Differential Equations and Discrete Mathematics

Simon Salamon

Mathematical Institute
University of Oxford
October 1996

Mathematical Institute
24–29 St Giles', Oxford, OX1 3LB

Preface

These notes accompany the first-year Oxford course of the same title. They are not of course meant to substitute the lectures themselves, which are likely to provide a less theoretical approach to the subject, with more emphasis on simple applications and problem-solving. Material from the course is covered in roughly the order of the lectures synopsis, though a number of subsections (at least those starred) could be omitted on a first reading.

The notes are designed to be read at leisure during the course of the entire academic year, and some of the explanations will make more sense after exposure to other series of lectures. For example, no effort has been made to elaborate on the concept of a limit, which recurs at various points. The notes also touch on a number of major topics which will be more properly covered elsewhere, such as partial differential equations and probability. Moreover, Euclid's algorithm is described in the Institute lecture notes [9].

There is an initial temptation to regard the course as split into two utterly distinct parts, with calculus (§§1–6) forming part one, and combinatorics and algorithms (§§7–12) part two. I believe this to be mistaken, since there is a definite interchange of ideas between these two parts, and this has been emphasized here. The designers of the course were well aware of links between difference equations and generating functions, the algorithmic nature of Euler's method for approximating solutions to differential equations, and so on. On the other hand, for learning and revision, the contents might profitably be sliced into quarters (§§1–3, §§4–6, §§7–9, §§10–12).

Much of the material provides an ideal setting for experimenting with computer packages and, conversely, a greater understanding of the theory can be gained with the help of a machine. For this reason, most exercise sections include one or two problems with MAPLE, although the latter does not form part of the course and (with one exception in §12.2) has been excluded from the main body of text. For best results, precede each exercise with the command `restart;` to wipe out earlier definitions. Most of the graphics survived from earlier years of the course, and were plotted with MATHEMATICA.

I am grateful to Richard Bird for some helpful comments, and to Chris Prior for his patience in reading and correcting parts of an earlier draft.

S.M.S.
September 1996

Contents

1 Elementary Calculus

1.1 Differentiation

Let $y = y(x)$ be a function expressing y in terms of x. Its derivative, written $\dfrac{dy}{dx}$ or y', is the new function whose value at $x = a$ equals the gradient of the graph of y at a. This value, written $\dfrac{dy}{dx}(a)$ or $\dfrac{dy}{dx}\Big|_a$ or $y'(a)$, can be expressed by a limit:

$$y'(a) = \lim_{x \to a} \frac{y(x) - y(a)}{x - a}. \tag{1.1}$$

This formula can be used to work out the derivatives of some simple functions from first principles. For example, if $y = \sqrt{x}$ then

$$y'(a) = \lim_{x \to a} \frac{\sqrt{x} - \sqrt{a}}{x - a} = \lim_{x \to a} \frac{1}{\sqrt{x} + \sqrt{a}} = \frac{1}{2\sqrt{a}}.$$

Understanding when limits exist and what properties they enjoy is accomplished in another course, but since limits will occur several times in these notes, we include a precise definition.

Let a be a fixed number. If $f(x)$ equals $(y(x) - y(a))/(x - a)$, or indeed any other function, the *limit of f as x tends to a* is said to exist and equal $\ell \in \mathbb{R}$ if for any number $\varepsilon > 0$ (however small it may be) there exists $\delta > 0$ such that

$$0 < |x - a| < \delta \quad \Rightarrow \quad 0 \leq |f(x) - \ell| < \varepsilon.$$

It is important to note that in this test x is not allowed to equal a; indeed in many situations (including the one at hand) $f(a)$ is not even defined. On the other hand, if $f(a)$ *is* defined and equal to the limit ℓ then the function f is said to be *continuous at* a. For example, $(\sin x)/x$ is undetermined when $x = 0$, though

$$\lim_{x \to 0} \frac{\sin x}{x} = \sin'(0) \tag{1.2}$$

is known to equal $\cos 0 = 1$. If we define $f(x)$ to equal $(\sin x)/x$ for $x \neq 0$ and set $f(0) = 1$ then f is continuous at all points of $\mathbb{R}$, as indicated by the unbroken graph in Figure 1.

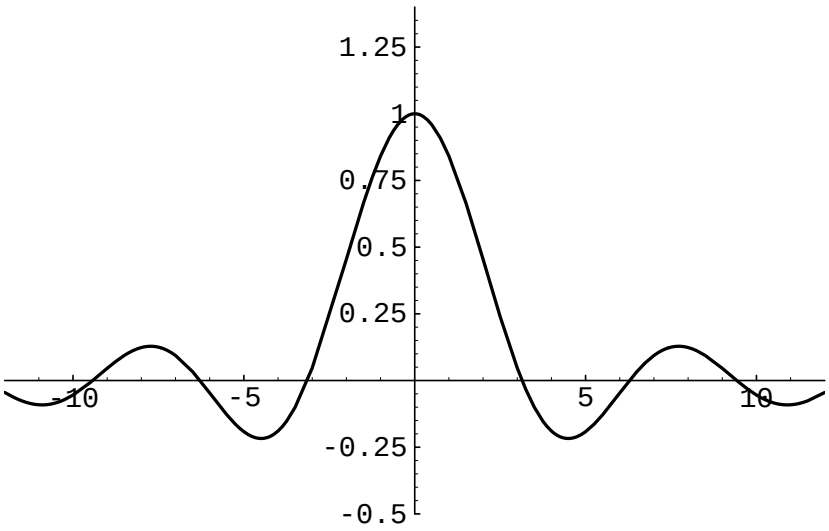
There are a number of rules that enable one to differentiate complicated functions quickly. The first is so obvious that it is often not stated explicitly, namely that the derivative of a sum of two functions is the sum of the individual derivatives. A slightly more general version of this rule is the

Linearity Property If a_1, a_2 are constants and y_1, y_2 are functions then

$$\frac{d}{dx}(a_1 y_1 + a_2 y_2) = a_1 \frac{dy_1}{dx} + a_2 \frac{dy_2}{dx}.$$

This will be of fundamental importance in seeking solutions to differential equations.

Figure 1: Extending $\dfrac{\sin x}{x}$ to $x = 0$



Product Rule If u, v are functions, then

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

Since $\dfrac{d}{dx}(x) = 1$, the product rule may be used repeatedly to prove that for any positive integer n,

$$\frac{d}{dx}(x^n) = nx^{n-1}. \tag{1.3}$$

Of course, this result is valid when n is replaced by any $r \in \mathbb{R}$, and is relevant to the generalized binomial theorem, discussed in the sequel.

Chain Rule If u, v are functions then

$$\frac{d}{dx}(v(u(x)) = v'(u(x)).u'(x). \tag{1.4}$$

The function that assigns x to $v(u(x))$ is called the *composition* of u with v, and is sometimes written $v \circ u$ so that (1.4) translates into the neat identity $(v \circ u)' = (v' \circ u)u'$. The chain rule will take on further significance in the discussion of partial derivatives in §6.2.

The above rules are powerful in combination. They are all needed, for example, to compute higher derivatives of the function e^{x^2}, given that e^x is a function equal to its own derivative:

$$\begin{aligned}\left(\frac{d}{dx}\right)^2(e^{x^2}) &= \frac{d}{dx}(e^{x^2}.2x)\\ &= \frac{d}{dx}(e^{x^2}).2x + e^{x^2}.\frac{d}{dx}(2x)\\ &= 2(2x^2+1)e^{x^2}.\end{aligned}$$

1.2 Higher derivatives

Assuming that $y'(a)$ exists for all a, one may differentiate the function y' to get the second derivative

$$y'' = (y')' = \frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{d^2y}{dx^2}$$

of y. When this process is iterated, the kth derivative of y (for k any positive integer) is written in one of the ways

$$y^{(k)}, \quad \left(\frac{d}{dx}\right)^k y, \quad \frac{d^k y}{dx^k}, \quad D^k y.$$

Some of this notation dates back to the seventeenth century, although 'D' is common in computer packages. For consistency, one also defines the '0th derivative' $y^{(0)}$ to equal the original function y.

Let k, n be integers with $1 \leq k \leq n$. Applying (1.3) repeatedly shows that

$$\left(\frac{d}{dx}\right)^k x^n = n^{\underline{k}} x^{n-k}, \tag{1.5}$$

where the coefficient on the right is defined by

$$n^{\underline{k}} = n(n-1)(n-2)\ \cdots\ (n-k+1). \tag{1.6}$$

In particular, $n^{\underline{n}} = n!$ is the factorial that equals the product of the positive integers from 1 to n inclusive. In this context, recall

Definition 1.2.1 The binomial coefficient $\binom{n}{k}$ equals $\dfrac{n!}{k!(n-k)!} = \dfrac{n^{\underline{k}}}{k!}$.

The definition (1.6) (due to [4, §2.6]) has the great advantage that it makes sense when n is replaced by any real number r; $r^{\underline{k}}$ is read 'r to the k falling', and will be used again in §7.3.

A *polynomial* is a function of the form

$$y(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0, \tag{1.7}$$

where the a_k are real (or possibly complex) constants. It has degree n if the 'leading coefficient' a_n is non-zero, and is called *monic of degree* n if $a_n = 1$. The polynomial (1.7) has $y^{(k)} = 0$ for all $k > n$, whereas

$$y^{(k)}(0) = k!a_k, \quad 1 \leq k \leq n. \tag{1.8}$$

Thus, a polynomial of degree n is completely determined by the values of itself and first n derivatives at 0. For functions that are not polynomials though, higher derivatives tend to become more and more complicated.

The expression $(1+x)^n$ clearly defines a monic polynomial of degree n in x. Evaluating its higher derivatives using the chain rule and (1.5) yields the binomial theorem:

Proposition 1.2.2 If n is a positive integer, then

$$(1+x)^n = \sum_{k=0}^{n} \binom{n}{k} x^k.$$

Well-known extensions of this result to the case in which the exponent n is no longer a positive integer will be discussed in §7.4.

1.3 Integration

The word 'integration' is often used simply to mean the reverse of differentiation, and a first exposure to the subject generally consists of taking a function f and trying to find a function y such that $f(x) = y'(x)$. If one is confident that y exists one writes

$$y(x) = \int f(x)dx + c.$$

This is an 'indefinite integral' specified only up to the addition of a constant c, even though the latter is often omitted.

The following scheme, in which an arrow represents the process of differentiation, indicates that the natural logarithm function, $\ln x$, fills in the 'gap' when integrating powers of x.

$$\begin{array}{ccccccccccccccc} & & & & & & & & \ln x & \leftarrow & x\ln x - x & \leftarrow & & & \\ & & & & & & & \swarrow & & & & & & & \\ \leftarrow & -\frac{6}{x^4} & \leftarrow & \frac{2}{x^3} & \leftarrow & -\frac{1}{x^2} & \leftarrow & \frac{1}{x} & & 1 & \leftarrow & x & \leftarrow & \frac{1}{2}x^2 & \leftarrow & \frac{1}{6}x^3 & \leftarrow \\ & & & & & & & \swarrow & & & & & & & \\ & & & & & & & 0 & & & & & & & \end{array}$$

Like $1/x$, the function $\ln x$ is not defined for $x = 0$. In fact both $\ln x$ and $\ln(-x)$ are valid integrals of $1/x$, though only one will be real depending on the sign of x. For this reason, one often writes

$$\int \frac{1}{x}\,dx = \ln|x| + c, \tag{1.9}$$

though at times the absolute value signs can be a hindrance (see Example 2.1.2).

The rules given in §1.1 give rise to corresponding ones for computing integrals. If a, b are constants and u, v functions then certainly

$$\int (a\,u(x) + b\,v(x))dx = a\int u(x)dx + b\int v(x)dx.$$

Integration by parts is a restatement of the product rule. If u, v are functions of x then

$$\int \frac{du}{dx} v(x)\,dx \;=\; u(x)v(x) \;-\; \int u(x)\frac{dv}{dx}\,dx.$$

Setting $du/dx = y$ and abbreviating the notation gives the more practical version

$$\int y\,v = (\int y)\,v - \int (\int y)v'.$$

For example, taking y to be the sine function,

$$\int x\sin x\,dx = (-\cos x)x - \int (-\cos x).1\,dx = -x\cos x + \sin x + c. \tag{1.10}$$

This last function is then the 'general solution' of the equation $y'(x) = x\sin x$.

Substitution allows one to 'cancel' dx's in the sense that

$$\int f(u(x))\frac{du}{dx}dx = \int f(u)du$$

when integrating a 'function of a function' of x.

This is a re-interpretation of the chain rule in which $f(u)$ plays the role of dv/du. As an example, taking $u = \cos$ gives

$$\int \tan x\,dx = \int \frac{\sin x}{\cos x}dx = -\int \frac{1}{u}\frac{du}{dx}dx = -\int \frac{1}{u}du$$
$$\Rightarrow \quad \int \tan x\,dx = -\ln|u| + c = -\ln|\cos x| + c.$$

Problem 1.3.1 Use the substitution $u(x) = \tan(\frac{1}{2}x)$ to evaluate $\int \csc x\,dx$, where $\csc x = \operatorname{cosec} x = 1/(\sin x)$.

Solution. Standard trigonometric identities imply that

$$\sin x = \frac{2u}{1+u^2}, \qquad \cos x = \frac{1-u^2}{1+u^2}.$$

Using the first of these, and the fact that

$$\frac{du}{dx} = \tfrac{1}{2}\sec^2(\tfrac{1}{2}x) = \tfrac{1}{2}(1+u^2)$$

gives

$$\begin{aligned}\int \csc x\,dx = \int \frac{1}{u}\frac{du}{dx}dx &= \int \frac{1}{u}du \\ &= \ln|u| + c \\ &= \ln|\tan(\tfrac{1}{2}x)| + c.\end{aligned}$$

□

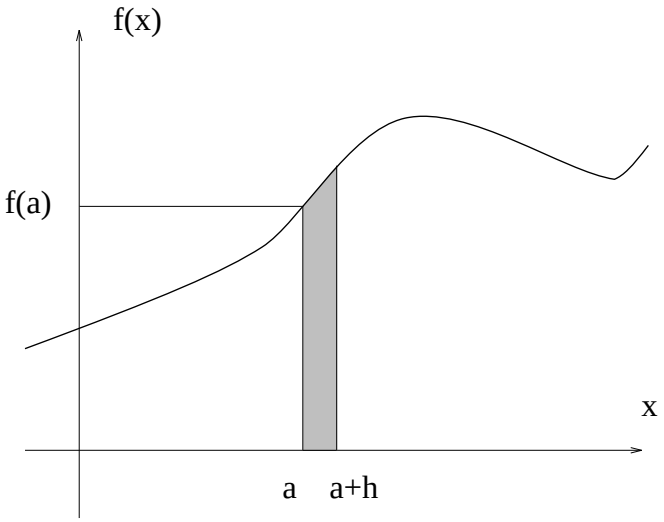
1.4 Definite integrals

To avoid the nuisance of constants of integration, one sets

$$y(x) = \int_a^x f(t)dt \tag{1.11}$$

to represent the unique function whose derivative is $f(x)$ and which satisfies the extra condition that $y(a) = 0$. For fixed x, the right-hand side of (1.11) is a 'definite integral' which may be interpreted geometrically as the area lying under the graph of f between a and x.

Figure 2:



To see why computing such areas acts as the inverse of differentiation, first apply (1.1) to obtain

$$\frac{d}{dx}\left(\int_0^x f(t)dt\right)\Big|_a = \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left(\int_a^{a+\varepsilon} f(t)dt\right).$$

The second integral is interpreted as the area of a tiny strip of width ε and height roughly $f(a)$ (see Figure 2), so the right-hand limit equals $f(a)$. But this must also be the value of the left-hand side if differentiation is to be the inverse of integration. (Strictly speaking, these arguments are only valid if f is a continuous function at $x = a$.)

Inequalities are preserved by definite integrals in the sense that

$$u(x) \le v(x) \Rightarrow \int_a^b u(x)dx \le \int_a^b v(x)dx, \quad a \le b.$$

This can be useful in showing that a given area is finite. For example, since $e^{-t^2} \le e^{-t}$ for all $t \ge 1$,

$$\int_1^x e^{-t^2}dt \le \int_1^x e^{-t}dx = -e^{-x} + e^{-1}, \quad x \ge 1. \tag{1.12}$$

Taking the limit of both sides as $x \to \infty$ gives

$$\int_1^\infty e^{-t^2} dt \leq \lim_{x\to\infty} (e^{-1} - e^{-x}) = e^{-1} = 0.367\ldots$$

where the 'infinite integral' represents the total area under the graph of e^{-t^2} to the right of the line $t = 1$ (see Figure 3). On the other hand, it is known that

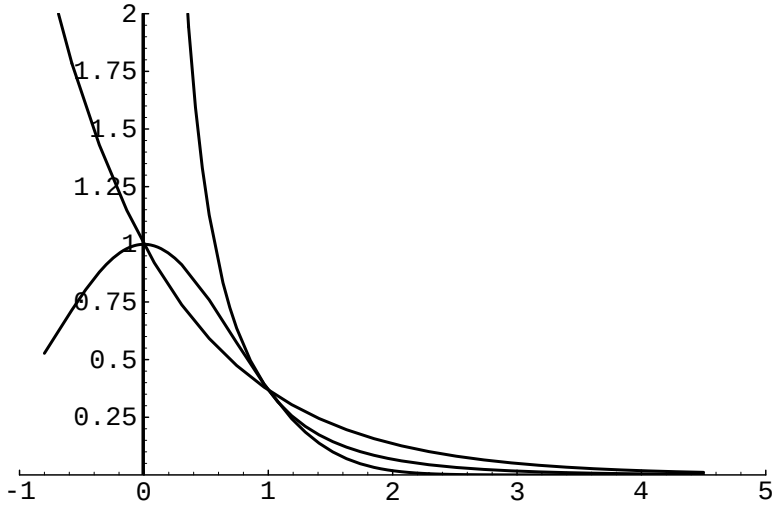
$$\int_0^\infty e^{-x^2} = \tfrac{1}{2}\sqrt{\pi} = 0.886\ldots$$

and the function

$$y(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \tag{1.13}$$

plays a special role in probability theory.

Figure 3: e^{-t^2}, e^{-t} and e^{-t}/t



Finding indefinite integrals is much harder than finding derivatives, in the sense that it is impossible to express the integrals of many simple functions in terms of familiar functions. A similar example of this type is the integral

$$\int_1^x \frac{e^{-t}}{t} dt, \tag{1.14}$$

whose value is sandwiched between each side of the inequality in (1.12).

1.5 Exercises

1. (i) Prove that (1.3) holds when n is replaced by a positive fraction p/q by setting $y^q = x^p$.

(ii) Deduce the quotient rule $(u/v)' = (u'v - uv')/v^2$ from the chain and product rules and (1.3) for $n = -1$.

2. (i) Prove that the derivative of $\ln(\csc x + \cot x)$ is $-\csc x$. How do you reconcile this with the answer in Problem 1.3.1? Find $\int \sec x\, dx$.

(ii) Simplify the function $\arctan\left(\dfrac{\sin x}{1+\cos x}\right)$ by differentiation, or otherwise.

3. Let $x > 0$. Find the derivative of $y = x^x$, by first taking logs, or otherwise. The minimum value of x^x occurs when x is the unique solution a of the equation $y'(x) = 0$; find a and sketch the graph of x^x.

4. Express each of the following integrals in terms of the function (1.14):

(i) $\displaystyle\int \frac{1}{\ln x} dx$, (ii) $\displaystyle\int \frac{e^{-x^2}}{x} dx$, (iii) $\displaystyle\int \ln(\ln x) dx$.

5. Using the notation of (1.6), show that $(r+1)^{\underline{n}} - r^{\underline{n}} = n r^{\underline{n-1}}$. In what way is this analogous to (1.3)?

6. Tables of standard derivatives and integrals are stored in MAPLE. Check to see that the following give expected results:

```
D(tan);              int(tan(x),x);
D(sec);              int(sec(x),x);
D(csc);              int(csc(x),x);
D(ln);               int(ln(x),x);
D(arcsin);           int(arcsin(x),x);
D(arctan);           int(arctan(x),x);
D(arctanh);          int(arctanh(x),x);
diff(E^x,x);         int(x*exp(x),x);
diff(ln(ln(x)),x);   int(ln(ln(x)),x);
```

7. Run separately the three 1-line programs

```
y:= x->x^3*ln(x)^2: for k to 9 do (D@@k)(y) od;
ln: for k to 5 do y:=int("(x),x): x->y od: y;
for k to 5 do int(1/(1+x^k),x) od;
```

and explain what these accomplish.

8. Investigate values of the function (1.13) by

```
for k from 0 to 5 by .1 do
  evalf(int(exp(-x^2),x=0..k))
od;
```

Carry out a similar analysis for (1.14).

2 Introduction to Differential Equations

2.1 First examples

The viewpoint of this course is that the term 'integration' actually refers to the process of solving any equations involving derivatives. For example, given the two equations

$$\text{(i)} \quad y' = \cot x,$$
$$\text{(ii)} \quad y' + 2y = y^2,$$

not only can one say that

$$\text{'}\ln|\sin x| \text{ is an integral of } \cot x\text{'},$$

but also

$$\text{'}y \equiv 2 \text{ is an integral of (ii)'}.$$

The symbol '$\equiv$' is used to emphasize that the '2' is being regarded as a function rather than just a number; the constant function 2 (or for that matter 0) is an obvious solution of (ii). In (i) any other solution is obtained by adding on a constant, though in (ii) the non-constant solutions are harder to discern.

The following equation can be rapidly solved by undoing double differentiation.

Example 2.1.1

$$\frac{d^2y}{dx^2} + \sin x = 0 \quad \text{or} \quad y''(x) = -\sin x.$$

Integrating gives

$$y'(x) = \cos x + c_1,$$
$$\Rightarrow y(x) = \sin x + c_1 x + c_2.$$

Here c_1 and c_2 are constants, included to give the most general solution; in fact $c_1 = y'(0) - 1$ and $c_2 = y(0)$. □

In this example, we might interpret x as time and y as the distance travelled by some particle. Then $y' = \dot{y}$ represents velocity and $y'' = \ddot{y}$ acceleration, the latter perhaps specified by some applied sinusoidal force. (Differentiations with respect to time are conventionally denoted by dots, following Newton.) If the particle has an initial velocity of 5 units, then the appropriate 'initial conditions' are

$$y(0) = 0, \quad \dot{y}(0) = 5,$$

which give rise to the 'particular solution' $y(x) = \sin x + 4x$.

Slightly less straightforward, but still much simpler than (ii) above, is

Example 2.1.2

$$y' + 2y = 0 \quad \text{or} \quad \frac{dy}{dx} = -2y$$

One approach to solving this is to separate the variables and plonk down integral signs without thinking what one is doing, so as to give

$$\int \frac{1}{y} dy = -2 \int dx.$$

One deduces that

$$\ln y = -2x + c, \quad \text{or} \quad y(x) = be^{-2x},$$

where c and $b = e^c$ are constants. In fact, be^{-2x} is the *general solution* in the sense that any solution of the differential equation must equal this for some value of b. On the other hand, the use of the logarithm is a bit artificial, and it is not completely obvious that dividing by y does not eliminate a solution. Also, the constant b can certainly be negative even though c must then be a complex number; this explains why the absolute signs of (1.9) are inappropriate. □

A more convincing way of obtaining the general solution in Example 2.1.2 is to spot that the positive function e^{-2x} is one solution, and then suppose that $y(x) = u(x)e^{-2x}$ is another. The equation becomes

$$0 = y' + 2y = (u' - 2u)e^{-2x} + 2ue^{-2x} = u'e^{-2x},$$

and is equivalent to $u' \equiv 0$, which means that u is a constant. Thus be^{-2x} is indeed the general solution, and we shall use this technique again and again in cases where one solution of an equation is already known. One should note that the implication

$$u' \equiv 0 \quad \Rightarrow \quad u \text{ constant} \tag{2.1}$$

actually underlies all the integration steps we have already made; although it appears obvious it is strictly speaking a consequence of the

Mean Value Theorem for differentiable functions. This states that for any a, b, the 'average' rate of change $(u(b)-u(a))/(b-a)$ equals $u'(c)$ for at least one point c between a and b.

2.2 Classification of equations

Definition 2.2.1 An *ordinary differential equation* (ODE) of order n is an equation of the form

$$F(x, y, y', y'', \ldots, y^{(n)}) = 0,$$

where n is the order of the highest derivative actually appearing. The equation is said to be *linear* if it has the form

$$g(x) + f_0(x)y(x) + f_1(x)y'(x) + \cdots + f_n(x)y^{(n)}(x) = 0$$

for suitable functions $g, f_0, f_1, \ldots, f_n$.

In a linear equation, if one pretends that x is a constant, F is a just a sum of multiples of y and its derivatives. A linear equation is called *homogeneous* if $g \equiv 0$, and the linearity property of §1.1 implies the important

Proposition 2.2.2 If y_1, y_2 are solutions of a linear homogeneous ODE then $c_1y_1+c_2y_2$ is also a solution for any constants c_1, c_2. □

Example 2.1.2 is defined by the function $F(x, y, y') = 2y + y'$, and is as nice an ODE as one can imagine: first order, linear and homogeneous. By comparison, here are some nastier ones:

(i) $y(1 + (y')^2) = 1$ is non-linear first order,

(ii) $y'' + \sin y = 0$ is non-linear second order,

(iii) $y''' + y = x \sin x$ is linear third order.

Equation (ii) is much harder to solve than the superficially similar Example 2.1.1, and is the ODE that governs the oscillations of a simple pendulum. For small values of y, its solutions are approximated by those of the linear equation $y'' + y = 0$, namely

$$y(x) = a \sin x + b \cos x = k \sin(x + \phi), \tag{2.2}$$

that give rise to what is referred to as *simple harmonic motion*.

The word 'ordinary' distinguishes these equations from partial differential equations, involving functions of more than one variable (see §6.4).

Definition 2.2.3 A first-order ODE is *separable* if it can be written in the form

$$g(y)\frac{dy}{dx} = h(x), \quad \text{or more informally} \quad g(y)dy = h(x)dx,$$

for some functions g, h.

In this case, the substitution rule implies that

$$\int g(y)dy = \int h(x)dx + c,$$

and solutions are obtained provided one can evaluate the two integrals. However, such solutions may be 'implicit', that is to say they relate x and y but do not express y directly as a function of x.

Problem 2.2.4 Solve the equations

(i) $(1 + y^2)y' = 1 - x^2$,

(ii) $(1 - x^2)y' = 1 + y^2$.

Solution. Equation (i) implies

$$\int (1 + y^2)dy = \int (1 - x^2)dx + c$$
$$\Rightarrow \quad y + \tfrac{1}{3}y^3 = x - \tfrac{1}{3}x^3 + c.$$

The last line is quite acceptable as a conclusion; in theory it could be used to express y directly in terms of x using square and cube roots but this might not help much in practice. By contrast, the separable equation (ii) leads to

$$\int \frac{1}{1+y^2}dy = \int \frac{1}{1-x^2}dx + c$$

which, at least for $|x| < 1$, has an explicit solution

$$y = \tan\left(\ln\sqrt{\frac{1+x}{1-x}} + c\right) = \tan(\operatorname{arctanh} x + c)$$

($u = \operatorname{arctanh} x$ is equivalent to $x = \tanh u = (e^{2u}-1)/(e^{2u}+1)$.) □

Sometimes a first-order ODE can be rendered separable by an easy change of variable. An important class consists of equations of the form

$$\frac{dy}{dx} = f\Big(\frac{y}{x}\Big), \tag{2.3}$$

where f is an arbitrary function. The right-hand side is an example of a *homogeneous* function of x, y (this more precise use of the word 'homogeneous' is explained in §6.3). Setting $u = y/x$ converts (2.3) into the separable equation

$$x\frac{du}{dx} + u = f(u),$$

and allows one to tackle a variety of examples (see §2.5).

2.3 First-order linear equations

This section will undertake a systematic investigation of the pair of linear equations in 'standard form'

$$\boxed{y' + f(x)y = 0} \tag{2.4}$$

$$\boxed{y' + f(x)y = g(x)} \tag{2.5}$$

In both cases, the coefficient of y' equals 1 and in the notation of Definition 2.2.1, $f_0 = -f$ and $f_1 \equiv -1$. This is no real restriction, as one may in general divide through by the coefficient function of y' provided one does not get upset if f or g are then undefined at particular points (see consequences of this in Problem 2.3.3).

The first equation is said to be the homogeneous equation *associated* with the second, and is obviously separable. It implies that

$$\ln y = \int \frac{1}{y}dy = -\int f(x)dx + a.$$

Whence

Proposition 2.3.1 The general solution of (2.4) is

$$y(x) = ce^{-\int f(x)dx},$$

where c is an arbitrary constant. □

Next, consider the non-homogeneous equation (2.5). Guided by the argument at the end of §2.1, we seek a solution in the form $y = ue^{-\int f}$ for some function u (we have suppressed the x's to save energy). Using all the calculus rules from §1.1,

$$\begin{aligned} u' = (ye^{\int f})' &= y'e^{\int f} + y(e^{\int f}.f) \\ &= e^{\int f}.(y' + fy) \\ &= e^{\int f}.g, \end{aligned}$$

and in theory we can integrate to determine u and so y.

To save having to remember the definition of u, to solve (2.5) in practice, one needs only remember to multiply both sides by the 'integrating factor'

$$\boxed{I(x) = e^{\int f(x)dx}} \tag{2.6}$$

For this converts the non-homogeneous equation into (in fuller notation)

$$\begin{aligned} & \frac{d}{dx}(I(x)y(x)) = I(x)g(x) \\ \Rightarrow \quad & I(x)y(x) = \int I(x)g(x)dx + c \end{aligned}$$

$$\Rightarrow \quad y(x) = I(x)^{-1} \int I(x)g(x)dx \; + \; cI(x)^{-1}. \tag{2.7}$$

Taking $c = 0$ we see that

$$\boxed{y_1(x) = I(x)^{-1} \int I(x)g(x)dx}$$

is a particular solution of (2.5), whereas $cI(x)^{-1}$ is the general solution of (2.4). In fact if y_1 and y_2 are both solutions of (2.5), then $y_1 - y_2$ is a solution of (2.4); conversely, $y_1 + ce^{-\int f}$ will solve (2.5) for any constant c. The same argument shows that

Proposition 2.3.2 The general solution of a linear non-homogeneous equation equals any particular solution of it plus the general solution of the associated homogeneous equation.

Problem 2.3.3 Solve the initial value problem (IVP for short)

$$\begin{cases} xy' + 2y = \sin x, \\ y(\pi) = 0 \end{cases}$$

Solution. First we put the equation into standard form by dividing through by x:

$$y' + \frac{2}{x}y = \frac{\sin x}{x}. \tag{2.8}$$

The integrating factor is

$$I(x) = e^{\int (2/x)dx} = e^{2\ln x} = x^2,$$

giving

$$(x^2 y)' = x^2 y' + 2xy = x \sin x.$$

One might have spotted that multiplying the original equation by x was all that was needed to transform the left-hand side into an exact derivative, though it is sometimes quicker to compute (2.6) methodically without thinking too hard. From (1.10),

$$x^2 y = \sin x - x\cos x + c,$$

giving the general solution

$$y = \frac{\sin x - x\cos x}{x^2} + \frac{c}{x^2}.$$

To solve the initial condition we need

$$0 = \frac{0 - \pi(-1)}{\pi^2} + \frac{c}{\pi^2} \quad \Rightarrow \quad c = -\pi,$$

so that finally

$$y(x) = (\sin x - x\cos x - \pi)/x^2.$$

□

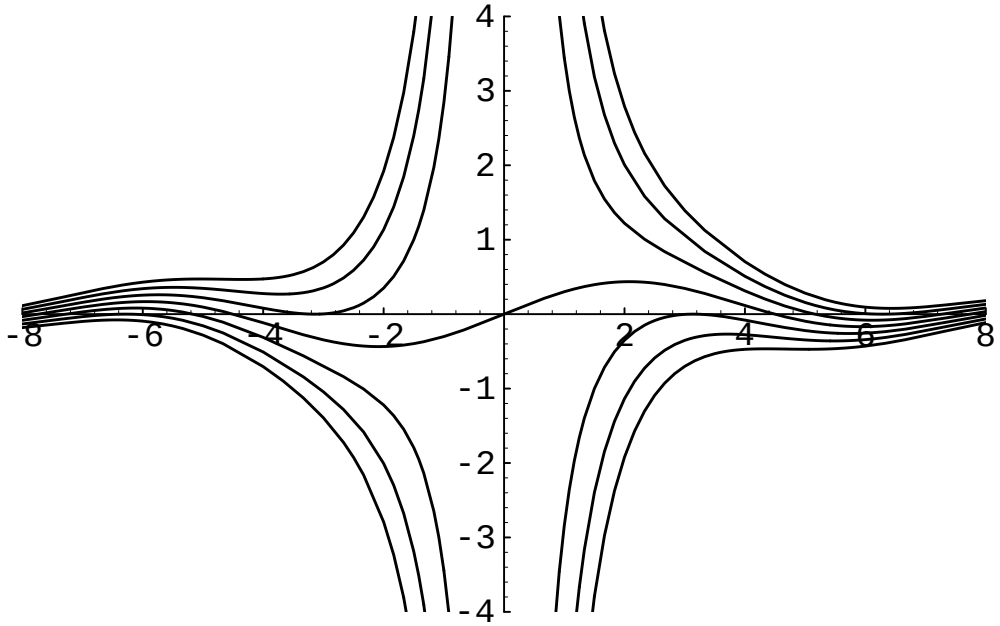
Writing the last ODE as

$$\frac{dy}{dx} = \frac{\sin x - 2y}{x}$$

shows that one is actually prescribing the gradient or slope of a curve at each point (x, y) in the plane. For example, our answer with $c = -\pi$ has slope 0 at $(\pi, 0)$, and diverges as $x \to 0$. Graphs of solutions for several different value of c illustrate that there is exactly one solution of the differential equation passing through a given point, but that this solution may not 'extend' for all values of x (see Figure 4). The assertion that a first-order differential equation possesses, under fairly general conditions, a unique solution passing through any point in the plane is an important theorem in more advanced treatments of the subject [2, §2.11].

