$$

so the set does span $\mathbb{R}^2$.

Linear independence This time we have to solve

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = a\begin{bmatrix} 1 \\ 0 \end{bmatrix} + b\begin{bmatrix} 0 \\ 1 \end{bmatrix} + c\begin{bmatrix} -2 \\ 3 \end{bmatrix}$$

Thus

$$\left.\begin{aligned} a - 2c &= 0 \\ b + 3c &= 0 \end{aligned}\right\} \Longrightarrow a = 2t, b = -3t, c = t,$$

and once again, choosing $t = 1$ say, we find

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

(To see this, just put $x = 0$ and $y = 0$ in (10.2) above.) The vectors are linearly dependent, so although they span the space $\mathbb{R}^2$, they do not form a basis.

In fact, any basis for $\mathbb{R}^2$ consists of two vectors, any basis for $\mathbb{R}^3$ consists of three vectors, and any basis for $\mathbb{R}^n$ will consist of exactly n vectors.

Dimension

It so happens that any basis for $\mathbb{R}^2$ consists of two vectors, any basis for $\mathbb{R}^3$ consists of three vectors, and more generally any basis for $\mathbb{R}^n$ will always have exactly n vectors. More formally:

Proposition 10.18 *Suppose $U = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ is one basis for $\mathbb{R}^n$ and that $W = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m\}$ is another basis for $\mathbb{R}^n$. Then $k = m = n$.*

Also:

Proposition 10.19

(i) *Every subset of $\mathbb{R}^n$ with more than n vectors is linearly dependent.*

(ii) *No subset of $\mathbb{R}^n$ with fewer than n vectors will span $\mathbb{R}^n$.*

In other words, the number of vectors in a basis for a vector space such as $\mathbb{R}^n$ is independent of any particular choice of basis: it's a fundamental property of the space itself, and happens to be equal to the geometric dimension of the space. In a more general context (which we won't really go into here) we can actually define the dimension of a given vector space in this way:

Definition 10.20 Let V be a finite dimensional vector space. The number of vectors in any basis for V is called the **dimension** of V, written $\dim V$.

11 Linear equations

In this section we study **equivalent matrices**, matrices which are related by finite sequences of certain operations, **similar matrices**, matrices which are conjugate to each other by some invertible transformation, and their applications to solving systems of simultaneous linear equations.

Simultaneous equations

Consider the following system of linear equations:

$$2x + 3y = 1 \tag{11.1}$$
$$5x + 7y = 3 \tag{11.2}$$

One method of solving this system is to multiply (11.1) by 5 to get

$$10x + 15y = 5 \tag{11.3}$$
$$5x + 7y = 3 \tag{11.4}$$

and multiply (11.2) by 2 to get

$$10x + 15y = 5 \tag{11.5}$$
$$10x + 14y = 6 \tag{11.6}$$

and then subtract (11.5) from (11.6) to get

$$10x + 15y = 5 \tag{11.7}$$
$$-y = 1 \tag{11.8}$$

and finally add 15×(11.8) to (11.7) to get

$$10x = 20 \tag{11.9}$$
$$-y = 1 \tag{11.10}$$

from which we can read off the solutions

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

as required. In solving this system, we have used two basic operations:

(i) Multiply an equation by a nonzero real number: $L_i \mapsto kL_i$.

(ii) Add a nonzero real multiple of one equation to another: $L_i \mapsto L_i + kL_j$.

Neither of these operations fundamentally alter the system of equations under investigation, in the sense that the solutions of the resulting system are the same as the solutions of the original system. A third operation which we didn't use in the example above, but which also doesn't fundamentally change the system, is:

(iii) Swap two equations: $L_i \leftrightarrow L_j$.

In practice, we will often combine the first two operations: $L_i \mapsto hL_i + kL_j$.

We can, in fact, write the original system of equations (11.1) and (11.2) in matrix form:

$$\begin{bmatrix} 2 & 3 \\ 5 & 7 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \tag{11.11}$$

The matrix formed from the coefficients is, reasonably, called the **coefficient matrix** of the system. For a system of m linear equations in n variables, this will be an $m \times n$ matrix. Instead, however, we can encode the system using a related object, the **augmented matrix** of the system:

$$\left[\begin{array}{cc|c} 2 & 3 & 1 \\ 5 & 7 & 3 \end{array}\right] \tag{11.12}$$

By rewriting a system of linear simultaneous equations in augmented matrix form, we have slightly changed the nature of the problem. Solving the original system required the application of three basic operations on equations, but having encoded the system as a matrix, we need to reformulate those three operations in the context of the rows of the augmented matrix.

Our aim, then, is to somehow reduce the augmented matrix to a simpler form that represents an equivalent system of equations which is more easily solved. To see this in action, consider the evolution of the augmented matrices corresponding to the above system of linear equations:

$$\left[\begin{array}{cc|c} 2 & 3 & 1 \\ 5 & 7 & 3 \end{array}\right] \longmapsto \left[\begin{array}{cc|c} 10 & 15 & 5 \\ 5 & 7 & 3 \end{array}\right] \longmapsto \left[\begin{array}{cc|c} 10 & 15 & 5 \\ 10 & 14 & 6 \end{array}\right]$$

$$\longmapsto \left[\begin{array}{cc|c} 10 & 15 & 5 \\ 0 & -1 & 1 \end{array}\right] \longmapsto \left[\begin{array}{cc|c} 10 & 0 & 20 \\ 0 & -1 & 1 \end{array}\right] \longmapsto \left[\begin{array}{cc|c} 1 & 0 & 2 \\ 0 & 1 & -1 \end{array}\right]$$

The fourth matrix in this chain (which is equivalent to equations (11.7) and (11.8)) is in a particularly useful form, because we can easily read off the solution $y = -1$ and then substitute it into the equation corresponding to the first line of the matrix to get the solution $x = 2$. The final matrix in the chain explicitly tells us the solutions of the original system of equations. Matrices which are in this particularly useful form have a special name, and we study them in generality next.

Echelon form

A matrix A is said to be in (**row**) **echelon form** if the number of zeros preceding the first nonzero entry of a row increases row-by-row until only zero rows remain. That is, if there exist nonzero entries

$$a_{1,j_1}, a_{2,j_2}, \ldots, a_{r,j_r} \quad \text{where } j_1 < j_2 < \cdots < j_r$$

with the property that

$$a_{i,j} = 0 \quad \text{for } i \leqslant r, j < j_i, \text{ and for } i > r$$

We call $a_{1,j_1}, \ldots, a_{r,j_r}$ the **distinguished elements** or **pivots** of the row echelon matrix A.

Example 11.1 The following matrices are in row echelon form, and the distinguished elements are in bold.

$$\begin{bmatrix} \mathbf{2} & 3 & 2 & 0 & 4 & 5 & -6 \\ 0 & 0 & \mathbf{7} & 1 & -3 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{6} & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} \mathbf{1} & 2 & 3 \\ 0 & 0 & \mathbf{4} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & \mathbf{1} & 3 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & \mathbf{1} & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{1} & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} \end{bmatrix}$$

In particular, an echelon matrix is called a **row reduced echelon matrix** or in **reduced echelon form** if the distinguished elements are

(i) the only nonzero entries in their respective columns, and
(ii) each equal to 1.

The third matrix in Example 11.1 is in reduced echelon form. Note also that the $m \times n$ zero matrix is also a row reduced echelon matrix.

Also, the fourth, fifth and sixth augmented matrices in the simultaneous equations example in the previous section are in row echelon form; in fact the last one is in reduced row echelon form. For the purposes of solving systems of linear simultaneous equations, then, it will make things much simpler if we can reduce the corresponding augmented matrix to row echelon (or preferably reduced row echelon) form.

So, we need to formulate a set of operations on the rows of a matrix with which we can reduce a given matrix to (reduced) row echelon form in a way that ensures the resulting matrices all correspond to systems of equations which are equivalent to (that is, have the same solutions as) the original system.

Elementary operations

A matrix A is said to be **row equivalent** to a matrix B if B can be obtained from A by a finite sequence of the following **elementary row operations**:

E$_1$ Interchange the ith row and the jth row: $R_i \leftrightarrow R_j$
E$_2$ Multiply the ith row by a nonzero scalar k: $R_i \mapsto kR_i$
E$_3$ Replace the ith row by k times the jth row plus the ith row: $R_i \mapsto kR_j + R_i$

In practice, we often apply E_2 and then E_3 in one step:

E Replace the ith row by h times the jth row plus (nonzero) k times the ith row: $R_i \mapsto hR_j + kR_i$

These operations are exactly the ones we want: they correspond to the allowed operations on systems of simultaneous linear equations.

Proposition 11.2 *Suppose A is the augmented matrix of a system L of simultaneous linear equations. If B can be obtained from A by a finite sequence of elementary operations of types E_1, E_2 and E_3, then the system K of simultaneous linear equations corresponding to B has the same solutions as L.*

Algorithm 11.3 (Reducing a matrix to row echelon form)

Step 1 Suppose the j_1 column is the first column with a nonzero entry. Interchange the rows so that this nonzero entry appears in the first row, that is, so that $a_{1,j_1} \neq 0$.

Step 2 For each $i > 1$ apply the operation

$$R_i \mapsto -a_{i,j_1} R_1 + a_{1,j_1} R_i$$

Repeat both these steps with the submatrix formed by all the rows excluding the first, until the matrix is in row echelon form.

Example 11.4 The following matrix A is reduced to echelon form by applying the operations $R_2 \mapsto -2R_1 + R_2$ and $R_3 \mapsto -3R_1 + R_3$ and then the operation $R_3 \mapsto -5R_2 + 4R_3$:

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

Applying the operations $R_1 \mapsto R_1 + \frac{3}{4}R_2$ and then $R_2 \mapsto R_2 - R_3$, and then the operations $R_2 \mapsto \frac{1}{4}R_2$ and $R_3 \mapsto \frac{1}{2}R_3$ converts A to reduced echelon form:

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

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

Proposition 11.2 ensures that elementary row operations don't affect the fundamental nature (that is, the solution set) of the system of linear equations represented by a matrix. The next proposition explores how these operations affect the determinant of a matrix.

Proposition 11.5 *Suppose A and B are square $n \times n$ matrices. Then:*

(i) *If A and B differ by an elementary row operation of type E_1: $R_i \leftrightarrow R_j$, then* $\det B = -\det A$.

(ii) *If A and B differ by an elementary row operation of type E_2: $R_i \mapsto kR_i$, then* $\det B = k \det A$.

(iii) *If A and B differ by an elementary row operation of type E_3, then* $\det B = \det A$.

In addition to the elementary row operations E_1, E_2 and E_3, we can also define the corresponding **elementary column operations**:

F$_1$ Interchange the ith column and the jth column: $C_i \leftrightarrow C_j$

F$_2$ Multiply the ith column by a nonzero scalar k: $C_i \mapsto kC_i$

F$_3$ Replace the ith column by k times the jth column plus the ith column: $C_i \mapsto kC_j + C_i$

In practice, we often apply F_2 and then F_3 in one step:

F Replace the ith column by h times the jth column plus (nonzero) k times the ith column: $C_i \mapsto hC_j + kC_i$

Unlike elementary row operations, elementary column operations are not useful for solving systems of simultaneous linear equations: if A and B are augmented matrices of systems of simultaneous linear equations, such that B may be obtained from A by a finite sequence of elementary column operations, it is not in general the case that the two systems will have identical solutions.

Elementary column operations do, however, have similar effects on determinants of square matrices as elementary row operations:

Proposition 11.6 *Suppose A and B are square $n \times n$ matrices.*

If A and B differ by an elementary column operation of type F_1: $C_i \leftrightarrow C_j$, then $\det B = -\det A$.

If A and B differ by an elementary column operation of type F_2: $C_i \mapsto kC_i$, then $\det B = k \det A$.

If A and B differ by an elementary column operation of type F_3, then $\det B = \det A$.

Definition 11.7 Two $n \times n$ square matrices A and B are said to be **similar** if there exists an invertible $n \times n$ square matrix P such that

$$A = P^{-1}BP.$$

We will study a particular class of similar matrices in Section 12, specifically those which are similar to a diagonal matrix. Note that similar matrices are not the same as row- or column-equivalent matrices.

Rank

In general, a system of simultaneous linear equations may fall into one of three categories, depending on the nature of its solutions (if any):

No solutions In this case, two or more of the individual equations are mutually inconsistent. For example, the system

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

has no consistent solutions. That is, there are no real values for the variables x, y and z which satisfy all three equations at the same time. Geometrically, we can interpret a linear equation in three variables as representing a plane in $\mathbb{R}^3$; in this case the planes do not all intersect at the same point.

One solution This is the case where there exists a single, unique solution to the system. For example, the system

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

has a single solution given by $x = y = z$. Geometrically, this corresponds to three planes intersecting at the single point $(1,1,1) \in \mathbb{R}^3$.

Infinitely many solutions The third case is that there may be infinitely many solutions. For example, solving the system

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

we find that any solution of the form $x = 0$, $y = 1 - z$ suffices. Geometrically, this corresponds to the case of three planes intersecting in more than a single point; in this specific example the planes intersect along the line $\{(x,y,z) : x = 0, y+z = 1\}$.

We can use the augmented matrix viewpoint to work out which category a given system of simultaneous linear equations falls into. The way we do this is by reducing the augmented matrix to reduced row echelon form and then counting the number of nonzero rows.

Definition 11.8 The **rank** of a matrix A is the number of nonzero rows when A is in row echelon form.

If this number is equal to the number of variables, then we have a single unique solution. If we have a row of the form $[0 \ \dots \ 0 \mid k]$ (where $k \neq 0$) then the system has no consistent solutions. If one or more nonzero rows have more than one nonzero number to the left of the vertical divider (such as $[0 \ \dots \ 0 \ 2 \ 1 \mid 3]$), then the system has an infinite number of solutions.

Equivalently, if $\operatorname{rank} A \neq \operatorname{rank} A'$ then the system has no consistent solutions, if $\operatorname{rank} A = \operatorname{rank} A'$ is equal to the number of variables, then the system has a single consistent solution, and if $\operatorname{rank} A = \operatorname{rank} A'$ is less than the number of variables, then the system has infinitely many solutions.

Example 11.9 The following matrix is in row echelon form and has rank 2.

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

Example 11.10 The following matrix is in (reduced) row echelon form and has rank 4.

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

12 Eigenvalues and eigenvectors

When we apply a map f, represented by matrix A, to $\mathbb{R}^2$, we expect points other than the origin to be mapped to new points in the plane. However, it may be that this is not always the case. Some straight lines may also remain fixed or invariant.

Scaling factors

Consider the matrix $A = \left[\begin{smallmatrix} 2 & 1 \\ 3 & 4 \end{smallmatrix}\right]$. Observe that

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

More generally,

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

So A maps any vector of the form $\left[\begin{smallmatrix} x \\ -x \end{smallmatrix}\right]$ to itself. All of these points lie on the line with equation $y = -x$.

Also,

$$\begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 5 \\ 15 \end{bmatrix} = 5\begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

And in general

$$\begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} x \\ 3x \end{bmatrix} = \begin{bmatrix} 5x \\ 15x \end{bmatrix} = 5\begin{bmatrix} x \\ 3x \end{bmatrix}.$$

So A also maps any point on the line $y = 3x$ to another point on the same line, specifically the one five times as far from the origin as the original point.

The lines $y = -x$ and $y = 3x$ are **invariant** (mapped to themselves) under the action of this matrix. Knowing these invariant lines and the associated scaling factors tells us pretty much everything about the geometric behaviour of the matrix transformation in question. We want to be able to find these invariant lines and scaling factors for any matrix.

Definition 12.1 Let A be a square $n \times n$ matrix. A vector $\mathbf{v} \in \mathbb{R}^n$ for which $A\mathbf{v} = \lambda\mathbf{v}$ for some scalar $\lambda \in \mathbb{R}$ is called an **eigenvector** of A, and λ is the associated **eigenvalue**.

We'll develop the theory for 2×2 matrices first, but it extends in an obvious way to $n \times n$ matrices. Consider a matrix $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$. We want to find the eigenvalues and corresponding eigenvectors.

First we write $A\mathbf{v} = \lambda\mathbf{v}$ in full:

$$\begin{aligned} ax + by &= kx \\ cx + dy &= ky \end{aligned}$$

Rearranging this gives

$$\begin{aligned} (a-\lambda)x + by &= 0 \\ cx + (d-\lambda)y &= 0 \end{aligned}$$

or, in matrix form,

$$\begin{bmatrix} a-\lambda & b \\ c & d-\lambda \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

We now have two homogeneous equations in the two variables x and y. Recall that these are always consistent; however there are two possibilities:-

(i) If the matrix on the left hand side has a nonzero determinant, then the equations have a unique solution $\mathbf{v} = \mathbf{0}$. *This is not the required solution to our problem* as we require $\mathbf{v} \neq 0$

(ii) If the determinant is zero, then the matrix is *not invertible* (rank is less than 2) and there are an infinite number of solutions. This is the case that we are going to investigate.