The right-hand side of (2.8) extends to a continuous function which has finite values for all x (see (1.2)). However, the coefficient of y is unbounded as $x \to 0$, which explains why most solutions also have this defect. Only when $c = 0$, do we obtain a solution, namely $(\sin x - x\cos x)/x^2$, which tends to a finite limit as $x \to 0$; indeed, l'Hôpital's

Figure 4: Solutions for $c = -3\pi, -2\pi, -\pi, 0, \pi, 2\pi, 3\pi$



rule implies that this limit is 0, so that the corresponding graph passes through the origin.

We shall return to the graphical interpretation of ODE's in §5.1.

2.4 Reduction of order

The next stage is to consider the second-order linear equations

$$\boxed{y''(x) + f_1(x)y'(x) + f_0(x)y(x) = 0} \tag{2.9}$$

$$\boxed{y''(x) + f_1(x)y'(x) + f_0(x)y(x) = g(x)} \tag{2.10}$$

and attempt an analysis based on Proposition 2.3.2. This time there is no systematic way to solve the homogeneous equation (2.9), but if a solution of it *is* known, a familiar technique can be used to find a solution of the non-homogeneous equation (2.10).

Lemma 2.4.1 Let y_1 be a solution of (2.9) and u a function such that $y_2 = uy_1$ is a solution of (2.10). Then $v = u'$ satisfies the first-order linear equation

$$y_1v' + (2y_1' + f_1y_1)v = g.$$

Proof. One has

$$\begin{aligned} y_2' &= u'y_1 + uy_1', \\ y_2'' &= u''y_1 + 2u'y_1' + uy_1'', \end{aligned}$$

and by assumption $y_1'' + f_1y_1' + f_0y_1 = 0$. Substituting into (2.10) gives the result. □

Example 2.4.2

$$y'' + y = \cot x.$$

A solution of the associated homogeneous equation is obviously given by $\sin x$, so let $y(x) = u(x)\sin x$, and $v(x) = u'(x)$. The lemma implies that

$$\begin{aligned} & (\sin x)v' + 2(\cos x)v = \cot x \\ \Rightarrow \quad & \frac{d}{dx}((\sin x)^2 v) = \sin x \cot x = \cos x \\ \Rightarrow \quad & v(x) = \csc x + c\csc^2 x. \end{aligned}$$

Setting $c = 0$ gives

$$u(x) = \int \csc x\, dx = \ln\tan(\tfrac{1}{2}x) = -\ln(\csc x + \cot x)$$

(see Problem 1.3.1), so that for $x > 0$ a solution is $\ln(\tan(\frac{1}{2}x))\sin x$. □

The lemma can also be applied to obtain a second solution of (2.9), given a first:

Problem 2.4.3 Verify that e^{2x} is a solution of the equation

$$xy'' - (x+1)y' - 2(x-1)y = 0,$$

and find a second solution of the form $y = ue^{2x}$.

Solution. The verification amounts to checking that $4x - 2(x+1) - 2(x-1) = 0$. Dividing by x and using Lemma 2.4.1,

$$\begin{aligned} & e^{2x}v' + \left(4e^{2x} - (1+\tfrac{1}{x})e^{2x}\right)v = 0 \\ \Rightarrow \quad & v' + \left(3 - \frac{1}{x}\right)v = 0. \end{aligned}$$

The integrating factor is $e^{3x-\ln x} = e^{3x}/x$, so

$$\begin{aligned} & v(x) = cxe^{-3x} \\ \Rightarrow \quad & u(x) = -\tfrac{1}{3}cxe^{-3x} + \tfrac{1}{3}c\int e^{-3x}dx = -\tfrac{1}{3}c(x+\tfrac{1}{3})e^{-3x}. \end{aligned}$$

Taking $c = -9$ gives the second solution $y(x) = e^{-x}(3x+1)$. □

Given two distinct solutions y_1, y_2 to the homogeneous second-order linear equation (2.9), it is important to know if these are proportional or 'linearly dependent' in the sense that

$$ay_1 + by_2 \equiv 0, \tag{2.11}$$

for some *constants* a, b, not both zero. For example there are many identities involving trigonometric functions that might obscure such a relationship. Differentiating (2.11) gives $ay_1' + by_2' \equiv 0$, and it follows that

$$y_1y_2' - y_2y_1' \equiv 0. \tag{2.12}$$

The function on the left, denoted $W(y_1, y_2)$ is called the *Wronskian* of y_1, y_2, and often assumes a simple form even if the individual solutions are complicated.

Conversely, it can be shown that if y_1, y_2 are solutions of (3.1) that are not proportional, then $W(y_1, y_2)$ is actually *nowhere* zero on any interval in which the functions f_0, f_1 are continuous [6, §2.7]. The latter condition is not satisfied by the equation in Problem 2.4.3 at $x = 0$, which tallies with the fact that

$$W(e^{2x}, e^{-x}(3x + 1)) = -9xe^x,$$

vanishes when $x = 0$.

2.5 Exercises

1. Verify that the given function is a solution of the corresponding ODE for $x > 0$:
(i) $\frac{1}{4}x^2(3 - 2\ln x)$, $\quad y'' + \ln x = 0$;
(ii) $\frac{1}{4}x^2 - 3$, $\quad y' = \sqrt{y + 3}$;
(iii) e^{-x}, $\quad y^{(iv)} = y + y' + y''$.
In each case, by inspection or otherwise, find a solution of the same equation not equal to the one given.

2. Find general solutions of the following first-order ODE's:
(i) $yy' = x^2$;
(ii) $(x^2 - 1)y' = x^3y$;
(iii) $y' + (\cot x)y = \csc x$;
(iv) $y' = 1 + x + y^2 + xy^2$;
(v) $(\sin y)y' = \cos x$.

3. By setting $y = e^u$, find a solution of $xy' = y(x^2 - \ln y)$ satisfying $y(1) = 1$. For which values of x is the solution valid?

4. Use the treatment of (2.3) to find general solutions of the equations
(i) $y' = \dfrac{x - y}{x + y}$;
(ii) $2x^2y' = x^2 + y^2$ (an obvious solution is $y(x) = x$);

(iii) $y' = \dfrac{x^2 + 3y^2}{2xy}$;

5. Verify that x^2 is a solution to $x^2y'' - 3xy' + 4y = 0$, and find another by setting $y = x^2u$. Equations of this type are discussed in §3.3.

6. The distance r of a planet from its sun satisfies

$$\ddot{r} = \frac{a}{r^3} - \frac{b}{r^2},$$

where a, b are positive constants, and a dot denotes differentiation with respect to time t. By considering $d(\dot{r}^2)/dt$, show that $s = r^2$ satisfies $\ddot{s} = 2b/\sqrt{s} + c$, where c is another constant. Try to spot a solution of this second equation when $c = 0$.

7. Verify that, for each fixed `c`, the function

```
f:= (x,c)->2/(c*exp(2*x)+1):
```

is a solution of the equation (ii) at the start of §2.1. In what sense is this the 'general' solution? Sketch the curves

```
seq(f(x,4*k),k=-2..2):
plot({"},x=-.5..1.5);
```

and try to enlarge the picture by modifiying the constants and range.

8. (i) Let $k \in \mathbb{R}$. Show that the substitution $u = y^{1-k}$ reduces the ODE $y' + y = y^k$ to a linear equation, and solve it.
(ii) Investigate the equation

```
eq:= D(y)(x)+y^j=y^k:
```

by assigning integers to `j` and `k` then solving, such as

```
j:=2: k:=3:
dsolve(eq,y(x));
```

3 Constant Coefficients

3.1 The characteristic equation

In §2.4, we saw that knowledge of one solution of a homogeneous linear equation led to others. In this section we shall explain how to solve the rather special equation

$$\boxed{y''(x) + py'(x) + qy(x) = 0} \tag{3.1}$$

in which p, q are real *constants*. In the next, we shall return to the non-homogeneous version in which the right-hand side is replaced by an assigned function $g(x)$.

We can express (3.1) in the somewhat abbreviated form

$$(D^2 + pD + q)y = 0,$$

where the symbol D denotes d/dx. The expression in parentheses can now be factored by treating D as an ordinary variable, and the outcome determines the type of solutions.

Definition 3.1.1 The auxiliary or *characteristic equation* associated with (3.1) is the equation $\lambda^2 + p\lambda + q = 0$.

This is a quadratic equation in λ, with 'characteristic roots'

$$\lambda_1 = \frac{-p + \sqrt{p^2 - 4q}}{2}, \quad \lambda_2 = \frac{-p - \sqrt{p^2 - 4q}}{2},$$

which coincide iff $p^2 = 4q$ and are non-real iff $p^2 < 4q$. The characteristic equation therefore takes the form

$$(\lambda - \lambda_1)(\lambda - \lambda_2) = 0, \quad \text{so} \quad \begin{cases} \lambda_1 + \lambda_2 = -p, \\ \lambda_1\lambda_2 = q. \end{cases}$$

Using the normal rules for expanding brackets, (3.1) can now be written as

$$(D - \lambda_1)(D - \lambda_2)y = 0;$$

for the left-hand side equals

$$DDy - D\lambda_2 y - \lambda_1 Dy + \lambda_1\lambda_2 y = D^2y - (\lambda_1 + \lambda_2)Dy + \lambda_1\lambda_2 y.$$

The last step only works because λ_2 is constant; this allows one to say $D\lambda_2 y = \lambda_2 Dy$.

Proposition 3.1.2 The general solution of (3.1) is given by

$$y = \begin{cases} c_1 e^{\lambda_1 x} + c_2 e^{\lambda_2 x}, & \text{if} \quad \lambda_1 \neq \lambda_2, \\ (c_1 x + c_2)e^{\lambda_1 x}, & \text{if} \quad \lambda_1 = \lambda_2, \end{cases}$$

where c_1, c_2 are arbitrary constants.

Proof. Let u denote the function $(D-\lambda_2)y$, so that

$$(D-\lambda_1)u = 0, \quad \text{or} \quad u' - \lambda_1 u = 0;$$

this is a first-order linear homogeneous equation with general solution $u(x) = ae^{\lambda_1 x}$, where a is a constant. From the definition of u,

$$y' - \lambda_2 y = ae^{\lambda_1 x},$$

which has integrating factor $I(x) = e^{-\lambda_2 x}$ and, from (2.7), general solution

$$\begin{aligned} y(x) &= I(x)^{-1} \int I(x) a e^{\lambda_1 x} dx + bI(x)^{-1} \\ &= ae^{\lambda_2 x} \int e^{(\lambda_1 - \lambda_2)x} dx + be^{\lambda_2 x}. \end{aligned}$$

The result follows from integrating the last line, in which one must distinguish the two cases $\lambda_1 \neq \lambda_2$ and $\lambda_1 = \lambda_2$. □

Note the consistency between Propositions 2.2.2 and 3.1.2: if y_1 and y_2 are solutions of (3.1) then so is $c_1 y_1 + c_2 y_2$.

We next illustrate what happens when the roots of the characteristic equation are not real.

Example 3.1.3 The displacement y of a mass on a spring moving in a fluid at time x is governed by the equation

$$y'' + 2y' + 5y = 0.$$

The characteristic equation $\lambda^2 + 2\lambda + 5 = 0$ has roots $(-2 \pm \sqrt{4-20})/2 = -1 \pm 2i$. The general solution is therefore

$$c_1 e^{(-1+2i)x} + c_2 e^{(-1-2i)x} = e^{-x}(c_1 e^{2ix} + c_2 e^{-2ix}).$$

Using the de Moivre formula

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

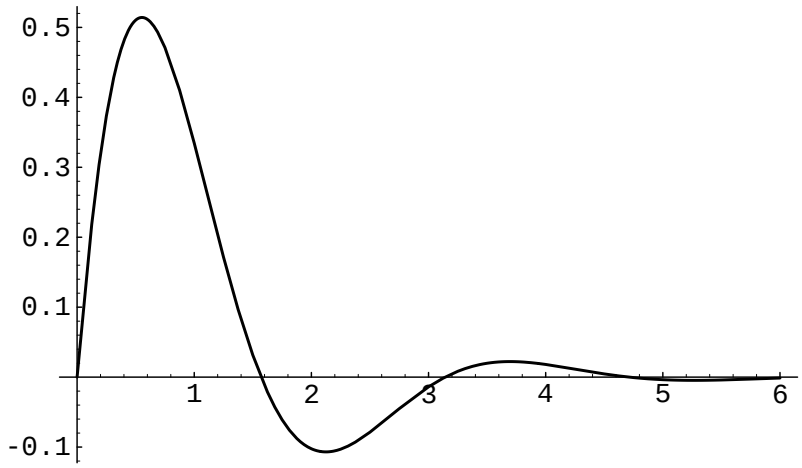
this may be rewritten as

$$e^{-x}(c_1(\cos 2x + i\sin 2x) + c_2(\cos 2x - i\sin 2x)) = e^{-x}(a_1\cos 2x + a_2 \sin 2x)), \qquad (3.2)$$

where $a_1 = c_1 + c_2$ and $a_2 = i(c_1 - c_2)$. Given the nature of the problem, we must restrict the 'general' solution to be real by taking $a_1, a_2 \in \mathbb{R}$ (and thereby allowing c_1, c_2 to be complex conjugates).

The coefficient of y' in the original equation arises from viscosity. If it were zero, the exponential term in the solution would be absent and the motion would be purely sinusoidal with continuing oscillations of equal magnitude as in (2.2). As it is, the e^{-x} factor causes the magnitude of the oscillations to decay, and this phenomenon is referred to as 'damping'. □

Figure 5: Underdamped motion



We shall always suppose that the constants p, q defining the ODE (3.1) are real. The behaviour of solutions then depends crucially on the sign of $\Delta = p^2 - 4q$. To explain this, suppose that $p > 0$. If $\Delta > 0$, no oscillations occur, and any motion represented by the solutions is said to be 'overdamped'. The 'critical' case $\Delta = 0$ corresponding to $\lambda_1 = \lambda_2$ is best regarded as a limiting case of overdamped motion. By contrast, if $\Delta < 0$ then $\lambda_1 = -\frac{1}{2}p \pm ir$ where $r = \frac{1}{2}\sqrt{4q - p^2}$ is a real number, and the general solution of (3.1) is

$$e^{-px/2}(a_1 \cos rx + a_2 \sin rx).$$

The motion is said to be 'underdamped'; this is illustrated in Figure 5 by the solution (3.2) with $a_1 = 0$, $a_2 = 1$.

3.2 Undetermined coefficients

Consider the non-homogeneous equation

$$\boxed{y''(x) + py'(x) + qy(x) = g(x),} \tag{3.3}$$

in which p, q are once again constants, and g is an assigned function. Proposition 2.2.2 applies to the pair of equations (3.1) and (3.3), so to fully solve (3.3) we need only find a particular solution of it. For many functions g, the most efficient way to do this is to make an informed guess involving constants (the so-called 'undetermined coefficients') which are subsequently evaluated by substitution.

Provided that the coefficients on the left-hand side are constant, this technique provides an easier alternative to the one based on Lemma 2.4.1, and will be illustrated in the series of problems below.

Problem 3.2.1 Solve the IVP

$$\begin{cases} y'' - 4y' + 4y = \sin x, \\ y(0) = 0 = y'(0). \end{cases}$$

Solution. Step 1: The characteristic equation is $\lambda^2 - 4\lambda + 4 = 0$ or $(\lambda - 2)^2 = 0$. Hence $\lambda_1 = \lambda_2 = 2$, and the associated homogeneous equation has general solution

$$y(x) = (c_1 x + c_2)e^{2x}.$$

Step 2: Seek a particular solution y_1 of the given non-homogeneous ODE. Given that repeated derivatives of $\sin x$ are multiples of itself or $\cos x$, it is natural to try $y_1 = A\sin x + B\cos x$, where A, B are constants to be determined from the equation

$$\begin{aligned} &(-A\sin x - B\cos x) - 4(A\cos x - B\sin x) + 4(A\sin x + B\cos x) = \sin x \\ \Rightarrow \quad &(3A + 4B)\sin x + (-4A + 3B)\cos x = \sin x. \end{aligned}$$

Since this must hold for all x, we get (for example, by taking x to equal $\frac{\pi}{2}$, 0 respectively)

$$\left.\begin{aligned} 3A + 4B &= 1 \\ -4A + 3B &= 0 \end{aligned}\right\} \quad \Rightarrow \quad A = \tfrac{3}{25},\ B = \tfrac{4}{25}.$$

Step 3: Find the values of c_1, c_2 that satisfy the initial conditions. In fact,

$$\begin{aligned} y(0) &= \tfrac{4}{25} + c_2 \quad \Rightarrow \quad c_2 = -\tfrac{4}{25}, \\ y'(0) &= \tfrac{3}{25} + c_1 + 2c_2 \quad \Rightarrow \quad c_1 = \tfrac{1}{5}, \end{aligned}$$

so the final answer is

$$y(x) = \tfrac{1}{25}\left(3\sin x + 4\cos x + (5x - 4)e^{2x}\right).$$

The graph of this function in Figure 6 shows how the exponential part dominates when $x > 0$ and the sinusoidal part when $x < 0$. □

We have already used the fact that $D = d/dx$ is a *linear operator* in the sense that

$$D(c_1 y_1 + c_2 y_2) = c_1 Dy_1 + c_2 Dy_2$$

whenever c_1, c_2 are constants and y_1, y_2 are functions. The same applies to the differential operator

$$L = D^2 + pD + q.$$

The function $u(x) = e^{\alpha x}$ (α a constant) satisfies

$$\boxed{Du = \alpha u}$$

which says that D transforms u to a multiple of itself. This is a very special situation, and one refers to u as an *eigenvector* of the linear operator D. An analogous situation characterizes the equations

$$R_1(\mathbf{v}_1) = \mathbf{v}_1,$$
$$R_2(\mathbf{v}_2) = -\mathbf{v}_2,$$

in which R_1 is a rotation of 3-dimensional space with axis parallel to a vector $\mathbf{v}_1$ and R_2 is a reflection in a plane perpendicular to a vector $\mathbf{v}_2$. Note that D^2 acts by multiplication by α^2 on u, so and

$$Lu = (D^2 + pD + q)u = (\alpha^2 + p\alpha + q)u,$$

showing that the exponential function u is also an eigenvector for L.

We can rewrite the last equation as

$$L\left(\frac{1}{\alpha^2 + p\alpha + q}e^{\alpha x}\right) = e^{\alpha x},$$

which tells us that $e^{\alpha x}/(\alpha^2 + p\alpha + q) = e^{\alpha x}/((\alpha - \lambda_1)(\alpha - \lambda_2))$ is a particular solution of the non-homogeneous equation

$$y'' + py' + qy = e^{\alpha x}.$$

This works provided that α is different from λ_1 and λ_2; for the other cases it is easily verified that $Ly = e^{\alpha x}$ has a particular solution

$$\begin{cases} \dfrac{1}{\alpha - \lambda_2}xe^{\alpha x}, & \text{if} \quad \alpha = \lambda_1 \neq \lambda_2, \\ \frac{1}{2}x^2e^{\alpha x}, & \text{if} \quad \alpha = \lambda_1 = \lambda_2. \end{cases} \tag{3.4}$$

Problem 3.2.2 Find a particular solution of

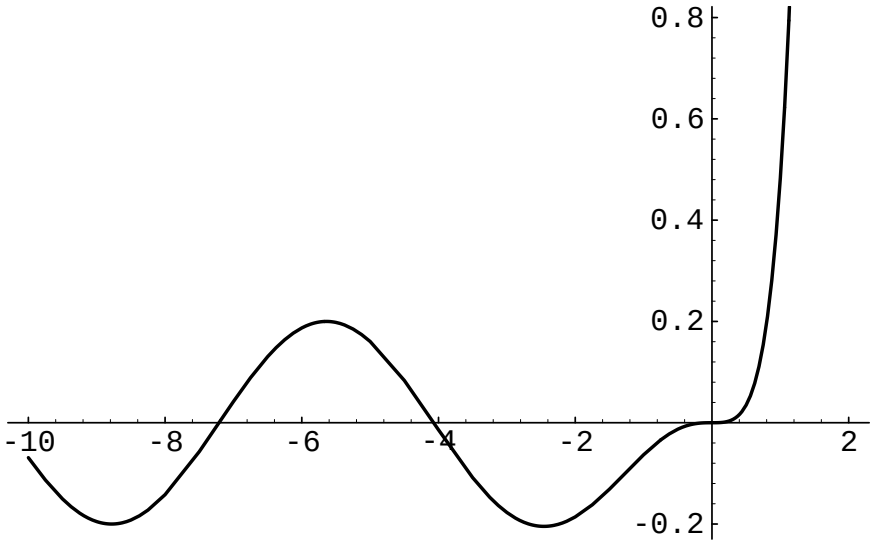
$$y'' - 5y' + 6y = e^x + e^{2x} + e^{3x}.$$

Solution. The characteristic equation is $0 = (\lambda^2 - 5\lambda + 6) = (\lambda - 2)(\lambda - 3)$. Let $L = D^2 - 5D + 6$. Then
(i) a particular solution of $Ly = e^x$ is $y_1 = \frac{1}{(1-5+6)}e^x$;
(ii) a particular solution of $Ly = e^{2x}$ is $y_2 = \frac{1}{2-3}xe^{2x}$;
(iii) a particular solution of $Ly = e^{3x}$ is $y_3 = \frac{1}{3-2}xe^{3x}$.
Since $L(y_1+y_2+y_3) = Ly_1+Ly_2+Ly_3$, the final solution is obtained by adding everything together and is

$$y_1 + y_2 + y_3 = \tfrac{1}{2}e^x - xe^{2x} + xe^{3x}.$$

□

Figure 6: $(3\sin x + 4\cos x + (5x-4)e^{2x})/25$



Illustrated above is the common practice of multiplying a 'first guess' at a particular solution by x if it involves a solution of the homogeneous equation. One needs to multiply by x again if the characteristic equation has repeated roots.

⋆ 3.3 Further techniques

The following result can save time in determining coefficients of particular solutions.

Lemma 3.3.1 If $L = D^2 + pD + q$ then $Ly = 0 \Rightarrow L(xy) = 2y' + py$.
Proof. $L(xy) = (xy)'' + p(xy)' + qxy = x(y'' + py' + qy) + 2y' + py$. □

Here is a sequel to Example 3.1.3:

Problem 3.3.2 Find a particular solution of

$$y'' + 2y' + 5y = x^2(1 + \sin x) + e^{-x}\cos 2x.$$

Solution. We take each of the three terms on the right separately.
(i) For x^2, try $A + Bx + Cx^2$. This gives

$$\begin{aligned} & 2C + 2(B + 2Cx) + 5(A + Bx + Cx^2) = x^2 \\ \Rightarrow \quad & 5C = 1, \quad 5B + 4C = 0, \quad 5A + 2B + 2C = 0 \\ \Rightarrow \quad & A = -\tfrac{2}{125},\ B = -\tfrac{4}{25},\ C = \tfrac{1}{5}. \end{aligned}$$

(ii) For $x^2 \sin x$, try $(A+Bx+Cx^2)\sin x+(D+Ex+Fx^2)\cos x$. In general, a polynomial of degree n times sin *or* cos necessitates trying a linear combination of sin *and* cos with coefficients that are also polynomials of degree n. A tedious calculation gives $A = \frac{29}{250}$, $B = -\frac{7}{25}$, $C = \frac{1}{5}$, $D = \frac{28}{250}$, $E = \frac{1}{25}$ and $F = -\frac{1}{10}$.
(iii) For $e^{-x}\cos 2x$, note that $y = e^{-x}(A\sin 2x + B\cos 2x)$ is no good as it is a solution of the associated homogeneous equation. Instead we try xy, and use the lemma:

$$L(xy) = 2e^{-x}(2A\cos 2x - 2B\sin 2x) + (-2+2)e^{-x}(A\sin 2x + B\cos 2x) = e^{-x}\cos 2x$$
$$\Rightarrow \quad A = \tfrac{1}{4}, \quad B = 0.$$

Adding everything together, a solution is

$$\tfrac{1}{250}\left(-4 - 40x + 50x^2 + (29-70x+50x^2)\sin x + (28+10x-25x^2)\cos x\right) + \tfrac{1}{4}xe^{-x}\sin 2x.$$

□

The reduction of constant coefficient equations to the algebra of the characteristic equation extends to arbitrary degree. This is because any ODE

$$y^{(n)} + p_{n-1}y^{(n-1)} + \cdots + p_1 y' + p_0 y = 0$$

with p_i constant can be factorized as

$$(D-\lambda_1)(D-\lambda_2)\ \cdots\ (D-\lambda_n)y = 0, \tag{3.5}$$

where $\lambda_i \in \mathbb{C}$ are the roots (in any order) of the associated monic polynomial with coefficients p_i. Any function y_i satisfying $(D-\lambda_i)y_i$ will therefore be a solution of (3.5), and the general solution will in fact be a linear combination of these y_i, plus any 'extra' solutions of the form $x^k y_i$ in the case of repeated roots.

Example 3.3.3

$$y^{(iv)} - 2y''' + 2y' - y = e^x.$$

The characteristic polynomial is

$$0 = \lambda^4 - 2\lambda^3 + 2\lambda - 1 = (\lambda-1)^3(\lambda+1),$$

so the roots are 1 (repeated thrice) and -1. In line with Proposition 3.1.2, the general solution of the homogeneous equation is therefore

$$(c_1 + c_2x + c_3x^2)e^x + c_4e^{-x}.$$

Futhermore, to find a particular solution of the non-homogeneous equation, we need to try $y(x) = Ax^3e^x$ which gives $12Ae^x = e^x$. Thus the general solution is

$$(c_1 + c_2x + c_3x^2 + \tfrac{1}{12}x^3)e^x + c_4e^{-x}.$$

□

With hindsight, it was obvious that $e^{\lambda x}$ is a solution of the constant coefficient equation $y'' + py' + qy = 0$ when $\lambda^2 + p\lambda + q = 0$. There is an analogous family of second-order linear ODE's for which x^λ is equally obviously a solution for suitable λ. These are the Euler-Cauchy equations

$$\boxed{x^2y'' + pxy' + qy = 0} \tag{3.6}$$

where p, q are again constants. Setting $y(x) = x^\lambda$ gives the 'new' characteristic equation

$$\lambda^2 + (p-1)\lambda + q = 0. \tag{3.7}$$

The question arises as to what happens when (3.7) has repeated roots, as for instance in the equation

$$x^2y'' + 3x\,y' + y = 0,$$

which has $1/x$ as one solution. We could find an extra solution of the form $u(x)x^\lambda$ with the aid of Lemma 2.4.1, but here we shall follow a much quicker route. If (3.7) has two roots $\lambda_1 = \lambda$ and $\lambda_2 = \lambda + \varepsilon$ then we might expect

$$\lim_{\varepsilon\to 0} \frac{x^{\lambda+\varepsilon} - x^\lambda}{\varepsilon} = x^\lambda \ln x$$

to be a solution, given that the quotient is, for any $\varepsilon \neq 0$. (The limit is evaluated by differentiating $x^\lambda = e^{\lambda \ln x}$ with respect to λ). It is easily verified that this is correct, and we have the following analogue of Proposition 3.1.2:

Proposition 3.3.4 The general solution of (3.6) is given by

$$y = \begin{cases} c_1x^{\lambda_1} + c_2x^{\lambda_2}, & \text{if} \quad \lambda_1 \neq \lambda_2, \\ (c_1 \ln x + c_2)x^{\lambda_1}, & \text{if} \quad \lambda_1 = \lambda_2, \end{cases}$$

where λ_1, λ_2 are the roots of (3.7) and c_1, c_2 are arbitrary constants.

3.4 Exercises

1. The equations
(i) $y'' - 2y' + 5y = e^{2x}\sin x$;
(ii) $y'' - 2y' + 5y = x\sin x$;
(iii) $y'' - 2y' + 5y = xe^x \sin 2x$
all have the same characteristic roots. Find a particular solution in each case. Why is your answer to (iii) simpler than might have been expected?

2. Solve the IVP

$$\begin{cases} y'' - 2y' + y = e^x + xe^x, \\ y(0) = 0 = y'(0), \end{cases}$$

given that the equation has a particular solution of the form $(Ax + B)x^2e^x$.

3. Find second-order ODE's with solutions
(i) $ae^x + be^{2x} + x\sin x$,
(ii) $a\sin 2x + b\cos 2x + e^x$,
in which a, b are arbitrary constants.

4. Verify that $y_1(x) = e^x$ is a solution of $xy'' - (x + 2)y' + 2y = 0$, and show that the equation also admits a solution of the form $y_2(x) = x^2 + Ax + B$. Compute the Wronskian (2.12) of these two solutions.

5. Solve the IVP

$$\begin{cases} x^2y'' + 2xy' - 6y = x, \\ y(1) = 0 = y'(1). \end{cases}$$

6. Without determining the coefficients, write down the general form of a particular solution of the ODE

$$y'' + 3y' - 4y = g(x)$$

when $g(x)$ equals (i) $(1+x)^3$, (ii) x^3e^x, and (iii) $x^3e^x\sin x$. Check your answers starting with

```
eq:= (D@@2)(y)(x)+3*D(y)(x)-4*y=(1+x)^3:
dsolve(eq,y(x));
```

7. Carry out the IVP computation

```
eq:= (D@@2)(y)(x)-2*D(y)(x)+5*y=4*exp(x):
dsolve({eq,y(0)=1,D(y)(0)=1},y(x));
```

Simplify the answer by hand into something more reasonable.

8. Find a particular solution of the equation $y'' + cy = \tan x$ for $c = 4$:

```
eq:= (D@@2)(y)(x)-4*y=tan(x):
dsolve(eq,y(x));
```

Experiment to see for which other values of the constant c there exists a solution in terms of familiar functions.

4 Difference Equations

4.1 Analogies with ODE's

Consider the array

$$\begin{array}{ccccccccccccccc} 0 & & 1 & & 4 & & 9 & & 16 & & 25 & & 36 & & 49 & \ldots\ldots \\ & 1 & & \underline{3} & & \underline{5} & & 7 & & 9 & & 11 & & 13 & & \\ & & 2 & & \underline{2} & & 2 & & 2 & & 2 & & 2 & & & \\ & & & 0 & & 0 & & 0 & & 0 & & 0 & & & & \end{array}$$

formed by listing perfect squares and taking successive differences. Let y_n denote the nth term in the top row, starting from 0 (so that $y_0 = 0, y_1 = 1, \ldots$). The fact that the third row is all 2's tells us that for example

$$2 = (y_3 - y_2) - (y_2 - y_1) = y_3 - 2y_2 + y_1,$$

or more generally that

$$y_{n+2} - 2y_{n+1} + y_n = 2. \tag{4.1}$$

This is an example of a second-order *difference equation*.

We shall soon see that the general solution of (4.1) is

$$y_n = n^2 + c_1 n + c_2, \tag{4.2}$$

where c_1, c_2 are arbitrary constants. The solution of a difference equation like (4.1) is a sequence of real numbers $(y_0, y_1, y_2, \ldots)$, that we may think of as a function defined on the set $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$ of natural numbers (including 0), so that

$$\begin{array}{rcl} f : \mathbb{N} & \longrightarrow & \mathbb{R} \\ n & \mapsto & y_n = y(n). \end{array}$$

The 'graph' of this function would consist of the points (n, y_n) for $n = 0, 1, 2, \ldots$. (It is important to understand the distinction between a *sequence* $(y_0, y_1, y_2, \ldots)$, and its underlying *set* $\{y_0, y_1, y_2, \ldots\}$ in which the order is irrelevant and any repetitions are ignored; thus a set does not define a function.)

Taking differences is in some ways analogous to differentiation, and one might define the 'derivative' of a sequence y_n to be the new sequence defined by $y'_n = y_{n+1} - y_n$. However, for our purposes, given the original sequence

$$y = (y_0, y_1, y_2, y_3, \ldots),$$

it will be more convenient to define the 'shifted sequences'

$$\begin{array}{rcl} Sy & = & (y_1, y_2, y_3, y_4, \ldots) \\ S^2 y & = & (y_2, y_3, y_4, y_5, \ldots) \\ & & \ldots\ldots \end{array}$$

by means of the operator S which replaces each term of y by its **S**uccessor, or equivalently **S**hifts the sequence one place to the left. Our original equation (4.1) becomes

$$S^2y - 2Sy + y = g, \quad \text{or} \quad (S-1)^2y = g,$$

where g stands for the constant sequence $(2, 2, 2, \ldots)$ (whose underlying set is $\{2\}$), and we are adding sequences and multiplying them by constants in an obvious term-wise fashion.

More generally, let p, q be real numbers and consider the difference equations

$$\boxed{y_{n+2} + py_{n+1} + qy_n = 0} \tag{4.3}$$

$$\boxed{y_{n+2} + py_{n+1} + qy_n = g_n,} \tag{4.4}$$

where $(g_0, g_1, g_2, \ldots)$ is an assigned sequence, and the problem is to find y_n. In the first example, g_n was equal to 2 for all n; more interesting might be the sequence defined by

$$g_n = 3 + (-1)^n = \begin{cases} 4, & n \text{ even}, \\ 2, & n \text{ odd}. \end{cases} \tag{4.5}$$

The equation (4.3) is the homogeneous equation assiociated to (4.4). The linearity property of the the operators S and $S^2 + pS + q$ ensures that, just as for differential equations,

Proposition 4.1.1 The general solution of (4.4) equals any particular solution plus the general solution of (4.3).