The determinant is

$$\chi_A = \det(A - kI) = (a-\lambda)(d-\lambda) - bc = \lambda^2 - (a+d)\lambda + (ad - bc)$$

which is a quadratic polynomial in λ, called the **characteristic polynomial** of A. The equation

$$\lambda^2 - (a+d)\lambda + (ad-bc) = 0,$$

is called the **characteristic equation** of A. The eigenvalues of A are the solutions of this equation.

Note that the sum of the eigenvalues is $(a+d)$, which is equal to the trace of A, and the product of the eigenvalues is $(ad-bc)$ which is the determinant of the original matrix A. Thus the characteristic equation for a 2×2 matrix is

$$\lambda^2 - \mathrm{tr}(A)\lambda + \det(A) = 0$$

(but the characteristic equation for 3×3 and larger matrices is slightly more complicated).

Having solved the characteristic equation and found the eigenvalues, we then have to find an eigenvector corresponding to each, by solving the equation

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix}$$

for x and y. Now if $\mathbf{v}_i$ is an eigenvector corresponding to eigenvalue λ_i, then so is $c\mathbf{v}_i$ for all nonzero constants c. So it follows that to each eigenvalue corresponds an infinite number of possible eigenvectors, all scalar multiples of each other. But this makes sense, because we're looking for the invariant lines of the matrix, and these are exactly the lines formed from all possible scalar multiples of a given eigenvector.

Example 12.2 Find the eigenvalues and eigenvectors for the matrix $A = \begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix}$.

The characteristic polynomial is $\chi_A = \det(A - kI) = \lambda^2 - 3\lambda - 4$. This factorises as $(\lambda - 1)(\lambda + 4)$, which means the roots are $\lambda = -1$ and $\lambda = -4$.

Considering $\lambda = -1$ first, we want to find vectors $\mathbf{v} = \begin{bmatrix} x \\ y \end{bmatrix}$ such that $A\mathbf{v} = -\mathbf{v}$. This yields the simultaneous equations

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

which we can solve to get $x = -y$. So any nonzero vector of the form $\begin{bmatrix} x \\ -x \end{bmatrix}$ is an eigenvector for A with eigenvalue -1, and we may choose a convenient representative: $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ will do.

Now considering $\lambda = 4$, we want to find vectors $\mathbf{v} = \begin{bmatrix} x \\ y \end{bmatrix}$ such that $A\mathbf{v} = 4\mathbf{v}$. As before, we get the equations

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

which give $2y = 3x$ or $y = \frac{3}{2}x$. Therefore any vector of the form $\begin{bmatrix} x \\ 3x/2 \end{bmatrix}$ will suffice, but for convenience we might as well pick one which avoids any untidy fractions, such as $\begin{bmatrix} 2 \\ 3 \end{bmatrix}$.

Geometrically, we now know that the invariant lines of the matrix A are $y = -x$ and $y = 3x/2$. But more than that, we know the scaling factors associated with these invariant lines. So A maps any point on the line $y = -x$ to its negative, the point the same distance from the origin but on the other side. And it maps any point on the line $y = 3x/2$ to another point four times further along in the same direction.

In this example, the characteristic polynomial had two distinct real roots. But this needn't always be the case. With a quadratic polynomial $ax^2 + bx + c$, the roots are given by the formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{12.1}$$

Depending on the sign of the **discriminant** $\Delta = b^2 - 4ac$, there are three possible outcomes:

(i) $\Delta > 0$: Two distinct real roots.

(ii) $\Delta = 0$: One real roots (or, equivalently, two identical real roots).

(iii) $\Delta < 0$: No real roots. (Actually, two complex roots, each a conjugate of the other.)

The third of these is beyond the scope of this course, so we'll only consider matrices with real eigenvalues (either distinct or repeated).

The other thing to note about this example is that we got one eigenvector per eigenvalue. More accurately, we got a single one-dimensional subspace (or **eigenspace**) of eigenvectors for each eigenvalue. This needn't always be the case:

Example 12.3 The matrix $A = \left[\begin{smallmatrix} 2 & 1 \\ 0 & 2 \end{smallmatrix}\right]$ has characteristic polynomial $\chi_A = \lambda^2 - 4\lambda + 4 = (\lambda - 2)^2$. This has a single (repeated) eigenvalue $\lambda = 2$, and it turns out (check this yourself) we can only find one eigenvector: $\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$.

But sometimes we can find more than one eigenvector corresponding to a repeated eigenvalue.

Some useful facts about eigenvalues and eigenvectors, that will be stated without proof.

The first one concerns the eigenvalues of the transpose of a matrix.

Proposition 12.4 *Let A be an $n \times n$ matrix. The transpose A^T has the same eigenvalues as A.*

There is a nice connection between the eigenvalues and the trace and determinant of a square matrix:

Proposition 12.5 *Let A be a real $n \times n$ matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$, some or all of which might be repeated. Then*

$$\begin{aligned} \lambda_1 + \cdots + \lambda_n &= \mathrm{tr}(A), \\ \lambda_1 \cdots \lambda_n &= \det(A). \end{aligned}$$

That is, the sum of the eigenvalues is equal to the trace, and the product of the eigenvalues is equal to the determinant.

An immediate consequence of this is that if one or more of the eigenvalues is zero, then the product of all the eigenvalues is zero too, and since this is equal to the determinant, the matrix must be singular. The converse holds as well: the only way the determinant can be zero is if at least one of the eigenvalues is zero.

Corollary 12.6 *An $n \times n$ matrix A is singular if and only if it has at least one eigenvalue equal to zero.*

Triangular and diagonal matrices

In general, to work out the eigenvalues and eigenvectors of a square matrix, we have to go through the procedure outlined in the previous section: calculate and solve the characteristic equation to get the eigenvalues, then use them to find the eigenvectors. But with triangular and diagonal matrices we can use a short cut:

Proposition 12.7 *If A is an $n \times n$ triangular or diagonal matrix, then its eigenvalues are the diagonal elements.*

Proof The characteristic polynomial $\det(A - kI)$ factorises neatly into the form $(\lambda - a_{11}) \ldots (\lambda - a_{nn})$, so the eigenvalues must be exactly the diagonal elements $a_{11}, \ldots, a_{nn}$. □

To illustrate and further justify this, we'll look at a couple of examples.

Example 12.8 The matrix $A = \left[\begin{smallmatrix} 2 & 0 \\ 0 & 3 \end{smallmatrix}\right]$ has characteristic polynomial $\chi_A = \det \left[\begin{smallmatrix} 2-\lambda & 0 \\ 0 & 3-\lambda \end{smallmatrix}\right] = (2-\lambda)(3-\lambda)$, which is already factorised to show that its roots (and hence the eigenvalues of A) are $\lambda = 2, 3$.

Example 12.9 The matrix $A = \left[\begin{smallmatrix} 2 & 1 \\ 0 & 1 \end{smallmatrix}\right]$ has characteristic polynomial $\chi_A = \det \left[\begin{smallmatrix} 2-\lambda & 1 \\ 0 & 1-\lambda \end{smallmatrix}\right] = (2-\lambda)(1-\lambda)$, which has roots $\lambda = 1, 2$.

Matrix diagonalisation

Remember that an $n \times n$ matrix A represents a linear map $f\colon \mathbb{R}^n \to \mathbb{R}^n$ relative to some choice of coordinate system. There are therefore (uncountably infinitely) many different ways of representing a given linear map with a matrix: we first have to choose a coordinate system (or basis) for $\mathbb{R}^n$ and then the entries of the matrix are determined.

Recall that two $n \times n$ matrices A and B are **similar** if there is an invertible matrix Q such that $A = QBQ^{-1}$.

Geometrically, similar matrices represent the same linear map relative to different coordinate systems (or bases). Another connection between similar matrices is given by the following proposition: similar matrices have the same eigenvalues.

Proposition 12.10 *If A and B are similar $n \times n$ matrices, then they have the same characteristic polynomial and the same eigenvalues.*

Now consider the matrix $\left[\begin{smallmatrix} 2 & 0 \\ 0 & 3 \end{smallmatrix}\right]$. Geometrically, this maps the plane $\mathbb{R}^2$ to itself, scaling everything by a factor of 2 parallel to the x–axis, and by a factor of 3 parallel to the y–axis. More generally, $\left[\begin{smallmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{smallmatrix}\right]$ scales everything by a factor of λ_1 horizontally, and by a factor of λ_2 vertically.

The matrix $A = \left[\begin{smallmatrix} 1 & 2 \\ 3 & 2 \end{smallmatrix}\right]$ scales everything by a factor of -1 along (and parallel to) the vector $\left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$, and by a factor of 4 along (and parallel to) the vector $\left[\begin{smallmatrix} 2 \\ 3 \end{smallmatrix}\right]$.

This matrix A represents a particular linear map relative to the standard basis. Now let's define a new basis consisting of the eigenvectors: $\left\{ \left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right], \left[\begin{smallmatrix} 2 \\ 3 \end{smallmatrix}\right] \right\}$. Relative to this new basis (check this yourself), the same linear map can be represented by the matrix $\left[\begin{smallmatrix} -1 & 0 \\ 0 & 4 \end{smallmatrix}\right]$.

The point of all this is that for many linear maps we can often find a suitable basis in which the corresponding matrix is diagonal. Or, phrased another way, many $n \times n$ matrix transformations are similar to diagonal matrices.

This is useful from a practical perspective, because diagonal matrices are often much easier to work with in calculations.

Definition 12.11 To **diagonalise** an $n \times n$ matrix A means to find a diagonal $n \times n$ matrix D and an invertible $n \times n$ matrix Q such that $A = QDQ^{-1}$.

From the above discussion, we expect D to have the eigenvalues of A as its diagonal elements. And from the earlier discussion about change of basis transformations, we expect Q to be formed from the new basis vectors. The following example demonstrates the general method.

Example 12.12 Given the transformation represented by the matrix

$$A = \begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix},$$

we now wish to analyse the transformation, by 'factorising' the matrix A.

The eigenvalues are 5 and 1 and the eigenvectors lie along the lines $y = 3x$ and $y = -x$ respectively, so essentially the matrix A produces scalings of amounts 5 and 1 along these skew lines – that is, not at right angles. Our strategy is to change coordinate axes so that we can apply the scaling matrix appropriately. We choose the u axis to be the line $y = 5x$ and the v axis to be $y = -x$. We can choose any point on the u axis to be a unit point, eg $(1, 3)$ and a unit point on the v axis to be $(1, -1)$, then the matrix Q is given by

$$Q = \begin{bmatrix} 1 & 1 \\ 3 & -1 \end{bmatrix}.$$

In order to change from (x, y) coordinates to (u, v) coordinates we multiply by Q^{-1}.

In the (u, v) plane we can now apply the appropriate scalings by multiplying by the diagonal matrix

$$D = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix},$$

and finally we change back to (x, y) coordinates by multiplying by Q.

The sequence is as follows: multiply by Q^{-1}, then by D, then by Q. In other words

$$A = QDQ^{-1}$$

or, equivalently,

$$Q^{-1}AQ = D.$$

We can't always do this, however. Construction of D is straightforward: we just need to find the eigenvalues of A, which (at least in principle) is just a matter of writing down and solving the characteristic equation. An $n \times n$ matrix will have n eigenvalues, and although some of them might be repeated, this won't be a problem. One potential problem is the existence of complex eigenvalues, but in this course we'll only be working with matrices with real eigenvalues so we can ignore that eventuality. (Complex eigenvalues don't actually stop us from diagonalising A, it's just that we end up with a complex diagonal matrix instead of a real one.)

But to construct Q we need exactly n eigenvectors (so that Q is an $n \times n$ matrix). We also need Q to be invertible, and it turns out that this is equivalent to requiring the eigenvectors of A to be linearly independent.

Proposition 12.13 *Let A be an $n \times n$ matrix. Then A is diagonalisable if and only if A has n linearly independent eigenvectors.*

If all the eigenvalues are distinct, this won't be a problem, since distinct eigenvalues have linearly independent eigenvectors:

Proposition 12.14 *If λ_1 and λ_2 are two distinct eigenvalues of an $n \times n$ matrix A, then their eigenvectors are linearly independent.*

With repeated eigenvalues this isn't necessarily the case (as we saw earlier) so if our matrix has repeated eigenvalues then it might not be diagonalisable.

But in general, to express an $n \times n$ matrix A with real eigenvalues in the form QDQ^{-1}, we use the following approach:

(i) Find the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A.

(ii) Define the matrix Λ by

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix}.$$

(iii) Find linearly independent eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$.

(iv) Define the matrix V by $V = \begin{bmatrix} \mathbf{v}_1 & \ldots & \mathbf{v}_n \end{bmatrix}$; that is, stack together the column vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ to form an $n \times n$ matrix. This matrix V is invertible.

(v) Then $A = V\Lambda V^{-1}$, so we can diagonalise A by setting $D = \Lambda$ and $Q = V$.

The choice of matrices D and Q aren't unique. We have $n!$ choices for what order to put the eigenvalues in when constructing Λ. And we have uncountably infinitely many choices of eigenvalues (since any nonzero scalar multiple of an eigenvalue is also an eigenvalue). But we have to make sure that whatever order we choose for the eigenvalues as the diagonal elements of Λ, we follow the same order when stacking the eigenvectors together to form V. As long as we do this, everything will work out fine. And if we multiply one or more of the eigenvectors by a nonzero scalar constant, then this will affect V, but it will also affect V^{-1} accordingly.

The main point of all this is that many problems in linear algebra involve finding higher powers of matrices: $A^2, A^3, \ldots, A^n$ and so on. In general, calculating higher powers of square matrices is cumbersome: there isn't an obvious short cut, you just have to do the requisite number of matrix multiplication operations.

But higher powers of diagonal matrices are an exception. In general, for any diagonal matrix D we have:

$$D = \begin{bmatrix} a_1 & 0 & \ldots & 0 \\ 0 & a_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a_n \end{bmatrix}^n = \begin{bmatrix} a_1^n & 0 & \ldots & 0 \\ 0 & a_2^n & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a_n^n \end{bmatrix}.$$

That is, to calculate the nth power of a diagonal matrix, we just have to calculate the nth power of each of the diagonal elements.

And if our matrix A is diagonalisable as $A = V\Lambda V^{-1}$ then $A^m = (V\Lambda V^{-1})^m = V\Lambda^m V^{-1}$. (This can be proved formally by induction.)

If we need to calculate A^{25}, say, then this simplifies things a lot: instead of doing 24 matrix multiplication operations, we just need to diagonalise A, calculate Λ^{25}, and then multiply V by Λ^{25} and V^{-1}.

13 Quadratic forms

Definition 13.1 A general **quadratic function** in n real variables is one of the form

$$\begin{aligned}&a_{11}x_1^2 + a_{22}x_2^2 + \cdots + a_{nn}x_n^2\\ &\qquad + 2a_{12}x_1x_2 + \cdots + 2a_{n-1,n}x_{n-1}x_n\\ &\qquad\qquad + b_1x_1 + b_2x_2 + \cdots + b_nx_n + c\end{aligned}$$

If we investigate the general nature of this function we find that it is not affected by the linear terms $b_1x_1 + b_2x_2 + \cdots + b_nx_n + c$. The remaining function

$$a_{11}x_1^2 + a_{22}x_2^2 + \cdots + a_{nn}x_n^2 + 2a_{12}x_1x_2 + \cdots + 2a_{n-1,n}x_{n-1}x_n,$$

all of whose terms are of degree 2, is called a **quadratic form**.

So, quadratic forms are functions $Q\colon \mathbb{R}^n \to \mathbb{R}$ of the form

$$\begin{aligned}Q(x) &= ax^2\\ Q(x,y) &= ax^2 + by^2 + 2cxy\\ Q(x,y,z) &= ax^2 + by^2 + cz^2 + 2dxy + 2exz + 2fyz\end{aligned}$$

and so on. We can write these in matrix form using a **symmetric matrix**:

$$\begin{aligned}Q(x,y) &= \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} a & c \\ c & b \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}\\ &= \begin{bmatrix} ax^2+by^2+2cxy \end{bmatrix}\\ Q(x,y,z) &= \begin{bmatrix} x & y & z \end{bmatrix} \begin{bmatrix} a & d & e \\ d & b & f \\ e & f & c \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}\\ &= \begin{bmatrix} ax^2+by^2+cz^2+2dxy+2exz+2fyz \end{bmatrix}\end{aligned}$$

Symmetric matrices

We can use nonsymmetric matrices too, but there are some advantages to using symmetric matrices.

Proposition 13.2 *The eigenvalues of a real symmetric matrix are all real.*

This means that we don't have to worry about the possibility of complex eigenvalues. (We've already decided not to worry about this for the scope of this module, but now we know we really don't have to worry about it.)

Proposition 13.3 *If A is a symmetric matrix, then any eigenvectors corresponding to distinct eigenvalues are orthogonal to each other.*

We know from earlier that distinct eigenvalues give linearly independent eigenvectors, but with a symmetric matrix we get something even better: orthogonal eigenvectors.