The general solution of (4.2) has precisely this form. Had we guessed that An^2 was a particular solution of (4.1) for some undetermined coefficient A, we could have substituted this to obtain

$$y_{n+2} - 2y_{n+1} + y_n = A(n+2)^2 - 2A(n+1)^2 + An^2 = 2A,$$

and confirm that $A = 1$.

4.2 Fibonacci type equations

Consider the homogeneous difference equation (4.3), in which p and q are constants. Note this this is really a *recurrence relation* in that it expresses the nth term

$$y_n = -py_{n-1} - qy_{n-2}$$

as a function of the preceding terms.

Proposition 4.2.1 The general solution to (4.3) is given by

$$y_n = \begin{cases} c_1\lambda_1{}^n + c_2\lambda_2{}^n, & \text{if} \quad \lambda_1 \neq \lambda_2, \\ (c_1 n + c_2)\lambda_1{}^n, & \text{if} \quad \lambda_1 = \lambda_2, \end{cases}$$

where c_1, c_2 are arbitrary constants and (as in §3.1) λ_1 and λ_2 are the roots of the characteristic equation $\lambda^2 + p\lambda + q = 0$.

Proof. Briefly, if λ_1 is a root of the characteristic equation, then $\lambda_1^{n+2}+p\lambda_1^{n+1}+q\lambda_1^n=0$, and $\lambda_1{}^n$ solves (4.3). If λ is a repeated root then $2\lambda^{n+2}+p\lambda^{n+1}$ vanishes, and this implies that $n\lambda^n$ is also a solution. By linearity, the sequences (y_n) defined above solve (4.3) for all values of c_1, c_2. Given any other solution $(\tilde{y}_n)$ of (4.3), one may choose the constants c_1, c_2 such that $(x_n = y_n - \tilde{y}_n)$ is a solution with $x_0 = x_1 = 0$, and it follows easily that $x_n = 0$ for all n. □

A more honest proof that works without prior knowledge of the solutions is given at the end of this section. Consider next two simple examples:

(i) The homogeneous equation $y_{n+2} - 2y_{n+1} + y_n = 0$ associated with the one in §4.1 has characteristic equation $(\lambda - 1)^2 = 0$ and therefore general solution

$$(c_1 n + c_2)1^n = c_1 n + c_2,$$

as claimed.

(ii) It is obvious that the general solution of $y_{n+2} + y_n = 0$ has the form

$$(a, b, -a, -b, a, b, -a, -b, \ \ldots)$$
$$\Rightarrow \quad y_{2n} = (-1)^n a, \quad y_{2n+1} = (-1)^n b,$$

where $y_0 = a$ and $y_1 = b$ are arbitrary. Alternatively, the roots of the characteristic equation are $\pm i$, and so according to Proposition 4.2.1, $y_n = c_1 i^n + c_2(-i)^n$, which amounts to the same thing if we set $a = c_1 + c_2$ and $b = i(c_1 - c_2)$. In general, if $\lambda_1, \lambda_2 = \rho e^{\pm i\theta}$ are complex (assuming always that p, q are real), the solution of (4.3) is better expressed in the form

$$\rho^n(a_1 \cos n\theta + a_2 \sin n\theta),$$

in analogy to (3.2).

A celebrated solution of a difference equation is the Fibonacci sequence

$$(0,\ 1,\ 1,\ 2,\ 3,\ 5,\ 8,\ 13,\ 21,\ 34,\ 55,\ \ldots)$$

which solves the 'initial value problem'

$$\begin{cases} F_{n+2} - F_{n+1} - F_n = 0, \\ F_0 = 0,\ F_1 = 1. \end{cases} \tag{4.6}$$

Thus, F_n denotes the nth Fibonacci number with the convention that $F_1 = 1 = F_2$. The characteristic equation is

$$\lambda^2 - \lambda - 1 = 0,$$

and has roots

$$\boxed{\phi = \frac{1+\sqrt{5}}{2}, \quad \widehat{\phi} = \frac{1-\sqrt{5}}{2}}$$

The positive root $\phi = 1.61803\ldots$ is the so-called *golden ratio*, and $\widehat{\phi} = 1 - \phi$. Many of the graphs of these notes (such as Figure 7) are automatically framed by an imaginary rectangle whose sides are in the proportion $\phi : 1$, as this is meant to be an especially pleasing shape. A rectangle of size $\phi \times 1$ can be divided into a smaller rectangle of the same shape plus a unit square. □

The previous proposition yields

Corollary 4.2.2 $F_n = \dfrac{1}{\sqrt{5}}(\phi^n - \widehat{\phi}^n)$.

Proof. To derive the Fibonacci numbers from the general solution $c_1\phi^n + c_2\widehat{\phi}^n$, we need to satisfy the initial conditions

$$\begin{cases} 0 = F_0 = c_1 + c_2 & \Rightarrow \quad c_2 = -c_1, \\ 1 = F_1 = c_1(\phi - \widehat{\phi}) & \Rightarrow \quad c_1 = 1/\sqrt{5}. \end{cases}$$

□

This corollary has some interesting consequences. Firstly, since $|\widehat{\phi}^n/\sqrt{5}| < 0.5$ for all $n \geq 0$, we have

$$F_n = \frac{1}{\sqrt{5}}\phi^n \quad \text{to the nearest integer.}$$

Also

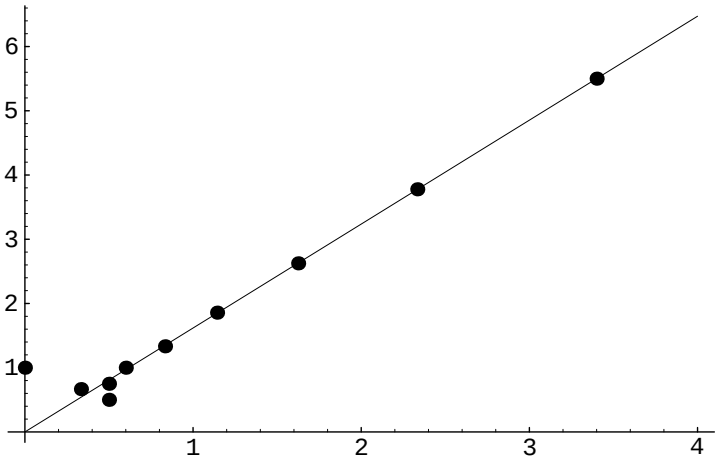
$$\lim_{n\to\infty} \frac{F_n}{F_{n-1}} = \lim_{n\to\infty} \frac{\phi^n - \widehat{\phi}^n}{\phi^{n-1} - \widehat{\phi}^{n-1}} = \lim_{n\to\infty} \frac{1 - \left(\widehat{\phi}/\phi\right)^n}{1 - \left(\widehat{\phi}/\phi\right)^{n-1}} \cdot \phi = \phi, \tag{4.7}$$

as $|\widehat{\phi}/\phi| < 1$. This result is illustrated in Figure 7 which plots the points

$$\{(\frac{F_{n-1}}{n}, \frac{F_n}{n}) : 1 \leq n \leq 10\},$$

and shows that the ratio of successive Fibonacci numbers converges very quickly to ϕ.

Figure 7: $y = \phi x$



Algebraic proof of Proposition 4.2.1. The homogeneous difference equation (4.3) can be written $(S^2 + pS + q)y = 0$, and as in §3.1, we split this into two first-order equations

$$(S - \lambda_1)u = 0, \quad \text{where} \quad u = (S - \lambda_2)y.$$

The general solution of the first is determined by writing

$$u_n = \lambda_1 u_{n-1} = {\lambda_1}^2 u_{n-2} = \cdots = {\lambda_1}^n u_0.$$

Since $(S - \lambda_2)y = u$, we now get

$$\begin{aligned} y_n &= \lambda_2 y_{n-1} + u_0 {\lambda_1}^{n-1} \\ &= \lambda_2(\lambda_2 y_{n-2} + u_0 {\lambda_1}^{n-2}) + u_0 {\lambda_1}^{n-1} \\ &= \lambda_2^2 y_{n-2} + u_0({\lambda_1}^{n-1} + {\lambda_1}^{n-2}\lambda_2) \\ &\qquad \cdots\cdots \\ &= \lambda_2^n y_0 + u_0({\lambda_1}^{n-1} + {\lambda_1}^{n-2}\lambda_2 + \cdots + {\lambda_2}^{n-1}). \end{aligned}$$

This is in the familiar form of the general solution of a first-order homogeneous equation, plus a particular solution of a non-homogeneous equation which we must now sum.

There are two cases. If $\lambda_1 = \lambda_2$, we get

$$y_n = {\lambda_2}^n y_0 + u_0.n{\lambda_1}^{n-1} = (c_1 n + c_2){\lambda_1}^n,$$

where c_1, c_2 are constants. If $\lambda_1 \neq \lambda_2$ then we can use the identity

$$x^n - 1 = (x - 1)(x^{n-1} + x^{n-2} + \cdots + x + 1) \tag{4.8}$$

with $x = \lambda_1/\lambda_2$ to get

$$y_n = {\lambda_2}^n y_0 + u_0 \frac{{\lambda_1}^n - {\lambda_2}^n}{\lambda_1 - \lambda_2} = c_1 {\lambda_1}^n + c_2 {\lambda_2}^n,$$

with different constants, but as required. □

4.3 Worked problems

Notice that the Fibonacci numbers $8, 13, 21$ satisfy $21.8 - 13^2 = 168 - 169 = -1$. This is generalized by

Problem 4.3.1 Prove Cassini's relations

$$F_{n+1}F_{n-1} - {F_n}^2 = (-1)^n. \tag{4.9}$$

Solution. Since $F_0 = 0$ and $F_1 = 1 = F_2$, (4.9) holds when $n = 1$. To complete a proof by induction we shall deduce from (4.9) the corresponding equation with n replaced by $n + 1$. To do this, we first use the equation in (4.6) twice, and then apply (4.9):

$$\begin{aligned} F_{n+2}F_n &= F_{n+1}(F_{n+1} - F_{n-1}) + {F_n}^2 \\ &= {F_{n+1}}^2 - ({F_n}^2 + (-1)^n) + {F_n}^2. \end{aligned}$$

Rearranging gives the required equation

$$F_{n+2}F_n - {F_{n+1}}^2 = (-1)^{n+1}.$$

□

We now return to applications of Proposition 4.2.1. The following should be compared to (3.4); it can be proved by straightforward verification.

Lemma 4.3.2 Let $L = S^2 + pS + q$, and let g denote the sequence with nth term $g_n = \alpha^n$. Then $Ly = g$ has a particular solution

$$y_n = \begin{cases} \dfrac{1}{(\alpha-\lambda_1)(\alpha-\lambda_2)}\alpha^n, & \text{if} \quad \lambda_1 \neq \alpha \neq \lambda_2, \\ \dfrac{1}{\alpha-\lambda_2} n\alpha^{n-1}, & \text{if} \quad \alpha = \lambda_1 \neq \lambda_2, \\ \frac{1}{2}n^2\alpha^{n-2}, & \text{if} \quad \alpha = \lambda_1 = \lambda_2. \end{cases}$$

□

Problem 4.3.3 Solve the IVP

$$\begin{cases} y_{n+2} - 5y_{n+1} + 6y_n = 1 + 2^n + 3^n, \\ y_0 = 0 = y_1, \end{cases}$$

and determine $\lim_{n\to\infty}(y_{n+1}/y_n)$.

Solution. The equation is analogous to the ODE of Problem 3.2.2, and its characteristic equation has roots 2 and 3. We use the lemma above:
(i) To get the term $1 = 1^n$ on the right, take $y_1 = \frac{1}{(1-2)(1-3)}$;
(ii) to get 2^n take $y_2 = \frac{1}{2-3}n2^{n-1}$;
(iii) to get 3^n take $y_3 = \frac{1}{3-2}n3^{n-1}$.
The general solution of the non-homogeneous equation is therefore

$$\begin{aligned} y_n &= y_1 + y_2 + y_3 + (\text{general solution of } Ly = 0) \\ &= \tfrac{1}{2} - 2^{n-1}n + 3^{n-1}n + c_2 2^n + c_3 3^n. \end{aligned}$$

Finally,

$$\begin{aligned} &y_0 = \tfrac{1}{2} + c_2 + c_3 \quad \Rightarrow \quad c_3 = -c_2 - \tfrac{1}{2} \\ &y_1 = \tfrac{1}{2} - 1 + 1 + 2c_2 + 3c_3 \quad \Rightarrow \quad c_2 = -1, \quad c_3 = \tfrac{1}{2}. \end{aligned}$$

Therefore,

$$\begin{aligned} y_n &= \tfrac{1}{2} - 2^{n-1}n + 3^{n-1}n - 2^n + \tfrac{1}{2}.3^n \\ &= \tfrac{1}{2} - (n+2)2^{n-1} + (n+\tfrac{3}{2})3^{n-1}. \end{aligned} \tag{4.10}$$

Proceeding exactly as in (4.7),

$$\frac{y_{n+1}}{y_n} = \frac{\frac{1}{2}(\frac{1}{3})^{n-1} - 2(n+3)(\frac{2}{3})^{n-1} + 3(n+\frac{5}{2})}{\frac{1}{2}(\frac{1}{3})^{n-1} - (n+2)(\frac{2}{3})^{n-1} + (n+\frac{3}{2})}.$$

Since $n\lambda^n \to 0$ whenever $|\lambda| < 1$, we obtain

$$\lim_{n\to\infty}\frac{y_{n+1}}{y_n} = \lim_{n\to\infty} 3\frac{1+\frac{5}{2n}}{1+\frac{3}{2n}} = 3,$$

which equals the characteristic root of greatest magnitude. □

Problem 4.3.4 Find an expression for the general term of the sequence defined by

$$y_{n+2} + 2y_{n+1} + y_n = \begin{cases} 4, & n \text{ even}, \\ 2, & n \text{ odd}, \end{cases}$$

and satisfying $y_0 = y_1 = y_2$.

Solution. First spot that the right-hand side of the equation equals g_n, in the notation of (4.5). The characteristic roots are both -1, so a general solution of the homogeneous equation is given by $(a_1 + a_2 n)(-1)^n$. Moreover,
(i) a particular solution of $y_{n+2} + 2y_{n+1} + y_n = 3$ is (obviously) $\frac{3}{4}$, and
(ii) a particular solution of $y_{n+2} + 2y_{n+1} + y_n = (-1)^n$ is (from Lemma 4.3.2) $\frac{1}{2}n^2(-1)^n$.
The general solution of the non-homogeneous equation is therefore

$$\tfrac{3}{4} + (a_1 + a_2 n + \tfrac{1}{2}n^2)(-1)^n.$$

The condition $y_0 = y_1 = y_2$ becomes

$$\tfrac{3}{4} + a_1 = \tfrac{3}{4} - (a_1 + a_2 + \tfrac{1}{2}) = \tfrac{3}{4} + (a_1 + 2a_2 + 2) \quad \Rightarrow \quad a_1 = \tfrac{1}{4},\ a_2 = -1,$$

and the final solution is

$$y_n = \begin{cases} 1 - n + \frac{1}{2}n^2, & n \text{ even}, \\ \frac{1}{2} + n - \frac{1}{2}n^2, & n \text{ odd}. \end{cases} \tag{4.11}$$

□

Having found the general formulae (4.10),(4.11) in the previous problems, one might be tempted to determine the y_n that satisfy the respective difference equations with n a *negative* integer. This amounts to 'completing' the solution backwards from y_0, and shows more clearly how the magnitude of the roots of the characteristic equation affects the behaviour of solutions. For the first problem, we obtain a solution

$$(\ldots, \tfrac{2147}{3888}, \tfrac{235}{432}, \tfrac{13}{27}, \tfrac{11}{36}, 0, 0, 3, 21, 101, 415, 1567, \ldots)$$

in which the negative terms are all fractional. By contrast, the second problem gives rise to a doubly-infinite sequence symmetric about the term y_1:

$$(\ldots, 25, -17, 13, -7, 5, -1, 1, 1, 1, -1, 5, -7, 13, -17, 25, -31, 41, \ldots).$$

Another difference between the two solutions is that the integer function (4.10) can be extended to a function of a real variable by merely replacing n by x. Finding such an explicit function assuming the values (4.11) at integer points is much harder.

4.4 Exercises

1. Make a list of the 'backwards Fibonacci numbers' F_n that satisfy (4.6) for $n \leq -1$. State and prove a formula that relates these to the ordinary Fibonacci numbers.

2. Find general solutions of the difference equations

(i) $y_{n+2} - 5y_{n+1} + 6y_n = 1$;
(ii) $y_{n+2} - 4y_{n+1} + 4y_n = 2^n n$.
In each case, determine also the unique solution satisfying $y_0 = 0 = y_1$.

3. Find solutions with $y_0 = 1$ for the first-order difference equations
(i) $y_n - 2y_{n-1} = 1$;
(ii) $y_n + 2y_{n-1} = 1$.
Does there exist a formula for the solution of $y_n - ny_{n-1} = 1$?

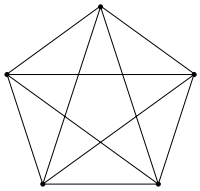
4. Find general solutions to the equations
(i) $y_{n+2} - y_n = f(n)$, where $f(n) = 5$ when n is even and $f(n) = 3$ when n is odd;
(ii) $py_{n+2} - y_{n+1} + qy_n + 2 = 0$, where p, q are constants such that $0 < p < q < 1$ and $p + q = 1$.

5. Prove the identity

$$F_{m+n} = F_{m+1}F_n + F_mF_{n-1} \tag{4.12}$$

for all $m, n \in \mathbb{Z}$, using the method of Problem 4.3.1.

6. Let δ be the length of the indicated diagonals of a regular pentagon with side 1 unit. Show that (i) $\delta\cos(\frac{2}{5}\pi) = \frac{1}{2}$, and (ii) $2\delta^2(1 - \cos(\frac{1}{5}\pi)) = 1$. Deduce that $\delta = \phi$. What is the length of each side of the smaller pentagon?



7. Cassini's equation (4.9) is an example of a *non-linear* difference equation. Investigate other solutions such as

```
y[1]:=1: y[2]:=3:
for n from 2 to 10 do
  y[n+1]:= (y[n]^2+(-1)^n)/y[n-1]
od;
```

8. Find the general solution of the equation $y_{n+3} + 2y_{n+2} - y_{n+1} - 2y_n = n$ using the routine

```
eq:= y(n+3)+2*y(n+2)-y(n+1)-2*y(n)=n:
ic:= y(0)=a,y(1)=b,y(2)=c:
rsolve({eq,ic},y);
```

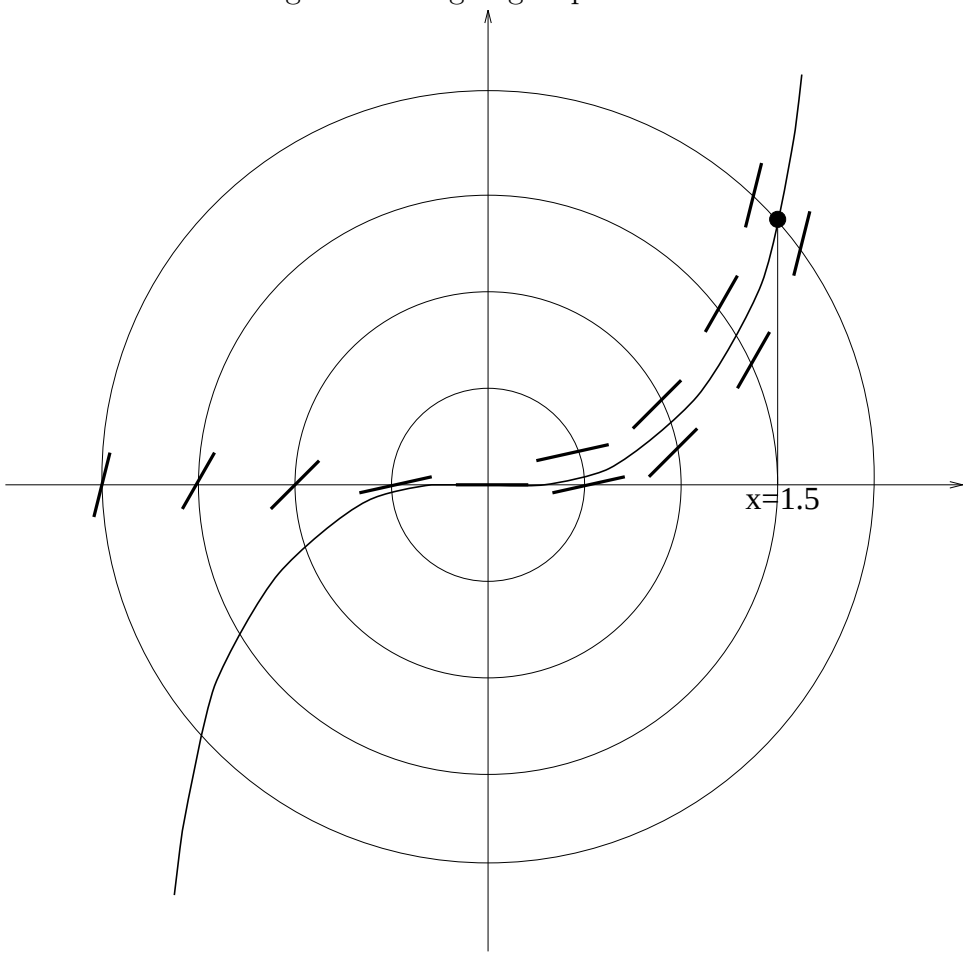
What is the simplest particular solution you can find?

5 Numerical Solutions

5.1 Euler's method

Many differential equations cannot be solved exactly, and in applied mathematical problems it is often necessary to make do with approximate or numerical solutions. In this section we shall investigate simple methods for finding these, but we first discuss the geometrical significance of an ODE.

Figure 8: Assigning slopes



Example 5.1.1

$$\frac{dy}{dx} = x^2 + y^2. \tag{5.1}$$

If the term 'x^2' were missing, the equation would be very easy and would admit solutions $y = 1/(c - x)$ where c is a constant. As it is, (5.1) is a difficult non-linear equation, though it does have a simple interpretation. The equation is asserting that the slope of

any solution passing through the point (x, y) equals the square of the distance of that point from the origin. An accurate sketch then enables one to make a reasonable guess at solutions, at least those near the origin. For example, Figure 8 indicates that the solution $y(x)$ to (5.1) satisfying the initial condition $y(0) = 0$ has $y(1.5)$ equal, at least roughly, to 1.5. Below, we shall explain a method that allows $y(1.5)$ to be approximated more accurately. □

Consider more generally the IVP

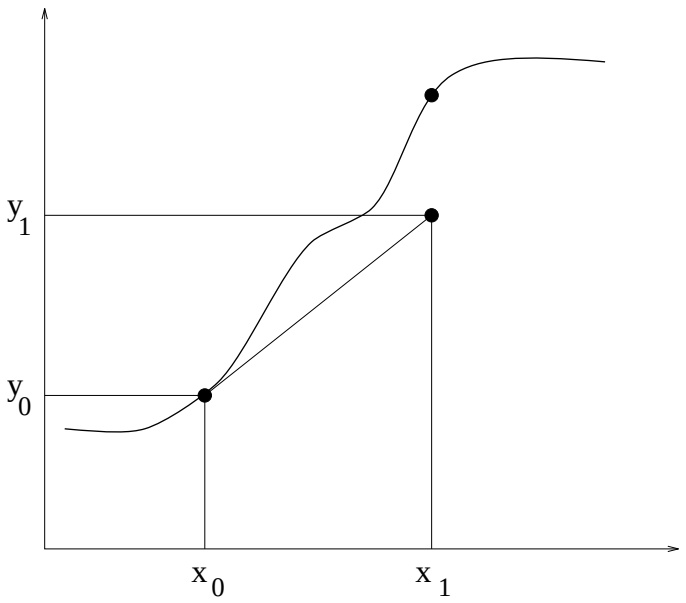
$$\begin{cases} y' = f(x, y), \\ y(x_0) = y_0, \end{cases}$$

where f is a given function and x_0, y_0 are given numbers. The solution is the curve passing through (x_0, y_0) whose slope at *any* point (x, y) equals $f(x, y)$. Choose a small distance h (called the 'step size') and let $x_1 = x_0 + h$. If we use the tangent line to the curve at (x_0, y_0) to approximate the curve, then for small h

$$y_1 = y_0 + h f(x_0, y_0)$$

should be close to the true value $y(x_1)$ of the solution at $x = x_1$ (see Figure 9).

Figure 9: Tangent approximation



Euler's method repeats the above idea with n steps, and calculates y_n recursively. To approximate the true value $y(x_0 + \ell)$ of the solution at $x = x_0 + \ell$ in n steps, one sets $h = \ell/n$ and puts into action the scheme

Start with	x_0	y_0
Define	$x_1 = x_0 + h$	$y_1 = y_0 + hf(x_0, y_0)$
	$x_2 = x_1 + h$	$y_2 = y_1 + hf(x_1, y_1)$
	.....	
	$x_i = x_{i-1} + h$	$\boxed{y_i = y_{i-1} + hf(x_{i-1}, y_{i-1})}$
	.....	
	$x_n = x_{n-1} + h$	$y_n = y_{n-1} + hf(x_{n-1}, y_{n-1})$

To determine the approximation y_2, we repeat the first step replacing x_0, y_0 by x_1, y_1. Whilst y_0 was the *true* value of the solution at x_0, y_1 is only an approximation to the true value $y(x_1)$. What is worse, $f(x_1, y_1)$ is the slope of a solution passing through (x_1, y_1), not the value $f(x_1, y(x_1))$ we should really use. Nonetheless the method can work quite well if h is small and n correspondingly large.

A word about notation: The subscript i is used above to indicate the general step, and the key formula is the one in the box. We shall always use n to denote the *total* number of steps, so that $x_n = x_0 + \ell$ and y_n denotes the final approximation.

The application of Euler's method is an example of an *algorithm*, a methodical procedure that takes input, carries out a finite number of steps with a clearly-defined stop, and produces output. The input consists of the values of h, n, x_0, y_0, and the most relevant output is the value of y_n. The intermediate steps are unambiguous and the whole process can be readily converted into a computer program that is capable of determining the approximation y_n with n very large.

Returning to Example 5.1.1, we shall first approximate the value of the solution satisfying $y(0) = 0$ at $x = 1.5$ using Euler's method with 15 steps. So take $n = 15$ and $h = 0.1$.

Start with	$x_0 = 0$	$y_0 = 0$
Define	$x_1 = 0.1$	$y_1 = 0 + 0.1(0 + 0) = 0$
	$x_2 = 0.2$	$y_2 = 0 + 0.1(0.01 + 0) = 0.001$
	$x_3 = 0.3$	$y_3 = 0.001 + 0.1(0.04 + 0.000001) = 0.005\ldots$
	$x_4 = 0.4$	$y_4 = 0.014\ldots$
	.....	
	$x_{10} = 1.0$	$y_{10} = 0.292\ldots$
	.....	
	$x_{15} = 1.5$	$y_{15} = 1.213\ldots$

Data in the table below was obtained using the MAPLE program given in §5.4, though it took an office machine a few minutes to report back y_{1500}.

n	1	15	150	1500	15000
a_n	0	$1.213\ldots$	$1.479\ldots$	$1.513\ldots$	$1.517\ldots$

Actually, the true solution satisfies $1.5174 < y(1.5) < 1.5175$. As it happens, solutions of (5.1) can be expressed in terms of so-called Bessel functions, although numerical methods are still needed to evaluate these. The rather naïve plot in Figure 8 conceals a much more complicated picture further out from the origin. In fact, the solution with $y(0) = 0$ tends to infinity as x approaches about 2, and solutions of the ODE to the right of this become increasingly steep and rapidly divergent (see §5.4).

5.2 Theoretical examples

The simplest type of initial value problem on which to try out Euler's method is

$$\begin{cases} y' = f(x), \\ y(a) = 0, \end{cases} \tag{5.2}$$

where a is a constant chosen to suit the definition of the function f. In this case, one is merely attempting to approximate the definite integral (1.11). With step size h, we obtain

$$x_i = a + ih, \quad y_i = y_{i-1} + h f(a + ih),$$

and with a total of n steps each of size ℓ/n,

Lemma 5.2.1 The true value $y(a + \ell)$ is approximated by

$$y_n = \frac{\ell}{n} \sum_{i=0}^{n-1} f(a + \frac{i\ell}{n}).$$

This amounts to approximating the area under the graph of f by means of n rectangles of width h. Although a silly application of Euler's method, proceeding blindly can give some intriguing summation formulae.

For example, taking $f(x) = 1/x$, $a = 1$, $\ell = 1$ gives

$$y_n = \sum_{i=n}^{2n-1} \frac{1}{i} \quad \text{as an approximation to} \quad y(2) = \ln 2. \tag{5.3}$$

For example, $y_{10} = \frac{1}{10} + \frac{1}{11} + \frac{1}{12} + \cdots + \frac{1}{19} = 0.718\ldots$, compared to the true value of $\ln 2 = .6931\ldots$. It is well known that the infinite series $\sum_{i=1}^{\infty} \frac{1}{i}$ *diverges*, which means that given any number M (however large) the finite sum starting from $i = 1$

$$\boxed{H_n = \sum_{i=1}^{n} \frac{1}{i}}$$

is greater than M for some n. (Infinite series are discussed a bit more in §7.4.) The sum H_n is called the nth *harmonic number* and it is easy to show that

$$\frac{1}{n} < H_n - \ln n < 1, \quad n \geq 2 \tag{5.4}$$

(see §5.4). This means that H_n, though unbounded, grows extremely slowly; incredibly for example, $H_{1000} < 8$. The validity of (5.3) is an easy consequence of the following much stronger statement that is beyond the scope of the course:

Theorem 5.2.2 The limit $\lim_{n\to\infty}(H_n - \ln n)$ exists and equals a number $\gamma = 0.5772\ldots$ (called Euler's constant). □

The most popular illustration of Euler's formula is to the IVP

$$\begin{cases} y' - y = 0, \\ y(0) = 1, \end{cases} \tag{5.5}$$

which has the exact solution $y(x) = e^x$. Similar techniques can be successfully applied to non-homogeneous ODE's like $y' + y = e^{\alpha x}$.

Problem 5.2.3 Apply Euler's method to (5.5) to approximate $y(1) = e$.

Solution. With step size h, the key formula is

$$y_i = y_{i-1} + hf(x_{i-1}, y_{i-1}) = y_{i-1}(1 + h),$$

which is an easy first-order difference equation. Working backwards, without expressing h in terms of n at this stage to avoid confusion, we get

$$\begin{aligned} y_n &= (1+h)y_{n-1} \\ &= (1+h)^2 y_{n-2} \\ &\cdots\cdots \\ &= (1+h)^n y_0. \end{aligned}$$

Now fix n and take $h = 1/n$ to give

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

as the sought-after approximation to e. □

This time, provided we quote some results from calculus, it is relatively easy to prove that the method works:

Proposition 5.2.4 Let $a_n = \left(1 + \frac{1}{n}\right)^n$. Then $\lim_{n\to\infty} a_n = e$.

Proof. We have

$$\ln a_n = n \ln(1 + \frac{1}{n}) = \frac{\ln(1+\varepsilon) - \ln 1}{\varepsilon},$$

where $\varepsilon = 1/n$. This means that

$$\lim_{n\to\infty} \ln a_n = \ln'(1) = \frac{1}{1} = 1$$

(using the definition of the derivative, or a special case of l'Hôpital's rule). Hence

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} \exp(\ln a_n) = \exp\left(\lim_{n\to\infty} \ln a_n\right) = e.$$

This last step is valid because the function $x \mapsto e^x$ is continuous (see §1.1). □

Like the previous approximation (5.3), this is inaccurate unless we take n to be very large. The table in the next section shows that n needs to be about 10^4 to get a_n correct to only 3 decimal places.

⋆ 5.3 An improvement

The relation between Euler's method and approximating definite integrals extends to the general case. Integrating the equation $y' = f(x, y)$ between x_{i-1} and $x_i = x_{i-1} + h$ gives

$$y(x_i) - y(x_{i-1}) = \int_{x_{i-1}}^{x_i} \frac{dy}{dx} dx = \int_{x_{i-1}}^{x_i} f(x, y(x)) dx, \tag{5.6}$$

where $y(x)$ is the true solution. The right-hand integral in (5.6) is approximated by the area $hf(x_{i-1}, y(x_{i-1}))$ of the rectangle of width h and height equal to the value of the integrand at x_{i-1}. However, we do not know $y(x_{i-1})$ and need to replace it by the previous approximation y_{i-1} to give the new approximation

$$y(x_i) - y_{i-1} \approx hf(x_{i-1}, y_{i-1}).$$

This turns out to be the Euler formula.