Proposition 13.4 *If $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset \mathbb{R}^n$ is an orthogonal (or orthonormal) set, then the $\mathbf{v}_i$ are linearly independent.*

Corollary 13.5 *Any n nonzero orthogonal (or orthonormal) vectors in $\mathbb{R}^n$ form a basis for $\mathbb{R}^n$.*

The canonical examples of these are the standard bases $\{\mathbf{i}, \mathbf{j}\}$ and $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$ for $\mathbb{R}^2$ and $\mathbb{R}^3$. It turns out that a diagonalising matrix formed from orthonormal eigenvectors is orthogonal:

Proposition 13.6 *If A is a real symmetric matrix, then it can be diagonalised by an orthogonal matrix. That is, there is a diagonal matrix D and an orthogonal matrix Q such that $A = QDQ^{-1} = QDQ^T$.*

Remember from the last section that not all matrices are diagonalisable. In particular, if a matrix doesn't have a full set of linearly independent eigenvectors (which might happen if we have a repeated eigenvalue) then we can't form the diagonalising matrix Q. But if our original matrix A is symmetric, then we don't have to worry: this proposition guarantees that, whether or not we have any repeated eigenvalues, we will still have enough linearly independent eigenvectors to form an invertible diagonalising matrix Q. Even better, we can find an orthogonal diagonalising matrix Q, which makes calculating the inverse $Q^{-1} = Q^T$ much easier.

To summarise, the advantages of using symmetric matrices are:

- their eigenvalues are real (no complex eigenvalues),
- they have **orthogonal** (not just linearly independent) eigenvectors corresponding to distinct eigenvalues,
- they can be diagonalised by means of an **orthogonal matrix**.

Classifying quadratic forms

For any quadratic form Q, clearly $Q(0, \ldots, 0) = 0$, but what happens when we use nonzero values of the variables $x, y, \ldots$?

Examples 13.7

- If $Q(x,y) = 3x^2 + 2y^2$ then $Q(x,y) > 0$ for all $(x,y) \neq (0,0)$.
- If $Q(x,y) = -3x^2 - 2y^2$ then $Q(x,y) < 0$ for all $(x,y) \neq (0,0)$.
- If $Q(x,y) = 3x^2 + 0y^2 = 3x^2$ then $Q(x,y) \geqslant 0$ for all $(x,y) \neq (0,0)$.
- If $Q(x,y) = -3x^2 + 0y^2 = -3x^2$ then $Q(x,y) \leqslant 0$ for all $(x,y) \neq (0,0)$.
- If $Q(x,y) = -3x^2 + 2y^2$ then $Q(x,y)$ can be < 0, $= 0$ or > 0 for some $(x,y) \neq (0,0)$.

We have five categories of quadratic forms:

Positive definite if $Q(x,y,\ldots) > 0$ for $(x,y,\ldots) \neq (0,\ldots,0)$.
Positive semidefinite if $Q(x,y,\ldots) \geqslant 0$ for $(x,y,\ldots) \neq (0,\ldots,0)$.
Negative definite if $Q(x,y,\ldots) < 0$ for $(x,y,\ldots) \neq (0,\ldots,0)$.
Negative semidefinite if $Q(x,y,\ldots) \leqslant 0$ for $(x,y,\ldots) \neq (0,\ldots,0)$.
Indefinite if $Q(x,y,\ldots)$ can be <0, $=0$ and >0 for values of $(x,y,\ldots) \neq (0,\ldots,0)$.

Sylvester's Criteria

We'll now look at two ways to classify quadratic forms into one of these five types. The first of these, **Sylvester's Criteria**, involves examining the signs of the form's **minors**: the determinants of certain submatrices of the coefficient matrix. This method is more convenient for matrices of functions, for example when studying the Hessian matrix to decide whether a given function is convex, concave, etc.

Let A be an $n \times n$ matrix.

Definition 13.8 A **minor** is the determinant of a submatrix of A, obtained by deleting equally many (or possibly no) rows and columns. The **order** of the minor is the order of the determinant: $1,\ldots,n$

Definition 13.9 A **principal minor** is a minor obtained by deleting only rows and columns of the same index. So, if row i is deleted, so is column i. Any number can be deleted (including zero, giving $\det A$) up to $(n-1)$, so there will be principal minors of all orders.

Definition 13.10 A **leading principal minor** is a principal minor obtained by deleting all rows and columns $j,\ldots,n$ for $j = 2,\ldots,n$; also $\det A$ is considered to be a leading principal minor.

These definitions are explained using the following 3×3 matrix. Suppose

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & k \end{bmatrix}$$

The minors of order 1 are

$$a, b, c, d, e, f, g, h, k;$$

the minors of order 2 are

$$\begin{vmatrix} a & b \\ d & e \end{vmatrix}, \quad \begin{vmatrix} a & c \\ d & f \end{vmatrix}, \quad \begin{vmatrix} b & c \\ e & f \end{vmatrix}, \quad \begin{vmatrix} a & b \\ g & h \end{vmatrix}, \ldots;$$

and the minor of order 3 is $\det A$ itself.

The principal minors of order 1 are

$$a, e, k;$$

the principal minors of order 2 are

$$\begin{vmatrix} a & b \\ d & e \end{vmatrix}, \quad \begin{vmatrix} a & c \\ g & k \end{vmatrix}, \quad \begin{vmatrix} e & f \\ h & k \end{vmatrix};$$

and the principal minor of order 3 is $\det A$ itself.

The leading principal minors are

$$a, \quad \begin{vmatrix} a & b \\ d & e \end{vmatrix}, \quad \det A.$$

We can use minors to determine the definiteness of a square matrix, such as the Hessian, or the coefficient matrix of a quadratic form:

Positive definite if *all leading principal minors* are strictly positive, (> 0). That is,

$$a_{11} > 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0, \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} > 0, \ldots, \det A > 0.$$

Negative definite if the leading principal minors satisfy the alternating pattern:

$$a_{11} < 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0, \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} < 0, \ldots$$

The last of these, $\det A$, will be positive or negative according to whether the order of the matrix A is even or odd.

Positive semidefinite if *all principal minors* are non-negative $(\geqslant 0)$.

Negative semidefinite if each principal minor of order k is either zero or has the same sign as $(-1)^k$.

Indefinite if $\det(A) \neq 0$ and the matrix is neither positive definite nor negative definite.

Diagonalisation

The second method we will look at uses matrix diagonalisation. This is more involved in some ways, and less useful for matrices of functions, but gives more information about the quadratic form in the process: it gives us an explicit change of variables in which the type of the form may be easily seen.

The quadratic form $Q(x, y) = 3x^2 + 2y^2$ corresponds to the diagonal matrix $\left[\begin{smallmatrix} 3 & 0 \\ 0 & 2 \end{smallmatrix}\right]$; it's easy to see that this is positive definite.

What about $Q(x, y) = 3x^2 + 4xy + 6y^2$? This corresponds to the non-diagonal symmetric matrix $\left[\begin{smallmatrix} 3 & 2 \\ 2 & 6 \end{smallmatrix}\right]$, and it's not so obvious which of the five categories Q belongs to.

If we **complete the square**, we get

$$\begin{aligned} Q(x, y) &= 3x^2 + 4xy + 6y^2 \\ &= 3\big(x + \tfrac{2}{3}y\big)^2 + \tfrac{14}{3}y^2 \\ &= \tfrac{7}{3}x^2 + 6\big(\tfrac{1}{3}x + y\big)^2 \qquad > 0 \qquad \text{for } (x, y) \neq (0, 0) \end{aligned}$$

so this form is positive definite.

The form

$$Q(x, y, z) = x^2 + 3y^2 + 9z^2 + 4xy + 6xz + 10yz$$

corresponds to the symmetric matrix $\left[\begin{smallmatrix} 1 & 2 & 3 \\ 2 & 3 & 5 \\ 3 & 5 & 9 \end{smallmatrix}\right]$. By completing the square, we get

$$Q(x, y, z) = (x + 2y + 3z)^2 - (y + z)^2 + z^2$$

which is indefinite. This method becomes fiddly and tedious with more variables, so we want a generally-applicable method which is easier.

Look at the quadratic form $Q(x,y) = 3x^2 + 4xy + 6y^2$. This corresponds to the matrix $\left[\begin{smallmatrix} 3 & 2 \\ 2 & 6 \end{smallmatrix}\right]$, which has eigenvalues 2 and 7 (the roots of its characteristic polynomial $k^2 - 9k + 14 = (k-2)(k-7)$). Its eigenvalues are (nonzero scalar multiples of)

$$\mathbf{v}_1 = \left[\begin{smallmatrix} 2 \\ -1 \end{smallmatrix}\right] \qquad\qquad \mathbf{v}_2 = \left[\begin{smallmatrix} 1 \\ 2 \end{smallmatrix}\right].$$

These are orthogonal, since

$$\mathbf{v}_1 \cdot \mathbf{v}_2 = 2 \times 1 + (-1) \times 2 = 0.$$

Note Eigenvectors corresponding to distinct eigenvalues of a matrix A will be linearly independent; if A is symmetric, then these eigenvectors will be orthogonal.

Eigenvectors corresponding to a repeated eigenvalue of a symmetric matrix A will be linearly independent but not orthogonal. For orthogonal diagonalisation we need a complete set of orthogonal (not just linearly independent) eigenvectors, so we must do a bit more work at this point. We won't go into the details here, but the basic idea is that for a repeated eigenvalue we don't just get a one-dimensional family of possible eigenvectors, but a two- or higher-dimensional subspace (or **eigenspace**) of eigenvectors. We then choose the required number of orthogonal eigenvectors from this subspace.

Having got a complete set of orthogonal eigenvectors, we must turn them into an orthonormal set by normalising them:

$$\|\mathbf{v}_1\| = \sqrt{5} \qquad \text{so } \hat{\mathbf{v}}_1 = \tfrac{1}{\sqrt{5}}\left[\begin{smallmatrix} 2 \\ -1 \end{smallmatrix}\right]$$

$$\|\mathbf{v}_2\| = \sqrt{5} \qquad \text{so } \hat{\mathbf{v}}_2 = \tfrac{1}{\sqrt{5}}\left[\begin{smallmatrix} 1 \\ 2 \end{smallmatrix}\right]$$

We now use these normalised eigenvectors (rather than the unnormalised ones) to make our 2×2 diagonalising matrix:

$$\hat{P} = \begin{bmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ -\frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{bmatrix} = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix}$$

This matrix $\hat{P}$ is orthogonal (check this by verifying that $\hat{P}\hat{P}^T = I = \hat{P}^T\hat{P}$). In fact, this process (using an orthogonal set of normalised eigenvectors, rather than just a linearly independent set as previously) will always yield an orthogonal matrix. So,

$$A = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix} = \frac{1}{5} \begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 7 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix} = \hat{P}D\hat{P}^T.$$

(Check this too.)

Before, when we diagonalised matrices corresponding to linear maps on vector spaces, what we were doing was changing into a new basis (coordinate system) where the linear map could be represented by a diagonal matrix. Similarly, when we diagonalise a quadratic form, we're choosing a new set of variables in which the quadratic form can be represented by a diagonal matrix.

So, the matrix $\hat{P}$ and its inverse $\hat{P}^{-1} = \hat{P}^T$ that we constructed in the above example can be thought of as representing a change of variables between the original variables x and y, and new variables X and Y in which Q is represented by the matrix $\left[\begin{smallmatrix} 2 & 0 \\ 0 & 7 \end{smallmatrix}\right]$ and hence has the form $2X^2 + 7Y^2$.

More precisely, $\hat{P}^T$ transforms from the original variables x, y to the new variables X, Y:

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \hat{P}^T \begin{bmatrix} x \\ y \end{bmatrix}$$

And $\hat{P}$ goes back the other way, transforming from the new variables X, Y into the original variables x, y:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \hat{P} \begin{bmatrix} X \\ Y \end{bmatrix}$$

For our worked example,

$$\begin{bmatrix} x \\ y \end{bmatrix} = \hat{P} \begin{bmatrix} X \\ Y \end{bmatrix} = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \frac{1}{\sqrt{5}} \begin{bmatrix} 2X + Y \\ 2Y - X \end{bmatrix}$$

and so

$$x = \frac{2X + Y}{\sqrt{5}} \qquad \text{and} \qquad \frac{2Y - X}{\sqrt{5}}.$$

Similarly,

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \hat{P}^T \begin{bmatrix} x \\ y \end{bmatrix} = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \frac{1}{\sqrt{5}} \begin{bmatrix} 2x - y \\ x + 2y \end{bmatrix}$$

and so

$$X = \frac{2x - y}{\sqrt{5}} \qquad \text{and} \qquad \frac{x + 2y}{\sqrt{5}}.$$

Substituting this into the diagonal form $Q(X, Y) = 2X^2 + 7Y^2$ represented by the diagonal matrix $D = \left[\begin{smallmatrix} 2 & 0 \\ 0 & 7 \end{smallmatrix}\right]$ we get

$$Q(x, y) = \tfrac{2}{5}(2x - y)^2 + \tfrac{7}{5}(x + 2y)^2.$$

(Check this by multiplying out the brackets and verifying that you get the original expression for $Q(x, y)$.)

This form is obviously positive definite.

The eigenvalue test

Given that we can easily determine the type of a diagonal quadratic form by looking at the coefficients of the terms $X^2, Y^2, \ldots$, and since those coefficients are exactly the eigenvalues of the original matrix, we can determine the type of the quadratic form just by calculating the eigenvalues. This is called the **eigenvalue test** and is as follows:

A quadratic form $Q(x, y, \ldots)$ represented by a symmetric $n \times n$ matrix A is:

- **Positive definite** if all the eigenvalues of A are positive.
- **Negative definite** if all the eigenvalues of A are negative.
- **Positive semidefinite** if all the eigenvalues of A are $\geqslant 0$ with at least one $= 0$.
- **Negative semidefinite** if all the eigenvalues of A are $\leqslant 0$ with at least one $= 0$.
- **Indefinite** if A has at least one positive eigenvalue and at least one negative eigenvalue.

III

Multivariate Calculus

14 Calculus and Optimisation

Now we want to extend and generalise the calculus of single-variable functions, in order to be able to study optimisation questions of multivariate functions. To start with, we need to generalise the concept of differentiation.

Partial differentiation

Recall that for a univariate function, the first derivative $f'(x)$ or $\frac{df}{dx}$ measures the rate of change of the value of $f(x)$ with respect to the change in x. That is, if we vary x by some small amount, how does that affect the value of $f(x)$?

Things are, as you might expect, more complicated with functions of more than one variable.

Suppose that we have a function $f\colon \mathbb{R}^2 \to \mathbb{R}$, where $f(x,y)$ is determined by some mathematical expression involving x and y. Then we might want to know how $f(x,y)$ varies with respect to either x or y. For example, if we have profit function $p(K,L)$ that depends on capital K and labour L, we may want to see how this varies depending on changes in either of those variables independently of the other.

That is, we keep one variable y fixed and see what happens to $f(x,y)$ as we vary x, or keep x fixed and see what happens when we vary y. Essentially, what we're proposing is to choose one variable and measure the variation in the value of our function, while treating all the other variables as constants.

Definition 14.1 Suppose $f\colon \mathbb{R}^n \to \mathbb{R}$, where $f(\mathbf{x}) = f(x_1, \ldots, x_n)$.

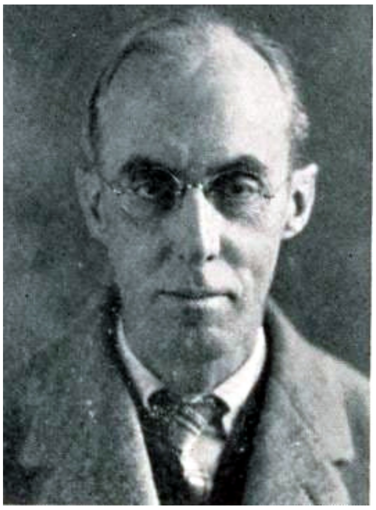
The **partial derivative** $\frac{\partial f}{\partial x_i}$ or f_{x_i} is the first-order derivative of the function $f(x_1, \ldots, x_n)$ with respect to x_i, with all the other variables x_j (for $j \neq i$) held constant.

As with ordinary differentiation, there are various forms of notation for partial differentiation. For a function $f\colon \mathbb{R}^n \to \mathbb{R}$, we can denote the partial derivative of f with respect to the variable x_i using modified versions of either the Leibniz or Lagrange notations.[1] Common forms are:

$$\frac{\partial f}{\partial x_i} \qquad f_{x_i} \qquad f'_{x_i} \qquad f_i \qquad f'_i$$

With the Leibniz-like notation, observe that we use curly ∂ symbols rather than an ordinary d.

[1] See page 12.



Charles Wiggins Cobb (1875–1949)

To illustrate this, we'll look at a couple of examples:

Example 14.2 Suppose that $f\colon \mathbb{R}^2 \to \mathbb{R}$ with

$$f(x,y) = x^2y + y^2x^2 + y^3 + 2x - 4.$$

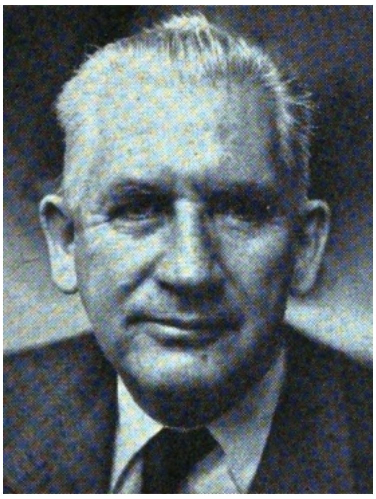
Then:

$$f_x = \frac{\partial f}{\partial x} = 2xy + 2xy^2 + 2, \qquad f_y = \frac{\partial f}{\partial y} = x^2 + 2x^2y + 3y^2$$

Example 14.3 Let $f(K,L) = K^\alpha L^\beta$ for $\alpha, \beta \in \mathbb{R}$ and $K, L > 0$.[2] Then:

$$f_K = \frac{\partial f}{\partial K} = \alpha K^{\alpha-1}L^\beta, \qquad f_L = \frac{\partial f}{\partial L} = \beta K^\alpha L^{\beta-1}$$

Definition 14.4 If a function $f\colon D \to \mathbb{R}$, where $D \subseteq \mathbb{R}^n$, has continuous partial derivatives of first order everywhere in D, then we say that f is **continuously differentiable**. In this case, f is said to be a C^1 **function**.



Paul Howard Douglas (1892–1976)

We can also define higher-order partial derivatives by partially differentiating the first-order derivatives, then partially differentiating again, and so on. But with multivariate functions we have more choices of variables to differentiate by.

Let's look at the two-variable case to start with. Suppose we have a function $f\colon \mathbb{R}^2 \to \mathbb{R}$ denoted $f(x,y)$. Then our first-order partial derivatives are

$$f_x = \frac{\partial f}{\partial x} \qquad \text{and} \qquad f_y = \frac{\partial f}{\partial y}.$$

We can define the second-order partial derivatives

$$f_{xx} = (f_x)_x = \frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}\frac{\partial f}{\partial x} \quad \text{and} \quad f_{yy} = (f_y)_y = \frac{\partial^2 f}{\partial y^2} = \frac{\partial}{\partial y}\frac{\partial f}{\partial y}.$$

But we can also differentiate partially with respect to the other variable, to get two other second-order derivatives

$$f_{xy} = (f_x)_y = \frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial y}\frac{\partial f}{\partial x} \quad \text{and} \quad f_{yx} = (f_y)_x = \frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial x}\frac{\partial f}{\partial y}.$$

William Henry Young (1863–1942)

This introduces a potential complication, in that we apparently need to distinguish between the two mixed second-order derivatives $f_{xy} = \frac{\partial^2 f}{\partial x \partial y}$ and $f_{yx} = \frac{\partial^2 f}{\partial y \partial x}$. In practice, however, we can safely gloss over this point, thanks to the following theorem:

Theorem 14.5 (Young's Theorem) *If a function $f\colon \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, then*

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

for $i \leqslant i, j \leqslant n$.

This theorem is usually attributed to the British mathematician William Henry Young, the German mathematician Hermann Schwartz,

[2] This is sometimes called a **Cobb–Douglas function** in economics.

or the French mathematician Alexis Clairaut, although the question had been studied by earlier mathematicians as far back as the early 18th century. In economics, we will usually work with relatively well-behaved functions that are at least twice continuously differentiable, so we can assume that the mixed second-order partial derivatives are equal.

Alexis Claude Clairaut (1713–1765)

Example 14.6 The function $f\colon \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \begin{cases} x^2 \sin\left(\frac{1}{x}\right) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

is not twice continuously differentiable. The first derivative $f'(x)$ exists everywhere, but it oscillates infinitely many times as $x \to 0$, so the second derivative doesn't exist at $x = 0$.

Karl Hermann Amandus Schwarz (1843–1921)

Gradients

For univariate functions, the first derivative measures the gradient of the graph of a function, and we would like to extend this idea to multivariate functions as well.

The partial derivative $\frac{\partial f}{\partial x_i}$ measures the rate of change of the value of f relative to the variable x_i; geometrically we can interpret this as the gradient of the cross-section through the graph of f, in the direction of the ith coordinate axis.

So, taken in combination, the first-order partial derivatives of f encode all the information about the rate of change of f with respect to its variables, or in a geometric sense, the gradient of the graph in all possible directions.

Definition 14.7 The **gradient** of a function $f\colon \mathbb{R}^n \to \mathbb{R}$, denoted ∇f, is the vector of partial derivatives:[3]

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

For example:

Example 14.8 Considering the Cobb–Douglas function $f(K, L) = K^\alpha L^\beta$, we have

$$\nabla f(K, L) = \left(\alpha K^{\alpha-1} L^\beta, \beta K^\alpha L^{\beta-1}\right).$$

What does this mean geometrically? The gradient vector $\nabla f(\mathbf{x})$ is perpendicular to the level curve of f passing through the point $\mathbf{x}$. It is the vector that points in the direction of steepest ascent (that is, the most rapid change of the value of $f(\mathbf{x})$) from the point $\mathbf{x}$.

The hyperplane tangent to the level curve at a point $\mathbf{a}$ is the set of all points $\mathbf{x}$ satisfying $\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{a} \rangle$.

Recall that for a univariate function $f\colon \mathbb{R} \to \mathbb{R}$, if we consider the input variable x to be itself a function of some other variable t (so

[3] The symbol ∇ is often pronounced 'del', although some older books refer to it as 'nabla', the latter after an ancient Greek or Phoenician harp of approximately triangular shape.

$x = x(t)$) then the **chain rule** tells us that

$$\frac{df}{dt} = \frac{df}{dx}\frac{dx}{dt}.$$

What is the analogous result for multivariate functions?

Definition 14.9 Suppose $f\colon \mathbb{R}^n \to \mathbb{R}$, with $f(\mathbf{x}) = f(x_1, \ldots, x_n)$, such that $x_1, \ldots, x_n$ are differentiable functions of some variable t. Then the **total derivative** is

$$\frac{df(\mathbf{x})}{dt} = \frac{\partial f}{\partial x_1}\frac{dx_1}{dt} + \cdots + \frac{\partial f}{\partial x_n}\frac{dx_n}{dt}.$$

Example 14.10 Consider the Cobb–Douglas function $f(K, L) = K^\alpha L^\beta$. Suppose that capital ($K$) and labour ($L$) both change with respect to time (t). Then

$$\frac{df}{dt} = \frac{\partial f}{\partial K}\frac{dK}{dt} + \frac{\partial f}{\partial L}\frac{dL}{dt} = \alpha K^{\alpha-1}L^\beta \frac{dK}{dt} + \beta K^\alpha L^{\beta-1}.$$

Now, suppose that $f\colon \mathbb{R} \to \mathbb{R}$ is a univariate, continuously differentiable function, and let $x^* \in \mathbb{R}$. Given δx small, we have

$$\begin{aligned} & \frac{\delta f(x^*)}{\delta x} \approx \frac{f(x^* + \delta x) - f(x^*)}{\delta x} \\ \implies \qquad & \frac{df(x^*)}{dx}\delta x \approx \underbrace{f(x^* + \delta x) - f(x^*)}_{\delta f(x^*)}. \end{aligned}$$

This tells us that we can approximate the change in $f(x)$ close to some value x^*, by taking the first derivative $\frac{df}{dx}$ at x^* and multiplying it by the small distance δx. This is essentially the rewritten form of the Mean Value Theorem[4] given in Corollary 5.14; equivalently it's the first degree approximation derived from Taylor's Theorem.[5]

In the case of a multivariate function $f\colon \mathbb{R}^n \to \mathbb{R}$, we use the **total differential** of f at $\mathbf{x}^*$, which is the following linear approximation:

$$\delta f(\mathbf{x}^*) \approx \left\langle \nabla f(\mathbf{x}^*), \delta\mathbf{x} \right\rangle = \frac{\partial f(\mathbf{x}^*)}{\partial x_1}\delta x_1 + \cdots + \frac{\partial f(\mathbf{x}^*)}{\partial x_n}\delta x_n \qquad (14.1)$$

(Here, $\delta\mathbf{x} = (\delta x_1, \ldots, \delta x_n)$.) This gives the total change in f, close to $\mathbf{x}^*$, due to the changes in $x_1, \ldots, x_n$.

Example 14.11 Consider the Cobb–Douglas function $f(K, L) = K^\alpha L^\beta$. Then the total differential is

$$\delta f(K^*, L^*) \approx \alpha(K^*)^{\alpha-1}(L^*)^\beta \delta K + \beta(K^*)^\alpha(L^*)^{\beta-1}\delta L.$$

For $\alpha = 2$, $\beta = 3$, $K^* = 2$, $L^* = 3$, $\delta K = 0.2$ and $\delta L = 0.1$ this gives

$$f(2.2, 3.1) - f(2, 3) \approx 2 \cdot 2^{2-1} \cdot 3^3 \cdot (0.2) + 3 \cdot 2^2 \cdot 3^{3-1} \cdot (0.1) = 32.4.$$

The exact value is

$$f(2.2, 3.1) - f(2, 3) = 2.2^2 \cdot 3.1^3 - 2^2 \cdot 3^3 = 36.18844,$$

so the approximation given by the total differential isn't too bad.

[4] Theorem 5.13, page 30.

[5] Theorem 4.2, page 21.

Homogeneous functions and Euler's Theorem

Suppose we have a function $f \colon \mathbb{R}^2 \to \mathbb{R}$ defined where $f(K, L)$ models the output of some process determined by capital (K) and labour (L). Sometimes, we may want to study what happens to the output if we increase the capital and labour by some constant multiple.

In order to study this sort of question, we introduce a special class of functions:

Definition 14.12 A function $f \colon D \to \mathbb{R}$, where $D \subseteq \mathbb{R}^n$, is said to be **homogeneous of degree** k if, for any $\mathbf{x} \in D$ and $t > 0$ we have

$$f(t\mathbf{x}) = t^k f(\mathbf{x}).$$

For example:

Example 14.13 Let $f \colon \mathbb{R}^2 \to \mathbb{R}$, such that $f(x, y) = 3x^2y - y^3$. This is homogeneous of degree 3, since:

$$\begin{aligned} f(tx, ty) = 3(tx)^2(ty) - (ty)^3 &= 3t^2x^2ty - t^3y^3 \\ &= t^3(x^2y - y^3) = t^3 f(x, y). \end{aligned}$$

Informally, every term in a degree–k homogeneous function has degree k. We can in some sense regard linear maps as homogeneous functions of degree 1.

Theorem 14.14 (Euler's Theorem) *Let $f \colon \mathbb{R}^n_+ \to \mathbb{R}$ be continuously differentiable and homogeneous of degree k. Then*

$$\langle \nabla f(\mathbf{x}), \mathbf{x} \rangle = k f(\mathbf{x}).$$

Proof Since f is homogeneous of degree k, we have

$$f(t\mathbf{x}) = t^k f(\mathbf{x}).$$

Differentiating with respect to t, the left-hand side becomes

$$\frac{\partial f(t\mathbf{x})}{\partial x_1} x_1 + \cdots + \frac{\partial f(t\mathbf{x})}{\partial x_n} x_n = \langle \nabla f(t\mathbf{x}), \mathbf{x} \rangle$$

and the right-hand side becomes

$$kt^{k-1} f(\mathbf{x}).$$

Setting $t = 1$, this becomes

$$\langle \nabla f(\mathbf{x}), \mathbf{x} \rangle = k f(\mathbf{x})$$

as claimed. □

Leonhard Euler (1701–1783)

Hicksian demand functions are homogeneous of degree 0.

The Hessian

In some sense, the gradient ∇f is the multivariate analogue of the first derivative of a univariate function. We now want to devise a multivariate analogue of the second derivative. This is the **Hessian matrix**, or often just the **Hessian**, which was originally developed by the German mathematician Otto Hesse. It is a matrix formed from all the second-order partial derivatives of the function in question.

Ludwig Otto Hesse (1811–1874)

Definition 14.15 Suppose that $f\colon \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable. Then the **Hessian** (or **Hessian matrix**) of f is the matrix

$$H_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} x_n \end{bmatrix} = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{nn} \end{bmatrix}$$

The entries of the Hessian are formulæ, and will typically assume different values at different points $\mathbf{x} \in \mathbb{R}^n$.

Thanks to Young's Theorem,[6] if f is twice continuously differentiable, then the Hessian H_f will be a symmetric matrix.

Example 14.16 Let $f(K, L) = K^\alpha L^\beta$. Then

$$\frac{\partial^2 f}{\partial K^2} = \alpha(\alpha-1)K^{\alpha-2}L^\beta \qquad \text{and} \qquad \frac{\partial^2 f}{\partial L^2} = \beta(\beta-1)K^\alpha L^{\beta-2},$$

and also

$$\frac{\partial^2 f}{\partial K \partial L} = \alpha\beta K^{\alpha-1}L^{\beta-1} = \frac{\partial^2 f}{\partial L \partial K}.$$

Thus the Hessian is

$$H_f = \begin{bmatrix} \alpha(\alpha-1)K^{\alpha-2}L^\beta & \alpha\beta K^{\alpha-1}L^{\beta-1} \\ \alpha\beta K^{\alpha-1}L^{\beta-1} & \beta(\beta-1)K^\alpha L^{\beta-2} \end{bmatrix}.$$

Example 14.17 Suppose that $f(x, y) = x^2y + x^2y^2 + y^3 + 2x - 4$. Then

$$\frac{\partial^2 f}{\partial x^2} = 2y + 2y^2, \quad \frac{\partial^2 f}{\partial y^2} = 2x^2 + 6y, \quad \frac{\partial^2 f}{\partial x \partial y} = 2x + 4xy = \frac{\partial^2 f}{\partial y \partial x}.$$

The Hessian is therefore

$$H_f = \begin{bmatrix} 2y + 2y^2 & 2x + 4xy \\ 2x + 4xy & 2x^2 + 6y \end{bmatrix}.$$

Stationary points

Now we want to use all of this to understand how to find and classify stationary points for multivariate functions, as we did for univariate functions.

Recall that for a univariate function $f\colon D \to \mathbb{R}$, a point $x^* \in D$ is a **stationary point** if the first order condition $f'(x^*) = 0$ holds. We want to extend this to multivariate functions $f\colon D \to \mathbb{R}$ where $D \subseteq \mathbb{R}^n$, and the way to do this is by using the first-order partial

[6] Theorem 14.5, page 94.

derivatives.

Definition 14.18 Let $f\colon D \to \mathbb{R}$, where $D \subseteq \mathbb{R}^n$. A point $\mathbf{x}^* \in D$ is a **stationary point** of f if $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

We're using the gradient ∇f as the analogue of f' here. This is equivalent to requiring that all the partial derivatives are simultaneously zero at the point $\mathbf{x}^*$ in question, since

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right) = (0, \ldots, 0) = \mathbf{0}$$

if and only if

$$\frac{\partial f}{\partial x_1} = 0, \quad \ldots \quad, \frac{\partial f}{\partial x_n} = 0.$$

Definition 14.19 Let $f\colon D \to \mathbb{R}$, where $D \subseteq \mathbb{R}^n$. A point $\mathbf{x}^*$ is a **local maximum** of f if, for all $\mathbf{x}$ in a neighbourhood (that is, an open ball $B_r(\mathbf{x}^*)$) of $\mathbf{x}^*$ we have $f(\mathbf{x}) \leqslant f(\mathbf{x}^*)$.

Similarly, a point $\mathbf{x}^*$ is a **local minimum** of f if, for all $\mathbf{x}$ in a neighbourhood of $\mathbf{x}^*$ we have $f(\mathbf{x}) \geqslant f(\mathbf{x}^*)$.

A **local extreme point** is a point which is either a local minimum or a local maximum.

The following is the multivariate analogue of Proposition 3.9, the **First Order Condition** (**FOC**):

Proposition 14.20 *For any function $f\colon D \to \mathbb{R}$, where $D \subseteq \mathbb{R}^n$, if $\mathbf{x}^*$ is a local extreme point, then it is a stationary point; that is, $\nabla f(\mathbf{x}^*) = \mathbf{0}$.*

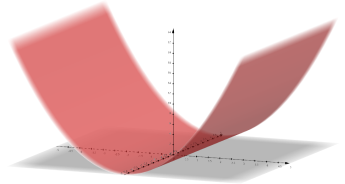


Figure 14.1: Graph of the function $f(x,y) = x^2$

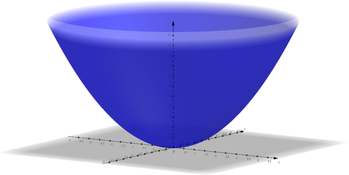


Figure 14.2: Graph of the function $f(x,y) = x^2 + y^2$

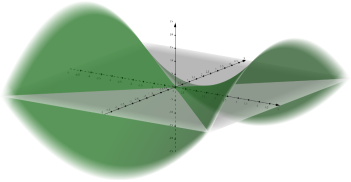


Figure 14.3: Graph of the function $f(x,y) = x^2 - y^2$

Figure 14.1 shows the graph of the function $f\colon \mathbb{R}^2 \to \mathbb{R}$ given by $f(x,y) = x^2$. This has a line of local minima along the y–axis. Figure 14.2 shows the function $f(x,y) = x^2 + y^2$, which has a single local minimum at $(0,0)$.