Taking a trapezium instead of a rectangle in an attempt to evaluate (5.6) gives a seemingly better approximation

$$y(x_i) - y_{i-1} \approx \tfrac{1}{2}h\left(f(x_{i-1}, y_{i-1}) + f(x_i, y(x_i))\right).$$

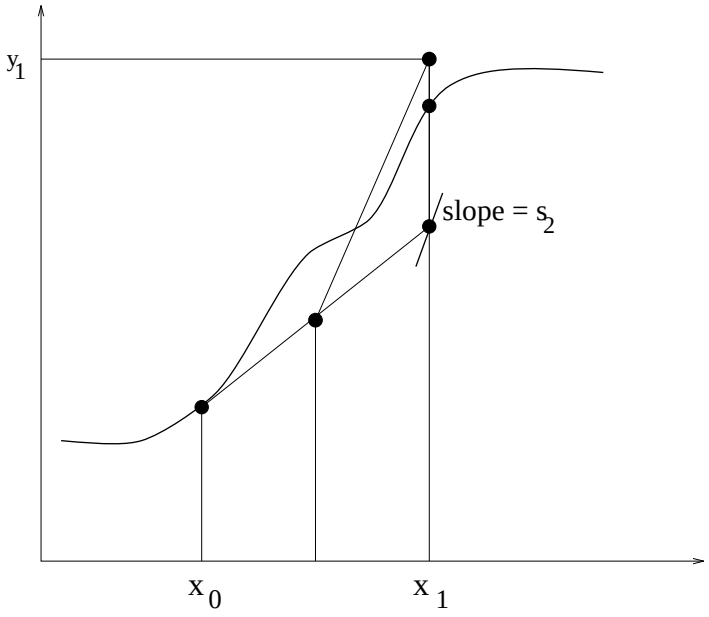
But it is $y(x_i)$ we are after, and if we are unable to solve for it explicitly, it makes sense to replace its occurrence on the right by the old Euler approximation $y_{i-1} + hf(x_{i-1}, y_{i-1})$. This yields

Heun's formula

$$y_i = y_{i-1} + \tfrac{1}{2}h(s_1 + s_2), \quad \text{where} \quad \begin{cases} s_1 = f(x_{i-1}, y_{i-1}), \\ s_2 = f(x_i, y_{i-1} + hs_1). \end{cases} \tag{5.7}$$

Figure 10 shows the first step graphically. The old approximation is first applied for the whole step in order to find the value of f at its endpoint. This value is then used to change the slope of the tangent line for the second half of the step.

Figure 10: Improved first step



Returning to the equation $y' = y$, we have $s_1 = y_{i-1}$, $s_2 = y_{i-1} + hy_{i-1}$, and

$$\begin{aligned} y_i &= y_{i-1} + \tfrac{1}{2}h(y_{i-1} + y_{i-1} + hy_{i-1}) \\ &= y_{i-1}(1 + h + \tfrac{1}{2}h^2). \end{aligned}$$

This leads to

$$b_n = \left(1 + \frac{1}{n} + \frac{1}{2n^2}\right)^n$$

as an approximation to e. The difference $|b_n - e|$ is referred to as the 'absolute error', and the ratio $|b_n - e|/e$ as the 'relative error', of the approximation. In §9.3, we shall prove that

$$\lim_{n\to\infty} 6n^2 \frac{|b_n - e|}{e} = 1, \tag{5.8}$$

which means that the relative error is decreasing roughly like $1/(6n^2)$ as n gets large.

The formula corresponding to (5.8) for the old approximation a_n in Proposition 5.2.4 would have an n instead of the quadratic term n^2, which makes it poorer.

n	a_n	b_n
1	2	2.5
5	2.48832	2.702708...
10	2.593742...	2.714080...
100	2.704813...	2.718236...
1000	2.716923	2.718281376
10000	2.718145	2.718281824

This is apparent in the table, in which the correct digits of e are underlined; if n is made 10 times as big, a_n generally improves by 1 decimal place, but b_n improves by 2.

The improved Euler method is still too inaccurate for serious use, although it is the prototype of more complicated 'predictor-corrector' methods. Most computer packages perform numerical integration with the so-called Runge-Kutta method, which is a refinement yielding quartic rather than quadratic accuracy [6, §20.1]. One version of this is defined below.

5.4 Exercises

1. The function e^{-5x} is the unique solution of the equation $y'+5y = 0$ satisfying $y(0) = 1$. Approximate e^{-5} by applying Euler's method to this equation on the interval $0 \leq x \leq 1$; take step size equal to (i) 0.5, (ii) 0.2, (iii) 0.1, and comment on the outcomes.

2. Apply Euler's method with 10 steps to the ODE $y' = (\ln x)y$, in order to approximate the value $y(2)$ of the solution satisfying $y(1) = 1$. Compare your answer with the true value.

3. Apply Euler's method to the IVP

$$\begin{cases} y' = \dfrac{y}{x} + x^3(y-x)^2, \\ y(1) = -4, \end{cases}$$

on the interval $1 \leq x \leq 2$. Calculate the resulting approximation to the value of the solution at $x = 2$, using step size equal to (i) 0.5 and (ii) 0.2.

4. Find the exact value of $y(2)$ in the previous question by first substituting $y = u + x$ and then applying a technique from §2.5.

5. (i) By considering the area under the graph of $1/x$, prove that

$$\sum_{j=2}^{n} \frac{1}{j} < \int_1^n \frac{1}{x}dx < \sum_{j=1}^{n-1} \frac{1}{j}, \quad n \geq 2,$$

and deduce (5.4).

(ii) Assuming Theorem 5.2.2, prove that $\lim\limits_{n\to\infty} \sum\limits_{i=n}^{2n-1} \frac{1}{i} = \ln 2$.

6. Euler's method can easily be run in MAPLE. Approximate the solution to $y' = x^2+y^2$ by first defining

```
f:= (x,y)->x^2+y^2:
x(0):=0: y(0):=0:
h:=0.001: n:=1500:
```

and computing

```
for i from 0 to n-1 do
  x(i+1):= x(i)+h:
  y(i+1):= y(i)+h*f(x(i),y(i))
od;
```

7. Let $y(x)$ denote the solution to (5.1) satisfying $y(0) = 0$, plotted in Figure 8. By comparing $y(x)$ to $1/(\frac{5}{2} - x)$, and given that $y(1.5) > 1$, show that y becomes infinite for $x < \frac{5}{2}$. Investigate the solution starting with

```
eq:= D(y)(x)=x^2+y^2:
dsolve({eq,y(0)=0},y(x)):
assign("):
plot(y(x),x=0..4);
```

8. The Runge-Kutta method is a generalization of (5.7) involving a combination of four approximations to $y(x_i)$, as specified by the program

```
for i from 0 to n-1 do
  s1:= h*f(x(i),y(i)):
  s2:= h*f(x(i)+h/2,y(i)+s1/2):
  s3:= h*f(x(i)+h/2,y(i)+s2/2):
  s4:= h*f(x(i)+h,y(i)+s3):
  x(i+1):= x(i)+h:
  y(i+1):= y(i)+(s1+2*s2+2*s3+s4)/6
od;
```

Try this out on (5.5) by supplying appropriate input.

6 Partial Derivatives

6.1 Functions of two variables

Let

$$f(x,y) = \frac{y}{x}.$$

We may think of f as a quantity which depends on *two* variables x, y, each of which can change independently. A physical example of such dependence arises when x and y are the volume and temperature of a gas, and $f(x,y)$ represents its pressure. For the ideal equation of state asserts that volume times pressure is proportional to temperature.

Definition 6.1.1 The *partial derivative of f with respect to x* is defined by treating y temporarily as a constant and differentiating $f(x,y)$ as a function of x.

One writes

$$\frac{\partial f}{\partial x} \text{ (or } f_x) \ = -\frac{y}{x^2};$$

this is a new function of two variables, and in the above example measures the rate of change of pressure with respect to volume when the temperature is maintained constant. The precise mathematical definition of the partial derivative with respect to x is

$$\frac{\partial f}{\partial x}\Big|_{(a,b)} = f_x(a,b) = \lim_{x\to a} \frac{f(x,b) - f(a,b)}{x-a}$$

(equal in the example to $-b/a^2$). Similarly, the partial derivative with respect to y is given by

$$\frac{\partial f}{\partial y}\Big|_{(a,b)} = f_y(a,b) = \lim_{y\to b} \frac{f(a,y) - f(a,b)}{y-b}$$

(equal to $1/a$).

We may also define various higher partial derivatives:

$$\begin{aligned}
&\frac{\partial}{\partial x}(\frac{\partial f}{\partial x}) = \frac{\partial^2 f}{\partial x^2} = (f_x)_x = f_{xx} = \frac{2y}{x^3},\\
&\frac{\partial}{\partial y}(\frac{\partial f}{\partial x}) = \frac{\partial^2 f}{\partial y \partial x} = (f_x)_y = f_{xy} = -\frac{1}{x^2},\\
&\frac{\partial}{\partial x}(\frac{\partial f}{\partial y}) = \frac{\partial^2 f}{\partial x \partial y} = (f_y)_x = f_{yx} = -\frac{1}{x^2},\\
&\frac{\partial^2 f}{\partial y^2} = f_{yy} = 0.
\end{aligned}$$

This illustrates an important result, namely that for all common functions one has

$$\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial^2 f}{\partial x \partial y}, \quad \text{or} \quad f_{xy} = f_{yx};$$

i.e., the 'mixed' second-order partial derivatives are equal. The precise statement is

Theorem 6.1.2 Let $f(x,y)$ be a function of two variables such that $f, f_x, f_y, f_{xy}, f_{yx}$ are defined and continuous on some rectangle with centre (a,b). Then

$$f_{xy}(a,b) = f_{yx}(a,b).$$

□

We shall not prove this result which relies on the concept of continuity in the context of $\mathbb{R}^2$. Instead, we remark that the function

$$g(x,y) = \begin{cases} \dfrac{x^3y - y^3x}{x^2+y^2}, & \text{if} \quad (x,y) \neq (0,0), \\ 0, & \text{if} \quad x = 0 = y \end{cases} \tag{6.1}$$

furnishes a counterexample with $g_{xy}(0,0) \neq g_{yx}(0,0)$ (see §6.5).

6.2 The chain rule

Suppose that $x = \cos t$ and $y = 2\sin t$. We may regard x, y as coordinates of a particle moving around the ellipse $x^2 + (y^2/4) = 1$. Then

$$f(x,y) = \frac{y}{x} = \frac{2\sin t}{\cos t} = 2\tan t$$

becomes a function of t and we are at liberty to compute the ordinary derivative of f with respect to t, which is of course $2\sec^2 t$. The proposition immediately below asserts that this derivative with respect to t equals

$$\frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}, \quad \text{or} \quad f_x.x'(t) + f_y.y'(t),$$

which evaluates to

$$\frac{-y}{x^2}(-\sin t) + \frac{1}{x}(2\cos t) = \frac{2\sin^2 t + 2\cos^2 t}{\cos^2 t} = 2\sec^2 t,$$

as expected.

Proposition 6.2.1 If $h(t) = f(x(t), y(t))$ then $h'(t) = f_x\, x'(t) + f_y\, y'(t)$.

Proof of a special case. Suppose that

$$\begin{cases} x(t) = a + ct, \\ y(t) = b + dt, \end{cases}$$

so that $h(t) = f(a+ct, b+dt)$. Then

$$h'(0) \;=\; \lim_{\varepsilon \to 0} \frac{f(a+c\varepsilon, b+d\varepsilon) \;\; [-f(a, b+d\varepsilon) + f(a, b+d\varepsilon)] \;\; - f(a,b)}{\varepsilon}.$$

Now,

$$f(a+c\varepsilon, b+d\varepsilon) - f(a, b+d\varepsilon) = c\varepsilon.f_x(a+\lambda_1 c\varepsilon, b+d\varepsilon), \quad \lambda_1 \in (0,1),$$

by applying the mean value theorem to the function $t \mapsto f(t, b+d\varepsilon)$ (λ_1 depends on ε). Therefore

$$\begin{aligned} h'(0) &= \lim_{\varepsilon \to 0} [f_x(a+\lambda_1 c\varepsilon, b+d\varepsilon)c + f_y(a, b+\lambda_2 d\varepsilon)d], \quad \lambda_2 \in (0,1), \\ &= f_x(a,b)c + f_y(a,b)d, \end{aligned}$$

provided f_x, f_y are continuous as functions of two variables. This is the required answer, as $c = x'(t)$ and $d = y'(t)$. The general proof of the proposition is not so different as $x(t)$ can be approximated by the linear function $x(0)+x'(0)t$, and $y(t)$ similarly, though one would need to use Proposition 6.3.1 below and continuity of x', y'. □

The following more complicated situation can be omitted on a first reading. Suppose now that x and y both depend themselves on three variables s, t, u, so that

$$h(s,t,u) = f(x(s,t,u), y(s,t,u)).$$

Then in the language of partial derivatives, Proposition 6.2.1 gives

$$\begin{cases} h_s = f_x x_s + f_y y_s, \\ h_t = f_x x_t + f_y y_t, \\ h_u = f_x x_u + f_y y_u. \end{cases}$$

These equations are best interpreted using a matrix scheme, regarding the function h as a composition of mappings from right to left:

$$\begin{array}{ccccc} \mathbb{R}^1 & \xleftarrow{f} & \mathbb{R}^2 & \longleftarrow & \mathbb{R}^3 \\ h(s,t,u) & \leftarrow & \begin{pmatrix} x(s,t,u) \\ y(s,t,u) \end{pmatrix} & \leftarrow & \begin{pmatrix} s \\ t \\ u \end{pmatrix} \end{array}$$

The chain rule then expresses the partial derivatives of h as the matrix product

$$(h_s, h_t, h_u) = (f_x, f_y) \begin{pmatrix} x_s & x_t & x_u \\ y_s & y_t & y_u \end{pmatrix}.$$

6.3 Homogeneous functions

Consider the function

$$\theta = \arctan(\frac{y}{x})$$

that represents the angle in polar coordinates of a point (x, y) in the plane. Then

$$\theta_x = \frac{1}{1+(\frac{y}{x})^2}.\frac{-y}{x^2} = \frac{-y}{x^2+y^2}. \tag{6.2}$$

To see this, put $u = y/x$ and use the chain rule of §1.1 and the fact that $d\theta/du = 1/(du/d\theta) = \sec^2\theta = 1+u^2$. Similarly,

$$\theta_y = \frac{1}{1+(\frac{y}{x})^2}.\frac{1}{x} = \frac{x}{x^2+y^2}. \tag{6.3}$$

This example illustrates a more modest type of chain rule involving more than one variable:

Proposition 6.3.1 If $h(x,y) = g(f(x,y))$ then

$$\begin{cases} h_x = g'(f(x,y))f_x, \\ h_y = g'(f(x,y))f_y. \end{cases}$$

In contrast to Proposition 6.2.1, no proof is required beyond that of the ordinary chain rule in §1.1.

Definition 6.3.2 A function f of two variables is said to be *homogeneous of weight* w if

$$f(tx,ty) = t^w f(x,y), \quad t \in \mathbb{R}$$

(valid at least for all x, y, t for which $f(tx,ty)$ is defined).

For example, $ax+by$ is homogeneous of weight 1, as is $r(x,y) = \sqrt{x^2+y^2}$ if one restricts to $t > 0$. On the other hand, y/x and θ are homogeneous of weight 0. Further,

$$x^n + x^{n-1}y + x^{n-2}y^2 + \cdots + y^n = \frac{x^{n+1} - y^{n+1}}{x - y} \tag{6.4}$$

is homogeneous of weight n (cf. (4.8)). More generally, a *homogeneous polynomial* of degree n is a sum of the form $\sum_{i=0}^{n} a_i x^i y^{n-i}$.

Equations (6.2),(6.3) together imply that $x\theta_x + y\theta_y = 0$. A similar result emerges if we let $p(x,y) = x^i y^{n-i}$ denote one of the terms in (6.4), and compute

$$xp_x + yp_y = x.ix^{i-1}.y^{n-i} + y.x^i(n-i)y^{n-i-1} = nx^i y^{n-i}.$$

These formulae are special cases of what is appropriately called *Euler's formula*:

Proposition 6.3.3 If $f(x,y)$ is homogeneous of weight w then

$$xf_x + yf_y = wf.$$

Proof. Fix c, d and let $h(t) = f(tc,td) = t^w f(c,d)$. From Proposition 6.2.1,

$$h'(t) = f_x(tc,td)c + f_y(tc,td)d = wt^{w-1}f(c,d),$$

and putting $t = 1$ gives

$$cf_x(c,d) + df_y(c,d) = wf(c,d).$$

This is a more precise statement of Euler's formula which avoids the confusion between subscripts and variables. □

An analogous formula hold for functions of 3 or more variables. Let $f(x,y,z)$ be a function defined at each point (x,y,z) of 3-dimensional space. The corresponding homogeneity condition $f(tx,ty,tz) = t^w f(x,y,z)$ has a geometrical interpretation: it implies that the values of f on the radial line joining the point (x,y,z) to the origin are

determined by the single number $f(x, y, z)$. Using the dot product for vectors, Euler's equation for a homogeneous function of 3 variables can be expressed in the neat form

$$(x, y, z) \cdot (\nabla f) = wf, \tag{6.5}$$

where

$$\nabla f = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}), \tag{6.6}$$

is the *gradient* of f [6, §8.9]. The left-hand side of (6.5) represents a 'directional derivative' of f in the radial direction.

⋆ 6.4 Some partial differential equations

Let c be a constant and g a function of 1 variable. Set $f(x, y) = x + cy$, and

$$h(x, y) = g(f(x, y)) = g(x + cy). \tag{6.7}$$

This is exactly the situation of Proposition 6.3.1, in which there is a composition of mappings:

$$h : \mathbb{R} \xleftarrow{g} \mathbb{R} \xleftarrow{f} \mathbb{R}^2.$$

Thus,

$$\begin{cases} h_x = g'.1 \\ h_y = g'.c \end{cases} \quad \text{and} \quad \begin{cases} h_{xx} = (g')_x = g'' \\ h_{yy} = (g'c)_y = g''c^2, \end{cases}$$

and, for *any* choice of g, $h(x, y)$ is a solution of

$$\boxed{h_{yy} - c^2 h_{xx} = 0.} \tag{6.8}$$

Equation (6.8) is the 1-dimensional wave equation with y representing time, and x displacement. It is easy to prove that its general solution has the form

$$h(x, y) = g_1(x + cy) + g_2(x - cy) \tag{6.9}$$

(see §6.5). Taking for example $g_1 = g_2 = \sin$ gives a 'stationary wave' $2 \sin x \cos(cy)$. Similar solutions can also be found by substituting $h(x, y) = h_1(x)h_2(y)$ into (6.8); the variables 'separate' leaving the two ODE's

$$h_1'' = a\, h_1, \quad h_2'' = ac^2 h_2,$$

where a is a constant. If $a < 0$ then these admit trigonometric solutions as in (2.2).

We have been tacitly assuming that $c \in \mathbb{R}$. But now let $c = i = \sqrt{-1}$ so that $f(x, y) = x + iy = \zeta$, say. Then (6.7) provides a solution to Laplace's equation

$$\boxed{h_{xx} + h_{yy} = 0} \tag{6.10}$$

provided the derivative $g'(\zeta)$ make sense with ζ a complex variable, and the method works when g is one of many standard functions:

Proposition 6.4.1 If $\zeta = x + iy$, the real and imaginary parts of ζ^n (for any $n \in \mathbb{Z}$), e^ζ and $\ln \zeta$ all satisfy (6.10). □

We do not justify this theoretically, since it is an easy matter to verify the solutions in each case. For example, the real and imaginary parts of ζ^2 are $x^2 - y^2$ and $2xy$ and obviously solve (6.10). For ζ^n with $n > 2$, see §6.5. Also

$$e^\zeta = e^{x+iy} = e^x(\cos y + i \sin y),$$

giving solutions $e^x \cos y$, $e^x \sin y$.

The situation for

$$\ln \zeta = \ln(re^{i\theta}) = \ln r + i\theta$$

is more tricky, as $\theta = \arctan(y/x)$ is ambiguous and $\ln \zeta$ is actually a 'multivalued' function. But there is no ambiguity about

$$\ln r = \tfrac{1}{2} \ln(x^2 + y^2)$$

which does solve (6.10). The quickest way to see this is to note (e.g. by differentiating the equation $r^2 = x^2 + y^2$ partially with respect to x) that $r_x = x/r$. Then $(\ln r)_x = x/r^2$ and

$$(\ln r)_{xx} + (\ln r)_{yy} = \Big(-\frac{2x^2}{r^4} + \frac{1}{r^2} \Big) + \Big(-\frac{2y^2}{r^4} + \frac{1}{r^2} \Big) = 0.$$

An extension of the last calculation concerns a problem relevant to the theory of gravitation:

Problem 6.4.2 Find a function g such that if r now denotes $\sqrt{x^2 + y^2 + z^2}$ then $h(x, y, z) = g(r)$ solves the 3-dimensional Laplace equation $h_{xx} + h_{yy} + h_{zz} = 0$.

Solution. To start, we have

$$h_x = g'(r) r_x = g' \frac{x}{r},$$

and so

$$h_{xx} = (g')_x \frac{x}{r} + g' \frac{1}{r} + g'(-\frac{x}{r^2}) \frac{x}{r} = g'' \frac{x^2}{r^2} + g'(\frac{1}{r} - \frac{x^2}{r^3}),$$

with similar expressions for h_{yy} and h_{zz}. Adding all three,

$$h_{xx} + h_{yy} + h_{zz} = g'' + g'(\frac{3}{r} - \frac{1}{r}) = g'' + \frac{2}{r} g';$$

the '3' here really arises as the dimension of the space we are considering. To solve Laplace's equation, we need

$$\begin{aligned} 0 = r^2(g'' + \frac{2}{r} g') = (r^2 g')' \quad &\Rightarrow \quad g' = \frac{c_1}{r^2} \\ &\Rightarrow \quad g(r) = -\frac{c_1}{r} + c_2. \end{aligned}$$

In particular $1/r$ is a solution. □

6.5 Exercises

1. (i) Compute the partial derivatives $f_x, f_y, f_{xx}, f_{xy}, f_{yx}, f_{yy}$ of the function $f(x,y) = e^{x \sin y}$, and verify that $f_{xy} = f_{yx}$.
(ii) Compute the partial derivatives g_x, g_y of (6.1), being careful to adopt the correct definition to find their values at $(0,0)$. Then prove that $g_{xy}(0,0) = -1$ and $g_{yx}(0,0) = 1$.

2. Cartesian coordinates x, y and polar coordinates r, θ are related by the equations $x = r\cos\theta$, $y = r\sin\theta$, with $r \geq 0$ and $0 \leq \theta < 2\pi$.
(i) Compute the partial derivatives $x_r, x_\theta, y_r, y_\theta$.
(ii) Express each of r, θ as a function of x, y and compute the corresponding partial derivatives $r_x, r_y, \theta_x, \theta_y$.
What is the relationship between the matrices $\begin{pmatrix} x_r & x_\theta \\ y_r & y_\theta \end{pmatrix}$, $\begin{pmatrix} r_x & r_y \\ \theta_x & \theta_y \end{pmatrix}$?

3. The coordinates of a particle moving in the plane are $x(t) = 2\cos t$ and $y(t) = \sin t$, where t denotes time. Verify that the particle moves along the ellipse $f \equiv 0$ where $f(x,y) = x^2 + 4y^2 - 4$, and sketch this curve.
Compute the partial derivatives f_x, f_y and verify that $f_x x'(t) + f_y y'(t) = 0$ for points on the curve. Evaluate the vectors $(x'(t), y'(t))$ and (f_x, f_y) when (i) $t = 0$, (ii) $t = \pi/4$, and represent them as arrows emanating from the corresponding point $(x(t), y(t))$ on your sketch.

4. Let $u = x + cy$ and $v = x - cy$. Show that (6.8) transforms into $h_{uv} = 0$, and deduce the general solution (6.9).

5. Let $\zeta = x + iy$. Show that the real part of ζ^n equals $\sum_{k=0}^{\lfloor n/2 \rfloor} (-1)^k \binom{n}{2k} x^{n-2k} y^{2k}$, where $\lfloor n/2 \rfloor$ means the largest integer less than or equal to $n/2$. Deduce that this function is a solution of Laplace's equation (6.10).

6. (i) Let $f(x,t) = y(\dfrac{x}{2\sqrt{t}})$, where y is defined by (1.13). Prove that f satisfies the heat equation

$$f_{xx} = f_t.$$

If f is defined as before, but with y arbitrary, what ODE needs to be satisfied by y for f to satisfy the same PDE?
(ii) Find solutions to the heat equation in the form $h_1(x)h_2(t)$.

7. Carry out the computation

```
x:= r*cos(t): y:= r*sin(t):
diff(f(x,y),r$2)+diff(f(x,y),r)/r+diff(f(x,y),t$2)/r^2:
simplify(");
```

and explain its relevance to Laplace's equation (6.10).

8. Figure 11 exhibits Dirac's electron equation in an unusual setting. It actually represents a *system* of four first-order PDE's, as ψ stands for complex-valued functions $\psi_i(x_1, x_2, x_3, t)$, $i = 1, 2, 3, 4$, each depending on spatial coordinates x_1, x_2, x_3 and time t. The symbols σ, ρ_1, ρ_3 denote 4×4 matrices, m is a constant, and ∇ encodes the operators $\partial_i = \partial/\partial x_i$ as in (6.6). Written out in full the equations become

$$i \begin{pmatrix} \frac{\partial \psi_1}{\partial t} \\ \frac{\partial \psi_2}{\partial t} \\ \frac{\partial \psi_3}{\partial t} \\ \frac{\partial \psi_4}{\partial t} \end{pmatrix} = \left\{ i \begin{pmatrix} 0 & 0 & \partial_3 & \partial_1 - i\partial_2 \\ 0 & 0 & \partial_1 + i\partial_2 & -\partial_3 \\ \partial_3 & \partial_1 - i\partial_2 & 0 & 0 \\ \partial_1 + i\partial_2 & -\partial_3 & 0 & 0 \end{pmatrix} + \begin{pmatrix} m & 0 & 0 & 0 \\ 0 & m & 0 & 0 \\ 0 & 0 & -m & 0 \\ 0 & 0 & 0 & -m \end{pmatrix} \right\} \begin{pmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \\ \psi_4 \end{pmatrix}$$

Find a solution in which ψ_1 and ψ_2 are both non-zero.

Figure 11

7 Binomial Coefficients

7.1 Pascal's triangle

Consider the expansion

$$(1+x_1)(1+x_2)(1+x_3)$$

$$= \underset{\emptyset}{1} + \underbrace{x_1+x_2+x_3}_{\{x_i\}} + \underbrace{x_2x_3+x_3x_1+x_1x_2}_{\{x_i,x_j\}} + \underset{X}{x_1x_2x_3}$$

consisting of 2^3 terms, each corresponding to an indicated subset of $X = \{x_1, x_2, x_3\}$. Putting $x_1 = x_2 = x_3 = x$ gives the familiar identity

$$(1+x)^3 = 1 + 3x + 3x^2 + x^3 = x^0 + 3x^1 + 3x^2 + x^3.$$

Thus, the coefficient of x^k counts the number of subsets of size k, for $0 \le k \le 3$.

From the binomial theorem (Proposition 1.2.2) we know more generally that the coefficient of x^k in $(1+x)^n$ equals

$$\frac{n^{\underline{k}}}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}. \tag{7.1}$$

The above argument condenses into

Lemma 7.1.1 The binomial coefficient $\binom{n}{k}$ equals the number of subsets of size k in a set of size n. □

This result may be regarded as giving an alternative definition of the coefficients $\binom{n}{k}$. Expressed in the more usual way, $\binom{n}{k}$ is the number of ways of choosing k objects from a total of n, without regard to order. In older books this number is denoted ${}^nC_k$. By contrast, the number of ways of choosing k objects from n in order (sometimes called 'ordered selection without repetition') equals $n^{\underline{k}}$, and this is also denoted ${}^nP_k$ or $(n)_k$.

A more convincing proof of the binomial theorem can be given directly by induction on n. The following homogeneous form of the theorem is easily seen to be equivalent to Proposition 1.2.2.

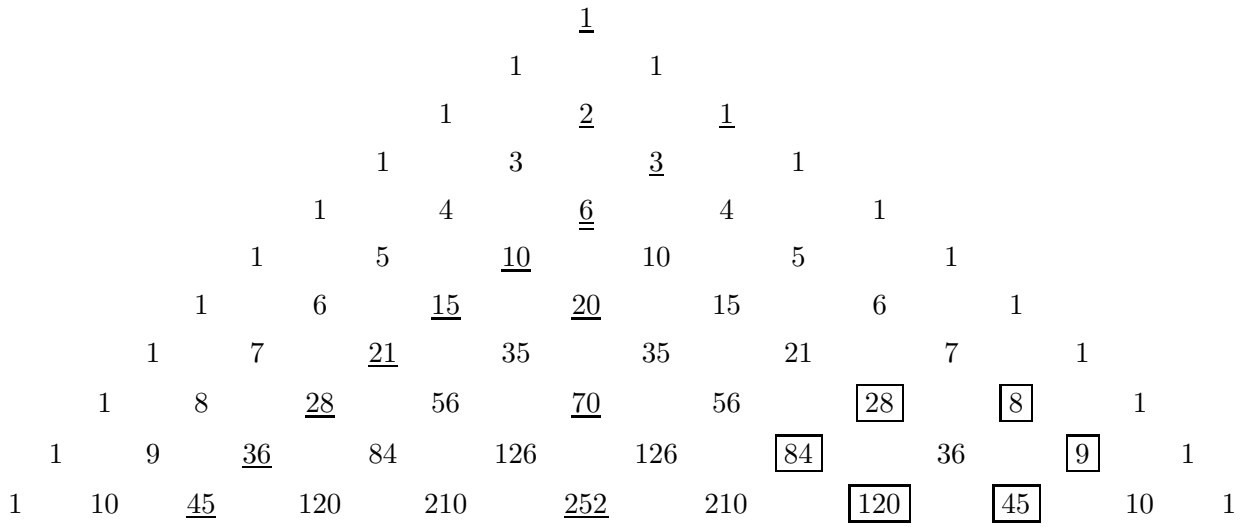
Proposition 7.1.2

$$(x+y)^n = \sum_{k=0}^{n} \frac{n!}{(n-k)!k!} x^k y^{n-k} \tag{7.2}$$

Proof. First observe that when $n = 0$, (7.2) becomes the equation $1 = \frac{0!}{0!0!}$. We deem this to be true by defining

$$0! = 1 \quad \text{and} \quad 0^0 = 1 \tag{7.3}$$

Figure 12:



(the latter is needed in case $x + y = 0$). Now suppose that (7.2) is true for a particular value of n, and consider

$$\begin{aligned}
(x+y)^{n+1} &= \sum_{k=0}^{n} \frac{n!}{(n-k)!k!} x^k y^{n-k}(x+y) \\
&= \sum_{j=1}^{n} \left(\frac{n!}{(n-j)!j!} + \frac{n!}{(n-j+1)!(j-1)!} \right) x^j y^{n+1-j} + x^{n+1} + y^{n+1} \\
&= \sum_{j=1}^{n} n! \frac{n-j+1+j}{(n-j+1)!j!} x^j y^{n+1-j} + x^{n+1} + y^{n+1} \\
&= \sum_{j=0}^{n+1} \frac{(n+1)!}{(n+1-j)!j!} x^j y^{n+1-j}.
\end{aligned}$$

The calculations are consistent with (7.3), and the last sum is simply (7.2) with n replaced by $n+1$. Starting with $n = 0$ and proceeding by induction therefore establishes the validity of (7.2) for all non-negative integers n. □

Many properties of the binomial coefficients are apparent from their triangular display in Figure 12 (discussed by Pascal in about 1650). The significance of the underlined and boxed numbers are explained in §7.4 and §7.5 respectively.

Lemma 7.1.3 The following identities are valid for $0 \le k \le n$:
(i) $\binom{n}{k} = \binom{n}{n-k}$;
(ii) $\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k}$;
(iii) $\sum_{j=0}^{n} \binom{n}{j} = 2^n$.

Proof. Each identity can be justified in two ways, depending whether the binomial coefficient is defined by factorials, or using Lemma 7.1.1.

(i) is obvious from the symmetry in the factorial formula. From the point of view of subsets, the result reflects the fact that we may associate to a subset S of $X = \{1, 2, \ldots, n\}$ of size k its complement $S^c = X \setminus S$ of size $n - k$.

(ii) is the addition formula that formed the crucial step in the proof of Proposition 7.1.2, and is extended by the convention that $\binom{n}{k} = 0$ whenever $k < 0$ or $k > n$. It corresponds to the fact that when we choose k elements from the set

$$\{1, 2, 3, \ldots, n + 1\} = X \cup \{n + 1\},$$

we may first decide whether or not to choose the element $n + 1$.

(iii) follows by setting $x = y = 1$ in Proposition 7.1.2, and reflects the fact that the total number of subsets of X is obviously 2^n as each element is either 'in' or 'out'. □

7.2 Probabilities

Motivated by Lemma 7.1.3(iii), suppose now that $x + y = 1$, so that

$$\sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} = 1. \tag{7.4}$$

In this section we wish to regard x and y as the probabilities of complementary events. Consider a situation consisting of a finite number of equally-likely outcomes to an experiment, so that if the experiment is repeated n times the number of possible outcomes will be raised to the power n. The probability of a particular result is defined informally to be

$$\frac{\text{number of outcomes yielding that result}}{\text{total number of outcomes}}.$$

It provides a way of 'measuring' the size of a subset of the set of all outcomes.

Suppose, for example, that a pair of dice is used to determine moves in a certain board game. If the dice have different colours, we can specify 36 different outcomes, each with probability 1/36. If we are desperate for two numbers that add up to 7, then the probability of success is easily seen to be $x = 6/36 = 1/6$, and of failure $y = 5/6$. If the dice are now thrown n times, there is a probability of $x^k y^{n-k}$ that the outcomes will follow a particular sequence such as

$$(7,*,*,7,7,\ldots,7,*)$$

with k 7's and $n - k$ totals different from 7 (indicated by the symbol $*$).

The probability that there will be a total of exactly k 7's amongst the n throws (let us call this eventuality E_k) equals $\binom{n}{k} x^k y^{n-k}$, since the binomial coefficient counts the number of ways of choosing k from n. Seen in this light, the binomial formula (7.4) simply asserts that the probabilities of the 'mutually exclusive' events E_k sum to 1.