And Figure 14.3 shows the function $f(x,y) = x^2 - y^2$, which has a **saddle point** at $(0,0)$. This is an example of a point for which $\nabla f(\mathbf{x}^*) = \mathbf{0}$, but which is neither a local maximum nor a local minimum.

This last example demonstrates that, as in the univariate case, a stationary point doesn't have to be a local extreme point: the First Order Condition $\nabla f(\mathbf{x}^*) = \mathbf{0}$ isn't enough: it's a necessary but not sufficient condition.

Second Order Conditions

We want to find the appropriate multivariate Second Order Condition. Suppose that $f\colon D \to \mathbb{R}$, for $D \subseteq \mathbb{R}^n$ is a multivariate, twice continuously differentiable function. Let $\mathbf{x}$ and $\mathbf{x}{+}\mathbf{h}$ lie in some ball $B_r(\mathbf{x}) \subset \mathring{D}$.[7] Here, $\mathbf{h}$ is relatively short; that is, $\|\mathbf{h}\|$ is small.

The multivariate counterpart of $f'(x)$ is the gradient ∇f. What is the analogue of the second derivative $f''(x)$? It's the Hessian H_f.

The second-order Taylor expansion around $\mathbf{x}$ of f is:

$$f(\mathbf{x}{+}\mathbf{h}) = f(\mathbf{x}) + \mathbf{h}^T \nabla f(\mathbf{x}) + \tfrac{1}{2}\mathbf{h}^T H_f(\mathbf{x})\mathbf{h} + R_2(\mathbf{x}, \mathbf{h})$$

[7] Here, $\mathring{D}$ is the **interior** of D; that is, D without its boundary ∂D.

Some remarks:

- The term $\mathbf{h}^T \nabla f(\mathbf{x}) = \langle \mathbf{h}, \nabla f(\mathbf{x}) \rangle$.
- Compare the term $\frac{1}{2}\mathbf{h}^T H_f(\mathbf{x})\mathbf{h}$ with the discussion on quadratic forms.
- The term $R_2(\mathbf{x}, \mathbf{h})$ is the analogue of the remainder term in the univariate version of Taylor's Theorem. It behaves like a cubic (degree–3) polynomial in $\|\mathbf{h}\|$, with no constant, first- or second-order terms. That is, approximately equal to $b\|\mathbf{h}\|^3$ for some constant $b \in \mathbb{R}$.
- If $\mathbf{h}$ is short, then the remainder term becomes neglibly small, and can in practice be ignored.
- If, in addition, $\nabla f(\mathbf{x}) = \mathbf{0}$, we have
$$f(\mathbf{x}+\mathbf{h}) \approx f(x) + \tfrac{1}{2}\mathbf{h}^T H_f(\mathbf{x})\mathbf{h}$$
and this approximation is fairly precise.

We want to know the appropriate second-order conditions to determine if a stationary point is a local maximum or minimum. This will involve the Hessian H_f, as the multivariate analogue of the second derivative $f''(x)$.

The univariate Second Order Condition[8] depends on the sign of the second derivative f'' at the point in question. For multivariate functions, we look at the definiteness of the Hessian.

Recall that H_f is:

Negative semidefinite at $\mathbf{x}^*$ if $\mathbf{h}^T H_f(\mathbf{x}^*)\mathbf{h} \leqslant 0$ for all $\mathbf{h} \neq \mathbf{0}$, and
Positive semidefinite at $\mathbf{x}^*$ if $\mathbf{h}^T H_f(\mathbf{x}^*)\mathbf{h} \geqslant 0$ for all $\mathbf{h} \neq \mathbf{0}$.

Then we have the following **Second Order Condition** (**SOC**):

Proposition 14.21 *Let $f\colon D \to \mathbb{R}$ be twice continuously differentiable.*

(i) *If $\mathbf{x}^*$ is a local maximum, then $H_f(\mathbf{x}^*)$ is negative semidefinite.*
(ii) *If $\mathbf{x}^*$ is a local minimum, then $H_f(\mathbf{x}^*)$ is positive semidefinite.*

Furthermore, if $\mathbf{x}^* \in D$ is a stationary point:

(iii) *If $H_f(\mathbf{x}^*)$ is negative definite, then $\mathbf{x}^*$ is a local maximum.*
(iv) *If $H_f(\mathbf{x}^*)$ is positive definite, then $\mathbf{x}^*$ is a local minimum.*

The Hessian is also related to questions of convexity and concavity:

Proposition 14.22 *Let $f\colon D \to \mathbb{R}$ be twice continuously differentiable. Then:*

(i) *$H_f(\mathbf{x})$ is negative semidefinite for all $\mathbf{x} \in D$ if and only if f is concave.*
(ii) *If $H_f(\mathbf{x})$ is negative definite for all $\mathbf{x} \in D$, then f is strictly concave.*
(iii) *$H_f(\mathbf{x})$ is positive semidefinite for all $\mathbf{x} \in D$ if and only if f is convex.*
(iv) *If $H_f(\mathbf{x})$ is positive definite for all $\mathbf{x} \in D$, then f is strictly convex.*

Note that the Hessian of $-f$ is $H_{-f}(\mathbf{x}) = -H_f(\mathbf{x})$. To verify that f is concave, we can check whether $-H_f(\mathbf{x})$ is positive semidefinite.

[8] Proposition 3.10, page 15.

15 Constrained Optimisation

In economics, we often want to *maximise* some quantity (wealth, happiness, health outcomes, profit, etc) or *minimise* some other quantity (suffering, effort, loss, etc) and we want solutions that are the *best possible* ones in the circumstances; that is, subject to certain *constraints*. In this chapter, we will study an important and versatile method for doing this.

Lagrangian optimisation

We want to formulate a given problem as an *optimisation* problem, where we *maximise* an **objective function** subject to **constraints** that limit our choices.

One example problem might be as follows:

Example 15.1 Suppose we want to maximise the value of $f\colon \mathbb{R}^3 \to \mathbb{R}$ such that $f(x,y,z) = x^2 + y^2 + z^2$, subject to the constraints

$$\begin{cases} x + 2y + z & = & 30, \\ 2x - y - 3z & = & 10. \end{cases}$$

That is, we want to find the values of $x, y, z \in \mathbb{R}$ that give the maximum value of $f(x,y,z)$ such that the given linear equations are satisfied.

The following is a more general example:

Example 15.2 Suppose we have n goods with prices $p_1, \ldots, p_n$, and the quantity we buy of each good is given by variables $x_1, \ldots, x_n$. Furthermore, suppose we have some utility function $U(x_1, \ldots, x_n)$ modelling our preferences of the various bundles of goods we might purchases. We want to buy the correct quantity of each good in order to maximise U, but we only have a budget of £m.

We can state this as a **constrained optimisation** problem:

$$\max_{x_1,\ldots,x_n} U(x_1, \ldots, x_n) \quad \text{subject to} \quad p_1 x_1 + \cdots + p_n x_n \leqslant m.$$

We're now going to study methods of solving this kind of problem.

First of all, we need to know that a given problem actually has a solution. Fortunately, the multivariate version of the Extreme Value Theorem[1] helps set our minds at rest here:

[1] Theorem 5.11, page 29.

Theorem 15.3 (Extreme Value Theorem) *Let $f\colon D \to \mathbb{R}$ be a continuous function defined on a compact domain $D \subset \mathbb{R}^n$. Then f has both a global maximum and a global minimum in D.*

So, on a compact domain,[2] a maximisation problem *always* has a solution.

The following is a consequence of Propositions 14.21 and 14.22:

Proposition 15.4 *Let $f\colon D \to \mathbb{R}$ have a stationary point $\mathbf{x}^*$.*

(i) *If f is concave, then $\mathbf{x}^*$ is a global maximum.*

(ii) *If f is convex, then $\mathbf{x}^*$ is a global minimum.*

So, for a concave function, a solution to the First Order Conditions $\nabla f(\mathbf{x}^*) = \mathbf{0}$ is always a maximum.

The standard maximisation problem

Let $D \subset \mathbb{R}^n$ be convex, and consider a function $f\colon D \to \mathbb{R}$. We study the following problem:

$$\max_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad \begin{cases} g_1(\mathbf{x}) & \geqslant & 0, \\ & \vdots & \\ g_m(\mathbf{x}) & \geqslant & 0, \end{cases} \tag{15.1}$$

where $f, g_1, \ldots, g_m\colon D \to \mathbb{R}$ are all continuously differentiable inside D, and quasiconcave on D.

We call f the **objective function** and $g_1, \ldots, g_m$ the **constraint functions**.

Definition 15.5 The **Lagrangian** of the problem (15.1) is the function

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{j=1}^{m} \lambda_j g_j(\mathbf{x}).$$

The variables $\lambda_1, \ldots, \lambda_m$ are called **Lagrange multipliers**.

- In the standard maximisation problem (15.1) we have m different resources.
- The function g_j accounts for the stock of resource j, which is being **depleted** in the maximisation process.
- We denote the optimal point by $\mathbf{x}^*$.
- If we use the entire stock of resource j to reach $\mathbf{x}^*$, then $g_j(\mathbf{x}^*) = 0$ and we say the constraint is **binding**.
- If $g_j(\mathbf{x}^*) > 0$ then at the optimum $\mathbf{x}^*$ there is still some of resource j left, and we say the constraint is **slack** or **not binding**.

[2] Recall that, for our purposes, $D \subset \mathbb{R}^n$ is **compact** if it is **closed** (includes its boundary) and **bounded** (can be contained within a finite-radius ball).

The Karush–Kuhn–Tucker Conditions

The **Karush–Kuhn–Tucker** (or **KKT**) **Conditions** for the maximisation problem (15.1) are:

$$\frac{\partial \mathcal{L}(\mathbf{x},\boldsymbol{\lambda})}{\partial x_i} = \frac{\partial f(\mathbf{x})}{\partial x_i} + \sum_{j=1}^{m} \lambda_j \frac{\partial g_j(\mathbf{x})}{\partial x_i} = 0 \qquad \text{for } i = 1,\dots,n \qquad (15.2)$$

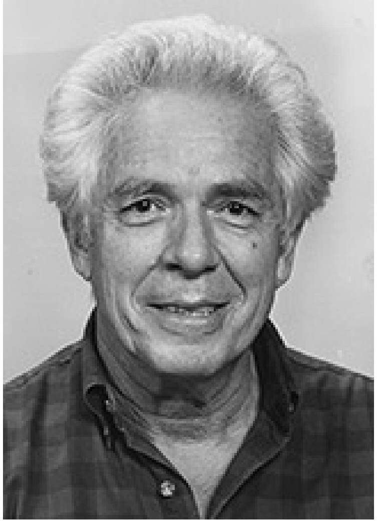
$$\frac{\partial \mathcal{L}(\mathbf{x},\boldsymbol{\lambda})}{\partial \lambda_j} = g_j(\mathbf{x}) \geqslant 0 \qquad \text{for } j = 1,\dots,m \qquad (15.3)$$

$$\lambda_j g_j(\mathbf{x}) = 0 \qquad \text{for } j = 1,\dots,m \qquad (15.4)$$

$$\lambda_j \geqslant 0 \qquad \text{for } j = 1,\dots,m \qquad (15.5)$$

William Karush (1917–1997)

The idea is that we use these conditions to derive a system of (possibly nonlinear) simultaneous equations, which we solve to find the optimal solution(s) for the problem. To illustrate this, let's look at an example:


Harold William Kuhn (1925–2014)

Example 15.6 Suppose we want to maximise the function

$$f(K,L) = K^{1/3}L^{1/3}$$

subject to the constraint

$$3K + 4L \leqslant 100.$$

This problem has two real variables $K, L > 0$ and a single constraint.

First of all, we have to turn our constraint into a function $g(K,L)$ such that $g(K,L) \geqslant 0$. We can do this by setting

$$g(K,L) = 100 - 3K - 4L \geqslant 0.$$

The Lagrangian of the problem is then

$$\mathcal{L}(K,L;\lambda) = K^{1/3}L^{2/3} + \lambda(100 - 3K - 4L).$$

We use this to write down the KKT conditions:

$$\begin{aligned}
\frac{\partial \mathcal{L}(K,L;\lambda)}{\partial K} &= \tfrac{1}{3}K^{-2/3}L^{2/3} - 3\lambda = 0 \\
\frac{\partial \mathcal{L}(K,L;\lambda)}{\partial L} &= \tfrac{2}{3}K^{1/3}L^{-1/3} - 4\lambda = 0 \\
\frac{\partial \mathcal{L}(K,L;\lambda)}{\partial \lambda} &= 100 - 3K - 4L \geqslant 0 \\
\lambda(100 - 3K - 4L) &= 0 \\
\lambda &\geqslant 0
\end{aligned}$$

In principle, we now solve this system of simultaneous equations and inequalities to find the optimal values K^* and L^*.

Albert William Tucker (1905–1995)

However, we haven't checked all the assumptions here: f and g must be continuously differentiable and quasiconcave for the KKT conditions to work. So there is no guarantee that the conditions in this example actually make sense.

Rationale for the KKT Conditions

Why do the KKT conditions work? Why do they ensure maximisation of the objective function subject to the various constraints? We'll look at each of them in turn:

(i) First let's look at condition (15.5); that is, $\lambda_j \geqslant 0$ for $j = 1, \ldots, m$. It can be shown that the response of the objective function f (evaluated at the optimum $\mathbf{x}^*$) to a small increase dy_j in the stock of resource j is $\frac{d}{df(\mathbf{x}^*)} y_j = \lambda_j$. So, increasing the stock of resource j by a small amount dy_j causes f to increase by $\lambda_j dy_j$.
A *rational agent* would be indifferent between staying at the optimum $\mathbf{x}^*$ or purchasing dy_j units of resource j for the price $p_j = \lambda_j$. So λ_j is called the **shadow price** of resource j.

(ii) Condition (15.4) says that $\lambda_j g_j(\mathbf{x}^*) = 0$. We know that if we use the entire supply of resource j to reach the optimum $\mathbf{x}^*$, then $g_j(\mathbf{x}^*) = 0$; that is, the constraint is **binding**.
We also know that the marginal prices of all the resources are non-negative (that is, $\lambda_j \geqslant 0$). If, in addition, $\lambda_j g_j(\mathbf{x}^*) = 0$, then every resource that isn't entirely used up (so $\lambda_j g_j(\mathbf{x}^*) > 0$ and the condition is *slack*) will have a marginal price equal to zero: $\lambda_j = 0$.
These are called **complementary slackness** conditions. They imply that, at the optimum $\mathbf{x}^*$,

$$\sum_{j=1}^{m} \lambda_j g_j(\mathbf{x}^*) = 0$$

and therefore $\mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = f(\mathbf{x}^*)$.

(iii) Condition (15.3) says that

$$\frac{\partial \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})}{\partial \lambda_j} = g_j(\mathbf{x}) \geqslant 0$$

for $j = 1, \ldots, m$. These are **inventory constraints**. They imply that you can't use more than a given resource than you actually have. They also ensure that the initial constraints hold.

(iv) Finally, the first condition, (15.2) says that

$$\frac{\partial \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})}{\partial x_i} = 0$$

for $i = 1, \ldots, n$. These are the first order conditions for the Lagrangian. However, we have already established that if the complementary slackness conditions hold, then the Lagrangian $\mathcal{L}$ is equal to the objective function f. So these conditions guarantee maximisation of f.

Technically, the solution to the maximisation problem (15.1) is $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$, a vector of both the xs and λs.

Importance of concavity

Why do we need the objective function to be concave? Well, if it is, the following result holds.

Proposition 15.7 *Let $f\colon D \to \mathbb{R}$ be concave. If there exists $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ satisfying the KKT conditions for the maximisation problem* (15.1)*, then $\mathbf{x}^*$ maximises* (15.1)*.*

If f is not concave, but it is at least quasiconcave, then we need to check another condition:

Proposition 15.8 *Let $f\colon D \to \mathbb{R}$ be quasiconcave. Then there exists $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ such that:*

(i) $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ *satisfies the KKT conditions for* (15.1)*, and*

(ii) $\nabla f(\mathbf{x}^*) \neq 0$,

then $\mathbf{x}^*$ maximises (15.1)*.*

Rewriting the problem

It may be necessary to rewrite the problem into the form of (15.1).

(i) If you need to minimise a quasiconvex objective function f, then maximising a quasiconcave function $-f$ yields the same outcome.

(ii) If g_j is quasiconvex and we have a constraint $g_j(\mathbf{x}) \leqslant 0$, then we can replace it with a (quasiconcave) constraint $h_j(\mathbf{x}) = -g_j(\mathbf{x}) \geqslant 0$.

(iii) If you have a constraint of the form $g_j(\mathbf{x}) \geqslant b$, where $b \neq 0$, then replace it with $h_j(\mathbf{x}) = g_j(\mathbf{x}) - b \geqslant 0$. Since vertical shifts don't affect gradients, we still have $\nabla g_j(\mathbf{x}) = \nabla h_j(\mathbf{x})$.

(iv) A constraint $g_j(\mathbf{x}) = 0$ can often be replaced with an inequality.
For example, suppose we have a budget constraint $\langle \mathbf{p}, \mathbf{x} \rangle = y$ (that is, you must spend all your income). This can be replaced by a constraint of the form $\langle \mathbf{p}, \mathbf{x} \rangle \leqslant y$ (that is, you can't spend more than your income).
Alternatively, you can replace a constraint $g_j(\mathbf{x}) = 0$ with two constraints

$$g_j(\mathbf{x}) \geqslant 0 \quad \text{and} \quad - g_j(\mathbf{x}) \geqslant 0.$$

If the initial constraint is linear, then so is $-g_j(\mathbf{x})$, and thus both concave and quasiconcave. Both constraints must bind, and the problem simplifies.

(v) If you have a constraint where $x_i \geqslant 0$, then you can introduce another constraint $g_{m+1}(\mathbf{x}) = x_i$.

KKT conditions for non-negative variables

If we know that our variables $x_1, \ldots, x_n$ are all non-negative, then we can use the following modified versions of the KKT conditions:

$$\frac{\partial \mathcal{L}(\mathbf{x},\boldsymbol{\lambda})}{\partial x_i} = \frac{\partial f(\mathbf{x})}{\partial x_i} + \sum_{j=1}^{m} \lambda_j \frac{\partial g_j(\mathbf{x})}{\partial x_i} \leqslant 0 \quad \text{for } i = 1, \ldots, n \tag{15.6}$$

$$\frac{\partial \mathcal{L}(\mathbf{x},\boldsymbol{\lambda})}{\partial \lambda_j} = g_j(\mathbf{x}) \geqslant 0 \quad \text{for } j = 1, \ldots, m \tag{15.7}$$

$$\lambda_j g_j(\mathbf{x}) = 0 \quad \text{for } j = 1, \ldots, m \tag{15.8}$$

$$x_i \frac{\partial \mathcal{L}(\mathbf{x},\boldsymbol{\lambda})}{\partial x_i} = 0 \quad \text{for } i = 1, \ldots, n \tag{15.9}$$

$$\lambda_j \geqslant 0 \quad \text{for } j = 1, \ldots, m \tag{15.10}$$

$$x_i \geqslant 0 \quad \text{for } i = 1, \ldots, n \tag{15.11}$$

(Alterations are highlighted in red.)

KKT conditions cookbook

To summarise all the above, this is the procedure for using the KKT conditions to solve an optimisation problem:

(i) Check that the domain $D \subseteq \mathbb{R}^n$ is convex.
(ii) Check that f (or $-f$, for a minimisation problem) is:
  (a) concave, or
  (b) quasiconcave (and in this case we also need $\nabla f(\mathbf{x}^*) \neq 0$).
(iii) Check that after you rewrite the initial conditions and define the constraints $g_1, \ldots, g_m$, they are all quasiconcave.
(iv) Check that f (or $-f$) and $g_1, \ldots, g_m$ are all continuously differentiable inside D.
(v) Write down the KKT conditions – either the original form, or the modified ones, or some mixture of the two if some variables are non-negative.

Example 15.9 Suppose that $\alpha \in (0, 1)$. Consider the maximisation problem

$$\max_{x,y} x^\alpha y^{1-\alpha} \quad \text{subject to} \quad \begin{cases} x + y & \leqslant & 1 \\ x & \leqslant & b \\ x & \geqslant & 0 \\ y & \geqslant & 0 \end{cases}$$

To start with, we need to define the constraints. The first two can be rewritten as

$$g_1(x, y) = 1 - x - y \geqslant 0$$
$$g_2(x, y) = b - x \geqslant 0$$

Now we need to check the assumptions:

(i) The set $D = \mathbb{R}^2$ is convex. (The same is true for $\mathbb{R}^2_+$ as well.)
(ii) The objective function is a Cobb–Douglas function with

exponents $\alpha + (1-\alpha) = 1$, and hence it is concave (and thus concave). To see this, check the definiteness of H_f.

(iii) The functions g_1 and g_2 are linear, and hence concave and quasiconcave.

(iv) The functions f, g_1 and g_2 are continuously differentiable inside D.

The Lagrangian is

$$\mathcal{L}(x, y; \lambda_1, \lambda_2) = x^\alpha y^{1-\alpha} + \lambda_1(1 - x - y) + \lambda_2(b - x)$$

Now we write down the KKT conditions. Since $x, y \geqslant 0$ we can use the modified versions:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial x} = \alpha\left(\tfrac{y}{x}\right)^{1-\alpha} - \lambda_1 - \lambda_2 &\leqslant 0 & \frac{\partial \mathcal{L}}{\partial y} = (1-\alpha)\left(\tfrac{x}{y}\right)^{\alpha} - \lambda_1 &\leqslant 0 \\
\frac{\partial \mathcal{L}}{\partial \lambda_1} = 1 - x - y &\geqslant 0 & \frac{\partial \mathcal{L}}{\partial \lambda_2} = b - x &\geqslant 0 \\
\lambda_1(1 - x - y) &= 0 & \lambda_2(b - x) &= 0 \\
x\frac{\partial \mathcal{L}}{\partial x} &= 0 & y\frac{\partial \mathcal{L}}{\partial y} &= 0 \\
\lambda_1 &\geqslant 0 & \lambda_2 &\geqslant 0 \\
x &\geqslant 0 & y &\geqslant 0
\end{aligned}$$

We now solve these to find the optimal values for x and y.

Exogenous parameters

In economics, we often want to solve optimisation problems that include certain parameters that we have no direct control over, such as prices, inflation, interest rates, exchange rates, tax rates, etc. These are called **exogenous parameters**.

We have a couple of questions to consider:

(i) How does the solution to the optimisation problem change in relation to changes in the underlying parameters?

(ii) How does the optimal value change when the parameters vary?

Both of these are related to stability.

Let $\mathbf{p} \in \mathbb{R}^l$ be a vector of parameters. Then our maximisation problem (15.1) becomes

$$\max_{\mathbf{x}} f(\mathbf{x}; \mathbf{p}) \quad \text{subject to} \quad \begin{cases} g_1(\mathbf{x}; \mathbf{p}) & \geqslant & 0 \\ & \vdots & \\ g_m(\mathbf{x}; \mathbf{p}) & \geqslant & 0 \end{cases} \tag{15.12}$$

Suppose we find a solution. It will typically depend on $\mathbf{p}$. Let:

- $\lambda^*(\mathbf{p}) = \lambda^*(\mathbf{x}(\mathbf{p}); \mathbf{p})$ denote the Lagrange multiplier associated with the maximum,
- $\mathbf{x}^*(\mathbf{p})$ and $\lambda^*(\mathbf{x}^*(\mathbf{p}); \mathbf{p})$ stand for the full solution of the problem,
- $V(\mathbf{p}) = f(\mathbf{x}^*(\mathbf{p}); \mathbf{p})$ represent the value of f at its solution determined by $\mathbf{p}$ – This is known as the **value function**.

Theorem 15.10 (The Envelope Theorem) *For any $k = 1, \dots, l$,*

$$\frac{\partial V(\mathbf{p})}{\partial p_k} = \left(\frac{\partial \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})}{\partial p_k}\right)\Bigg|_{\substack{\mathbf{x}=\mathbf{x}^*(\mathbf{p}),\\ \boldsymbol{\lambda}=\boldsymbol{\lambda}(\mathbf{x}^*(\mathbf{p}),\mathbf{p}).}}$$

If $n = m = l = 1$, this simplifies to

$$\frac{\partial V(p)}{\partial p} = \left(\frac{\partial f(x;p)}{\partial p} + \lambda \frac{\partial g(x;p)}{\partial p}\right)\Bigg|_{\substack{x=x^*(p),\\ \lambda=\lambda^*(x^*(p),p).}}$$

We will just prove the simpler case ($n = m = l = 1$) here:

Proof The KKT conditions require the solution to satisfy

$$\frac{\partial f(x^*(p);p)}{\partial x} + \lambda^* \frac{\partial g(x^*(p);p)}{\partial x} = 0$$

and hence

$$\frac{\partial f(x^*(p);p)}{\partial x} = -\lambda^* \frac{\partial g(x^*(p);p)}{\partial x}. \tag{15.13}$$

Now differentiate $V(p)$ with respect to p:

$$\frac{\partial V(p)}{\partial p} = \frac{\partial f(x^*(p);p)}{\partial x}\frac{\partial x^*(p)}{\partial p} + \frac{\partial f(x^*(p);p)}{\partial p}$$

Substituting (15.13) for $\frac{\partial f}{\partial x}$ we get

$$\frac{\partial V(p)}{\partial p} = -\lambda^* \frac{\partial g(x^*(p);p)}{\partial x}\frac{\partial x^*(p)}{\partial p} + \frac{\partial f(x^*(p);p)}{\partial p}. \tag{15.14}$$

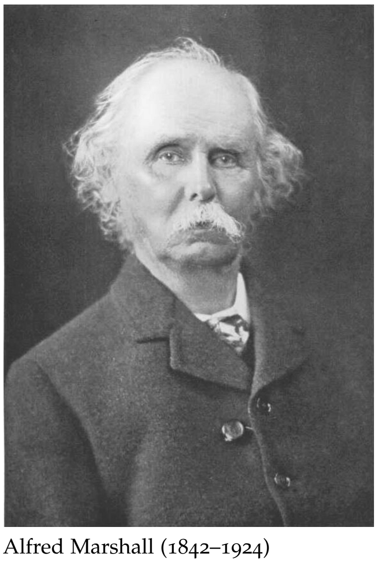
If $\lambda^* = 0$ then we have finished. If $\lambda^* > 0$ then $g(x^*(p), p) = 0$ (the constraint holds). Differentiate it with respect to p to get:

$$0 = \frac{\partial g(x^*(p);p)}{\partial x}\frac{\partial x^*(p)}{\partial p} + \frac{\partial g(x^*(p);p)}{\partial p}$$

and hence

$$\frac{\partial g(x^*(p);p)}{\partial x}\frac{\partial x^*(p)}{\partial p} = -\frac{\partial g(x^*(p);p)}{\partial p}.$$

Substitute this into (15.14) and the result follows. □



Alfred Marshall (1842–1924)

Example 15.11 We want to maximise a utility function $u(\mathbf{x})$ where $\mathbf{z}$ is a vector of prices, and y is income:

$$\max_{\mathbf{x}} u(\mathbf{x}) \quad \text{subject to} \quad y - \langle \mathbf{z}, \mathbf{x} \rangle \geqslant 0.$$

The vector of parameters is $\mathbf{p} = \left[\begin{smallmatrix}\mathbf{z}\\ y\end{smallmatrix}\right]$.

Let $\mathbf{x}^*(\mathbf{z}, y)$ be the solution (**Marshallian demand**) and $V(\mathbf{z}, y) = u(\mathbf{x}^*(\mathbf{z}, y))$ be the value function for this problem (indirect utility). How does the utility change when prices or income change?

To solve this problem, first we set up the Lagrangian:

$$\mathcal{L}(\mathbf{x}, \lambda; \mathbf{z}, y) = u(\mathbf{x}) + \lambda(y - \langle \mathbf{z}, \mathbf{x} \rangle).$$

Then we apply the Envelope Theorem to obtain:

$$\frac{\partial V(\mathbf{z},y)}{\partial z_i} = \left(\frac{\partial f(\mathbf{x};\mathbf{z},y)}{\partial z_i} + \lambda\frac{\partial g(\mathbf{x};\mathbf{z},y)}{\partial z_i}\right)\Bigg|_{\mathbf{x}=\mathbf{x}^*,\lambda=\lambda^*} = -\lambda^* x_i^*$$

and

$$\frac{\partial V(\mathbf{z},y)}{\partial y} = \left(\frac{\partial f(\mathbf{x};\mathbf{z},y)}{\partial y} + \lambda\frac{\partial g(\mathbf{x};\mathbf{z},y)}{\partial y}\right)\Bigg|_{\mathbf{x}=\mathbf{x}^*,\lambda=\lambda^*}$$

We combine these two to get Roy's Identity.

René François Joseph Roy (1894–1977)

Summary

- To obtain the derivative of a value function with respect to a given parameter, calculate the partial derivative of the Lagrangian $\mathcal{L}$ with respect to this parameter. Next evaluate it at the optimal $\mathbf{x}^*$ and $\boldsymbol{\lambda}^*$.
- If the complementary slackness conditions hold:

$$\lambda_j^* g_j(\mathbf{x};\mathbf{p}) = 0 \quad \text{for } j = 1,\ldots,m$$

  then the Lagrangian $\mathcal{L}$ at the optimum $\mathbf{x}^*$ will not depend on $\boldsymbol{\lambda}^*$. However, its partial derivatives might depend on $\boldsymbol{\lambda}^*$. These derivatives are equal to the partial derivatives of V.
- Interpretation: The Envelope Theorem is a direct impact theorem. At the maximum only, the direct impact of a change in parameter $\mathbf{p}$ matters. The indirect effect through $\mathbf{x}^*(\mathbf{p})$ cancels out. It matters how changes in price affects your budget, but its impact on optimal consumption is negligible.

16 Systems of Difference Equations

Earlier, we learned how to solve a single first-order difference equation. Now we'll develop a technique for solving systems of simultaneous first-order difference equations.

Matrix notation

Suppose we have a system

$$\begin{aligned} x_{t+1} &= ax_t + by_t + r \\ y_{t+1} &= cx_t + dy_t + s \end{aligned}$$

of coupled difference equations. That is, two sequences of real numbers (x_t) and (y_t) where each element in either sequence depends linearly on the previous element in both sequences, and also on some fixed constant. We can rewrite this in matrix form:

$$\begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} + \begin{bmatrix} r \\ s \end{bmatrix}.$$

This is a system of first order, linear, inhomogeneous difference equations.

It is first order because each variable depends only on the values of the variables in the previous time period. It is linear, because each variable is a linear combination of previous values. And it is inhomogeneous because at least some of the equations have an additional constant term.

We can write this in more compact form as

$$\mathbf{v}_{t+1} = A\mathbf{v}_t + \mathbf{b}$$

where A is the **coefficient matrix** of the system. If $\mathbf{b} = \mathbf{0}$ the system is **homogeneous**; if not, it's **inhomogeneous**.

We want to solve this by means of a (linear) change of variables.

Homogeneous difference equations

We will start by looking at the homogeneous case ($\mathbf{b} = \mathbf{0}$) and then progress to the inhomogeneous case later.

Diagonal coefficient matrix

Suppose we have a diagonal coefficient matrix

$$A = \begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix}.$$

Then the system we get is

$$\begin{aligned} x_{t+1} &= ax_t \\ y_{t+1} &= dx_t \end{aligned}$$

Here, the two sequences (x_t) and (y_t) are completely unrelated, so essentially we just have two separate homogeneous, first-order equations to solve. We know how to do this already.

More generally, if

$$\mathbf{v}_t = \begin{bmatrix} x_{1,t} \\ \vdots \\ x_{n,t} \end{bmatrix} \qquad \text{and} \qquad A = \begin{bmatrix} a_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & a_n \end{bmatrix}$$

then we have the following system:

$$\begin{aligned} x_{1,t+1} &= a_1 x_{1,t} \\ &\vdots \\ x_{n,t+1} &= a_n x_{n,t} \end{aligned}$$

Diagonalisable coefficient matrix

Now suppose we have a coefficient matrix A that is diagonalisable. That is, there exists some diagonal matrix D and invertible matrix Q such that $A = QDQ^{-1}$.

Then our system $\mathbf{v}_{t+1} = A\mathbf{v}_t$ can be rewritten as $\mathbf{v}_{t+1} = QDQ^{-1}\mathbf{v}_t$. Why is this helpful?

We want a closed form expression for $\mathbf{v}_t$, and by substitution we find that

$$\mathbf{v}_t = A\mathbf{v}_{t-1} = A(A\mathbf{v}_{t-2}) = \cdots = A(A(\cdots(A\mathbf{v}_0)\cdots)) = A^t\mathbf{v}_0.$$

If $A = QDQ^{-1}$, then $A^t = QD^tQ^{-1}$, and so

$$\mathbf{v}_t = QD^tQ^{-1}\mathbf{v}_0$$

which in principle is relatively straightforward to calculate.

Example 16.1 Consider the system

$$\begin{aligned} x_{t+1} &= 4x_t + 2y_t \\ y_{t+1} &= -x_t + y_t \end{aligned}$$

with $x_0 = 1$ and $y_0 = 0$.

The coefficient matrix of this system is $A = \left[\begin{smallmatrix} 4 & 2 \\ -1 & 1 \end{smallmatrix}\right]$, which has characteristic polynomial $\chi_A = k^2 - 5k + 6 = (k-2)(k-3)$.