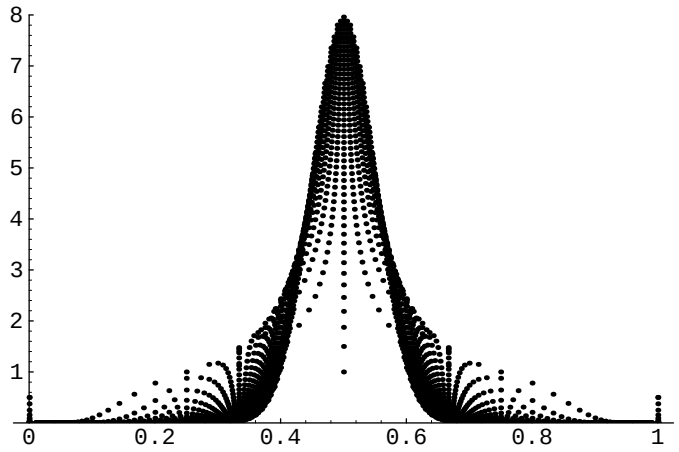
A simpler case is that in which $x = y = 1/2$, as with tossing an unbiased coin. The rows of Pascal's triangle confirm that, naturally enough, the most likely of the events E_k

are now those with k about half of n. As n gets large this favouring of middle values of k is represented by Figure 13 in which the points

$$\{(\frac{k}{n}, \frac{n}{2^n}\tbinom{n}{k}) \ : \ 0 \leq k \leq n\}$$

have been plotted for $1 \leq n \leq 100$.

Figure 13: Binomial plots



Problem 7.2.1 A poker player is dealt a hand of 5 cards from a conventional pack of 52, and the order of the 5 cards is immaterial. How many of the possible hands are
(i) flushes (meaning that all 5 cards belong to only one of the 4 suits)?
(ii) straights or runs (meaning that the cards are 5 consecutive ones in the list A,2,3,4,5,6, 7,8,9,10,J,Q,K,A)? Estimate the respective probabilities.

Solution. (i) To count the number of different (hands that are) flushes, we choose the suit and then the 5 cards from a possible 13. So the number is

$$4 \cdot \tbinom{13}{5} = 4 \cdot \frac{13 \cdot 12 \cdot 11 \cdot 10 \cdot 9}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 5148,$$

and the probability

$$\frac{5148}{\tbinom{52}{5}} = \frac{5148}{2598960} = 0.00198\ldots \tag{7.5}$$

is about 1 in 500. Alternatively, whatever the first card is, the next card must be one of 12 in the remaining 51, and so on; this gives directly the probability of a flush as

$$\frac{12}{51} \cdot \frac{11}{50} \cdot \frac{10}{49} \cdot \frac{9}{48} = \frac{4.13!47!}{52!8!},$$

which gives the same answer.

(ii) To count the number of runs, we choose a lowest card, which can be any one of A,2,3,...,10, and then suits for each of the 5. This gives a probability of

$$\frac{10 \cdot 4^5}{\binom{52}{5}} = \frac{10240}{2598960} = 0.00394\ldots,$$

almost twice (7.5). □

A famous example of a probability calculation with a surprising result is the so-called birthday paradox:

Example 7.2.2 Let $bd(k)$ be the probability that no two persons in a group of k share the same birthday. Then $bd(k) = 365^{\underline{k}}/365^k$.
This formula (that uses the notation (1.6)) is based on the ideal assumption that birthdays are equally distributed throughout the year and that none occur on 29 February. If the persons are placed in order and asked to call out their birthday one at a time, the total number of outcomes conceivable is 365^n. The number of these needed to prevent a common birthday is

$$365 \cdot 364 \cdot 363 \cdots (365 - k + 1) = 365^{\underline{k}},$$

whence the result.

In fact $bd(15) < 0.75$, $bd(23) < 0.5$ and $bd(32) < 0.25$. □

7.3 Generalized binomial coefficients

We know that the coefficient of x^k in $(1 + x)^n$ equals

$$\frac{n^{\underline{k}}}{k!} = \frac{n(n-1)\cdots(n-k+1)}{k!} \tag{7.6}$$

whenever $1 \le k \le n \in \mathbb{N}$. Given that the definition of the numerator $n^{\underline{k}}$ makes sense when n is replaced by an arbitrary real number r, we make

Definition 7.3.1 For any $r \in \mathbb{R}$ and $k \in \mathbb{N}$,

$$\binom{r}{k} = \begin{cases} \dfrac{r^{\underline{k}}}{k!} = \dfrac{r(r-1)(r-2)\cdots(r-k+1)}{k(k-1)(k-2)\cdots 1}, & k > 0, \\ 1, & k = 0, \\ 0, & k < 0. \end{cases}$$

These may be called 'binomial coefficients with real exponent r'. For instance,

$$\binom{4}{6} = \frac{4.3.2.1.0(-1)}{6.5.4.3.2.1} = 0,$$

$$\binom{1/2}{4} = \frac{\frac{1}{2}(-\frac{1}{2})(-\frac{3}{2})(-\frac{5}{2})}{4.3.2.1} = -\frac{5}{128}.$$

Note that $\binom{r}{0}$ equals 1 for all r by *definition*; this is consistent with (7.3).

Many of the familiar identities involving ordinary binomial coefficients have their counterparts in which the exponent is negative. This is often thanks to

Lemma 7.3.2 $\binom{-n}{k} = (-1)^k \binom{n+k-1}{k}$.

Proof. A straightforward application of the definition:

$$\frac{(-n)(-n-1)\cdots(-n-k+1)}{k!} = (-1)^k \frac{(n+k-1)(n+k-2)\cdots n}{k!}.$$

□

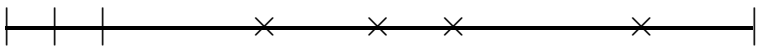
This lemma will be used in §7.4, together with

Lemma 7.3.3 $\binom{n+k-1}{k}$ equals the coefficient of x^k in $(1+x+x^2+\cdots+x^k)^n$.

Proof. This coefficient coincides with the coefficient of x^{k+n} in

$$x^n.(1+x+x^2+\cdots+x^k)^n = (x+x^2+x^3+\cdots+x^{k+1})^n,$$

and equals the number of ways of choosing n positive integers $a_1, \ldots, a_n$ with sum $k+n$ (so $1 \leq a_i \leq k+1$). This is also the number of ways of placing $n-1$ crosses strictly between 0 and $k+n$ (a_i being the distance between adjacent crosses or endpoints):



The answer is $\binom{n+k-1}{n-1} = \binom{n+k-1}{k}$. □

Somewhat more surprising is the way in which the generalized coefficients with $r = -1/2$ reduce to the 'middle' binomial coefficients:

Lemma 7.3.4 $(-4)^k \binom{-1/2}{k} = \binom{2k}{k}$.

Proof.

$$\begin{aligned}
\binom{-1/2}{k} &= \frac{(-\frac{1}{2})(-\frac{3}{2})\cdots(\frac{-2k+1}{2})}{k!} \\
&= (-1)^k \frac{1\cdot 3\cdot 5\cdots(2k-1)}{2^k k!} \cdot \frac{2\cdot 4\cdot 6\cdots 2k}{(1\cdot 2\cdot 3\cdots k)2^k} \\
&= (-1)^k \frac{(2k)!}{2^{2k}(k!)^2} \\
&= (-\tfrac{1}{4})^k \binom{2k}{k}.
\end{aligned}$$

□

7.4 Infinite series

First, a digression. Let $a = (a_0, a_1, a_2, \ldots)$ be a sequence of real numbers, and let $s_n = \sum_{i=0}^{n} a_i$ be the so-called *partial sums*. Then $s = (s_0, s_1, s_2, \ldots)$ is a new sequence. If s converges to ℓ, which simply means that $\lim_{n\to\infty} s_n$ exists and equals ℓ, one writes

$$\sum_{i=0}^{\infty} a_i = \ell,$$

and says that the infinite series *converges* to ℓ. As an example, suppose that $|x| < 1$ so that $x^n \to 0$ as $n \to \infty$. Taking $a_i = x^i$ and using the formula

$$1 + x + x^2 + \cdots + x^n = \frac{1 - x^{n+1}}{1 - x} = \frac{1}{1 - x} - \frac{x^{n+1}}{1 - x},$$

that follows from (4.8), yields

Lemma 7.4.1 $\displaystyle\sum_{i=0}^{\infty} x^i = \frac{1}{1 - x}$ for $|x| < 1$. □

In general, $a_n = s_n - s_{n-1}$, and if s converges it follows that $\lim_{n\to\infty} a_n = 0$. The converse is false though; an obvious counterexample is given by the harmonic series $\sum_{i=1}^{\infty} \frac{1}{i}$, which diverges in a way made precise by Theorem 5.2.2. In fact, for $|x| < 1$ it is legitimate to integrate the last lemma to get

$$x + \tfrac{1}{2}x^2 + \tfrac{1}{3}x^3 + \cdots = \sum_{i=1}^{\infty} \frac{1}{i} x^i = \int_0^x \frac{1}{1 - t} dt = -\ln(1 - x),$$

and as x approaches 1 the logarithm becomes infinite. Incidentally, one can also differentiate the lemma to obtain (after multiplying both sides by x) the useful formula

$$\sum_{i=1}^{\infty} i x^i = \frac{x}{(1 - x)^2}. \tag{7.7}$$

Lemmas from §7.3 show that $(-1)^k \binom{-n}{k}$ is the coefficient of x^k in the expansion

$$(1 + x + x^2 + x^3 + \cdots + x^k + \cdots)^n = \Big(\sum_{i\geq 0} x^i\Big)^n = \left(\frac{1}{1 - x}\right)^n.$$

It follows that if $|x| < 1$ then

$$(1 - x)^{-n} = \sum_{k=0}^{\infty} \binom{-n}{k} (-x)^k, \quad n \geq 1,$$

and combined with Proposition 1.2.2,

$$\boxed{(1+x)^n = \sum_{k=0}^{\infty} \binom{n}{k} x^k, \quad \text{for all } n \in \mathbb{Z}}$$

This is the binomial theorem for integer exponents.

Some simple instances are worth memorizing:

$$\begin{aligned} \frac{1}{(1-x)^2} = \sum_{k=0}^{\infty} \binom{-2}{k}(-x)^k &= \sum_{k=0}^{\infty} \binom{2+k-1}{k} x^k \\ &= \sum_{k=0}^{\infty} (k+1)x^k \\ &= 1 + 2x + 3x^2 + 4x^3 + 5x^4 + \cdots \end{aligned}$$

Similarly, the expansion

$$\frac{1}{(1-x)^3} = 1 + 3x + 6x^2 + 10x^3 + 15x^4 + 21x^5 + \cdots \tag{7.8}$$

involves the so-called triangle numbers.

In fact, Taylor's theorem may be used to prove the following 'generalized binomial theorem':

Theorem 7.4.2 Let $r \in \mathbb{R}$ and $|x| < 1$. Then $\sum_{k \geq 0} \binom{r}{k} x^k$ converges to $(1+x)^r$, i.e.

$$(1+x)^r = \sum_{k=0}^{\infty} \binom{r}{k} x^k.$$

□

As an illustration, Lemma 7.3.4 yields

$$\frac{1}{\sqrt{1-4x}} = \sum_{k=0}^{\infty} \binom{2k}{k} x^k = 1 + 2x + 6x^2 + 20x^3 + 70x^4 + 252x^5 + \cdots$$

Referring back to Figure 12 shows what has been accomplished in this section. The binomial coefficents with which we began specified the row entries of the triangle, whereas the most recent formulae provide (in the language of §8.1) generating functions for columns and diagonals.

7.5 Exercises

1. Let $n \geq 3$ be a positive integer.

(i) Use Proposition 1.2.2 to show that $\sum_{j=0}^{n} (-1)^j \binom{n}{j} = 0$.

(ii) By first differentiating, show that $\sum_{j=1}^{n}(-1)^j j\binom{n}{j} = 0$.

(iii) Is it true that $\sum_{j=1}^{n}(-1)^j j^2\binom{n}{j} = 0$?

2. Further to Problem 7.2.1, how many hands are running flushes (both (i) and (ii))? How many hands are 'full houses', meaning three cards are of one value and the remaining two of another (such K9KK9)?

3. The National Lottery consists in individuals choosing 6 numbers between 1 and 49, after which 7 numbers in the same range are drawn at random. A first prize is given for having chosen (in any order) the first 6 numbers drawn, and a second prize for having chosen (in any order) the 7th number drawn and any 5 of the first 6. Find the respective probabilities p_1, p_2 of success.

4. (i) Let $k, n \in \mathbb{N}$. Prove the 'negative' version of Lemma 7.1.3(ii), namely that

$$\binom{-n}{k} = \binom{-n-1}{k} + \binom{-n-1}{k-1}.$$

(ii) Applying Lemma 7.3.2 (with $m = n + k$) and (i) repeatedly, show that

$$\sum_{j=0}^{k}(-1)^j\binom{m}{j} = (-1)^k\binom{m-1}{k}.$$

5. A construction toy consists of 100 bricks of the same size but with varying numbers of the colours black, red and yellow. Show that there are 4851 different ways of putting together such a pack so that there is at least one brick of each colour. (It may help to know that 4851 is a binomial coefficient.) How many packs are there if it is no longer necessary that all colours be represented?

6. Let k, n be integers with $1 \leq k \leq n$. Prove the property

$$\binom{n-1}{k-1}\binom{n}{k+1}\binom{n+1}{k} = \binom{n-1}{k}\binom{n}{k-1}\binom{n+1}{k+1},$$

illustrated by the boxed numbers in Figure 12.

7. A total of n straight lines are drawn in the plane such that no two are parallel, and no three meet in a single point. Find the resulting number (i) a of points of intersection, (ii) b of line segments (finite or infinite), and (harder) (iii) c of regions (finite or infinite) the plane is divided into. Deduce that $a - b + c = 1$.

8. (i) Let $bd(k)$ be as in Example 7.2.2. Use the inequality $1 - x \leq e^{-x}$ (valid for all $x \in \mathbb{R}$) to prove that $\ln bd(k) \leq -\binom{k}{2}/365$.
(ii) Compare the resulting estimate for $bd(k)$ with its true value as follows:

```
for k to 30 do
  evalf([k,product(365-i,i=0..k-1)/365^k,exp(-k*(k-1)/730)])
od;
```

8 Generating Functions

8.1 Closed forms for sequences

Consider the following problem: Given a sequence $(a_0, a_1, a_2, \ldots)$, try to express

$$\sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + \cdots$$

as a function $g(x)$ in 'closed form'. If this is possible, the series is likely to converge for $|x|$ sufficiently small, though in any case $g(x)$ is called the *generating function* (GF) for the sequence. The theory of such generating functions is more concerned with formal manipulations, rather than worrying about questions of convergence which we shall return to in §9.3.

The concept is illustrated by some common GF's that arise from §7.4:

$$\begin{array}{lll} \dfrac{1}{1-x} & \text{is the GF of} & (1,1,1,1,1,\ldots), \\ \dfrac{1}{1+\lambda x} & " & (1,-\lambda,\lambda^2,-\lambda^3,\lambda^4,\ldots), \\ 1/\sqrt{1-4x} & " & (1,2,6,20,70,252,\ldots). \end{array}$$

To these we may add

$$\begin{array}{lll} e^x & \text{is the GF of} & (1,1,\frac{1}{2!},\frac{1}{3!},\frac{1}{4!},\ldots), \\ -\ln(1-x) & " & (0,1,\frac{1}{2},\frac{1}{3},\frac{1}{4},\ldots), \\ \sin x & " & (0,1,0,-\frac{1}{3!},0,\frac{1}{5!},\ldots). \end{array}$$

The above examples illustrate

Lemma 8.1.1 If $g(x)$ is the GF of $(a_0, a_1, a_2, \ldots)$ then

$$\begin{array}{lll} g'(x) & \text{is the GF of} & (a_1, 2a_2, 3a_3, \ldots), \text{ and} \\ \displaystyle\int g(x) & " & (?, a_0, \tfrac{1}{2}a_1, \tfrac{1}{3}a_2, \ldots). \end{array}$$

□

It is often convenient to write e^x as $\exp x$. The equation $\exp' = \exp$ reflects the fact that the sequence $(1, 1, \frac{1}{2!}, \frac{1}{3!}, \frac{1}{4!}, \ldots)$ is unchanged by the 'differentiation' operation described in the lemma. Rewriting Proposition 7.1.2 in the form

$$\frac{1}{n!}(x+y)^n = \sum_{k=0}^{n} \left(\frac{1}{k!}x^k\right)\left(\frac{1}{(n-k)!}y^{n-k}\right) \tag{8.1}$$

corresponds to the equation

$$\exp(x+y) = \exp x \cdot \exp y.$$

Problem 8.1.2 Use generating functions to show that $\sum_{k=0}^{n} \binom{n}{k}^2 = \binom{2n}{n}$.

Solution. We have

$$(1+x)^n = \sum_{j=0}^{n} \binom{n}{j} x^j, \qquad (1+\tfrac{1}{x})^n = \sum_{k=0}^{n} \binom{n}{k} \frac{1}{x^k}.$$

It follows that $\sum_{k=0}^{n} \binom{n}{k}^2$ equals the constant term in

$$(1+x)^n(1+\tfrac{1}{x})^n = \frac{1}{x^n}(1+x)^{2n},$$

which gives the required answer. □

Just how 'closed' a GF needs to be is a matter for debate, and therein lies the usefulness of the concept. Consider for example the sum

$$g(x) = \sum_{j=1}^{\infty} \frac{x^j}{1-x^j} \tag{8.2}$$

This may be expanded as a power series $a_0 + a_1x + a_2x^2 + \cdots$ by applying Lemma 7.4.1 to each denominator. The coefficient a_k is then obtained by summing from $j = 1$ to $j = k$ (there is no need to go any further), and equals the number of ways of writing x^k as $(x^j)^i$, or k as ij, for $1 \le i \le k$. We may therefore usefully regard (8.2) as the GF for the sequence

$$(0, 1, 2, 2, 3, 2, 4, 1, 4, 3, 4, 2, \ldots)$$

whose nth term is the number of positve-integer divisors of n (starting with $a_0 = 0$). The entry 2 characterizes the prime numbers, and 3 the squares of primes. The symbol ∞ in (8.2) is purely symbolic; it does not matter that the series does not converge since we may get as many coefficients as we need with a finite sum. More complicated sums of this type are discussed in [5, §17.10].

On the other hand, one can often guarantee convergence of a sequence, and increase the likelihood of finding a closed form as follows. One says that $(a_0, a_1, a_2, \ldots)$ is *bounded* if there exists $C > 0$ such that $|a_k| \le C$ for all $k \ge 0$. In this case, by comparison with the exponential series in which $a_k = 1$ for all k, the corresponding series

$$h(x) = \sum_{k=0}^{\infty} \frac{a_k}{k!} x^k$$

converges for *all* $x \in \mathbb{R}$. The function $h(x)$ is then called the *exponential generating function* of the original sequence $(a_0, a_1, a_2, \ldots)$. Thus,

e^x	is the exponential GF of	$(1, 1, 1, 1, \ldots)$,
$\cos x$	"	$(1, 0, -1, 0, 1, 0, \ldots)$,
$(1+x)^r$	"	$(1, r, r^{\underline{2}}, r^{\underline{3}}, \ldots)$, $\quad r \in \mathbb{R}$.

Example 8.1.3 The exponential GF for the sequence $(B_0, B_1, B_2, B_3, \ldots)$ of Bernoulli numbers is $x/(e^x - 1)$.

This defines the kth Bernoulli number by the formula

$$\boxed{\frac{x}{e^x - 1} = \sum_{k=0}^{\infty} \frac{B_k}{k!} x^k} \tag{8.3}$$

Now,

$$e^x - 1 = x(1 + \tfrac{1}{2!}x + \tfrac{1}{3!}x^2 + \tfrac{1}{4!}x^3 + \cdots),$$

and so the left-hand side of (8.3) may be expanded as

$$\begin{aligned} \left(1 + \tfrac{1}{2!}x + \tfrac{1}{3!}x^2 + \tfrac{1}{4!}x^3 + \cdots\right)^{-1} &= 1 + \sum_{j=1}^{\infty} (-1)^j \left(\tfrac{1}{2!}x + \tfrac{1}{3!}x^2 + \cdots\right)^j \\ &= 1 - \tfrac{1}{2}x + \tfrac{1}{12}x^2 - \tfrac{1}{720}x^4 + \cdots \end{aligned}$$

It follows that $B_0 = 1$, $B_1 = -\frac{1}{2}$, $B_2 = \frac{1}{6}$, $B_3 = 0$ and $B_4 = -\frac{1}{30}$.

In fact, one can show that $B_k = 0$ whenever $k \geq 3$ is odd. To see this, it suffices to observe that

$$\frac{x}{e^x - 1} + \frac{x}{2}$$

is unchanged when x is replaced by $-x$. □

8.2 Derangements

A permutation of $X = \{1, 2, \ldots, n\}$ is a reordering of its elements, i.e. a bijective map from the set to itself. An element $i \in X$ is called a *fixed point* of f if $f(i) = i$.

Definition 8.2.1 A *derangement* is a permutation in which no object stays in the same place, or a bijection of X with no fixed points. The number of derangements of n objects will be denoted by d_n.

It is easy to see that $d_1 = 0$, $d_2 = 1$, $d_3 = 2$, and $d_4 = 9$. The last number may be thought of as the number of ways of changing the places of 4 persons at a dinner table so that no-one stays in the same seat (similar problems tax the minds of Oxford tutors).

An amusing application is the following. Two 52-card decks of playing cards are shuffled and placed next to each other face up on a table. The two top cards (one from each deck) are inspected together to see if they coincide or 'snap' (such as 4♣, 4♣), and then discarded. This process is repeated until all cards have been looked at in pairs, and one asks:

What is the probability of getting through the decks with no snaps?

We need regard only the second deck as shuffled (in one of 52! ways). The number of shuffles with no snaps is then by definition d_{52}, so the answer is $d_{52}/52!$. We shall see that this is almost exactly $1/e$, which being less than $\frac{1}{2}$ means (perhaps surprisingly) that at least one snap is more likely than not.

Proposition 8.2.2 The exponential GF of the sequence $(d_0, d_1, d_2, \ldots)$ is $e^{-x}/(1-x)$.

Proof. Let $0 \le k \le n$. The number of permutations of $\{1, 2, \ldots, n\}$ with exactly k fixed points equals $\binom{n}{k} d_{n-k}$, though for this to work for $k = n$ we need to define $d_0 = 1$. Therefore

$$\begin{aligned} n! &= d_n + n d_{n-1} + \tbinom{n}{2} d_{n-2} + \cdots + \tbinom{n}{n-2} 1 + 0 + 1 \\ &= \sum_{k=0}^{n} \tbinom{n}{n-k} d_k \\ \Rightarrow \quad 1 &= \sum_{k=0}^{n} \frac{d_k}{k!} \cdot \frac{1}{(n-k)!}. \end{aligned}$$

The right-hand side equals the coefficient of x^n in

$$\Big(\sum_{k=0}^{\infty} \frac{d_k}{k!} x^k \Big) \left(\sum_{j=0}^{\infty} \frac{1}{j!} x^j \right) = h(x) e^x,$$

where $h(x)$ is the sought-after exponential GF. Thus,

$$1 + x + x^2 + x^3 + \cdots = h(x) e^x, \tag{8.4}$$

and the result follows from Lemma 7.4.1. □

Combining the expansion $e^{-x} = 1 - x + \frac{1}{2!}x^2 - \frac{1}{3!}x^3 + \cdots$ with (8.4) gives the key formula

$$\boxed{\frac{d_n}{n!} = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!} = \sum_{k=2}^{n} \frac{(-1)^k}{k!}} \tag{8.5}$$

For example,

$$\tfrac{1}{120} d_5 = \tfrac{1}{2} - \tfrac{1}{6} + \tfrac{1}{24} - \tfrac{1}{120} \quad \Rightarrow \quad d_5 = 60 - 20 + 5 - 1 = 44,$$

and in this way we may fill in some more values:

n	d_n	$d_n/n!$
0	1	1
1	0	0
2	1	0.5
3	2	0.333333...
4	9	0.375
5	44	0.366666...
6	265	0.368055...
7	1854	0.367857
8	14833	0.367881
9	133496	0.367879
10	1334961	0.367879
52	$\sim 3.10^{67}$	$1/e$ to 69 places

Remark. In the proof of Proposition 8.2.2, we used the following useful fact:

$$\left.\begin{array}{l} g(x) \text{ is the GF of } a = (a_0, a_1, a_2, \ldots) \\ h(x) \text{ is the GF of } b = (b_0, b_1, b_2, \ldots) \end{array}\right\}$$

$$\Rightarrow \quad g(x)h(x) \text{ is the GF of } c = (c_0, c_1, c_2, \ldots), \text{ where } c_n = \sum_{k=0}^{n} a_k b_{n-k}.$$

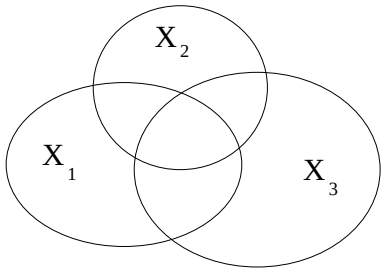
The sequence c is called the *convolution* of a and b; another example is (8.1).

8.3 The inclusion-exclusion principle

If $X_1, \cdots, X_n$ are subsets of a set X, we shall use the notation

$$X_{ij} = X_i \cap X_j, \quad X_{ijk} = X_i \cap X_j \cap X_k, \quad \text{etc.},$$

provided that $i < j < k < \cdots$. For $n = 3$, the Venn diagram



shows that

$$|X_1 \cup X_2 \cup X_3| = |X_1| + |X_2| + |X_3| - |X_{23}| - |X_{31}| - |X_{12}| + |X_{123}|.$$

This equation generalizes in an obvious way:

Proposition 8.3.1

$$|X_1 \cup \cdots \cup X_n| = \sum_i |X_i| - \sum_{i<j} |X_{ij}| + \sum_{i<j<k} |X_{ijk}| - \ \cdots \ - (-1)^n |X_{12\cdots n}|.$$

Proof. We must check that each element x in the union $X_1 \cup \cdots \cup X_n$ is counted exactly once on the right-hand side. This is correct for the following reason. Suppose, for sake of argument, that

$$\begin{array}{lllll} & x & \text{belongs to exactly} & 5 & \text{of the } X_i \\ \Rightarrow & x & " & \binom{5}{2} & " \ X_{ij} \\ \Rightarrow & x & " & \binom{5}{3} & " \ X_{ijk} \\ & & \cdots\cdots & & \end{array}$$

Then on the right-hand side of the equation we are trying to prove, the number of times x is counted equals

$$\binom{5}{1} - \binom{5}{2} + \binom{5}{3} - \binom{5}{4} + \binom{5}{5}$$

$$= 1 - \left(\binom{5}{0} - \binom{5}{1} + \binom{5}{2} - \binom{5}{3} + \binom{5}{4} - \binom{5}{5}\right) = 1 - (1-1)^5 = 1.$$

□

In practice, X_i may denote the number of objects of the set X having a particular property 'P_i'. Then X_{ij} denotes the set of objects possessing at least property P_i and property P_j, and so on. The number of objects in the set having none of the n properties equals $|X| - |X_1 \cup \cdots \cup X_n|$, which can be calculated by means of the proposition.

As an application, let X_i denote the set of permutations f of $\{1, 2, \ldots, n\}$ such that $f(i) = i$. Then

$$|X_i| = (n-1)!, \quad |X_{ij}| = (n-2)!, \quad \text{etc.}$$

The number of non-derangements of $\{1, 2, \ldots, n\}$ is then

$$\begin{aligned} n! - d_n = |X_1 \cup \cdots \cup X_n| &= \tbinom{n}{1}(n-1)! - \tbinom{n}{2}(n-2)! + \tbinom{n}{3}(n-3)! - \ \cdots \ \mp 1 \\ &= n!\left(1 - \tfrac{1}{2!} + \tfrac{1}{3!} - \ \cdots \ \mp \tfrac{1}{n!}\right). \end{aligned}$$

This is consistent with (8.5), either of which provide the formulae

$$d_n = \sum_{i=2}^{n} (-1)^i n^{\underline{n-i}} = n^{\underline{n-2}} - n^{\underline{n-3}} + n^{\underline{n-4}} - \ \cdots \ \pm 1$$

and

$$d_n = n d_{n-1} + (-1)^n. \tag{8.6}$$

An instance of the latter is clearly visible near the bottom of the table in §8.2.

Problem 8.3.2 How many anagrams of ABCDEFGH contain at least one of the pairs AB,CD,EF,GH (such as BACHDEFG)?

Solution. To proceed with the help of Proposition 8.3.1, let

$$\begin{array}{rcl} X_1 &=& \{\text{ permutations in which } \ AB \ \text{ persists }\}, \\ X_2 &=& \{ \qquad\qquad " \qquad\quad CD \qquad " \quad \}, \\ X_3 &=& \{ \qquad\qquad " \qquad\quad EF \qquad " \quad \}, \\ X_4 &=& \{ \qquad\qquad " \qquad\quad GH \qquad " \quad \}. \end{array}$$

To belong to X_1 a permutation must move AB as a single entity (think of scrabble letters glued together), so $|X_i| = 7!$. Similarly,

$$|X_{ij}| = 6!, \qquad |X_{ijk}| = 5!, \qquad |X_{ijk\ell}| = 4!$$

The answer is therefore

$$\tbinom{4}{1}7! - \tbinom{4}{2}6! + \tbinom{4}{3}5! - \tbinom{4}{4}4! = 16296.$$

□

8.4 Difference equations revisited

In this section, we shall find the generating function $F(x)$ for the Fibonacci sequence

$$(F_0, F_1, F_2, F_3, \ldots) = (0, 1, 1, 2, 3, 5, 8, \ldots)$$

defined by the difference equation $F_{n+2} - F_{n+1} - F_n = 0$, or

$$S^2F - SF - F = 0, \tag{8.7}$$

where SF and S^2F are the shifted sequences which start with F_1 and F_2 respectively (see §4.1).

Since $F(x) = F_0 + F_1x + F_2x^2 + \cdots$, with $F_0 = 0$ and $F_1 = 1$, the GF's of SF and S^2F are respectively

$$\begin{aligned} \frac{1}{x}F(x) &= F_1 + F_2x + F_3x^2 + \cdots \\ \frac{1}{x^2}F(x) - \frac{1}{x} &= F_2 + F_3x + F_4x^2 + \cdots \end{aligned}$$

Thus (8.7) becomes

$$\begin{aligned} &\frac{1}{x^2}F(x) - \frac{1}{x} - \frac{1}{x}F(x) - F(x) = 0 \\ \Rightarrow \quad &F(x) - x - xF(x) - x^2F(x) = 0, \end{aligned}$$

and so

$$\boxed{F(x) = \frac{x}{1 - x - x^2}} \tag{8.8}$$

We may check this answer with the aid of Lemma 7.4.1, in the same sort of way that we treated (8.2). For

$$\begin{aligned} F(x) = \frac{x}{1 - (1+x)x} &= x + (1+x)x^2 + (1+x)^2x^3 + (1+x)^3x^4 + \cdots \\ &= x + x^2 + 2x^3 + 3x^4 + 5x^5 + \cdots \end{aligned}$$

as expected. Notice that the x on top in (8.8) is not terribly important; it merely ensures that F_2 equals 1 rather than 2.

Equation (8.8) also leads to an independent proof of Corollary 4.2.2, since

$$\frac{x}{1 - x - x^2} = \frac{x}{(1 - \phi x)(1 - \hat{\phi}x)} = \frac{1}{\sqrt{5}}\left(\frac{1}{1 - \phi x} - \frac{1}{1 - \hat{\phi}x}\right),$$

where $\phi, \hat{\phi}$ are the roots of $x^2 - x - 1 = 0$. This gives

$$\begin{aligned} F(x) &= \frac{1}{\sqrt{5}}\Big((1 + \phi x + \phi^2x^2 + \cdots) - (1 + \hat{\phi}x + \hat{\phi}^2x^2 + \cdots)\Big) \\ \Rightarrow \quad F_n &= \frac{1}{\sqrt{5}}(\phi^n - \hat{\phi}^n). \end{aligned}$$

A similar technique can be applied effectively to solve the other difference equations discussed in §4.3. It is also possible to tackle some difference equations with non-constant coefficients using generating functions, and Lemma 8.1.1 can be useful for this purpose.

Problem 8.4.1 A positive integer consisting of e even digits and o odd digits is assigned the 'value' $2e + o$. Find the generating function $v(x) = \sum_{k=1}^{\infty} v_k x^k$, where v_k denotes the number of positive integers with value k.

Solution. Since there are 5 even digits and 5 odd ones, the number of k-digit 'integers' with value n equals the coefficient of x^n in $(5x+5x^2)^k$ (which is of course zero if $n < k$). But this includes numbers which start with the digit 0 which are not permitted. To correct for this, we need to subtract x^2 times the coefficient of x^{n-2} in $(5x + 5x^2)^{k-1}$. Hence,

$$v(x) = \sum_{k=1}^{\infty}(5x + 5x^2)^k - x^2 \sum_{k=0}^{\infty}(5x + 5x^2)^k = \frac{5x + 4x^4}{1 - 5x - 5x^2},$$

by Lemma 7.4.1. □

8.5 Exercises

1. Let $n \geq 2$. By multiplying both sides of the defining equation (8.3) for the Bernoulli numbers by $e^x - 1$, prove that $\sum_{j=0}^{n-1} \binom{n}{j} B_j = 0$. Use this to check that $B_4 = -\frac{1}{30}$ and to find B_6 and B_8.

2. If the decks in §8.2 each contain only 10 cards to start with (e.g. A to 10 of clubs), is the chance of at least one snap less or greater than with the 52-card decks?

3. Observe that (8.6) implies the formula $d_n = (n-1)(d_{n-1}+d_{n-2})$. Prove this directly by considering the effect of a permutation f of $\{1, 2, \ldots, n\}$ on the last digit n, considering two cases: (i) $f(f(n)) = n$, (ii) $f(f(n)) \neq n$.