Its eigenvalues are thus $k = 2$ and $k = 3$. The corresponding eigenvectors are $\left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 2 \\ -1 \end{smallmatrix}\right]$, which are linearly independent, and hence A is diagonalisable. So we can set

$$D = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}, \qquad Q = \begin{bmatrix} 1 & 2 \\ -1 & -1 \end{bmatrix}, \qquad Q^{-1} = \begin{bmatrix} -1 & -2 \\ 1 & 1 \end{bmatrix}.$$

Then $\mathbf{v}_t = A^t \mathbf{v}_0$, and so

$$\begin{aligned} \begin{bmatrix} x_t \\ y_t \end{bmatrix} &= \begin{bmatrix} 4 & 2 \\ -1 & 1 \end{bmatrix}^t \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 2 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} 2^t & 0 \\ 0 & 3^t \end{bmatrix} \begin{bmatrix} -1 & -2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} \\ &= \begin{bmatrix} -2^t + 2(3^t) & -2(2^t) + 2(3^t) \\ 2^t - 3^t & 2(2^t) - 3^t \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} \\ &= \begin{bmatrix} (-2^t + 2(3^t))x_0 + (-2(2^t) + 2(3^t))y_0 \\ (2^t - 3^t)x_0 + (2(2^t) - 3^t)y_0 \end{bmatrix} \end{aligned}$$

The general solution is therefore:

$$\begin{aligned} x_t &= (-2^t + 2(3^t))x_0 + (-2(2^t) + 2(3^t))y_0 \\ y_t &= (2^t - 3^t)x_0 + (2(2^t) - 3^t)y_0 \end{aligned}$$

The specific solution when $x_0 = 1$ and $y_0 = 0$ is:

$$\begin{aligned} x_t &= -2^t + 2(3^t) \\ y_t &= 2^t - 3^t \end{aligned}$$

More generally, suppose we have n difference equations, and an $n \times n$ diagonalisable coefficient matrix $A = QDQ^{-1}$ for some $n \times n$ invertible matrix Q and $n \times n$ diagonal matrix D.

Suppose that A has eigenvalues $k_1, \ldots, k_n$ (which might not all be distinct), so that

$$D = \begin{bmatrix} k_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & k_n \end{bmatrix}.$$

Let $\mathbf{w}_t = Q^{-1}\mathbf{v}_t$. This is our linear change of variables. Then the system $\mathbf{w}_{t+1} = D\mathbf{w}_t$ has solutions

$$\mathbf{w}_t = \begin{bmatrix} c_1 k_1^t \\ \vdots \\ c_n k_n^t \end{bmatrix}$$

where the constants $c_1, \ldots, c_n$ are the initial values of $\mathbf{w}_t$; that is,

$$\mathbf{w}_0 = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}.$$

Now we can solve the original system $\mathbf{v}_{t+1} = A\mathbf{v}_t$ by observing that $\mathbf{v}_t = Q\mathbf{w}_t$. Hence

$$\mathbf{v}_t = Q\mathbf{w}_t = Q \begin{bmatrix} c_1 k_1^t \\ \vdots \\ c_n k_n^t \end{bmatrix} = c_1 k_1^t \mathbf{u}_1 + \cdots + c_n k_n^t \mathbf{u}_n,$$

where $\mathbf{u}_i$ is the eigenvector of A corresponding to the eigenvalue k_i for $i = 1, \ldots, n$.

Furthermore, $\mathbf{v}_0 = Q\mathbf{w}_0$, so $\mathbf{w}_0 = Q^{-1}\mathbf{v}_0$ gives the values of the constants $c_1, \ldots, c_n$.

Applying this approach to Example 16.1 above, we have

$$\mathbf{w}_t = Q^{-1}\mathbf{v}_t = \begin{bmatrix} -1 & -2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} -x_t - 2y_t \\ x_t + y_t \end{bmatrix}.$$

Hence if $\mathbf{v}_0 = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$ we have

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \mathbf{w}_0 = \begin{bmatrix} -1 - 2(0) \\ 1 + 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

Then the solution is

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = (-1)(2^t) \begin{bmatrix} 1 \\ -1 \end{bmatrix} + (1)(3^t) \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} -2^t + 2(3^t) \\ 2^t - 3^t \end{bmatrix}$$

which is the solution we found above.

Inhomogeneous difference equations

Now we want to solve a system of the form

$$\mathbf{v}_{t+1} = A\mathbf{v}_t + \mathbf{b} \tag{16.1}$$

where A is diagonalisable and $\mathbf{b} \neq 0$.

As in the single equation case, we want a **particular solution** $\overline{\mathbf{v}}_t$. First we'll look for a **steady state solution** – one for which $\mathbf{v}_t = \mathbf{v}^*$ for all t; that is, a constant solution that doesn't change as t varies.

So let $\mathbf{v}^* = \mathbf{v}_t = \mathbf{v}_{t+1}$ and substitute into (16.1):

$$\mathbf{v}^* = A\mathbf{v}^* + \mathbf{b} \quad \Longrightarrow \quad (I - A)\mathbf{v}^* = \mathbf{b} \quad \Longrightarrow \quad \mathbf{v}^* = (I - A)^{-1}\mathbf{b}.$$

For this to work, we need the matrix $(I - A)$ to be invertible; that is, $\det(I - A) \neq 0$.[1] If $\det(I - A) = 0$ then the system doesn't have a steady-state solution. In this case, there are other things we can try: we can simplify the problem in some way, or we can try low-degree polynomials.[2]

When we've found a particular solution $\overline{\mathbf{v}}_t$ we can use the fact that the general solution to an inhomogeneous system of linear difference equations is the sum of:

(i) *any* particular solution to the system (such as a steady-state solution if one exists), and

(ii) the general solution to the corresponding homogeneous system $\mathbf{v}_{t+1} = A\mathbf{v}_t$.

We now know how to solve both of these.

Proof Suppose that

$$\mathbf{v}_{t+1} = A\mathbf{v}_t + \mathbf{b}.$$

Let $\overline{\mathbf{v}}_t$ be a particular solution of this, so

$$\overline{\mathbf{v}}_{t+1} = A\overline{\mathbf{v}}_t + \mathbf{b}.$$

[1] This is the analogue of the case $a \neq 1$ when we solved the single difference equation.

[2] Recall the solution $\overline{x}_t = bt$ in the single-equation case.

Let

$$\mathbf{u}_{t+1} = A\mathbf{u}_t$$

be the corresponding homogeneous system. Set $\mathbf{v}_t = \overline{\mathbf{v}}_t + \mathbf{u}_t$. Then

$$\begin{aligned} A\mathbf{v}_t + \mathbf{b} &= A(\overline{\mathbf{v}}_t + \mathbf{u}_t) + \mathbf{b} \\ &= (A\overline{\mathbf{v}}_t + \mathbf{b}) + A\mathbf{u}_t \\ &= \overline{\mathbf{v}}_{t+1} + \mathbf{u}_{t+1} \\ &= \mathbf{v}_{t+1} \end{aligned}$$

as claimed. □

The general solution to the system

$$\mathbf{v}_{t+1} = A\mathbf{v}_t + \mathbf{b}$$

is

$$\mathbf{v}_t = c_1 k_1^t \mathbf{u}_1 + \cdots + c_n k_n^t \mathbf{u}_n + (I - A)^{-1}\mathbf{b}$$

where $k_1, \ldots, k_n$ are the eigenvalues of the coefficient matrix A, and $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are the corresponding eigenvectors. If $(I - A)$ is singular (that is, it has zero determinant and is thus not invertible) then the steady state solution doesn't exist, so the last term above will need to be replaced by some other particular solution.

To summarise, the general method is as follows:

(i) Find any particular solution $\overline{\mathbf{v}}_t$ of the inhomogeneous system. (First try the steady state solution, if one exists.)

(ii) Calculate the eigenvalues $k_1, \ldots, k_n$ and eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ of the coefficient matrix A, and substitute them into

$$\mathbf{v}_t = c_1 k_1^t \mathbf{u}_1 + \cdots + c_n k_n^t \mathbf{u}_n + (I - A)^{-1}\mathbf{b}.$$

(iii) If we have initial conditions, we can calculate the coefficients $c_1, \ldots, c_n$ from

$$\mathbf{v}_0 = c_1 \mathbf{u}_1 + \cdots + c_n \mathbf{u}_n + \overline{\mathbf{v}}_0.$$

Stability

The linear system

$$\mathbf{v}_{t+1} = A\mathbf{v}_t + \mathbf{b}$$

is **globally asymptotically stable** if, for all initial values of $\mathbf{u}_0$, the corresponding homogeneous system

$$\mathbf{u}_{t+1} = A\mathbf{u}_t$$

converges to $\mathbf{0}$ as $t \to \infty$.

Proposition 16.2 *Consider the linear system*

$$\mathbf{v}_{t+1} = A\mathbf{v}_t + \mathbf{b}.$$

If every eigenvalue of A has an absolute value strictly less than 1, then the system is globally asymptotically stable, and for every $\mathbf{v}_0$ we have $\mathbf{v}_t \to (I - A)^{-1}\mathbf{b}$ as $t \to \infty$ if $(I - A)$ is invertible.

Lemma 16.3 *Let A be an $n \times n$ matrix with $\sum_{j=1}^{n} |a_{ij}| < 1$ for all $i = 1, \ldots, n$.*[3] *Then every eigenvalue of A has absolute value strictly less than* 1.

Example 16.4 Consider the system

$$\begin{aligned} x_{t+1} &= a y_t \\ y_{t+1} &= \tfrac{1}{2} x_t \end{aligned}$$

with $x_0 = y_0 = 1$. The coefficient matrix is $A = \left[\begin{smallmatrix} 0 & a \\ 1/2 & 0 \end{smallmatrix}\right]$ with characteristic polynomial $\chi_A = k^2 - \frac{1}{2}a = (k + \sqrt{a/2})(k - \sqrt{a/2})$. This system is stable if $0 \leqslant a < 2$.

Higher order difference equations

How do we solve a higher order difference equation? One method is to rewrite them as a system of first-order difference equations, and then use the approach we've just developed.

Example 16.5 The **Fibonacci Sequence** is the second-order homogeneous linear difference equation

$$x_{t+1} = x_t + x_{t-1}$$

with $x_0 = x_1 = 1$. We can rewrite this as a system of two first-order linear difference equations by introducing a new sequence y_t that is the original sequence shifted along by one.

Set $y_t = x_{t+1}$. So now we have a system

$$\begin{aligned} x_{t+1} &= y_t \\ y_{t+1} &= x_{t+2} = x_{t+1} + x_t = y_t + x_t \end{aligned}$$

This has coefficient matrix $A = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right]$.

Example 16.6 Consider the third-order inhomogeneous linear difference equation

$$x_{t+3} = 2x_{t+2} + 3x_{t+1} - x_t + 1.$$

Define $y_t = x_{t+1}$ and $z_t = y_{t+1} = x_{t+2}$. Then we get

$$\begin{aligned} x_{t+1} &= y_t \\ y_{t+1} &= z_t \\ z_{t+1} &= x_{t+3} = 2x_{t+2} + 3x_{t+1} - x_t + 1 \\ &= -x_t + 3y_t + 2z_t + 1 \end{aligned}$$

which can be written in matrix form as

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}_{t+1} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 3 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}_t + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

and solved using the method discussed earlier.

[3] Here, a_{ij} is the (i, j)th element of A.

17 *Differential Equations*

Differential equations often turn up in economics, and can be regarded as a continuous analogue of difference equations, for when we want to model some process as time varies continuously rather than discretely.

Definitions

A **differential equation** is an equation involving first or higher derivatives. The **order** of a differential equation is the highest order of derivative included in the equation.

Example 17.1 The equation

$$\frac{dx}{dt} + 7x^2 = e^{2t}$$

is a first-order differential equation, while

$$\frac{d^2y}{dx^2} + 3\frac{dy}{dx} - 4y = \sin(3x)$$

is a second-order differential equation.

These are **ordinary differential equations** (sometimes abbreviated as **ODEs**) because they only involve ordinary differentiation; **partial differential equations** (**PDEs**) include partial derivatives, and are beyond the scope of this module.

Example 17.2 Examples of partial differential equations include **Laplace's Equation**

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0$$

and the 1–dimensional **wave equation**

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2},$$

both of which are relevant in physics.

Pierre-Simon, Marquis de Laplace (1749–1827)

A differential equation is **linear** if it has only linear terms in the main variable and its derivatives.

Example 17.3 The equations

$$\begin{aligned} 3\frac{dy}{dx} + 2y &= e^x \\ 3t\frac{dx}{dt} + 2\sin(t)x &= e^t \\ \frac{d^2y}{dx^2} - 3\frac{dy}{dx} - 4y &= \sin(3x) \end{aligned}$$

are linear, while

$$\left(\frac{dy}{dx}\right)^2 + \cos(y) = e^x$$

is nonlinear.

We will learn how to solve three common types of differential equation: first order separable equations, first order linear equations, and second order linear equations with constant coefficients.

First-order separable equations

Definition 17.4 An ordinary differential equation of the form

$$\frac{dx}{dt} = f(x)g(t),$$

where f is a function of x and g is a function of t, is said to be **separable**.

Given a separable equation

$$\frac{dx}{dt} = f(x)g(t),$$

we can rearrange it to give

$$\frac{1}{f(x)}\frac{dx}{dt} = g(t).$$

Integrating both sides with respect to t, we get

$$\int \frac{1}{f(x)}\frac{dx}{dt}\,dt = \int g(t)\,dt.$$

The left hand side of this can be rewritten using the chain rule:

$$\int \frac{1}{f(x)}\,dx = \int f(x)g(t)\,dt.$$

If possible, we solve both of these integrals, being careful not to forget the constant of integration, and attempt to rearrange the resulting equation to give an expression for x in terms of t. This is the **general solution** to the equation.

If we then have initial conditions $x = x_0$ when $t = t_0$, we can then substitute these in to eliminate arbitrary constants and find the **specific solution** corresponding to those **initial conditions** or **boundary conditions**.

However, if $f(x_0) = 0$ then the specific solution will be a constant function $x(t) = c$ for some particular $c \in \mathbb{R}$.

We'll illustrate this with a worked example.

Example 17.5 Consider the equation

$$\frac{dx}{dt} + 2x^2 t = 0$$

We can rearrange this to

$$\frac{dx}{dt} = -2x^2 t$$

and then to

$$-\frac{1}{2x^2}\frac{dx}{dt} = t.$$

Now we can integrate both sides with respect to t:

$$\begin{aligned} -\frac{1}{2}\int x^{-2}\frac{dx}{dt}\,dt &= \int t\,dt \\ \implies \qquad -\frac{1}{2}\int x^{-2}\,dx &= \int t\,dt \end{aligned}$$

which gives

$$\tfrac{1}{2}x^{-1} = \tfrac{1}{2}t^2 + c$$

which we can simplify and rearrange to give

$$x = \frac{1}{t^2 + b}.$$

Here, b is an arbitrary constant. Any solution of this form will satisfy the original equation (check this), so we now have a family of solutions, each determined by some constant b. This is the **general solution** of the equation. Some example solutions are shown in Figure 17.1.

Now suppose that we have the initial conditions $x = -\frac{1}{2}$ when $t = 0$. We substitute these values into the general solution and solve for b, to get $b = -2$, and the specific solution

$$x = \tfrac{1}{t^2 - 2}.$$

This is the function whose graph passes through the point $(t, x) = (0, -\frac{1}{2})$; it is the orange curve in Figure 17.1.

We must also consider something else. Given $\frac{dx}{dt} = -2x^2 t$, what happens when $x = 0$? In that case, the equation becomes $\frac{dx}{dt} = 0$, which means that x is a constant ($x = 0$) for all $t \in \mathbb{R}$.

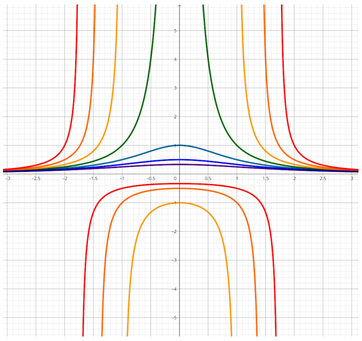


Figure 17.1: Graphs of some solutions for the differential equation in Example 17.5

First order linear differential equations

We'll start with an example.

Example 17.6 Consider the equation

$$x^2\frac{dy}{dx} + 2xy = b(x)$$

for some function $b(x)$. How do we solve this? We want an expression for y as a function of x, but it's not immediately obvious how we'd find one. But notice that

$$\frac{d}{dx}(x^2y) = 2xy + x^2\frac{dy}{dx},$$

which happens to be the left hand side of the original equation, which we can then rewrite as

$$\frac{d}{dx}(x^2y) = b(x).$$

Then integrate both sides with respect to x, to get

$$x^2y = \int b(x)\,dx + c$$

and hence

$$y = \frac{1}{x^2}\int b(x)\,dx + \frac{c}{x^2}.$$

This is the general solution of the equation.

Now let's try solving the equation

$$x^2\frac{dy}{dx} + 2xy = \cos(x)$$

where $y = 1$ when $x = \pi$.

Using the method above, we get the general solution

$$y = \frac{c}{x^2} + \frac{\sin(x)}{x^2}.$$

Some examples are shown in Figure 17.2. Substituting in $y = 1$ and $x = \pi$ and solving for c we get $c = \pi^2$, so the specific solution to this problem is

$$y = \frac{\pi^2 + \sin(x)}{x^2}.$$

This is the solution that passes through the point $(x, y) = (\pi, 1)$, and is the purple graph in Figure 17.2.

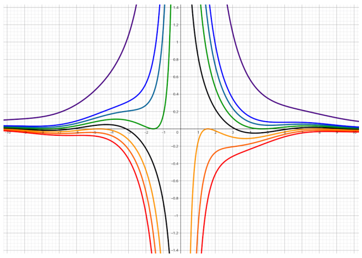


Figure 17.2: Graphs of some solutions for the differential equation in Example 17.6

In this case we were fortunate that the left hand side of the equation happened to be the derivative of x^2y via the product rule for differentiation. This won't always be the case, so we want a more general and widely-applicable method for solving equations of this type.

The idea is that given a first order linear differential equation

$$\frac{dy}{dx} + a(x)y = b(x) \qquad (17.1)$$

we want to rewrite the left hand side as $\frac{d}{dx}(c(x)y)$ for some function $c(x)$. In general, the equation (17.1) won't be in this form already, so we want an **integrating factor**: a function $c(x)$ such that

$$c(x)\frac{dy}{dx} + a(x)c(x)y = \frac{d}{dx}(c(x)y). \tag{17.2}$$

By the product rule for differentiation:

$$\frac{d}{dx}(c(x)y) = \frac{dc}{dx}y + c\frac{dy}{dx}$$

Setting this equal to the left hand side of (17.2) we get

$$c\frac{dy}{dx} + acy = \frac{dc}{dx}y + c\frac{dy}{dx}.$$

So we need a function $c(x)$ such that $\frac{dc}{dx} = ca$. If we set $c(x) = e^{\int a(x)\,dx}$ then

$$\frac{dc}{dx} = a(x)e^{\int a(x)\,dx} = a(x)c(x)$$

as required. This function

$$c(x) = e^{\int a(x)\,dx} \tag{17.3}$$

is the **integrating factor**.

Example 17.7 We want to solve the equation

$$\frac{dy}{dx} + xy = x$$

with $y = 3$ when $x = 0$.

The integrating factor is $c(x) = e^{\int x\,dx} = e^{x^2/2}$. So our original equation becomes

$$\begin{aligned}
& e^{x^2/2}\frac{dy}{dx} + xe^{x^2/2}y = xe^{x^2/2} \\
\implies \quad & \frac{d}{dx}(e^{x^2/2}y) = xe^{x^2/2} \\
\implies \quad & e^{x^2/2}y = \int xe^{x^2/2}\,dx = e^{x^2/2} + k \\
\implies \quad & y = 1 + ke^{-x^2/2}
\end{aligned}$$

This is the general solution, and some examples are shown in Figure 17.3. To find the specific solution, we substitute $y = 3$ and $x = 0$, then solve for k:

$$3 = 1 + k$$

Hence $k = 2$ and the specific solution is

$$y = 1 + 2e^{x^2/2}.$$

This is the pink graph in Figure 17.3

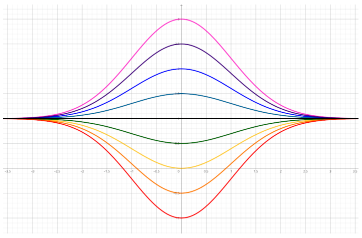


Figure 17.3: Graphs of some solutions to the differential equation in Example 17.7

So, to summarise:

(i) Rearrange the equation into the form $\frac{dy}{dx} + a(x)y = b(x)$ shown in (17.1).

(ii) Calculate the integrating factor $c(x) = e^{\int a(x)\,dx}$ as in (17.3). You can ignore the constant of integration at this stage. (If you don't, it will just end up combining with the constant of integration in step (iv).)

(iii) Multiply both sides of the equation by this factor. The left hand side can then be rewritten in the form $\frac{d}{dx}(c(x)y)$.

(iv) Integrate both sides with respect to x, taking care not to forget the constant of integration.

(v) Solve for y. This is the general solution.

(vi) Substitute in any initial or boundary conditions to get the specific solution.

Second order linear differential equations

First, we will look at the homogeneous case:

$$a\frac{d^2y}{dx^2} + b\frac{dy}{dx} + cy = 0 \tag{17.4}$$

In particular, we will assume that $a, b, c \in \mathbb{R}$ are constants.

Directly integrable equations

If $c = 0$ then we have the equation

$$a\frac{d^2y}{dx^2} + b\frac{dy}{dx} = 0.$$

Integrating this equation with respect to x gives the first order linear equation

$$a\frac{dy}{dx} + by = k$$

which we can solve using the integrating factor method.

Linearity

Let

$$\mathcal{L} = a\frac{d^2}{dx^2} + b\frac{d}{dx} + c$$

be a **linear differential operator**. Then the equation (17.4) can be wriutten as

$$\mathcal{L}y = 0.$$

Suppose we have two functions u_1 and u_2. Then

$$\begin{aligned}\mathcal{L}(u_1 + u_2) &= a\frac{d^2}{dx^2}(u_1 + u_2) + b\frac{d}{dx}(u_1 + u_2) + c(u_1 + u_2) \\ &= \left(a\frac{d^2u_1}{dx^2} + b\frac{du_1}{dx} + cu_1\right) + \left(a\frac{d^2u_2}{dx^2} + b\frac{du_2}{dx} + cu_2\right) \\ &= \mathcal{L}u_1 + \mathcal{L}u_2\end{aligned}$$

Also, for any function u and constant $k \in \mathbb{R}$ we have

$$\begin{aligned}\mathcal{L}(ku) &= a\frac{d^2}{dx^2}(ku) + b\frac{d}{dx}(ku) + c(ku)\\ &= ka\frac{d^2u}{dx^2} + kb\frac{du}{dx} + kcu\\ &= k\left(a\frac{d^2u}{dx^2} + b\frac{du}{dx} + cu\right)\\ &= k\mathcal{L}u\end{aligned}$$

These are **linearity properties**. Compare with the explanation and definition of linear differential equations given earlier, and the discussion of linear maps in the material on linear algebra.

What this means is that if u_1 and u_2 are solutions of (17.4), that is, $\mathcal{L}u_1 = 0$ and $\mathcal{L}u_2 = 0$, then so is any function of the form $Au_1 + Bu_2$, where $A, B \in \mathbb{R}$.

The complementary solution

The **general solution** of (17.4) is of the form $Au_1 + Bu_2$. Unless $u_1 = ku_2$ for some $k \in \mathbb{R}$, in which case we say that u_1 and u_2 are **linearly dependent** and we need to do a bit more work.

The approach we will use is to try a solution of the form $y = e^m x$ for some $m \in \mathbb{R}$. Then

$$\frac{dy}{dx} = me^{mx} \qquad \text{and} \qquad \frac{d^2y}{dx^2} = m^2e^{mx}.$$

So

$$\begin{aligned}\mathcal{L}y &= am^2e^{mx} + bme^{mx} + ce^{mx}\\ &= e^{mx}(am^2 + bm + c)\end{aligned}$$

For a homogeneous equation we have $\mathcal{L}y = 0$, so

$$e^m x(am^2 + bm + c) = 0.$$

We know that $e^{mx} \neq 0$ for any $x \in \mathbb{R}$, so this means that

$$am^2 + bm + c = 0. \tag{17.5}$$

This is the **characteristic equation** of the problem (17.4) and we can solve it by one of the usual methods. There are three cases to consider:

Case 1 (two distinct real roots): Suppose the characteristic equation (17.5) has two distinct real roots m_1 and m_2. Then $y = e^{m_1x}$ and $y = e^{m_2x}$ are both solutions to (17.4) and they are both linearly independent, since $e^{m_1x} \neq e^{m_2x}$ if $m_1 \neq m_2$. So the general solution is

$$y = Ae^{m_1x} + Be^{m_2x}. \tag{17.6}$$

Case 2 (one repeated real root): Suppose we have one (repeated) real root m. Then $y = e^{mx}$ is a solution to (17.4) but we need another independent solution. Try $y = xe^{mx}$ (check this works). Then the general solution is

$$y = Ae^{mx} + Bxe^{mx} = (A + Bx)e^{mx} \tag{17.7}$$

Case 3 (no real roots / two distinct complex roots): Suppose we have two complex roots $m = p \pm qi$ (where $i^2 = -1$).[1] This gives solutions $y = e^{(p+qi)x}$ and $y = e^{(p-qi)x}$. But[2]

$$e^{ix} = \cos(x) + i\sin(x)$$

so the general solution is

$$y = e^{px}(A\cos(qx) + B\sin(qx)) \tag{17.8}$$

Example 17.8 Solve

$$\frac{d^2y}{dx^2} - 3\frac{dy}{dx} + 2y = 0.$$

The characteristic equation is

$$m^2 - 3m + 2 = (m-1)(m-2) = 0$$

This is case 1, so the general solution is

$$y = Ae^x + Be^{2x}$$

since $m = 1$ and $m = 2$ are the distinct real roots. See Figure 17.4.

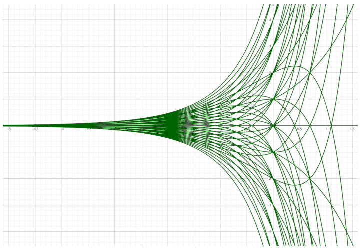


Figure 17.4: Some solutions to the equation in Example 17.8

Example 17.9 Solve

$$\frac{d^2x}{dt^2} + 4\frac{dx}{dt} + 4x = 0.$$

The characteristic equation is

$$m^2 + 4m + 4 = (m+2)^2 = 0.$$

This is case 2, so the general solution is

$$x = (A + Bt)e^{-2t},$$

since we have a single root $m = -2$. See Figure 17.5.

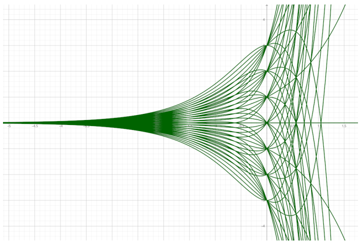


Figure 17.5: Some solutions to the equation in Example 17.9

Example 17.10 Solve

$$\frac{d^2y}{dt^2} - 6\frac{dy}{dt} + 13y = 0.$$

The characteristic equation is

$$m^2 - 6m + 13 = 0.$$

Using the quadratic formula,

$$m = \frac{6 \pm \sqrt{36 - 52}}{2} = 3 \pm 2\sqrt{-1} = 3 \pm 2i.$$

This is case 3, so the general solution is

$$y = e^{3t}(A\cos(2t) + B\sin(2t)),$$

since we have two complex roots. See Figure 17.6.

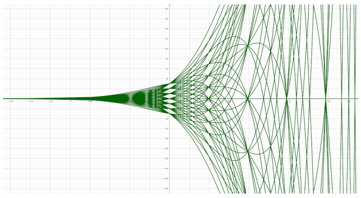


Figure 17.6: Some solutions to the equation in Example 17.10

[1] If you're not familiar with complex numbers, then don't worry, just accept this on faith.

[2] This requires x to be in radians.

The particular solution

We now know how to solve homogeneous second order linear differential equations of the form (17.4). Now we want to solve *inhomogeneous* second order linear differential equations of the form

$$a\frac{d^2y}{dx^2} + b\frac{dy}{dx} + cy = f(x) \qquad (17.9)$$

where $a, b, c \in \mathbb{R}$ are constants, and f is some function of x.

As with difference equations, we want to find a **particular solution** that accounts for the nonzero term on the right hand side of (17.9). If

$$\mathcal{L} = a\frac{d^2}{dx^2} + b\frac{d}{dx} + c$$

is the appropriate linear differential operator, and u_1 and u_2 are linearly independent solutions to the homogeneous equation (17.4), then suppose that v is a function satisfying the inhomogeneous equation (17.9). Then

$$\mathcal{L}(Au_1 + Bu_2 + v) = A\mathcal{L}u_1 + B\mathcal{L}u_2 + \mathcal{L}v = 0 + 0 + f(x)$$

so the general solution of (17.9) is the sum of the complementary solution (the general solution of the associated homogeneous equation) and the particular solution (the solution of the inhomogeneous equation).

So how do we find the particular solution? We'll adopt a trial-and-error approach.

Example 17.11 We want to solve the equation

$$\frac{d^2y}{dx^2} - 3\frac{dy}{dx} + 2y = 4x.$$

We found the complementary solution $y = Ae^x + Be^{2x}$ in Example 17.8. For the particular solution, we will try a solution of the form $v = Cx + D$. Then

$$\frac{dv}{dx} = C \qquad \text{and} \qquad \frac{d^2v}{dx^2} = 0.$$

Substituting this into the original equation we have

$$\begin{aligned} & 0 - 3C + 2(Cx + D) = 4x \\ \Longrightarrow \qquad & 2Cx + (2D - 3C) = 4x \end{aligned}$$

Now we compare the coefficients of x and the constant terms on either side of this equation to see that

$$2C = 4, \qquad\qquad 2D - 3C = 0.$$

Solving for C and D we get $C = 2$ and $D = 3$, so the particular solution is

$$v(x) = 2x + 3.$$

The general solution of the inhomogeneous equation is thus

$$y(x) = Ae^x + Be^{2x} + 2x + 3.$$

Now suppose we have initial conditions $y = 3$ and $\frac{dy}{dx} = 1$ when $x = 0$. The first derivative of the general solution is

$$\frac{dy}{dx} = Ae^x + 2Be^{2x} + 2$$

and substituting the initial conditions we find that

$$\begin{aligned} y = 3 = A + B + 3 = Ae^x + Be^{2x} + 2x + 3 \quad &\implies \quad A + B = 0 \\ \frac{dy}{dx} = 1 = A + 2B + 2 = Ae^x + 2Be^{2x} + 2 \quad &\implies \quad A + 2B = -1 \end{aligned}$$

We solve these to get $A = 1$ and $B = -1$. Then the specific solution satisfying the given initial conditions is

$$y = e^x - e^{2x} + 2x + 3.$$

See Figure 17.7 for a graph of this function.

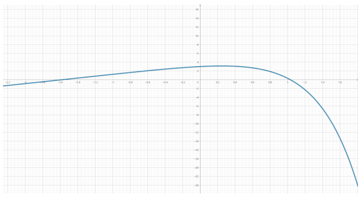


Figure 17.7: Graph of the specific solution in Example 17.11

Broadly speaking, to find the particular solution, we should try the most general function of the same type as $f(x)$. See Table 17.1 for a list of suggested trial solutions.

Table 17.1: Suggested trial solutions

$f(x)$	trial solution
constant	C
x	$Cx + D$
x^2	$Cx^2 + Dx + E$
degree–n polynomial in x	$A_n x^n + \cdots + A_1 x + A_0$ (general degree–n polynomial in x)
e^{kx}	Ce^{kx}
$\sin(kx)$ or $\cos(kx)$	$C\cos(kx) + D\sin(kx)$

In the last two cases, where $f(x) = e^{kx}$, $\sin(kx)$ or $\cos(kx)$, we must be a little careful. We need the particular solution to be linearly independent from the complementary solution, so be prepared to try Cxe^{kx} or Cx^2e^{kx} if e^{kx} is part of the complementary solution, and something like $Cx\cos(kx) + Dx\sin(kx)$ if $\cos(kx)$ or $\sin(kx)$ are part of the complementary solution.

Example 17.12 Solve

$$\frac{d^2y}{dx^2} - 3\frac{dy}{dx} + 2x = 3e^x.$$

The complementary solution is $y = Ae^x + Be^{2x}$ as before. To find the particular solution, normally we'd use $v = Ce^x$ as the trial solution, but this time we have to use $v = Cxe^x$ instead. (And if that doesn't work, try Cx^2e^x, Cx^3e^x and so on, until we find something that works.)

Here, then,

$$\frac{dv}{dx} = C(1 + x)e^x \qquad \text{and} \qquad \frac{d^2v}{dx^2} = C(2 + x)e^x.$$

Substituting these into the differential equation we get

$$\begin{aligned} C(2+x-3-3x+2x)e^x &= 3e^x \\ -Ce^x &= 3e^x \end{aligned}$$

So $C = -3$ and hence $v = -3e^x$. Therefore, the general solution to the inhomogeneous equation is

$$y = Ae^x + Be^{2x} - 3xe^x = (A-3x)e^x + Be^{2x}.$$

Summary

To solve a linear second order differential equation

$$a\frac{d^2y}{dx^2} + b\frac{dy}{dx} + cy = f(x)$$

subject to given initial conditions:

(i) Solve the characteristic equation $am^2 + bm + c = 0$.

(ii) Write down the complementary solution

$$\begin{aligned} u(x) &= Ae^{m_1 x} + Be^{m_2 x}, \\ u(x) &= (A+Bx)e^{mx}, \\ \text{or} \qquad & e^{px}(A\cos(qx) + B\sin(qx)) \end{aligned}$$

(iii) Find the particular solution $v(x)$ and write down the general solution $y = u + v$.

(iv) Now find the specific solution by substituting the boundary conditions into this general solution and solving the resulting simultaneous equations.