4. Let n, p, q be positive integers with p, q prime. Show that the number of integers between 1 and n inclusive that are not divisible by p or q equals

$$n - \lfloor \frac{n}{p} \rfloor - \lfloor \frac{n}{q} \rfloor + \lfloor \frac{n}{pq} \rfloor,$$

where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x. Does this formula work if p and q are not both prime? Use Proposition 8.3.1 to extend the formula to the case of three primes, and count the number of integers between 1 and 1000 that are divisible by none of 5,7,11.

5. Explain why there are 7! permutations of A,B,C,D,E,F,G,H,I,J in which the word BEAD appears. Show that there are exactly 3583560 permutations of the same ten letters in which neither of the words BEAD, JIG appear.

6. Let n be a positive integer. A *partition* of n is an equation of the form

$$n = a_1 + a_2 + \cdots + a_k, \quad 1 \leq k \leq n,$$

where $a_i \in \mathbb{N}$ and $1 \leq a_1 \leq a_2 \leq \cdots \leq a_k$. Let p_n denote the number of distinct partitions of n, and let $p_0 = 1$.

(i) Show that p_n equals the number of solutions $(b_1, b_2, \ldots, b_n)$ to the equation $\sum_{j=1}^{n} jb_j = n$ where each $b_j \geq 0$ is a non-negative integer.

(ii) Deduce that the generating function of $(p_0, p_1, p_2, \ldots)$ equals $\prod_{k=1}^{\infty} (1 - x^k)^{-1}$.

7. (i) Determine the first few values of p_n:

```
p:= product(1/(1-x^k),k=1..10):
series(p,x=0,10);
```

(ii) Compute

```
series((5*x+4*x^2)/(1-5*x-5*x^2),x=0,10);
```

and find a recursion relation satisfied by the integers v_k of Problem 8.4.1.

8. Further to (8.2), let

```
f(x):= sum(x^j/(1-x^j)^2,j=1..20):
g(x):= sum(j*x^j/(1-x^j),j=1..20):
```

Explain why

```
series(f(x),x=0,20);
series(g(x),x=0,20);
```

give the same answer. What do the coefficients of this power series represent?

9 Asymptotic Notation

9.1 'O' Terminology

In this section we shall be concerned with the behaviour of sequences of real numbers, denoted $(a_n), (b_n), (x_n)$ etc., as n becomes very large.

Definition 9.1.1 $a_n = \mathrm{O}(b_n)$ means that there exists $C > 0$ and $N \in \mathbb{N}$ such that $|a_n| \leq C|b_n|$ for all $n \geq N$, or more informally,

$$\text{'}|a_n|/|b_n| \text{ is bounded as } n \to \infty\text{'}.$$

This implies that

$$|a_n| \leq C'|b_n| \text{ for } \textit{all } n \geq 0, \text{ where } C' = \max\{C, \tfrac{|a_0|}{|b_0|}, \cdots, \tfrac{|a_{N-1}|}{|b_{N-1}|}\},$$

provided no b_i is zero, but C' may be much less convenient to find than C and N. For example $n^{20}/2^n \to 0$ as $n \to \infty$ since $20 \ln n - n \ln 2 \to -\infty$, so $n^{20} = \mathrm{O}(2^n)$. Indeed,

$$\begin{aligned} &n^{20} \leq 1.2^n, \quad \text{for } n \geq N, \qquad \text{where } N = 144, \quad \text{and} \\ &n^{20} \leq C'2^n, \quad \text{for } n \geq 0, \qquad \text{where } C' = 3.3 \times 10^{20}. \end{aligned}$$

Definition 9.1.2 $a_n = \Omega(b_n)$ means the same as '$b_n = \mathrm{O}(a_n)$', and $a_n = \Theta(b_n)$ means the same as '$a_n = \mathrm{O}(b_n)$ and $a_n = \Omega(b_n)$'.
It follows that $a_n = \Theta(b_n)$ if and only if there exist $c, C > 0$ and $N \in \mathbb{N}$ such that $c|b_n| \leq |a_n| \leq C|b_n|$, $n \geq N$.

Recall that, if $a > 1$, logarithms to base a are defined by

$$x = \log_a y \quad \Longleftrightarrow \quad y = a^x. \tag{9.1}$$

Of particular importance are the choices $a = 2, e, 10$. In these notes, we use

ln	to denote the	'natural logarithm' $\log_e$,
log	"	'common logarithm' $\log_{10}$, and
lg	"	'binary logarithm' $\log_2$.

One deduces from (9.1) that

$$\log_a x = \frac{\ln x}{\ln a}.$$

It follows from this formula that $\log_a n = \Theta(\ln n)$, and ln, log and lg may be used interchangeably in formulae involving O or Ω.

Definition 9.1.3 $a_n \sim b_n$ means $|a_n|/|b_n| \to 1$ as $n \to \infty$.

It is an exercise in the ε, δ language that $a_n \sim b_n$ implies $a_n = \Theta(b_n)$. For instance, let $a_n = 4n^3 - 3n + 1$, so that a_n is positive for $n \geq 0$. Then

$$\frac{a_n}{n^3} = |4 - \frac{3}{n^2} + \frac{1}{n^3}| \leq 4 + \frac{3}{n^2} + \frac{1}{n^3} \leq 8, \quad n \geq 1,$$

so $a_n = \mathrm{O}(n^3)$. More careful examination before applying the absolute value signs reveals that $a_n/n^3 < 4$ for $n \geq 1$. Moreover, $a_n/n^3 \to 4$ as $n \to \infty$, so actually $a_n \sim 4n^3$. A less simple example follows from Theorem 5.2.2, which implies that $(H_n - \ln n)/\ln n \to 0$ and

$$H_n \sim \ln n.$$

The above terminology is summarized by the scheme:

$$\begin{array}{ccccc} \mathrm{O} & \longleftarrow & \Theta & \longrightarrow & \Omega \\ & & \uparrow & & \\ & & \sim & & \end{array}$$

Its real importance lies in the disregard of the precise values of the constants involved and the behaviour of finitely many terms of the sequence. This philosophy is emphasized by the following

Problem 9.1.4 A sequence (x_n) is defined recursively by setting

$$x_n = 2x_{\lfloor n/2 \rfloor} + n, \qquad n \geq 1,$$

and $x_0 = 0$, so that it begins

$$(0, 1, 4, 5, 12, 13, 16, 17, 32, 33, \ldots).$$

(The notation $\lfloor\ \rfloor$ is defined in (10.2).) Prove that $x_n = \mathrm{O}(n \lg n)$.

Solution. We shall establish this result by showing by induction that $x_n \leq cn \lg n$ for some $c > 0$. Suppose firstly that this is true for all $n \leq N - 1$, with N a fixed integer. Then

$$\begin{aligned} x_N = 2x_{\lfloor N/2 \rfloor} + N &\leq 2c\lfloor \tfrac{N}{2} \rfloor \lg(\lfloor \tfrac{N}{2} \rfloor) + N \\ &\leq cN(\lg N - 1) + N \\ &= cN \lg N + N(1 - c). \end{aligned}$$

The induction is therefore successful provided that $c \geq 1$. Now we look at the first terms of the sequence: ignore x_0, x_1, but notice that $x_2 = 4$ satisfies $x_2 \leq c2 \lg 2 = 2c$ provided $c \geq 2$. In conclusion then, $x_n \leq 2n \lg n$ for all $n \geq 2$. □

9.2 Rates of convergence

The next result helps in the examples that follow it.

Lemma 9.2.1 $\lim_{x \downarrow 0} x \ln x = 0$.

Proof. The notation tells us quite reasonably to restrict to positive values of x in evaluating the limit. The derivative $1 + \ln x$ of $x \ln x$ is negative for $x < 1/e$, so there exists $c > 0$ such that $-c < x \ln x < 0$ for $0 < x < 1$. Putting $x = y^2$ with $y > 0$,

$$|x \ln x| = 2y|y \ln y| \leq 2cy, \quad 0 < x < 1,$$

and the right-hand side tends to 0 as $x \downarrow 0$. □

Examples 9.2.2 (i) Taking $x = 1/n$ (for $n \in \mathbb{N}$) in the lemma gives $(\ln n)/n \to 0$ as $n \to \infty$. This also gives the well-known result that $n^{1/n} = e^{(\ln n)/n} \to 1$ as $n \to \infty$. Now $(e^y - 1)/y$ tends to $\exp'(0) = 1$ as $y \to 0$, and taking $y = (\ln n)/n$ gives

$$\boxed{n^{1/n} - 1 = \mathrm{O}\left(\frac{\ln n}{n}\right)}$$

(ii) The sequence

$$b_n = \left(1 + \frac{1}{n} + \frac{1}{2n^2}\right)^n \tag{9.2}$$

was introduced in §5.3 as an approximation to e. Problem 9.3.2 below implies that

$$\boxed{|b_n - e| = \mathrm{O}(\frac{1}{n^2})} \tag{9.3}$$

(iii) Suppose that the mapping $f\colon [0,1] \to [0,1]$ is a *contraction*, i.e. there exists a constant $k < 1$ such that

$$|f(x) - f(y)| \leq k|x - y|, \quad \text{for all } x, y \in [0,1].$$

The contraction mapping theorem (or rather its proof) states that in these circumstances, if $x_0 \in [0,1]$ is chosen and x_n is defined inductively by $x_n = f(x_{n-1})$ for all $n \geq 1$, then x_n converges to the unique point $s \in [0,1]$ such that $f(s) = s$. It is also known that

$$|x_n - s| \leq \frac{k^n}{1-k}|x_1 - x_0|,$$

and we can extract the essence of this formula by stating that

$$\boxed{|x_n - s| = \mathrm{O}(k^n)}$$

□

Each of the boxed equations gives some estimate on how quickly the relevant sequence approaches its limit, and we may now compare these different rates of convergence. For example, $1/n^2 = \mathrm{O}((\ln n)/n)$ because $1/(n \ln n)$ is bounded (indeed it tends to 0) as $n \to \infty$. Dealing with k^n is slightly harder, but $k^n n^2 = n^2 e^{-n|\ln k|}$ is bounded as $n \to \infty$, and so $k^n = \mathrm{O}(1/n^2)$. These relations can be summarized by the single line

$$\mathrm{O}(k^n) = \mathrm{O}(\frac{1}{n^2}) = \mathrm{O}(\frac{\ln n}{n}), \tag{9.4}$$

in which we have begun to widen the scope of (some would say abuse) the 'O' notation.

In (9.4), the expression $\mathrm{O}(b_n)$ is best understood as representing the *set*

$$\mathrm{O}(b_n) = \{\, (a_n) : |a_n|/|b_n| \text{ is bounded as } n \to \infty \,\}$$

of sequences which are 'smaller or comparable' to b_n. Each equality sign in (9.4) is then understood as really standing for the subset symbol '$\subseteq$'. As a slight extension of this idea, we can rewrite (9.3) as

$$b_n = e + \mathrm{O}(\frac{1}{n^2}), \tag{9.5}$$

by interpreting the equality as roughly '$\in$' and

$$e + \mathrm{O}(\frac{1}{n^2}) = \{(e + c_n) : |c_n|n^2 \text{ is bounded as } n \to \infty\}.$$

The syntax (9.5) is very common and leads to no confusion, provided one remembers the secret meanings of '=' and *never* swaps the left and right-hand sides of an equation.

We can fill in (9.4) so as to form a whole scale of sequences that converge to 0. For example, in the array

$$\mathrm{O}(\frac{1}{n^2}) = \mathrm{O}(\frac{1}{n \ln n}) = \mathrm{O}(\frac{1}{n}) = \mathrm{O}(\frac{\ln n}{n}) = \mathrm{O}(\frac{1}{\ln n}) = \mathrm{O}(\frac{1}{\ln \ln n}),$$

the further to the right a sequence is, the more slowly it converges to zero.

⋆ 9.3 Power series estimates

The 'O' notation can be applied equally well to a continuous variable x in place of n, to describe the behaviour of a function $g(x)$ as x approaches a fixed value or ∞. All the previous conventions apply except that it is essential to include the limiting value of x, in order to avoid ambiguity.

In the following result, $\mathrm{O}(x^{p+1})$ means any function $f(x)$ (or set of functions) such that $f(x)/x^{p+1}$ is bounded as $x \to 0$.

Proposition 9.3.1 Suppose that $\sum_{n=0}^{\infty} a_n x^n$ converges to a function $g(x)$ whenever $|x|$ is sufficiently small. Then for any $p \in \mathbb{N}$,

$$g(x) = \sum_{k=0}^{p} a_k x^k + \mathrm{O}(x^{p+1}) \quad \text{as } x \to 0.$$

Proof. For this we need to know that a power series possesses a radius of convergence R with the property that it converges when $|x| < R$ and does not converge for $|x| > R$. The value of R is determined by the limiting behaviour of the coefficients of the series [6, §14.2]. In our case,

$$\frac{1}{x^{p+1}} \left(g(x) - \sum_{k=0}^{p} a_k x^k \right) = \sum_{j=0}^{\infty} a_{j+p+1} x^j,$$

and the right-hand side converges, as it has the same radius of convergence as the original sum. □

For example,

$$e^x = 1 + x + \tfrac{1}{2!}x^2 + \cdots + \tfrac{1}{p!}x^p + \mathrm{O}(x^{p+1}) \quad \text{as } x \to 0, \tag{9.6}$$

though the corresponding statement would be blatantly false for $x \to \infty$. The equation (9.6) is equivalent to asserting that

$$\frac{e^x - 1 - x - \cdots - \frac{1}{p!}x^p}{x^{p+1}} = \mathrm{O}(1),$$

i.e. that the left-hand side is bounded. This can also be deduced by applying l'Hôpital's rule $p+1$ times; each time the numerator is differentiated it remains zero when $x = 0$.

More generally, the hypotheses of the proposition imply that

$$a_k = \frac{g^{(k)}(0)}{k!}$$

in analogy to (1.8), and what results is the so-called Maclaurin series of $g(x)$. We may revert to a discrete variable by replacing x by $1/n$ to give

$$g(\frac{1}{n}) = \sum_{k=0}^{p} \frac{a_k}{n^k} + \mathrm{O}(\frac{1}{n^{p+1}}) \quad \text{as } n \to \infty. \tag{9.7}$$

We next use this as the basis for some more advanced manipulation that displays the power of the 'O' notation.

Problem 9.3.2 Prove that (9.2) satisfies

$$b_n = e\left(1 - \frac{1}{6n^2} + \mathrm{O}(\frac{1}{n^3})\right).$$

Solution. As a first step, write

$$b_n = e.\exp\,(\,-1 + n\ln(1 + \frac{1}{n} + \frac{1}{2n^2})). \tag{9.8}$$

The series $\ln(1+x) = -\sum_{k=1}^{\infty}(-1)^k x^k/k$ converges for $|x| < 1$, so from (9.7),

$$\begin{aligned} \ln\,(1 + \frac{1}{n} + \frac{1}{2n^2}) &= (\frac{1}{n} + \frac{1}{2n^2}) - \tfrac{1}{2}(\frac{1}{n} + \frac{1}{2n^2})^2 + \tfrac{1}{3}(\frac{1}{n} + \frac{1}{2n^2})^3 + \mathrm{O}(\frac{1}{n^4}) \\ &= \frac{1}{n} - \frac{1}{6n^3} + \mathrm{O}(\frac{1}{n^4}). \end{aligned}$$

Subsituting this into (9.8) gives

$$b_n = e.\exp\,(\,-\frac{1}{6n^2} + \mathrm{O}(\frac{1}{n^3})),$$

and the result follows from (9.6). □

In all the asymptotic expansions above, the left-hand side approaches a finite limit as $n \to \infty$; below we shall meet some examples where this is no longer true.

9.4 Stirling's formula

In this section, we shall examine the behaviour of large factorials. Obviously, $n! < n^n$ for all $n \geq 2$, so without effort one obtains the crude estimate

$$n! = \mathrm{O}(n^n).$$

The following result gives much more precise information about the behaviour of $n!$ for large n.

Theorem 9.4.1

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \quad \text{as} \quad n \to \infty.$$

□

This can be restated in the form

$$(n!)^2 \sim 2\pi n^{2n+1}e^{-2n}$$
$$\Rightarrow \quad \lim_{n\to\infty}(n!)^2 n^{-2n-1}e^{2n} = 2\pi.$$

Let

$$S_n = \sqrt{2\pi n}(n/e)^n$$

denote Stirling's 'approximation' to $n!$. This is a bit of a misnomer, as both S_n and $n!$ are unbounded as $n \to \infty$, though the theorem asserts that

$$\left|\frac{S_n}{n!} - 1\right| = \frac{|S_n - n!|}{n!} \to 0 \text{ as } n \to \infty. \tag{9.9}$$

In other words, the *relative* error tends to zero; this is shown in the table. By contrast, the absolute error $|S_n - n!|$ definitely does *not* tend to 0; it is in fact unbounded.

n	$n!$	S_n	$\lvert S_n - n!\rvert/n!$
1	1	0.92...	0.077...
2	2	1.9...	0.040...
3	6	5.8...	0.027...
4	24	23...	0.020
5	120	118	0.016
6	720	710	0.013
7	5040	4980	0.011
8	40320	39902	0.010
9	362880	359536	0.009
10	3628800	3598695	0.008

Moreover,

$$\ln\left(\frac{S_n}{n!}\right) = \ln S_n - \ln(n!) = \tfrac{1}{2}\ln(2\pi) + \tfrac{1}{2}\ln n + n\ln n - n - \ln(n!)$$

The left-hand side tends to 0 as $n \to \infty$, and dividing throughout by $n \ln n$,

$$1 - \frac{\ln(n!)}{n \ln n} \to 0,$$

whence

Corollary 9.4.2

$$\ln(n!) \sim n \ln n \quad \text{as} \quad n \to \infty.$$

It is a false step to exponentiate both sides of the corollary to conclude that $n! \sim n^n$. In fact, Theorem 9.4.1 implies that $n!/n^n \to 0$ as $n \to \infty$. □

Some motivation helps to relate Stirling's formula to a more elementary topic, namely the relatively well-known identities

$$\sum_{k=1}^{n} k = \tfrac{1}{2}n(n+1), \quad \sum_{k=1}^{n} k^2 = \tfrac{1}{6}n(2n+1)(n+1).$$

These are special cases of the more general formula

$$\sum_{k=1}^{n} k^p = \frac{1}{p+1}\sum_{k=0}^{p} \binom{p+1}{k} B_k (n+1)^{p+1-k},$$

valid for any $p \geq 1$, where B_k are the Bernoulli numbers defined in §8.1. Analogous summation formulae exist for sums of other functions, and Stirling's approximation can be deduced from such a formula for

$$\ln(n!) = \sum_{k=2}^{n} \ln k.$$

In fact,

$$\ln S_n - \ln(n!) = -\sum_{k=1}^{p} \frac{B_{2k}}{2k(2k-1)n^{2k-1}} + \mathrm{O}(\frac{1}{n^{2p+1}}), \tag{9.10}$$

an equation also valid for any $p \geq 1$ [4, §9.6].

9.5 Exercises

1. Show that, as $n \to \infty$,
(i) $n^{1/n} = 1 + \frac{\ln n}{n} + \mathrm{O}((\frac{\ln n}{n})^2)$, and
(ii) $n(n^{1/n} - 1) - \ln n \to 0$.

2. Explain why $a_n \sim b_n$ does not in general imply that $e^{a_n} \sim e^{b_n}$. Does it imply that $\sqrt{a_n}+1 \sim \sqrt{b_n}+1$?

3. Decide which of the following estimates are true or false:

(i) $\sum_{j=1}^{n} j = \mathrm{O}(n), \quad \sum_{j=1}^{n} j = \Omega(n)$;

(ii) $\lg(n+1) = \mathrm{O}(\lg n), \quad \lg(n+1) = \Omega(\lg n)$;

(iii) $\sum_{j=1}^{2^n-1} \frac{1}{j} = \mathrm{O}(n), \quad \sum_{j=2}^{2^n} \frac{1}{j} = \Omega(n)$.

4. Explain what the following statements mean and prove them:

(i) $\mathrm{O}(a_n + b_n) = \mathrm{O}(a_n) + \mathrm{O}(b_n)$;

(ii) $\mathrm{O}(a_n{}^2) = \mathrm{O}(a_n)^2$;

(iii) $\mathrm{O}(\mathrm{O}(a_n)) = \mathrm{O}(a_n)$.

5. A sequence (x_n) is defined recursively by setting $x_n = x_{\lfloor n/2 \rfloor} + 1$ for $n \geq 1$, and $x_0 = 0$. Determine x_n by hand for $1 \leq n \leq 20$, and prove that $x_n = \Theta(\lg n)$ by showing separately by induction that there exists

(i) a constant C such that $x_n \leq C \lg n$ for all $n \geq 2$;

(ii) a constant c such that $x_n \geq c \lg n$ for all $n \geq 2$.

6. Use Stirling's formula to show that $\binom{2n}{n} \sim 4^n/\sqrt{n\pi}$ as $n \to \infty$, and rearrange this so as to obtain

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \cdot \frac{2 \cdot 4 \cdot 6 \cdots (2n-2) \cdot 2n}{1 \cdot 3 \cdot 5 \cdots (2n-3) \cdot (2n-1)} = \sqrt{\pi}.$$

Deduce that π can be expressed as an infinite product $2 \cdot \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdot \frac{8}{7} \cdots$

7. Assuming (9.10), show that

$$n! = S_n \exp\left(\frac{1}{12n} + \mathrm{O}(\frac{1}{n^3})\right),$$

and deduce that $|S_n - n!|/n! = \mathrm{O}(1/n)$.

8. Determine the constants

(i) $k = \lim_{n\to\infty} n(H_n - \ln n - \gamma)$,

(ii) $\ell = \lim_{n\to\infty} n^2(H_n - \ln n - \gamma - \frac{k}{n})$,

by experiment. (γ is called `gamma` in MAPLE.)

10 Euclid's Algorithm

10.1 Integer division

Let $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$ denote the set of natural numbers including 0 (i.e. **N**on-negative integers). Given $a, b \in \mathbb{N}$, one writes $a|b$ to mean 'a divides b', i.e.

$$\text{there exists } c \in \mathbb{N} \text{ such that } ac = b.$$

If $a|b$, we also say that 'a is a divisor or *factor* of b'. It is then immediate that for any $a, b, c \in \mathbb{N}$,
(i) $a|a$,
(ii) $a|b,\ b|a \Rightarrow a = b$,
(iii) $a|b,\ b|c \Rightarrow a|c$.
A relation on a set, like division on $\mathbb{N}$ or inequality $\leq$ on $\mathbb{N}$ or $\mathbb{R}$, satisfying these properties is called a *partial order*.

The reflexivity (i) and transitivity (iii) are identical to the corresponding properties in the definition of an equivalence relation, though (ii) above is the 'opposite' of the symmetry property because it says that a and b can only be related to each other if they are equal. To prove this, suppose that $b = ac_1$ and $a = bc_2$; then $a = ac_1c_2$ and (unless $a = 0 = b$), $c_1c_2 = 1$ forcing $c_1 = 1 = c_2$.

Division amongst the numbers $\{1, 2, 3, \ldots, 16\}$ is indicated in Figure 14. In a diagram of this type, a line joining a (below) to b (above) indicates not only that $a|b$ but also that there is no c such that $a|c$ and $c|b$. Notice that the hierarchy is unrelated to the usual ordering $\leq$ on integers; indeed if we were to include 0 it would go at the *top* of the diagram, not the bottom, since $a|0$ for any $a \in \mathbb{N}$, whereas 0 divides only itself.

Figure 14: Division as a partial order

Definition 10.1.1 Let $a, b \in \mathbb{N}$. Then $d \in \mathbb{N}$ is called the highest common factor or *greatest common divisor* (written $d = \gcd(a, b)$) of a, b if
(i) $d|a$ and $d|b$;
(ii) $e|a,\ e|b \Rightarrow e|d$.

Because condition (ii) reads $e|d$ and not $e \leq d$, it is not completely obvious that any pair (a, b) has such a d. On the other hand, if $d_1, d_2 \in \mathbb{N}$ both satisfy (i) and (ii) then $d_1|d_2$ and $d_2|d_1$, so $d_1 = d_2$ and the definite article 'the' is justified – greatest common divisors are unique.

In Figure 14, the gcd of two numbers occurs when paths downwards from them *first* meet; for instance $3 = \gcd(12, 15)$. The row of numbers above 1 are the 'primes':

Definition 10.1.2 Let $p \in \mathbb{N}$ and $p \geq 2$. Then p is a *prime number* if and only if the only factors of p in $\mathbb{N}$ are 1 and p.

Thus, if p is prime then for any $a \in \mathbb{N}$ we must have $\gcd(p, a) = 1$ or $\gcd(p, a) = p$. It is a well-known fact (Theorem 10.3.1 below) that any positive integer can be uniquely factored as a product of prime numbers, and given such factorizations of a and b it may be easy to spot $\gcd(a, b)$. For example

$$\left.\begin{array}{rcl} 5432 & = & 2^3.7.97, \\ 2345 & = & 5.7.67 \end{array}\right\} \quad \Rightarrow \quad \gcd(5432, 2345) = 7.$$

However the precise steps in this implication need further analysis, and we do not wish to assume the factorization theorem at this stage, as its use does not provide an effective procedure for computing the gcd of large numbers.

Let us add that

$$\gcd(a, 0) = a \quad \text{for all} \quad a \in \mathbb{N}. \tag{10.1}$$

This follows immediately from Definition 10.1.1, since $a|a$ and $a|0$, and if $e|a$ and $e|0$ then it is a tautology that $e|a$.

Lemma 10.1.3 Given $a, b \in \mathbb{N}$ with $b > 0$, there exist unique $q, r \in \mathbb{N}$ such that

$$a = qb + r, \quad \text{with} \quad 0 \leq r < b.$$

Proof. We shall regard this result as fairly self evident, though a precise recipe for a computer to determine the quotient q is called the 'division algorithm' (see [9]). It can be reduced to the task of finding $\lfloor x \rfloor$ for $x \in \mathbb{R}$, where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x and is read 'the floor of x':

$$\lfloor x \rfloor \leq x < \lfloor x \rfloor + 1, \quad \lfloor x \rfloor \in \mathbb{Z}. \tag{10.2}$$

Given a, b, we may take $q = \lfloor a/b \rfloor$ and $r = a - qb$. The fact that $0 \leq r < b$ follows from setting $x = a/b$ in (10.2) and multiplying both sides by b. Given two solutions $(q, r), (q', r')$ to the problem,

$$(q - q')b = r - r' \quad \Rightarrow \quad b|(r - r'),$$

which contradicts the assumptions unless $r = r'$ and $b = b'$. □

If a and a' have the same remainder $r = r'$ when they are divided by b then one says that a is *congruent* to a' modulo b, and this is written

$$a \equiv a' \bmod b, \quad \text{or} \quad a \equiv a' \ (b). \tag{10.3}$$

For fixed b, (10.3) defines an equivalence relation on $\mathbb{Z}$. If the equivalence class containing a is denoted $\overline{a}$, it is easy to see that $\overline{x+y} = \overline{x}+\overline{y}$, and $\overline{x \cdot y} = \overline{x}\cdot\overline{y}$. These equations allow 'modular' addition and multiplication to be defined on the set $\mathbb{Z}_b = \{\overline{0}, \overline{1}, \overline{2}, \overline{3}, \ldots, \overline{b-1}\}$ of equivalence classes.

The next result is a consequence of the previous lemma.

Corollary 10.1.4 Let $a, b \in \mathbb{N}$. Then $\gcd(a, b)$ exists and equals the least positive integer in the set $\{ua + vb : \ u, v \in \mathbb{Z}\}$.

Proof. Let d denote the least positive integer of the form $ua + vb$ with $u, v \in \mathbb{Z}$. We claim that $d|a$. For the lemma allows us to write

$$a = qd + r, \quad 0 \leq r < d,$$

which implies that r too can also be expressed in the form $ua + vb$. Since $r < d$, the only possibility is that $r = 0$, as required. Similarly $d|b$. Finally, if $e|a$ and $e|b$ then e must divide any number of the form $ua + vb$, and d in particular. □

10.2 Computing greatest common divisors

Proposition 10.2.1 Given $a, b \in \mathbb{N}$, with $b > 0$ and r as in Lemma 10.1.3,

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

Proof. Let $d = \gcd(a, b)$. First note that not only is $d|b$ true, but also $d|r$ because $d|a$ and $r = a - qb$; thus d is at least a *common* divisor of b and r. To show that it is the greatest, suppose that $e|b$ and $e|r$. Then $e|(qb + r)$ so e is a common divisor of a and b; by definition of d we have $e|d$. Therefore $d = \gcd(b, r)$. □

This simple result provides the key step that enables one to compute the gcd of two numbers by repeatedly applying Lemma 10.1.3. The result is Euclid's algorithm which we now illustrate.

Problem 10.2.2 Compute the greatest common divisor of 5432 and 2345.

Solution. We begin with $a = 5432$, $b = 2345$, find r and then repeat the same process having first replaced a by b and b by r. The work can be laid out as follows, in which diagonal arrows represent the replacements:

$$
\begin{array}{rcrcrl}
a & & q.\underline{b} & & \underline{r} & \\
5432 & = & 2.\underline{2345} & + & \underline{742} & \gcd(5432,2345) \\
 & \swarrow & & \swarrow & & \| \\
2345 & = & 3.\underline{742} & + & \underline{119} & \gcd(2345,742) \\
 & \swarrow & & \swarrow & & \| \\
742 & = & 6.\underline{119} & + & \underline{28} & \gcd(742,119) \\
 & \swarrow & & \swarrow & & \| \\
119 & = & 4.\underline{28} & + & \underline{7} & \gcd(119,28) \\
 & \swarrow & & \swarrow & & \| \\
28 & = & 4.\underline{7} & + & \underline{0} & \gcd(28,7)=7.
\end{array}
$$

The process stops as soon as the remainder becomes 0, since in that line b must divide a and $\gcd(a,b)=\gcd(b,0)=0$. Tracing the equalities of the last column back we obtain $\gcd(5432,2345)=7$. □

Remark. Notice that if one interchanges the two numbers in the above example, the process is 'self-correcting' in the sense that it resorts to the above after one extra line. This expresses the obvious fact that $\gcd(a,b)=\gcd(b,a)$:

$$
\begin{array}{rcrcrl}
2345 & = & 0.\underline{5432} & + & \underline{2345} & \gcd(2345,5432) \\
 & \swarrow & & \swarrow & & \| \\
5432 & = & 2.\underline{2345} & + & \underline{742} & \gcd(5432,2345)
\end{array}
$$

The above algorithm is the world's oldest, dating from 300 BC. It can be described by a flow diagram (see §12.1) that illustrates the key features of an algorithm mentioned in §5.1. However, only in middle of this century, with the advent of electronic computers, was the study of algorithms taken seriously and developed further. Nowadays this study is a vital area on the borderline between mathematics and computation.

The existence of $u,v\in\mathbb{Z}$ such that $7=5432u+2345v$, guaranteed by Corollary 10.1.4, can also be inferred by tracing back the calculations that led to 7 as the gcd. Indeed, we see that

$$
\begin{array}{lllllllllll}
7 & + & 4.28 & - & 119 & & & & & & = 0 \\
 & - & 4(28 & + & 6.119 & - & 742) & & & & = 0 \\
 & & & & 25(119 & + & 3.742 & - & 2345) & & = 0 \\
 & & & & & & -79(742 & + & 2.2345 & - & 5432) = 0 \\
\hline
7 & & & & & & & & -183.2345 & + & 79.5432 = 0
\end{array}
$$

Thus one solution is $u=79$ and $v=-183$. A more systematic way of finding u and v can be incorporated into the algorithm itself; the result is the so-called 'extended Euclid algorithm'. We shall omit the details, which can be found in the MAPLE manual [11] and in [9].

Turning Figure 14 upside down leads to the following concept that is 'dual' to the gcd:

Definition 10.2.3 Let $a, b \in \mathbb{N}$. Then $m \in \mathbb{N}$ is called the *lowest common multiple* (written $m = \mathrm{lcm}(a, b)$) of a, b if
(i) $a|m$ and $b|m$;
(ii) $a|n,\ b|n \Rightarrow m|n$.

Proposition 10.2.4

$$\mathrm{lcm}(a, b) = \frac{ab}{\gcd(a, b)}.$$

Proof. Given $a, b \in \mathbb{N}$, let $d = \gcd(a, b)$ and let $m = ab/d$. Since a/d and b/d are whole numbers, we have $a|m$ and $b|m$ so m is at least a *common* multiple. Now recall that $d = ua + vb$ for some $u, v \in \mathbb{Z}$, and suppose that $a|n$ and $b|n$. Setting $ac_1 = n = bc_2$ gives

$$dn = (ua + vb)n = ab(uc_2 + vc_1),$$

and so m divides n. Therefore $m = \mathrm{lcm}(a, b)$. □

⋆ 10.3 Prime numbers

If p is a prime number, then for any $a \in \mathbb{N}$, $\gcd(p, a)$ equals p or 1. Combined with Corollary 10.1.4 this leads to what is effectively an alternative definition of prime number that is exploited in more advanced algebra:

Proposition 10.3.1 p is a prime number if and only if

$$p|(ab) \text{ with } a, b \in \mathbb{N} \quad \Rightarrow \quad p|a \text{ or } p|b. \tag{10.4}$$

Proof. If p is not prime then $p = ab$ where $a, b > 1$ and certainly (10.4) is false. So let p be prime, and suppose that $p|(ab)$. If $\gcd(p, a) = p$ then $p|a$. Otherwise $\gcd(p, a) = 1$ and there exist $u, v \in \mathbb{Z}$ such that

$$up + va = 1 \quad \Rightarrow \quad (ub)p + v(ab) = b.$$

The last equation implies that $p|b$. □

Two integers a, b are said to be relatively prime or *coprime* if $\gcd(a, b) = 1$.

Proposition 10.3.2 Suppose that a, b, c are integers with a, b coprime and $b > 0$. Then the congruence $xa \equiv c \bmod b$ has an integer solution x with $0 \leq x < b$.

Proof. We need to find $x, k \in \mathbb{Z}$ such that $xa = c + kb$. By assumption, there exist $u, v \in \mathbb{Z}$ such that

$$ua + vb = 1 \quad \Rightarrow \quad (cu)a = c - (cv)b,$$

and we may take x to equal cu minus a suitable multiple of b.

Theorem 10.3.3 Any integer $a \geq 2$ can be written in exactly one way as

$$a = p_1^{r_1} p_2^{r_2} \cdots p_k^{r_k}, \tag{10.5}$$

where $p_1 < p_2 < \cdots < p_k$ are primes and $r_i \geq 1$.

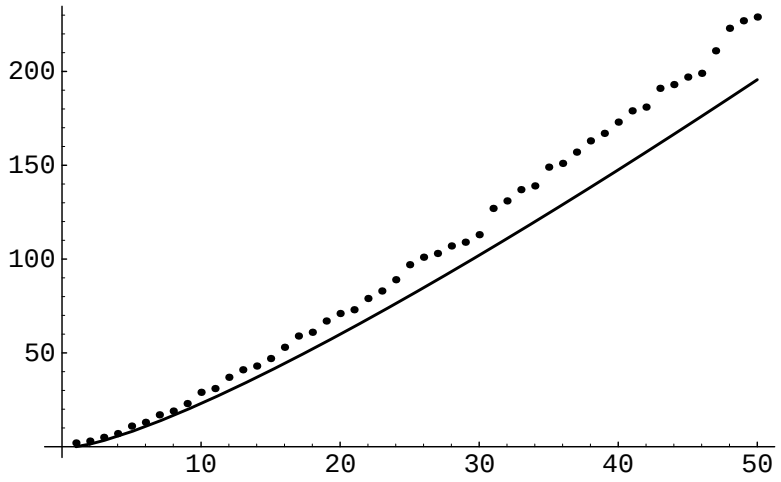
Proof. The existence of the factorization (10.5) is established by induction on a. It is obvious if $a = 2$. In general, if a is not itself prime it can be factored into the product of two smaller numbers for which factorization may be hypothesized. But then a itself is factored.

To prove the uniqueness of the factorization, we need property (10.4). For suppose that

$$p_1^{r_1} p_2^{r_2} \cdots p_k^{r_k} = q_1^{s_1} q_2^{s_2} \cdots q_\ell^{s_\ell},$$

where $p_1 < p_2 < \cdots < p_k$ and $q_1 < q_2 < \cdots < q_\ell$. Since p_1 divides the left-hand side, it must divide the right and writing the latter as a product in various ways one sees eventually that p_1 must divide at least one of $q_1, \ldots, q_\ell$. If $p_1 | q_i$ then since q_i is prime it must be that $p_1 = q_i$. But then the corresponding equation $q_1 = p_j$ is impossible unless $i = 1$. Hence $p_1 = q_1$, and the argument may be completed by induction. □

Figure 15: P_n and $x \ln x$



A corollary to Theorem 10.3.3 is the well-known fact that there are infinitely many prime numbers. For if $\{P_1, P_2, \ldots, P_n\}$ were a complete list of primes, the number

$$1 + \prod_{i=1}^{n} P_i$$

would have no prime factors. In any case, let P_n denote the nth prime number, so that $P_1 = 2$. It is a much more difficult problem to understand how fast the prime numbers grow, and thereby 'approximate' P_n in analogy to Stirling's formula in §9.4. This is accomplished by the so-called prime number theorem [5, §1.8], one version of which is

Theorem 10.3.4 $P_n \sim n \ln n$.

This is illustrated by Figure 15; the plots of primes and the graph of $x \ln x$ diverge sufficently slowly that the *ratio* $P_n/(n \ln n)$ approaches 1. The function $n \ln n$ will play an important role in the study of algorithms in §12.2. Observe also the first big 'gap': there are no prime numbers between $P_{30} = 113$ and $P_{31} = 127$.

10.4 Polynomial division

In this section we shall consider polynomials

$$A = A(x) = A_0 + A_1 x + A_2 x^2 + \cdots + A_n x^n$$

in which the coefficients A_i are real or complex numbers. Recall (cf. (1.7)) that A has *degree* n if $A_n \neq 0$, and is *monic* if $A_n = 1$. Polynomials of degree 0 are just numbers or *constants*, and polynomials of degree 1 are called *linear*.

Lemma 10.4.1 Given monic polynomials A, B, there exist unique polynomials Q, R such that

$$A = QB + R, \quad \text{with} \quad 0 \leq \deg R < \deg B.$$

If A, B are real, so are Q, R.

Proof. If $\deg A < \deg B$, one may simply take $R = A$ and $Q = 0$. Suppose inductively that the result is true whenever $\deg A < \alpha$, irrespective of B. Now take $\deg A = \alpha$ and $\deg B = \beta$ with $\alpha \geq \beta$. Since $\deg(A - Bx^{\alpha-\beta}) < \alpha$, there exist by assumption Q', R such that

$$A - Bx^{\alpha-\beta} = Q'B + R, \quad \text{or} \quad A = QB + R, \text{ with } 0 \leq \deg R < \deg B,$$

as required. Uniqueness of Q, R follows from the fact that A and B are monic. □

A number λ (real or complex) is called a *root* of a polynomial A if it satisfies the equation $A(\lambda) = 0$. In this situation, A is really being regarded as a function from $\mathbb{R}$, or $\mathbb{C}$, to itself. Lemma 10.4.1 allows us to write

$$A(x) = Q(x)(x - \lambda) + R,$$

where R is constant and zero if λ is a root of A. Whence the 'remainder theorem': λ is a root of A if and only if $A(x)$ is divisible by $x - \lambda$.

The fundamental theorem of algebra asserts that any polynomial A with complex coefficients has a root $\lambda = \lambda_1 \in \mathbb{C}$. It follows by induction that

$$A(x) = (x - \lambda_1)(x - \lambda_2) \cdots (x - \lambda_n), \quad \lambda_i \in \mathbb{C}, \tag{10.6}$$

where $n = \deg A$, though the roots λ_i need not of course be distinct. Using (10.6) and taking complex conjugates, it is easy to see that a polynomial with real coefficients can always be factored into linear and quadratic polynomials with real coefficients. It follows that there are no 'prime' (the correct word is *irreducible*) real polynomials of degree

greater than 2, though the situation is very different if one restricts to polynomials with integer coefficients (see §10.5).

A version of Euclid's algorithm works for real or complex polynomials, and also monic polynomials with integer coefficients. Strictly speaking one should first discuss each case separately, though for simplicity we restrict now to the case of real coefficients. One may then define the greatest common divisor of A, B following Definition 10.1.1, but if D_1, D_2 satisfy (i) and (ii) then all one may say is that $D_1 = aD_2$ for some $a \in \mathbb{R}$. We shall therefore write $\gcd(A, B) = D$ only if D satisfies the additional property of being monic.

The proof of Proposition 10.4.1 shows that

Proposition 10.4.2 If A, B are real monic polynomials of degree $\alpha \geq \beta$ then

$$\gcd(A, B) = \gcd(A - Bx^{\alpha-\beta}, B).$$

□

This is a simpler analogue of Proposition 10.2.1 than one might expect, and gives rise to a painless method for computing the greatest common divisor of two polynomials that we now illustrate.

Problem 10.4.3 Find $\gcd(2x^8 + 3x^7 + 1, x^5 + x^4 - x^3 - 1)$.

Solution. The gcd is computed by applying the proposition repeatedly, swapping the order of A, B and multiplying by constants where appropriate. The answer is

$$\begin{aligned}
& \gcd(2x^8 + 3x^7 + 1 - 2x^3(x^5 + x^4 - x^3 - 1), x^5 + x^4 - x^3 - 1) \\
= \ & \gcd(x^7 + 2x^6 + 2x^3 + 1 - x^2(x^5 + x^4 - x^3 - 1), x^5 + x^4 - x^3 - 1) \\
= \ & \gcd(x^6 + x^5 + 2x^3 + x^2 + 1 - x(x^5 + x^4 - x^3 - 1), x^5 + x^4 - x^3 - 1) \\
= \ & \gcd(x^4 + 2x^3 + x^2 + x + 1, x^5 + x^4 - x^3 - 1 - x(x^4 + 2x^3 + x^2 + x + 1)) \\
= \ & \gcd(x^4 + 2x^3 + x^2 + x + 1, 0).
\end{aligned}$$

The answer is therefore $x^4 + 2x^3 + x^2 + x + 1$, which by the remainder theorem actually equals $(1 + x)(x^3 + x^2 + 1)$. □

10.5 Exercises

1. Use Euclid's algorithm to find the greatest common divisor of

(i) 682, 545; (ii) 5432, 8730; (iii) 6765, 10946.

Retrace your steps in (i) so as to find integers u, v such that $682u + 545v = 1$.

2. Let $A(x) = 2x^{12} + 3x^3 + 1$ and $B(x) = x^9 - x^8 + x^7 + x^3 + 1$. Find the greatest common divisor, D, of A, B, and polynomials U, V such that $UA + VB = D$.

3. Let p be a prime number, and k a positive integer. Prove that
(i) $\binom{p}{k}$ is divisible by p provided $1 \leq k \leq p-1$;
(ii) $\binom{kp}{p}$ is divisible by p if and only if k is.

4. Suppose that $a, b \in \mathbb{N}$ and that $d = \gcd(a,b) = ua + vb$. Show that the general solution of the equation $xa + yb = d$ with $x, y \in \mathbb{Z}$ is

$$x = u + \frac{m}{a}k, \; y = v - \frac{m}{b}k,$$

where k is an arbitrary integer and $m = \operatorname{lcm}(a,b)$.

5. Let $a, b, c \in \mathbb{N}$. Using the defining properties of gcd, prove that

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

Explain why this number can reasonably be called the greatest common divisor of a, b and c. Determine its value when $a = 5432$, $b = 8730$ and $c = 460$.

6. Euclid's algorithm is already encoded into MAPLE using the command `gcd`. Nonetheless, try out the homemade version

```
gcd1:= proc(a,b)
  if b=0 then a
  else print(b); gcd1(b,a-b*trunc(a/b)) fi
end:
```

Compute `gcd1(100!+1,100^100-1);` and modify the program to find out how many times `b` is printed.

7. Explain why the following algorithm also computes gcd's of positive integers:

```
gcd2:= proc(a,b)
  if a<b then gcd2(b,a)
  elif a=b then a
  else gcd2(a-b,b) fi
end:
```

Is it as effective as `gcd1`?

8. (i) Find out how many irreducible integer-polynomial factors $1 + x^k$ has:

```
for k to 20 do k; factor(1+x^k) od;
```

For which values of k are there only 1 or 2 factors? Try to answer the same question for $1 - x^k$ before running the corresponding program.
(ii) Interpret the output from

```
for k to 100 do
  if x^k+x+1=Factor(x^k+x+1) mod 2 then print(k) fi
od;
```

11 Graphical Optimization

11.1 Graphs

Graphs are basically networks of points joined by lines. Whilst this course is not concerned with graph theory as such, we need some basic notions to understand the algorithms discussed below and the following will suit our purposes.

Definitions 11.1.1 (i) A *graph* consists of a set $\mathcal{V}$ and a symmetric relation $\sim$ on $\mathcal{V}$. Thus $u \sim v \Rightarrow v \sim u$, and we interpret elements of $\mathcal{V}$ as vertices and join u, v by an edge if and only if $u \sim v$. An edge is therefore an unordered pair $\{u, v\}$ (also denoted e_{uv} or uv) of vertices with $u \sim v$, and the set of edges is denoted by $\mathcal{E}$.

This definition excludes the possibility of multiple edges, though if one were to drop the qualification 'symmetric' one would get a *digraph* in which each edge is assigned a direction. To keep things simple we also exclude loops by assuming that $v \sim v$ is false for all $v \in \mathcal{V}$, i.e. $\sim$ is 'anti-reflexive'. It therefore follows that $|\mathcal{E}| \leq \binom{|\mathcal{V}|}{2}$.
(ii) A *path* is a sequence of vertices $(v_0, v_1, v_2, \ldots, v_n)$ with $v_i \sim v_{i+1}$ $(0 \leq i \leq n-1)$ and $v_0, \ldots, v_n$ distinct, except that v_0 is allowed to equal v_n if $n \geq 3$. A *cycle* is a closed path, meaning $v_n = v_0$ (so $n \geq 3$).

Figure 16:

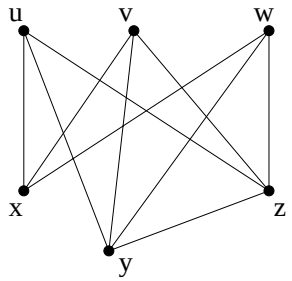


Figure 16 illustrates a graph with

$$\mathcal{V} = \{u, v, w, x, y, z\}, \quad \mathcal{E} = \{e_{ux}, e_{uy}, e_{uz}, e_{vx}, \ldots\},$$

$|\mathcal{V}| = 6$ and $|\mathcal{E}| = 10$. The sequence (y, v, z, y, u) is not a path, since it passes through an intermediate vertex twice (it's called a *trail*!), whereas (u, x, w, z, v, y, u) is a cycle.
(iii) A graph is *connected* if there exists a path between any two vertices. A *tree* is a connected graph with no cycles.

Proposition 11.1.2 A connected graph $\mathcal{G}$ has $|\mathcal{E}| \geq |\mathcal{V}| - 1$, and if $\mathcal{G}$ is a tree then $|\mathcal{E}| = |\mathcal{V}| - 1$.

Proof. Regard the edges of $\mathcal{G}$ as matchsticks or pieces of wire that are all separated, so that their endpoints form a total of $2|\mathcal{E}|$ vertices. Rebuild $\mathcal{G}$ by choosing an edge to start with, and adding back one edge at a time, so that the partially-constructed graph is connected at each stage. Each time one of the $|\mathcal{E}| - 1$ edges is added back, the total number of vertices drops by 1 or 2. Thus,

$$|\mathcal{V}| \leq 2|\mathcal{E}| - (|\mathcal{E}| - 1) = |\mathcal{E}| + 1. \tag{11.1}$$

If in the process of rebuilding, both ends of some edge need to be attached to the partially-constructed graph, then $\mathcal{G}$ must have a cycle. This cannot happen if $\mathcal{G}$ is a tree, so in that case there must be equality in (11.1). □

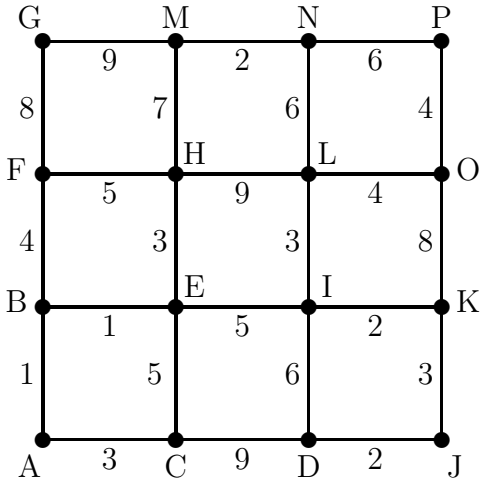
Any two vertices of a tree are connected by a unique path, because two such paths would together give rise to a cycle. Trees crop up in many instances in computing; consider for example a file directory structure. On the other hand, a human family tree need not be a tree in the above sense; for example, the parents of King George V were both descendants of George II, providing an Anglo-Danish cycle.

Definition 11.1.3 A graph is *weighted* if a positive number $w(e) > 0$ is assigned to each $e \in \mathcal{E}$. The weight of a subset $\mathcal{E}'$ of $\mathcal{E}$ is then the sum of the weights of its elements:

$$w(\mathcal{E}') = \sum_{e \in \mathcal{E}'} w(e).$$

There are many possible interpretations of weighted graphs. The most obvious are those in which the vertices represent geographical locations, and the edges correspond to existing transport routes, with w the time or distance associated to each. A variant is that in which edges represent potential routes under construction, with w the corresponding cost.

Figure 17: A weighted graph



Example 11.1.4 The graph in Figure 17 is weighted by the first 24 digits of π (which conveniently contain no zeros) in a snake-like fashion, commencing bottom left. This is somewhat artificial but it will serve well to illustrate the problems below. The square formed by $\mathcal{E}' = \{AC, CE, EB, BA\}$ has $w(\mathcal{E}') = 10$, whereas $w(\mathcal{E}) = 115$. □

11.2 Kruskal's algorithm

In this course, we shall be mostly concerned with the following problem that arises when one is given a connected weighted graph $\mathcal{G}$:

Of all the subsets of $\mathcal{E}$ that connect all the vertices together, find one of least weight.

A solution to this problem necessarily forms a tree, since if there were a cycle at least one of its edges could be removed without depriving any vertex. Thus, we shall call a solution to the problem a *minimum spanning tree* (MST); it is also called a *minimum connector*.

Applications that help visualize this problem include the construction of a network of irrigation canals or other utilities linking various sites, and electronic circuitry involving joint connections between a number of components.

Theorem 11.2.1 Given a connected weighted graph, an MST is obtained by successively choosing an edge of least weight not already selected, and discarding it if a cycle is formed.

This procedure is called *Kruskal's algorithm*, though its discovery is attributed to Boruvka in 1926. Choices may be involved if some edges have the same weight, and the solution is not then unique. To make the procedure trully methodical, one needs a strategy to select the successive edges of least weight, and in practice this might exploit the concept of a *heap* described in §12.3. However, on a modestly-sized graph, it suffices to place the edges in some linear order so that in the event of edges having equal weight one can choose the edge that comes first in this ordering.

Returning to Figure 17, we order the edges diagonally starting bottom left, following successive digits of π. The set of edges characterizing the MST constructed using Theorem 11.2.1 is then

$$\mathcal{E}_{\rm K} = \{\underbrace{AB, BE}_{1}, \underbrace{DJ, IK, MN}_{2}, \underbrace{AC, EH, IL, JK}_{3}, \underbrace{BF, LO, OP}_{4}, \underbrace{EI}_{5}, \underbrace{NL}_{6}, \underbrace{FG}_{8}\},$$

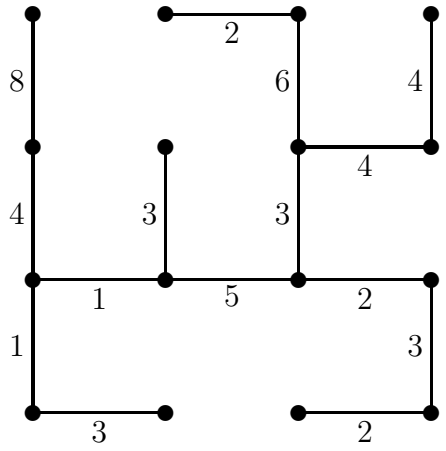
where the weights of each edge are indicated. Edges CE, FH, DI, NP, HM are discarded as they form cycles, and by the time FG is added all the vertices have been connected. The solution is illustrated in Figure 18, and has $w(\mathcal{E}_{\rm K}) = 51$. □

The original graph has $|\mathcal{V}| = 16$, $|\mathcal{E}| = 24$. It is an example of a 'planar graph' (meaning that when drawn on paper every intersection of edges is a vertex unlike in Figure 16), and in this case one always has

$$|\mathcal{E}| - |\mathcal{V}| = (\text{number of regions including the outside}) - 2. \qquad (11.2)$$

By contrast, the MST has $|\mathcal{E}_{\rm K}| = 15 = |\mathcal{V}| - 1$ as predicted by Proposition 11.1.2. The solution is not unique – we could replace the edge NL of weight 6 by PN, though NL happened to come first in our ordering.

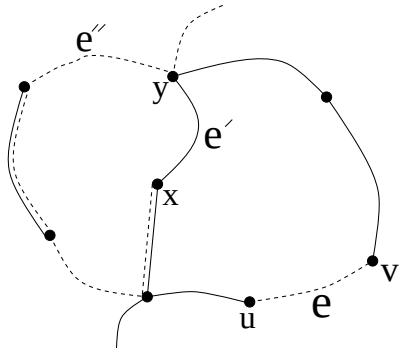
Figure 18: A minimum spanning tree



Kruskal's algorithm epitomizes a characteristic of a class of algorithms, namely it is 'greedy' in the sense that it makes choices that appear to be best at every stage. It is by no means obvious that these choices really do turn out to be best for the final solution, and thus one needs to verify the algorithm's 'correctness' by giving a

Proof of Theorem 11.2.1. Given the graph, the theorem constructs a subset $\mathcal{E}_{\mathrm{K}} \subseteq \mathcal{E}$ defining a spanning tree as in the example; the question is whether $\mathcal{E}_{\mathrm{K}}$ is minimal. A MST certainly exists, and if it is not unique we can choose one that has the largest number of edges in common with $\mathcal{E}_{\mathrm{K}}$; let $\mathcal{E}'$ be the set of edges of this MST.

Figure 19:



Choose an edge $e = e_{uv}$ in $\mathcal{E}_{\mathrm{K}} \setminus \mathcal{E}'$ with $w(e)$ as small as possible. The unique path in $\mathcal{E}'$ from u to v must contain an edge $e' = e'_{xy}$ in $\mathcal{E}' \setminus \mathcal{E}_{\mathrm{K}}$. (See Figure 19 where edges in $\mathcal{E}'$ are shown solid, and ones in $\mathcal{E}_{\mathrm{K}}$ dashed). Similarly, the path in $\mathcal{E}_{\mathrm{K}}$ from x to y contains an edge e'' in $\mathcal{E}_{\mathrm{K}} \setminus \mathcal{E}'$ (conceivably $e'' = e$).

Then $w(e') \geq w(e)$ for otherwise $w(e') < w(e) \leq w(e'')$, and e' would have been added before e in the construction of $\mathcal{E}_{\mathrm{K}}$ (as e'' is absent at that stage of the algorithm). Exchanging e' for e,

$$(\mathcal{E}' \setminus \{e'\}) \cup \{e\}$$

is a spanning tree with total weight no greater than $w(\mathcal{E}')$, and so an MST with more edges in common with $\mathcal{E}_{\mathrm{K}}$. This is a contradiction. □

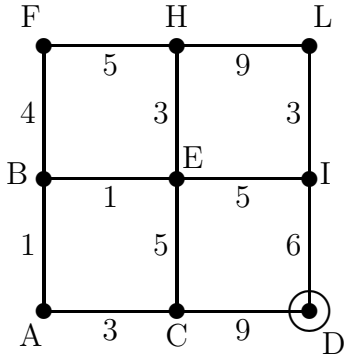
11.3 Prim's algorithm

In the construction of $\mathcal{E}_{\mathrm{K}}$ above, the graph that eventually turns into the spanning tree may be disconnected at intermediate stages. (Typically it consists of a disjoint union of trees, referred to as a *forest*.) The following method, due independently to Jarník and Prim, gets round this defect and gives a way of 'growing' an MST.

Theorem 11.3.1 Let $\mathcal{G}$ be a weighted graph. Choose any vertex to start, and successively add edges of least weight joining a vertex already in the tree to one 'outside'. The result defines an MST for $\mathcal{G}$. □

The procedure is effectively the same as for Kruskal's algorithm, except that one restricts attention to edges *connected* to ones already chosen. Because of this, Prim's algorithm has the big advantage that it can be performed from the *adjacency matrix*, without visualizing the graph. The adjacency matrix of a graph with n vertices is the symmetric $n \times n$ matrix with entries $w(e_{ij})$ where e_{ij} is the edge joining vertex i to vertex j. By convention, the entry 0 means $i \not\sim j$ (though it might be more accurate to use ∞ for this purpose).

Figure 20:



Example 11.3.2 Below is the adjacency matrix corresponding to the mini-version of Figure 17 illustrated in Figure 20. It is followed by an application of the matrix interpretation of Prim's algorithm, starting from vertex D.

$$
\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ H \\ I \\ L \end{array}
\begin{array}{c}
\begin{array}{ccccccccc} A & B & C & D & E & F & H & I & L \end{array} \\
\left(\begin{array}{ccccccccc}
0 & 1 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 4 & 0 & 0 & 0 \\
3 & 0 & 0 & 9 & 5 & 0 & 0 & 0 & 0 \\
0 & 0 & 9 & 0 & 0 & 0 & 0 & 6 & 0 \\
0 & 1 & 5 & 0 & 0 & 0 & 3 & 5 & 0 \\
0 & 4 & 0 & 0 & 0 & 0 & 5 & 0 & 0 \\
0 & 0 & 0 & 0 & 3 & 5 & 0 & 0 & 9 \\
0 & 0 & 0 & 6 & 5 & 0 & 0 & 0 & 3 \\
0 & 0 & 0 & 0 & 0 & 0 & 9 & 3 & 0
\end{array}\right)
\end{array}
$$

$$
\left\{\begin{array}{l} \text{Delete column D} \\ \text{Circle a smallest positive number in row D} \\ \text{Store the corresponding edge DI} \end{array}\right.
$$

$$
\left\{\begin{array}{l} \text{Delete column I} \\ \text{Circle a smallest positive number remaining in rows D,I} \\ \text{Store the corresponding edge IL} \end{array}\right.
$$

$$
\left\{\begin{array}{l} \text{Delete column L} \\ \text{Circle a smallest positive number remaining in rows D,I,L} \\ \text{Store the corresponding edge IE} \end{array}\right.
$$

.....

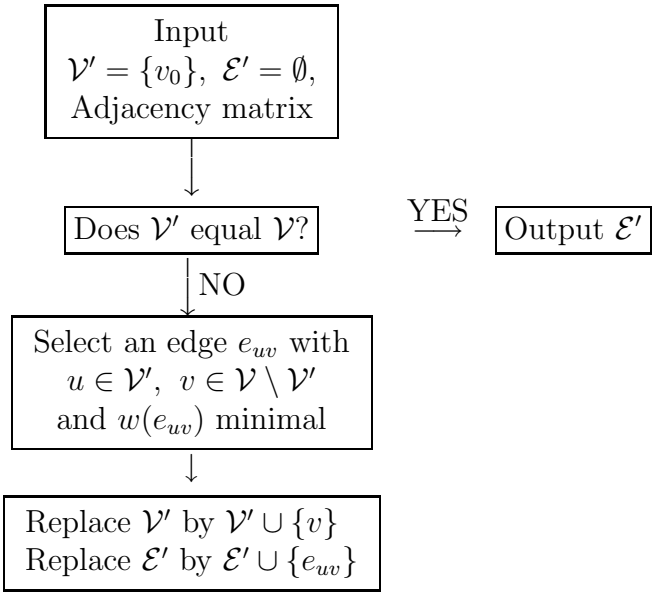
A MST is finally determined by the edge subset

$$\mathcal{E}_{\rm P} = \{DI, IL, IE, EB, BA, AC, EH, BF\},$$

with $w(\mathcal{E}_{\rm P}) = 6+3+5+1+1+3+3+4 = 26$. Whereas EB *had* to be chosen before BA, edge AC was chosen before EH only out of respect for the ordering we adopted in the previous section. □

We have not justified Prim's algorithm, which again exploits the greedy principle. The success of the latter can be extended to a wider class of mathematical objects called 'matroids'. This theory builds on the analogy between subsets of $\mathcal{E}$ (of a given graph) for which there are no cycles, and subsets of linearly independent vectors in a vector space [4, §17.4]. Both types of subset exhibit the 'exchange' property that is used in linear algebra to prove that any maximal set of linearly independent vectors has the same size, namely the dimension of the space. The analogue for a connected graph is that any MST has $|\mathcal{V}| - 1$ edges.

The general procedure for Prim's algorithm can be represented by a flow diagram, in which $\mathcal{V}'$ and $\mathcal{E}'$ are the vertex and edge subsets of the tree at each stage of its growth. To make it methodical, a strategy is once again required to select and extract a suitable edge of least weight at each stage.



⋆ 11.4 Other problems

Drawing by numbers

One of the most natural of all graphical optimization tasks is the *Shortest Path Problem*. Namely, given a weighted graph containing vertices u, v, find a path from one to the other of least total weight (also called *length*). In symbols,

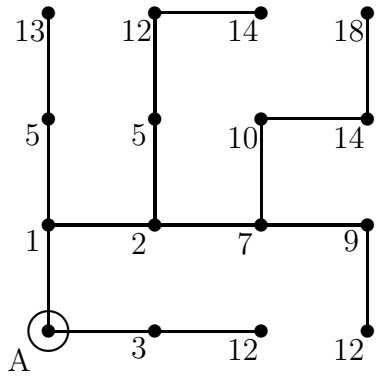
$$\text{Find } \{u = v_0, v_1, v_2, \ldots, v_n = v\} \text{ with } \sum_{i=0}^{n-1} w(v_i v_{i+1}) \text{ minimal.} \qquad (11.3)$$

As an example, the shortest path from A to P in Figure 17 has a length of 18, and is illustrated in Figure 21 together with the shortest paths (and their lengths) from A to all other vertices.

Given u, (11.3) is solved simultaneously for every v by Dijkstra's algorithm, which proceeds by assigning temporary labels to vertices indicating the lengths of shortest paths using only some of the edges. The algorithm starts by labelling u with 0 and all other vertices with ∞. At each intermediate stage, one 'examines' the vertex (or one of the vertices) x with the least label, n, of those that have not already been examined. The label of each vertex y such that $x \sim y$ is then replaced by $n + w(e_{xy})$ if this is smaller. When all the vertices have been examined, *every* vertex z is automatically labelled with the length of the shortest path from u to z.

The shortest paths themselves are found by including every edge e_{xy} for which $w(e_{xy}) = |\ell(x) - \ell(y)|$, where $\ell(x)$ denotes the final label assigned to x. (Hence the title of this subsection.) More explanation and verification can be found in [6, §22.3] and [12, §8.2].

Figure 21: Shortest paths from A



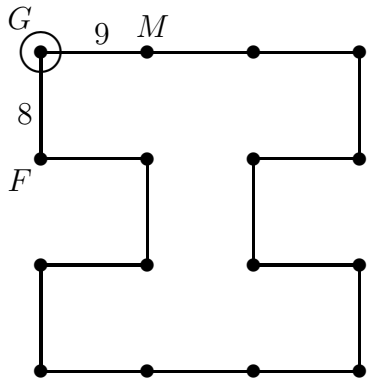
Selling by cycles

Given a connected weighted graph $\mathcal{G}$, the *Travelling Salesperson Problem* consists of finding a cycle that passes through every vertex exactly once (assuming that such a 'Hamiltonian' cycle exists) of least total weight or length. This an example of a problem for which there is no known algorithm, other than checking through a list of all potential solutions (that would be prohibitive for large graphs). However, an estimate on the length of a winning cycle can be derived from knowledge of minimal spanning trees of certain subgraphs as follows.

Let $\mathcal{C} = (v_0, v_1, \ldots, v_{n-1}, v_0)$ be a solution; thus $n \geq 3$ and we have distinguished one of the vertices, namely v_0, on the cycle. Let $\mathcal{G}'$ be the graph formed from $\mathcal{G}$ by removing v_0 and all edges joining v_0. Then $(v_1, \ldots, v_{n-1})$ is a spanning tree for $\mathcal{G}'$, so

$$\begin{aligned} w(\mathcal{C}) &\geq (\text{weight of an MST for } \mathcal{G}') + w(v_0v_1) + w(v_{n-1}v_0) \\ &\geq (\text{weight of an MST for } \mathcal{G}') + (\text{smallest two weights attached to } v_0). \end{aligned}$$

Figure 22: Salesperson's cycle



If $v_0 = G$ in our original example then from Figure 18 it is clear that $\mathcal{G}'$ has an MST of weight $51 - 8 = 43$. So without knowing $\mathcal{C}$ we can predict that

$$w(\mathcal{C}) \geq 43 + 8 + 9 = 60,$$

which is not far off the mark. Actually, Figure 17 possesses exactly 6 cycles that pass through every vertex once and only once (to see this, first note that all four 'corners' must be included). The shortest one is shown in Figure 22 and has length 65.

Working by zeros

A rather different class of problems can be illustrated by a weighted graph $\mathcal{G}$ in which the set of vertices is the disjoint union of two disjoint subsets $\mathcal{V}_1$, $\mathcal{V}_2$ with $|\mathcal{V}_1| = |\mathcal{V}_2| = n$, and each element of $\mathcal{V}_1$ is joined by exactly one edge to each element of $\mathcal{V}_2$. (Ignoring the edge from y to z and adding weights, the graph in Figure 16 is of this type.) The problem is then to find a subset $\mathcal{E}'$ of n edges providing a bijective correspondence between $\mathcal{V}_1$ and $\mathcal{V}_2$, and with $w(\mathcal{E}')$ minimal. This becomes a matter of optimal 'job assignment' if $\mathcal{V}_1$ represents a set of jobs to be done, $\mathcal{V}_2$ a set of workers available, and $w(e_{ux})$ measures the unsuitability of giving job u to worker x.

A successful algorithm for determining $\mathcal{E}'$ can be carried out in matrix form, in the same spirit as Prim's algorithm. The relevant part of the adjacency matrix of $\mathcal{G}$ is the block in which the rows are labelled by elements of $\mathcal{V}_1$ and the columns by elements of $\mathcal{V}_2$. As an example, let

$$\mathcal{V}_1 = \{F, B, A, C\}, \quad \mathcal{V}_2 = \{H, L, I, D\},$$

and let $w(e_{ux})$ equal the length of the shortest path from u to x in Figure 20. The new adjacency block is then

$$\begin{array}{c} \\ F \\ B \\ A \\ C \end{array} \begin{array}{c} \begin{array}{cccc} H & L & I & D \end{array} \\ \begin{pmatrix} 5 & 13 & 10 & 16 \\ 4 & 9 & 6 & 12 \\ 5 & 10 & 7 & 12 \\ 8 & 13 & 10 & 9 \end{pmatrix} \end{array}$$

The first step consists of subtracting the least element of a row from all the elements in that row. After this has been done for each row, one carries out the same procedure on the columns, subtracting the least element in each column from all the elements in that column. This reduces the above matrix to

$$\begin{array}{c} \\ F \\ B \\ A \\ C \end{array} \begin{array}{c} \begin{array}{cccc} H & L & I & D \end{array} \\ \begin{pmatrix} 0 & 3 & 3 & 10 \\ 0 & 0 & 0 & 7 \\ 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Because subtracting a constant from a given row or column does not affect the solution, any choice of four 0's such that no two are in the same row and no two in the same

column (these are called 'independent zeros') now determines the four edges whose total weight is least. There are two solutions above, namely

$$\mathcal{E}' = \{FH, BL, AI, CD\} \quad \text{and} \quad \{FH, AL, BI, CD\},$$

both with the smallest value $w(\mathcal{E}') = 30$ which happens to be the trace of the first matrix.

In general, there is no guarantee that after the row and column subtractions, the resulting matrix $\mathcal{M} = (m_{ij})$ possesses n independent zeros. In this case, further steps are required to complete the algorithm. Suppose that all the 0's lie in a union of rows and columns determined by respective subsets I and J of $\{1, 2, \ldots, n\}$ with $|I| + |J| < n$. The set of entries of $\mathcal{M}$ is then the disjoint union

$$F_+ \cup F_0 \cup F_-,$$

where $m_{ij} \in F_+$ iff bith $i \in I$ and $j \in J$, and $m_{ij} \in F_-$ iff both $i \notin I$ and $j \notin J$. A matrix $\mathcal{M}' = (m'_{ij})$ is formed by subtracting the least element f of F_- from all elements of F_- and adding f to all elements of F_+, and it is easy to see that

$$\sum_{i,j} m'_{ij} < \sum_{i,j} m_{ij}. \tag{11.4}$$

If $\mathcal{M}'$ has n independent zeros their position gives the solution; if not the process can be repeated, and (11.4) guarantees success after a finite number of steps. The correctness of this algorithm is established in [1, §3.3].

11.5 Exercises

1. Apply Kruskal's algorithm to Figure 17 but weighted instead
(i) by the first 24 digits 271828182845904523536029 of e, treating '0' as ∞ (which amounts to deleting the corresponding edges HM and MN);
(ii) equally, i.e. with all edges of weight 1.
In each case the resulting MST will depend (to a lesser or greater extent) on your ordering of the edges.

2. A graph consists of 8 vertices labelled by the numbers $2, 3, 4, 5, 6, 7, 8, 9$, and every pair $\{i, j\}$ (with $i \neq j$) of vertices is joined by an edge of weight
(i) $|i - j|$;
(ii) ij;
(iii) $ij + 7|i - j|$;
(iv) $\mathrm{lcm}(i, j)$.
In each of these four separate cases, find an MST by Prim's algorithm, and represent it by a sketch *without* attempting to draw the original graph.

3. Consider the following weighted graph. Its vertices are labelled BS, E, GP, H, KC, LS, NHG, OC, PC, SK, V, and correspond to 11 particular stations on intersections of

the Bakerloo, Central, Circle, Piccadilly, and Victoria underground lines in London. Its edges are the segments of these lines that join two vertices without passing through a third, and the weight of an edge is assigned by counting each passage between adjacent stations as 1 unit (e.g. $H \leftrightarrow LS$ is 4). Find an MST using (i) Kruskal's algorithm, and (ii) Prim's algorithm.

4. In the previous question, find the lengths of a shortest path from BS to each of the other vertices. What is the length of a shortest cycle that passes through each vertex exactly once?

5. Prove the formula (11.2) for planar graphs.

6. Is it a coincidence that the configuration of paths in Figure 21 forms a tree? Find the configuration when Figure 17 is weighted by the digits 57721566490153286060651 2 of γ (and deleting the edges corresponding to '0').

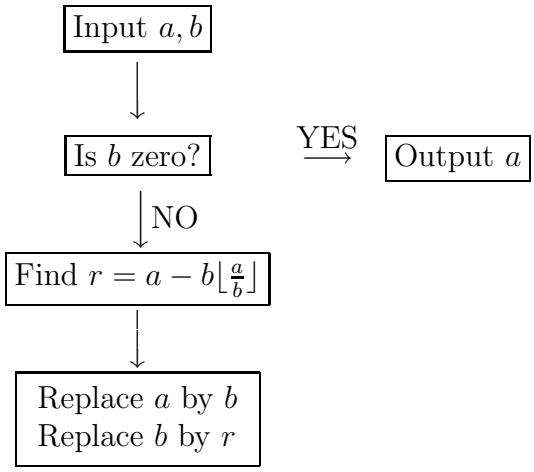
7. (i) Determine by trial and error the 16 shortest paths in Figure 17 between each vertex in $\mathcal{V}_1 = \{G, F, B, A\}$ and each vertex in $\mathcal{V}_2 = \{P, O, K, J\}$. Write down the corresponding lengths as a 4×4 matrix $\mathcal{M}$.
(ii) Let $\mathcal{G}$ be the weighted graph with vertex set $\mathcal{V}_1 \cup \mathcal{V}_2$ and 16 edges joining the elements of $\mathcal{V}_1$ and $\mathcal{V}_2$ and adjacency matrix $\mathcal{M}$ as in §11.4. Find an optimal assignment $\mathcal{E}'$.

8. Prove the inequality (11.4).

12 Algorithm Analysis and Sorting

12.1 Efficiency of Euclid's algorithm

Euclid's algorithm was exploited in §10.2 to find the greatest common divisor of two positive integers a, b. It consists in repeating what is essentially the same simple division step until the remainder becomes zero. The whole process can be summarized by the following flow diagram:



In a computer program, the replacement operations would be represented by instructions like `a:=b; b:=r;` and must be carried out in the correct order. For, if b is replaced by r first, the old value of b will be lost and a will get replaced by r too.

Let $t(a, b)$ denote the number of times the double replacement operation in the last box is carried out, or equivalently the number of 'loops' that are executed. Observe that $t(a, b)$ is necessarily finite since each time b is replaced by r its size reduces; also $t(a, b) = 0$ iff $b = 0$. Referring back to Problem 10.2.2, we see that $t(a, b)$ is simply the number of rows of the form $a = qb + r$ up to and including the first in which $r = 0$; thus

$$t(5432, 2345) = 5, \quad t(2345, 5432) = 6.$$

On the other hand, $t(a, b)$ can be deceptively small even when a and b are enormous; for instance (using `gcd1` from §10.5)

$$t(100! + 1, 100^{100} - 1) = 337. \tag{12.1}$$

Without wanting to investigate the technicalities of how a computer accomplishes the various steps of the algorithm, it is reasonable to suppose that $t(a, b)$ will provide a fair indication of the time taken to compute the greatest common divisor of a and b. We are not interested in knowing *exactly* what the function t is, but merely some idea of how it grows in size relative to a and b. The next result answers this question by bringing together two topics central to the course, namely the Fibonacci numbers F_n and asymptotic notation.

Proposition 12.1.1

$$t(a,b) = \mathrm{O}(\lg b).$$

Proof. We may suppose that $a > b > 0$, for if $a < b$ one extra loop reverses the roles of a, b as explained in §10.2. We first show by induction on n that

$$\boxed{t(a,b) = n \quad \Rightarrow \quad a \geq F_{n+2}, \quad b \geq F_{n+1}}$$

The statement is valid for $n = 1$ since if $t(a,b) = 1$ then $b \geq 1 = F_2$ and $a \geq 2 = F_3$. In general, write $a = qb + r$ with $0 \leq r < b$ so that $t(a,b) = t(b,r) + 1$ and

$$b \geq F_{n+1},\ r \geq F_n \quad \Rightarrow \quad a = qb + r \geq F_{n+1} + F_n = F_{n+2}, \quad b \geq F_{n+1},$$

and the induction is successful.

Given $b \geq F_{n+1} \geq 2$ with $n = t(a,b)$, using Corollary 4.2.2, we get

$$\begin{aligned} \phi^n < \phi^{n+1} < \sqrt{5}(b+1) \quad &\Rightarrow \quad n \lg \phi \leq \tfrac{1}{2} \lg 5 + 2 \lg b \\ &\Rightarrow \quad \frac{n}{\lg b} \leq \frac{\lg 5}{2 \lg \phi \lg b} + \frac{2}{\lg \phi}. \end{aligned}$$

This means that there exists $C > 0$ such that $t(a,b) \leq C \lg b$ for all $b \geq 2$ (in fact $C = 5$ works) and the result follows from Definition 9.1.1. □

A worst case scenario is exemplified by the Fibonacci numbers. If $a = F_{n+2}$ and $b = F_{n+1}$ for some n, then Euclid's algorithm requires n steps to arrive at a line with remainder 0. This is because F_k divides F_{k+1} once with remainder of F_{k-1} for all k. Consider, for instance, the frustrating 9-step calculation

$$\begin{aligned} \gcd(F_{11}, F_{10}) &= \gcd(89, 55) \\ &= \gcd(55, 34) \\ &= \gcd(34, 21) \\ &= \gcd(21, 13) \\ &= \gcd(13, 8) \\ &= \gcd(8, 5) \\ &= \gcd(5, 3) \\ &= \gcd(3, 2) \\ &= \gcd(2, 1) = 1. \end{aligned}$$

Nevertheless, the above proposition confirms that Euclid's algorithm is remarkably effective at computing the greatest common divisor of a pair of large numbers.

The relative performance or 'complexity' of algorithms is often assessed in terms of such an asymptotic upper bound on the number of basic operations needed to process data of 'size' n. Strictly speaking, such a bound gives no information since the constant C in Definition 9.1.1 could be enormous; in fact the choice of an optimal algorithm in given circumstances *will* often depend on the data size. However, in many cases the value of n dwarfs any hidden constant, and on an asymptotic scale we shall see that it is too much to expect of other algorithms that they need 'only' $\mathrm{O}(\lg n)$ operations.

12.2 The sorting problem

In attempting to solve optimization problems of the type discussed in §11.3, one of the most basic tasks required in practice is the ability to reorder a sequence of numbers numerically. Indeed, to apply Prim's algorithm on a computer one needs to take one step in this direction to find a smallest number in rows of the adjacency matrix. In everyday life, one often takes for granted the convenience of having data organized numerically, and the effect of an operation that achieves this can even be seen on BBC at about (currently) 8pm each Saturday night.

More precisely, the *sorting problem* consists in finding an algorithm or mechanical procedure to arrange a sequence $(a_1, a_2, \ldots, a_n)$ of n real numbers into ascending numerical order. This amounts to finding a permutation π of the set $\{1, 2, \ldots, n\}$ such that

$$a_{\pi(1)} \leq a_{\pi(2)} \leq \cdots \leq a_{\pi(n)}.$$

It is not a good idea though to embark on a list of all possible $n!$ permutations, because $n!$ is so large in comparison to n (and each permutation would require up to $n-1$ pairwise comparisons to check its validity). Instead, a brief study of some practical sorting algorithms will illustrate the quest for more efficiency.

Example 12.2.1 The following program is run in MAPLE, after first defining `y` to be a sequence of 10 real numbers:

```
for j from 1 to 9 do
  for k from 10-j to 9 do
    if y[k]>y[k+1] then y:=subsop(k=y[k+1],k+1=y[k],y) fi
  od:
  print(y)
od;
```

The rather complicated looking command on the third line after `then` merely redefines `y` to be the new sequence obtained by swapping the kth and $(k+1)$st terms `y[k]`,`y[k+1]` of the old one.

If we input the sequence

```
y:=[8,3,5,9,2,6,5,1,7,4]:
```

and then the program, MAPLE produces the lines

$$[8, 3, 5, 9, 2, 6, 5, 1, 4, 7]$$
$$[8, 3, 5, 9, 2, 6, 5, 1, 4, 7]$$
$$[8, 3, 5, 9, 2, 6, 1, 4, 5, 7]$$
$$[8, 3, 5, 9, 2, 1, 4, 5, 6, 7]$$
$$[8, 3, 5, 9, 1, 2, 4, 5, 6, 7]$$
$$[8, 3, 5, 1, 2, 4, 5, 6, 7, 9]$$
$$[8, 3, 1, 2, 4, 5, 5, 6, 7, 9]$$
$$[8, 1, 2, 3, 4, 5, 5, 6, 7, 9]$$
$$[1, 2, 3, 4, 5, 5, 6, 7, 8, 9]$$

The effect of the program is clear from examining the output. Each line is produced from the previous one by moving an appropriate digit as far to the right as needed so that there remains one less digit of the original sequence unsorted. □

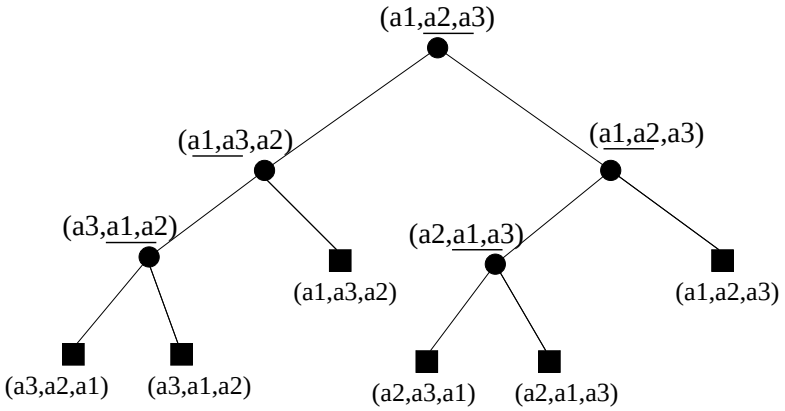
The above example requires a total of $1 + 2 + 3 + \cdots + 9 = 45$ single comparisons, though with the given sequence only 26 swaps were actually made. The maximum number of comparisons needed to sort a sequence of n numbers with the same method is

$$\tfrac{1}{2}n(n-1) = \mathrm{O}(n^2). \tag{12.2}$$

This is much 'worse' than Euclid's $\mathrm{O}(\lg n)$, and would require excessive time for large n, even on a very powerful computer. For example, if each comparison is done in 10^{-6} of a second, it would take well over an hour to order a modest 100000 numbers, neglecting storage and accessing time. In the remainder of this section, we shall address the question of to what extent there might exist other methods of sorting that improve on (12.2). This has wide implications since, as remarked above, sorting is a constituent of many more general algorithms.

For simplicity, we shall consider only those programs that are based on comparing two numbers at a time. This rules out other methods that can for example be used if all the numbers are positive integers, so that their own *values* can be exploited in the ordering process. To sort $n = 3$ numbers (which need not be integers) as above one actually mimics the following 'decision tree' in which a solid dot represents a comparison of the two numbers underlined. If the first is less than or equal to the second, descend to the right, if not descend to the left and interchange the two numbers. Eventually one arrives at a square which outputs the correctly-ordered sequence. The tree has 3 'levels', so at most 3 comparisons are needed irrespective of the original order of the sequence.

Figure 23: Decision tree



Given a sorting algorithm for n numbers that proceeds by repeatedly comparing two numbers at a time, let us denote the maximum number of comparisons needed by $h(n)$.

Proposition 12.2.2

$$h(n) = \Omega(n \lg n).$$

Proof. The procedure can always be represented by a decision tree as in Figure 23, but with the maximum number of edges from top to bottom equal to $h(n)$; thus there are $h(n)$ levels instead of 3. Even if all the branches of the tree are present the total number of outputs cannot exceed $2^{h(n)}$. Each of the possible $n!$ orderings may conceivably be output in more than one place, but in any case the key inequality is

$$\boxed{n! \leq 2^{h(n)}}$$

This implies that

$$h(n) \geq \lg(n!) \quad \Rightarrow \quad \frac{h(n)}{n \lg n} \geq \frac{\lg(n!)}{n \lg n} = \frac{\ln(n!)}{n \ln n}.$$

But the last term tends to 1 by Corollary 9.4.2, so certainly $h(n) \geq cn \lg n$ for some $c > 0$ and all n sufficiently large. □

This result gives a theoretical asymptotic lower bound on $h(n)$. Roughly speaking, it says that however clever an algorithm we design, the best we can hope for is that the time taken to sort n numbers will be proportional to $n \lg n$. The function $n \lg n$ lies between n and n^2 (see the graphs in §12.4), but is 'closer' to n in the sense that if one is prepared to wait 10^6 milliseconds $\sim$ 20 minutes for a result, one can probably wait $10^6 \lg(10^6)$ milliseconds $\sim 5\frac{1}{2}$ hours, but not 10^{12} milliseconds $\sim$ 32 years!

⋆ 12.3 MergeSort and HeapSort

In this section we shall show that there do exist sorting algorithms that achieve the lower bound of Proposition 12.2.2. In other words, it is possible to sort n numbers in such a way that the maximum number of comparisons needed is $\Theta(n \lg n)$. This estimate fits in snugly between those of Example 12.2.1 and Euclid's algorithm.

Example 12.3.1 The 'merging' operation, denoted μ, combines two ordered sequences of respective length m, n into a single ordered one of length $m + n$. For example,

$$\mu((2,3,5,8,9),(1,4,5,6,7)) = (1,2,3,4,5,5,6,7,8,9). \tag{12.3}$$

The MergeSort function, denoted M, is then defined recursively by

$$M(a_1, \ldots, a_n) = \mu(M(a_1, \ldots, a_{\lfloor n/2 \rfloor}), M(a_{\lfloor n/2 \rfloor + 1}, \ldots, a_n)), \quad n \geq 2,$$

and $M(a_1) = (a_1)$ for $n = 1$.

The operation μ can easily be carried out methodically. The right-hand side of (12.3) could have been arrived at by 'stripping away' numbers from the beginning of the sequences on the left, repeatedly comparing the two initial numbers remaining at each stage. After the half-way point

$$\mu((\ldots,8,9),(\ldots,5,6,7)) = (1,2,3,4,5,\ldots),$$

only the three comparisons between 8 and 5, 8 and 6, 8 and 7 are left to complete the job. In general, using this method, one may merge sequences of length m, n using a maximum of $m + n - 1$ comparisons.

The following scheme shows what happens when MergeSort is applied to the sequence $(8, 3, 5, 9, 2, 6, 5, 1, 7, 4)$ of Example 12.2.1; it is designed merely as a visual aid and is not an accurate reflection of the order a computer might carry out the steps. The sequence is first broken up into groups until all the subsequences have size 1, achieved in the middle row. Then the operation μ is applied to recombine the groups in the same way.

8359265174

83592 65174

83 592 65 174

8 3 5 92 6 5 1 74

8 3 5 9 2 6 5 1 7 4

38 5 29 56 1 47

38 259 56 147

23589 14567

1234556789

□

MergeSort uses the 'divide and conquer' principle common to many other algorithms; it breaks the problem up into smaller parts which can be tackled more easily. Related to this is the concept of 'modular' design, whereby an algorithm is built up from subroutines that can be inserted in various orders to form a whole. The merging operation μ, itself accomplished by an algorithm described above, forms the building block in this case.

Proposition 12.3.2 The number $h(n)$ of comparisons needed to MergeSort n numbers satisfies $h(n) = \mathrm{O}(n \lg n)$.

Proof. The MergeSort of n numbers can always be represented by the above scheme. If $2^{k-1} < n \leq 2^k$, there are k rows below the middle one, and the merging required to pass from one row to the next involves less than 2^k comparisons. Since $k - 1 < \lg n$, the total number of comparisons required is less than

$$k2^k < (1 + \lg n)2^{1+\lg n} = 2(1 + \lg n)n = \mathrm{O}(n \lg n).$$

□

Our final example has been included to illustrate an ingenious way of organizing data in the implementation of an algorithm. This is based on the concept of a binary tree which formalizes the structure of Figure 23, and has widespread applications. Further details can be found in [3], from where the following summary is taken.

A *binary tree* is not strictly speaking a tree in the sense of §11.1. Rather it is a tree with a finite number of vertices (also called nodes) and edges, *plus* some extra structure.

First and foremost, a binary tree has a special node, called the *root* and (despite its name) conventionally placed at the top of the diagram. The root is attached to 0, 1 or 2 edges which (provided at least one is present) are distinguished by the names 'left' and 'right'. Each of these edges is then attached to the root of another binary tree, and in this way the structure is defined inductively.

The existence of the root r allows one to define the *ancestors* of a given node x of a binary tree as all the nodes that lie on the unique path from r to x. Moreover, x is a *child* of y if y is an ancestor connected by a single edge to x, and each node has at most a left child and a right child. Two binary trees are deemed to be different if the same node has only a left child in one and a right child in the other (see §12.4). A node with no children is called a *leaf* (shown by a square in Figure 23). Given a node x, we shall denote by h_x the number of 'levels' below x, i.e. the greatest number of edges on a path from x to a leaf.

Example 12.3.3 A sequence of n numbers is first arranged as exemplified in Figure 24 for $n = 10$. The kth number is assigned to the kth node of a binary tree ordered a row at a time from left to right, starting from node 1, the root. The kth node gives rise to exactly two children, namely the $2k$th and $(2k + 1)$st nodes, until the data runs out.

Such a structure is said to be a *heap* if the number assigned to each node is at least as great as those assigned to its (0, 1 or 2) children. A heap may be regarded as a 'partially ordered display'; a maximum from the set of numbers will always be positioned at the root of the heap, and can be extracted immediately.

Just as MergeSort was based on the merging operation, so the fundamental process for HeapSort is 'heapifying' a given node x, which means moving numbers around so that the data assigned to the binary tree with root x forms a heap. In carrying out this process, it is assumed that *the binary trees whose roots are the children of x already form heaps*. Heapifying a node x then consists in moving its number as far down the tree as necessary, swapping it with the greatest of its two children until it dominates both. The process requires a maximium of $2h_x$ comparisons, where h_x is defined above.

The HeapSort Algorithm has two parts:

Part I. This consists in converting the data into a heap. Starting from the bottom of the tree and working backwards, nodes are heapified one at a time. In fact, node i has no children if $i > \lfloor n/2 \rfloor$, so one needs only start at node $\lfloor n/2 \rfloor$ and work backwards until reaching the root r. By the time r has itself been heapified, the whole tree structure is indeed a heap (see Figure 25).

Part II. For each i starting at n and finishing at 1, the number at the root is outputted and (provided $i \geq 2$) replaced by the number at the bottom remaining node i. This node is then deleted from the tree along with its upward edge, and the root immediately heapified. After part II is complete, all the elements of the original sequence have been extracted in numerical order. (At the half-way stage in Figure 26, the operation for $i = 6$ has just been executed by extracting a 5 and moving 1 up to the root, which is about to be heapified by interchanging 1 and the remaining 5.) □

Figure 24: Initial data

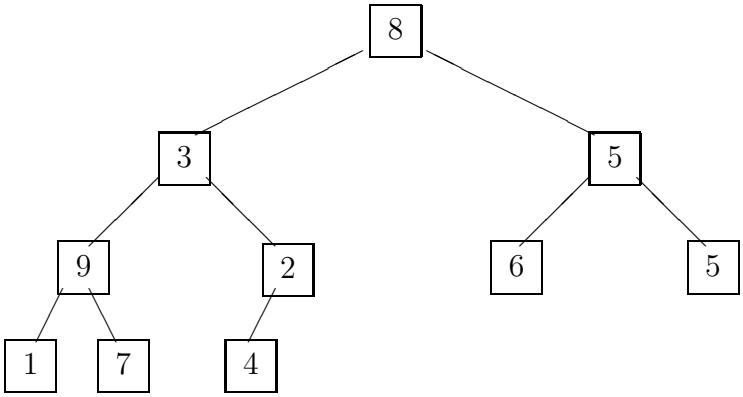


Figure 25: A heap

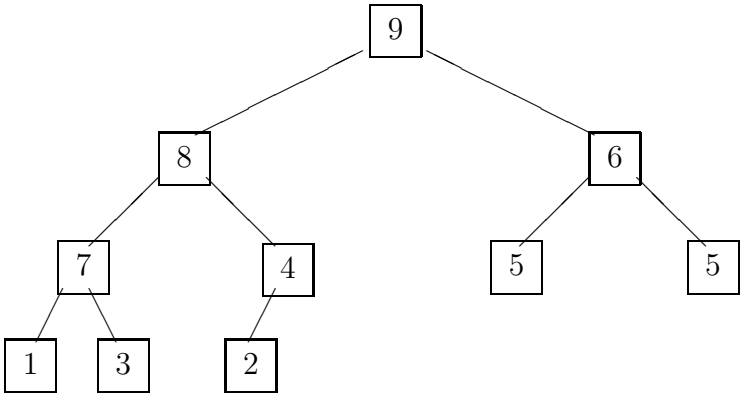
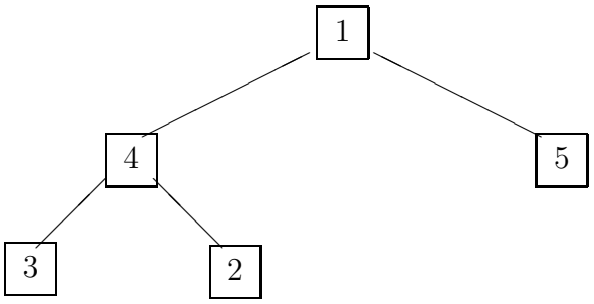


Figure 26: Output



. . . ,5,6,7,8,9.

Let $h = h_r$ denote the number of levels below node 1, the root, so that

$$2^h \leq n < 2^{h+1}. \tag{12.4}$$

The number of comparisons needed to carry out part I of HeapSort is certainly no greater than $\lfloor n/2 \rfloor .2h = \mathrm{O}(n \lg n)$, since each of the $\lfloor n/2 \rfloor$ nodes heapified has no more than h levels below it. The number of comparisons needed for part II is no more than $(n-1).2h$, so the overall bound for completing HeapSort is $\mathrm{O}(n \lg n)$, just as for MergeSort.

The first part of this estimate can be improved:

Proposition 12.3.4 Given n numbers, the number of comparisons required to complete part I of HeapSort is $\mathrm{O}(n)$.

Proof. In part I, the nodes that are heapified are (in reverse order): 1 (which has h levels below it); 2,3 (which have $h-1$ levels below); 4,5,6,7 ($h-2$ levels below), and so so. The maximum number of comparisons required to heapify a node x equals $2h_x$, so the sum of these maxima is no greater than twice

$$\begin{aligned} 1.h + 2.(h-1) + 4.(h-2) + \cdots + 2^{h-1}.1 &= \sum_{i=1}^{h} 2^{h-i}.i \\ &< 2^h . \sum_{i=1}^{\infty} i 2^{-i} \\ &\leq \tfrac{1}{2} n, \end{aligned}$$

using (7.7) and (12.4). □

This result is expressed by saying that a heap can be built in 'linear time'. It has applications for the graphical algorithms discussed earlier, since the problem of selecting an edge of least weight can be solved by organizing relevant data into a heap. With this technique, it can be shown (see [3, §23]) that Prim's algorithm can be implemented with complexity $\mathrm{O}(|\mathcal{E}| \lg |\mathcal{V}|)$, where $|\mathcal{E}|$ is the number of edges and $|\mathcal{V}|$ the number of vertices of the graph.

12.4 Exercises

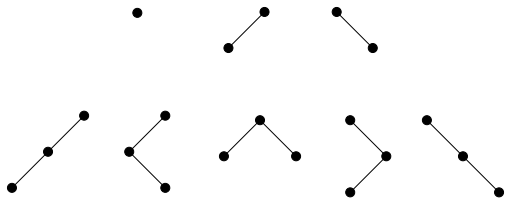
1. The following procedure computes the nth Fibonacci number:

```
f:= proc(n)
  if n<2 then n
  else f(n-1)+f(n-2) fi
end:
```

(i) How many times is the recursive definition after `else` applied to compute `f(8)`? Show that, if t_n denotes the number of times this is applied to compute `f(n)` then $t_n = t_{n-1} + t_{n-2} + 1$. Deduce from §4.2 that $t_n = \mathrm{O}(2^n)$.
(ii) Make a table of some large integers `n` and the times a computer takes to output `f(n)`. Explain why inserting `option remember:` as a new second line drammatically improves the results.

2. Let $\mathcal{G}$ be a weighted graph with n vertices. Show that the number of comparisons of pairs of numbers needed to complete Prim's algorithm from the adjacency matrix of $\mathcal{G}$ is $\mathrm{O}(n^2)$.

3. Let C_n denote the number of binary trees with n nodes, and let $f(x) = \sum_{k=0}^{\infty} C_k x^k$ be its GF (with $C_0 = 1$). The picture shows that $C_1 = 1$, $C_2 = 2$ and $C_3 = 5$. Verify that $C_4 = 14$. By considering the root and its two subtrees, prove that $x f(x)^2 - f(x) + 1 = 0$. (A related example is carried out in [1, §4.1], but without the distinction between left and right that exists for binary trees.)



4. Solve the quadratic equation for $f(x)$ in the previous question. Expand the square root, and prove that $C_n = \frac{1}{n+1}\binom{2n}{n}$. Deduce from §9.4 that $C_n = \mathrm{O}(2^{2n})$.

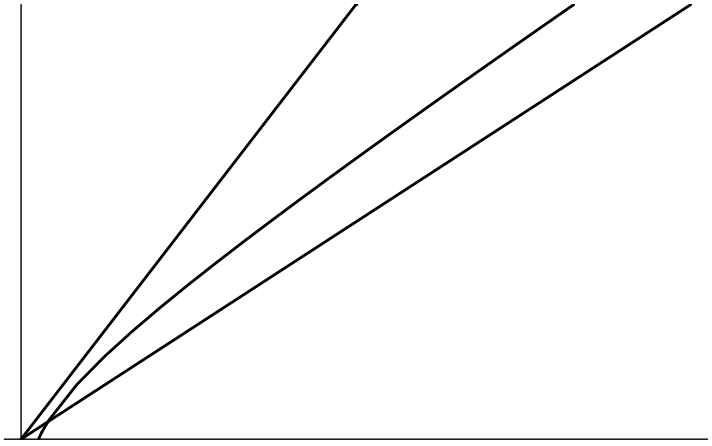
5. A comparison of the growth of the sequences n^2, $n \lg n$ and n is sometimes represented by plotting their respective graphs with 'log scales' as in Figure 27. But exactly what functions are being plotted here (there is a clue in §4.2)? And how would $n^2/\lg n$ and $n \lg\lg n$ fit into the picture?

6. Use the identity (4.12) to prove the remarkable equation

$$\gcd(F_m, F_n) = F_d, \quad \text{where} \quad d = \gcd(m, n).$$

7. Match up the following description of HeapSort in Maple with the vaguer English

Figure 27:



description in the text. Heapifying node `i` of a binary tree representing the `n` terms of a sequence `x` is given by the procedure

```
heapify:= proc(i,n,x)
  local m,y: y:=x:
  if 2*i<=n and y[2*i]>y[i] then m:=2*i else m:=i fi:
  if 2*i+1<=n and y[2*i+1]>y[m] then m:=2*i+1 fi:
  if m>i then y:=subsop(i=y[m],m=y[i],y); y:=heapify(m,n,y) fi:
  y
end:
```

Part I is then accomplished by

```
for i from trunc(n/2) by -1 to 1 do
  x:=heapify(i,n,x)
od;
```

and part II by

```
for i from n by -1 to 2 do
  x:=subsop(1=x[i],i=x[1],x):
  n:=n-1: x:=heapify(1,n,x)
od;
```

Try this out on the usual suspects

```
n:=10; x:=[8,3,5,9,2,6,5,1,7,4]:
```

8. MAPLE has its own built-in sorting command, which is in fact based on a variant of MergeSort. Read `?sort` to find out what this is capable of.

Bibliography

The following texts are quoted for reference purposes only. Many of them contain material that extends far beyond the Moderations course, but the chapters indicated incorporate valuable treatments of relevant topics.

1. I. Anderson: *A First Course in Combinatorial Mathematics*, Oxford University Press, reprinted second edition, 1992 [Chapters 2,4,5].
2. W.E. Boyce, R.C. DiPrima: *Elementary Differential Equations and Boundary Value Problems*, John Wiley & Sons, fifth edition, 1992 [Parts of chapters 1,2,3,8].
3. T.H. Corman, C.E. Leiserson, R.L. Rivest: *Introduction to Algorithms*, MIT Press and McGraw-Hill, eleventh printing, 1994 [§1–5, §7, §24, §33].
4. R.L. Graham, D.E. Knuth, O. Patashnik: *Concrete Mathematics*, Addison-Wesley, second edition, 1994 [Parts of chapters 5,6,7,9].
5. G.H. Hardy, E.M. Wright: *Introduction to the Theory of Numbers*, Oxford University Press, reprinted third edition, 1956 [Parts of chapters I,II,XIX].
6. E. Kreyszig: *Advanced Engineering Mathematics*, John Wiley & Sons, seventh edition, 1993 [Chapters 1,2,20,22].
7. E. Kreyszig, E.J. Norminton: *Maple Computer Manual for Advanced Engineering Mathematics*, John Wiley & Sons, 1994 [Chapters 1,2,14,20].
8. C.L. Liu: *Introduction to Combinatorial Mathematics*, McGraw-Hill, 1968 [Chapters 1,2,3,4]
9. W.B. Stewart: *Abstract Algebra*, Mathematical Institute, 1994.
10. D. Stirzaker: *Elementary Probability*, Cambridge University Press, 1994 [Chapters 1,2,3].
11. A.J. Wathen: *Exploring Mathematics with Maple*, Students' Guide, MT, Mathematical Institute, 1996.
12. R.J. Wilson, J.J. Watkins: *Graphs, an Introductory Approach*, John Wiley & Sons, 1990 [Chapters 1,2,8,10].